Production and Operations Management

Identifying Influential Users by Topic in Unstructured User-generated Content

Abstract: Social media users generate a variety of content that can potentially influence their followers' behaviors and preferences. However, most user-generated content, such as text and images, is unstructured, making it difficult to analyze how individuals impact the creation of such data as existing models primarily focus on the numerical aspects of behavioral data. In this paper, we propose a method to identify influential users who significantly affect others' interest in content topics, with a specific focus on text and images. Our study introduces a new variant of the topic model, incorporating a hierarchical structure and a vector auto-regressive approach. This method accounts for both the evolution of topic distribution and the social influence among users on their interest in content topics. The empirical application of our model to image-sharing social media data demonstrates that our model outperforms conventional topic models in terms of predictive accuracy and topic interpretability. Moreover, we illustrate how visualizing the estimated social influence within the network can provide valuable insights for seeded marketing campaigns and data-driven product development by identifying influential users in various content areas. This approach also offers a deeper understanding of evolving trends in content, preferences, and demand.

Key words: Social influence, User-generated content, Unstructured data, Networks, Topic models

1 Introduction

The concept of social influence, where consumers affect each other's behavior, has become a key area of research in marketing and demand management. Early studies, such as Bass (1969)'s model of product diffusion, recognized the role of imitators who are influenced by existing customers through interactions like word-of-mouth. Subsequent research has examined how customer review ratings impact product adoption and evaluation by potential customers (Godes and Mayzlin 2004). These studies primarily used aggregate data to explore group-level social influence. However, recent years have seen a growing interest in the individual-level impact of social interactions, particularly through the lens of disaggregated behavioral data available on social media platforms. For instance, Trusov et al. (2010) developed an econometric model to assess the influence of a user's login activity on their followers' behavior within a social media platform. Their model estimates the distinct social influence of each connection between users and identifies the most influential individuals in the network. This offers valuable insights for viral marketing seeding strategies and demand estimation based on the extent of influence.

Many existing studies often assume that an influencer's impact is consistent across all topics, content, and products. However, this assumption may not always hold true. Consider a hypothetical scenario involving

two social media users, A and B. User A follows user B and has diverse interests, as reflected in her posts about fashion and music. Initially, A's content is almost evenly divided between these topics, but over time, her posts increasingly focus on music. User B, an established authority in music, consistently posts content on this topic. If A's growing interest in music is influenced by B's posts, it would be reasonable to conclude that B's influence is stronger on music-related content than on other topics. Our study seeks to develop a methodology for estimating such topic-specific influences from user-generated content (UGC) data.

This study addresses three critical research gaps in the existing literature on social influence. First, despite the extensive body of literature on social influence, there has been limited exploration of its structure within unstructured user-generated content (UGC), such as text, images, and videos. Most research has focused on quantifiable aspects of consumer behavior, such as the number of adoptions or the volume of spending. However, on social media, users often draw information from non-numerical content found in text and images (Wang et al. 2021). Therefore, it is essential to analyze how this unstructured content influences other users and shapes their motivations for creating content. This study specifically investigates the impact of UGC on content topics, with a particular emphasis on image-based content generated by other users. Such insights could be valuable for understanding evolving trends in product and content design, ultimately contributing to the development of generative AI applications (Carlson et al. 2023).

Second, while previous studies have examined certain moderators that affect the strength of social influence, such as types of connections and individual attributes, this study takes a more comprehensive approach. It generalizes the heterogeneity of influences based on network locations by assuming distinct social influences for each network edge. This broad assumption captures all potential moderators, enabling us to identify who exerts strong influence on whom within the network, down to the level of individual relationships. Furthermore, our model accounts for the diversity of content topics generated by users, recognizing that a user's significant influence in one area may not necessarily extend to another. This approach aligns with the idea that the impact on purchasing behavior varies according to product characteristics, as suggested by Schulze et al. (2014) and Park et al. (2018).

Third, this study's identification of social influence on UGC by network edge and content topic makes significant contributions to seeded marketing campaigns (SMCs) and demand management in two distinct ways. First, our proposed method for estimating topic-specific influence enhances the efficiency of firms' SMCs by allowing them to adjust seed targets based on the content topics they aim to disseminate. Different users can serve as seeds depending on the content topic. While previous studies, such as Hinz et al. (2011), have explored various factors for successful SMCs, have largely overlooked the heterogeneity of influence associated with the content being transmitted. By identifying topic-specific seeds, this approach also allows for a more accurate estimation of potential demand for specific products or designs, moving away from the assumption of uniform influences.

Furthermore, our approach highlights the identification of key users who exert significant influence across multiple topics, a phenomenon we term 'topic-spillover influence.' For instance, consider users A and B. If user A initially follows user B for their music content but gradually develops an interest in fashion due to user B's posts on both topics, this spillover from music to fashion illustrates the potential for broader influence beyond the initial topic of interest. Such users with topic-spillover influence are valuable in SMCs, as firms can capture attention on secondary topics as a byproduct of marketing efforts focused on a primary topic, leading to demand spillover. Previous studies on spillover effects in SMCs, like Chae et al. (2017) and Kwark et al. (2021), have concentrated on product-based spillovers - where increased word-of-mouth (WOM) for one product can boost WOM or sales for another. In contrast, this study offers new insights by considering social influence that spills over into other topics when planning seeding strategies in SMCs.

In summary, this study aims to address the identified research gaps by introducing a new variation of the topic model. This model not only represents the content generation process but also integrates the impact of social influence on other users, taking into account the heterogeneity of each UGC topic and network edge. We empirically evaluate the performance of our proposed model against existing models using real data from an image-sharing social media platform.

The remainder of this paper is structured as follows: Section 2 provides a review of the literature relevant to this study. In Section 3, we detail the proposed model. Our empirical findings and their discussion are presented in Section 4. Finally, Section 5 concludes the paper with a summary of our results and proposes directions for future research.

2 Research Background and Positioning

2.1 User Behavior and Social Influence

Since the 1960s, social influence in marketing has been a key research area. The Bass model (Bass 1969) proposed that cautious consumers, or imitators, make purchase decisions based on early adopters' experiences, measuring this effect as the imitation coefficient. Research has also examined the effect of word-of-mouth (WOM) on consumer behavior in various sectors, including television (Godes and Mayzlin 2004) and movies (Liu 2006, Chintagunta et al. 2010). The rise of online shopping and social media since the 2010s has provided more data on individual behavior, prompting studies like Nair et al. (2010) to focus on the individual-level effects of social interactions using non-aggregated data. Early research predominantly aimed to confirm the existence of social influence, often assuming it was uniform across different individuals, regardless of data aggregation.

With the growing recognition of social influence, research has increasingly focused on its heterogeneity, exploring how connections among people affect its strength. Studies like Iyengar et al. (2011), Chen et al. (2017), and Rishika and Ramaprasad (2019) have analyzed behavioral data across multiple networks and different types of ties. Product characteristics as moderators of influence have also been a key focus, with

studies like Zhu and Zhang (2010), Schulze et al. (2014), and Park et al. (2018) showing how product traits like utilitarian value or hedonicity impact influence levels. Wang et al. (2013) expanded this to include consumer attributes, demonstrating how expertise or popularity can affect purchasing influence in different product domains. Despite these varied approaches, Trusov et al. (2010) presented a unique model assuming that each social media connection exerts a distinct influence, highlighting the heterogeneity of social influence across different user relationships. This model encapsulates all potential moderators, including connection types and node attributes, reflecting the complex nature of social influence in networks.

Research into the role of user-generated unstructured content in influencing others has been gaining attention. Archak et al. (2011) developed a model to estimate the influence of word-of-mouth (WOM) on digital camera sales, considering not just the numerical aspects of review texts (like volume, valence, and variance) but also the content, such as product attributes mentioned in the reviews and the reviewers' evaluations. Liu et al. (2019) found that the impact of customer reviews varies depending on the review topics, like aesthetics and conformance, influencing the purchasing decisions of review readers.

Additionally, studies have looked at the influence between social media users. Gong et al. (2017) demonstrated that company posts rebroadcasted by influential users can increase TV show viewership. Trusov et al. (2010), and Rishika and Ramaprasad (2019) further clarified that social media user behaviors, such as log-in activities and preference expressions, are influenced by their social network connections. However, these studies primarily focused on the numerical aspects of targeted behaviors. Despite recognizing the importance of unstructured content in social media, there is still a lack of methodological development in estimating its influence.

The unique aspect of our study is the focus on the influence of user-generated unstructured content in social media, emphasizing the diversity of this influence based on network connections and content topics. Previous studies predominantly concentrated on the social influence on numerical aspects of consumer behaviors. While the research by Trusov et al. (2010) and Liu et al. (2019) also addressed the variability of influence based on network edges and content topics, they only considered one of these factors. Our study, however, integrates both aspects in our proposed model. Additionally, Archak et al. (2011) looked at group-level influence by aggregating review-writing behaviors, whereas our study develops an individual-level behavioral model to identify influential users in social media and determine their specific impact across different content topics.

Table 1 Summary of literature review

Paper	Factors of influence	Outcome variables	Aggregation	Moderators of influence	Model
Bass (1969)	Purchase of product (volume)	Purchase of product (volume)	Aggregated	-	Diffusion model
Godes and Mayzlin (2004)	WOM for TV shows (volume, valence, variance)	Review rating for TV shows (volume)	Aggregated	-	Linear regression
Nair et al. (2010)	Prescriptions by opinion leader (volume)	Prescriptions by physician (volume)	Disaggregated	-	Linear regression
Trusov et al. (2010)	Log-in activity (volume)	Log-in activity	Disaggregated	Edges	Poisson regression
Iyengar et al. (2011)	Prescription (volume, binary)	Adoption of drug (binary)	Disaggregated	Types of network	Discrete-time hazard model
Archak et al. (2011)	WOM for digital cameras (volume, valence, variance, content)	Sales rank (volume)	Aggregated	Interaction of product attributes and their evaluation	Linear regression
Wang et al. (2013)	Choice of products (binary)	Choice of products (binary)	Disaggregated	Popularity, expertise, early adoption, gender, and product characteristics	Logistic regression
Wang et al. (2013)	Choice of products (binary)	Choice of products (binary)	Disaggregated	Popularity, expertise, early adoption, gender, and product characteristics	Logistic regression
Schulze et al. (2014)	Sharing of apps (volume)	Reach of apps in social media (volume)	Aggregated	Sharing mechanism and utilitarian	Linear regression
Chen et al. (2017)	Seed user in diffusion program (binary)	Reach of programs by seed (volume)	Disaggregated	Types of network	Linear regression
Gong et al. (2017)	Retweet by recruited users (binary)	Views of TV shows (volume)	Aggregated	-	Linear regression
Park et al. (2018)	Spending products in games (volume)	Spending (volume) and playing (binary) games	Both	Product characteristics	Linear regression
Ameri et al. (2019)	WOM (volume, valence) and adoption (volume) of anime	Adoption of anime (volume)	Disaggregated	Types of behaviors	Linear-probaility model
Liu et al. (2019)	WOM for products (content)	Purchase of products (volume)	Both	Dimensions of review content	Logistic regression
Rishika and Ramaprasad (2019)	Preference of music (binary)	Playing music (volume)	Disaggregated	Types of connections between users	Propensity score matching
This study	User generated image contents (content)	User generated image contents (content)	Disaggregated	Edges and topics of content	Structural dynamic topic model

2.2 Identifying Seeds and Spillovers

Research has delved into various factors that firms can improve to enhance the success of seeded marketing campaigns (SMCs). Godes and Mayzlin (2009) discovered that effective word-of-mouth (WOM) driving sales often originates from less loyal customers and occurs between acquaintances. They noted that while opinion leaders can spread WOM among very loyal customers, their impact on less loyal ones is limited. Stephen and Lehmann (2016) found that launching campaigns in networks with well-connected consumers can expedite diffusion by encouraging regular consumers to choose them as WOM receivers. Dost et al. (2019) provided insights for marketers on SMCs, emphasizing the need to consider interaction effects with other marketing mixes like advertising and sales promotions, especially when products are similar in design, size, and cost.

Additionally, identifying the right target or 'seed' users for SMCs is crucial. Research has examined various traits of effective target users, including their connectivity (Hinz et al. 2011), knowledge (Bao and Chang 2014, Jansen and Hinz 2022), status difference from followers (Lanz et al. 2019), and strength of connections (Moldovan et al. 2017). However, no previous study has explored how the level of influence varies with the topics of diffused content. Our approach identifies users with significant influence on specific content topics, offering new insights for firms on optimizing their SMCs. By aligning seed users' expertise with the desired content topics, firms can enhance both the effectiveness of their campaigns and the accuracy of demand estimation.

Chae et al. (2017) explored how campaigns on a focal product can inadvertently reduce word-of-mouth (WOM) for another product within the same brand or category. Sanchez et al. (2020) and Kwark et al. (2021) investigated the spillover effects of WOM for a product in online retail and social media on purchasing behaviors for other products. Krijestorac et al. (2020) estimated the spillover effects of firm-generated content on new platforms on the consumption levels of content on earlier platforms, using new platform content to generate WOM.

Despite these insights, discussions on spillover effects in seeded marketing campaigns (SMCs) have been limited, mostly focusing on numerical aspects like the impact of a campaign for one product on another. Our study, however, introduces a novel perspective by identifying 'topic-spillover': the effect of user-generated content (UGC) for one topic on followers' interests in another topic, especially when users exert positive influences across multiple topics. This approach broadens the understanding of spillover effects in SMCs beyond traditional numerical measures.

2.3 Estimating Social Influence in Unstructured UGC

Existing research in social influence has employed various metrics to gauge user behavior. A key metric is the volume of behavior, which includes measures like the number of customer reviews posted (Godes

and Mayzlin 2004), the adoption of products by peers (Bollinger and Gillingham 2012), and the length of review texts (Lu et al. 2018). Another important metric is valence, often related to the positive or negative quality of text. Examples include the average customer review ratings (Moe et al. 2011) and the emotional tone of review texts (Sonnier et al. 2011, Wu et al. 2015). These metrics provide a foundation for assessing the extent and nature of social influence based on user behaviors.

Unstructured user-generated content (UGC) offers more information than what numerical metrics alone can convey. Previous studies focusing on unstructured text data have developed various methods to extract semantic coherence, such as product attributes in review texts. Techniques include clustering (Lee and Bradlow 2011, Archak et al. 2011), ontology-learning (Moon and Kamakura 2017), and classification through machine learning using supervised data (Liu et al. 2019).

Our study, however, introduces a new variant of the topic model, specifically latent Dirichlet allocation (LDA, Blei et al. 2003). This approach, frequently applied in analyzing customer reviews (Tirunillai and Tellis 2014, Büschken and Allenby 2016) and product descriptions (Toubia et al. 2019), is scalable and does not require prior knowledge or classifiers based on supervised data. Moreover, social media encompasses not only text but also images and video content. Our models are designed to handle these varied data types through extensions and integrations within a Bayesian modeling framework. This allows for fitting models to different data formats and considering multiple data sources, including numerical information.

Our study enhances the conventional Latent Dirichlet Allocation (LDA) model by integrating dynamics and hierarchical regression, allowing it to depict how content generation is influenced by friends. LDA typically represents a user's interest in content topics as a topic distribution parameter. Our model, recognizing that these interests dynamically change under the influence of friends' content, extends LDA to capture both the dynamic shifts in topic distribution and the impact of friend-generated content within a hierarchical topic distribution structure.

Blei and Lafferty (2006) introduced a dynamic topic model (DTM), an LDA extension that models the evolution of topic distributions as a state-space model. Glynn et al. (2019) further developed this concept with the dynamic linear topic model (DLTM), which integrates a dynamic linear model to account for systemic changes like seasonality and trends. Additionally, Blei and Lafferty (2005) proposed a correlated topic model (CTM) capable of capturing topic correlations by assuming a normal distribution prior with a covariance structure for topic distribution. Roberts et al. (2016) then presented a structured topic model (STM), incorporating a regression model with the CTM's mean parameter to consider exogenous variables' impact on topic distribution. Building on these advancements, our study introduces a new variant of the topic model with a hierarchical structure. This model incorporates a vector auto-regressive approach, considering both the evolution of topic distribution and the social influence among users on their levels of interest in content topics. It combines the principles of DTM and STM to offer a more nuanced understanding of topic evolution and social influence in content generation.

Table 2 List of notations

Notation	Description
Indices	
u, u'	User
t	Time
i	Element of content (word, object, etc.)
k, k'	Topic
v	Element in a vocabulary set
f	Followee (a user who is followed by the focal user)
f'	Follower (a user who follows the focal user)
Variables	
w_{ut}	A multiset of elements of contents generated by user u at time t
$\mathcal{F}_u$	A set of users who are followed by user u
$\tilde{\mathcal{F}}_u$	A set of users who follow user u
U	The number of users
T	The number of time periods
N_{ut}	The number of elements in w_{ut}
K	The number of topics
V	The number of unique elements in the element vocabulary
N_{utk}	The number of elements to which topic k assigns in w_{ut}
x_{uf}	Explanatory variables in the hierarchical social influence prior
Parameters	
z_{uti}	Topic assignment for w_{uti}
ϕ_k	Element distribution conditioned on topic k
θ_{ut}	Topic distribution of user u at time t
η_{ut}	Topic natural parameter of user u at time t
α_k	Auto-regressive parameter for topic k
β_{ufk}	Lagged social influence of followee f on user u for topic k
γ_{tk}	Time-topic fixed effect
δ_{uk}	User-topic fixed effect
π_{ufk}	Binary indicator of the presence of social influence
ρ_k	Coefficients in the hierarchical social influence prior
ω_{ufk}, ξ_k	Parameters in the hierarchical social influence prior
ζ_{ufk}	Auxiliary variable following a Pólya-Gamma distribution
B_k	Matrix of coefficient containing α_k and β_{ufk}
C_{tk}	Vector of fixed effect parameters

3 Model

In this section, we present the model specification and the hierarchical social influence structure highlighting its sparsity and moderating effects.

3.1 Model Specification

We describe our proposed model that accounts for the dynamic evolution of users' interest levels and the social influence exerted by friend-generated content across various topics of user-generated content (UGC). Although our empirical analysis focuses on image data from a real social media platform, our model is versatile and can be applied to other contexts, such as text (e.g., Tirunillai and Tellis 2014) and purchase history (e.g., Jacobs et al. 2016). This adaptability stems from our model's ability to function independently of content type. Table 2 in this study summarizes the key notations used throughout.

Our observed user-generated content (UGC) data comprises basic unit constructs (such as words in text or objects extracted from images), where the order of these constructs is ignored, but their multiplicity is

preserved. We represent this UGC data as $W = \{w_{ut}\},\ u = 1,\ldots,U,\ t = 1,\ldots,T$, where U and T denote the number of users and time points, respectively. The content generated by user u at time t is denoted by w_{ut}, which is expressed as $w_{ut} = (w_{ut1},\ldots,w_{utN_{ut}})^\top$, with N_{ut} representing the number of constructs by user u at time t. If user u does not generate any content at time t, then w_{ut} is an empty set, i.e., $N_{ut} = 0$.

We also observe the following relationships among users: let $\mathcal{F}_u$ denote the set of users that user u follows, and $\tilde{\mathcal{F}}_u$ denote the set of users who follow user u. For simplicity, we assume that the network remains unchanged during the observation period, omitting the time indicator t for $\mathcal{F}_u$ and $\tilde{\mathcal{F}}_u$. While network changes, such as following new users or removing connections, may occur, these deviations from the static network assumption are minimal and can be disregarded (Ameri et al. 2019).

Given the above setting, we now illustrate the generative process of modeling UGC data. Figure 1 presents a commonly used plate diagram that provides an overview of our proposed model. We adopt the Bayesian framework of conventional topic modeling, which assumes a fixed number of latent topics (K) distributed across multiple documents (i.e., w_{ut}). The proportion of these topics within the content (w_{ut}) generated by user u at time t, is represented by θ_{ut}. This is a $K-1$ dimensional simplex vector that indicates the distribution of topics. The latent topic assignment of the i^{th} element in w_{ut} is determined by a categorical distribution. Given the topic assignment z_{uti}, the corresponding element w_{uti} is then generated from another categorical distribution.

$$z_{uti} \sim Categorical(\theta_{ut}), \qquad w_{uti} \mid z_{uti} = k \sim Categorical(\phi_k) \tag{1}$$

$\phi_k = (\phi_{k1},\ldots,\phi_{kV})^\top$ is an element distribution (i.e., similar to topic-word distribution in latent Dirichlet allocation), which represents the generative probability of elements (i.e., words or objects) in the context of topic k.

Next, we introduce the hierarchical structure of topic distribution when taking into account the dynamic changes in users' interests and the influence of friend-generated content on these interests. Following the existing literature on the extension of topic models (e.g., Blei and Lafferty 2006, Roberts et al. 2016), each dimension of our topic distribution, θ_{utk}, that is nonnegative and is on a simplex vector, is realized through a softmax transformation of a natural parameter, η_{utk}, that follows a normal distribution.

$$\theta_{utk} = \frac{\exp(\eta_{utk})}{\sum_{k'}\exp(\eta_{utk'})}, \qquad \eta_{utk} \sim N(\lambda_{utk}, 1) \tag{2}$$

The mean parameter, λ_{utk}, is obtained through a linear regression model (Roberts et al. 2016), which considers an autoregressive component, lagged social influence from the network, and time-topic and user-topic fixed effects.

$$\lambda_{utk} = \alpha_k \cdot \eta_{ut-1k} + \sum_{f\in\mathcal{F}_u} \beta_{ufk} \cdot \eta_{ft-1k} + \gamma_{tk} + \delta_{uk} \tag{3}$$

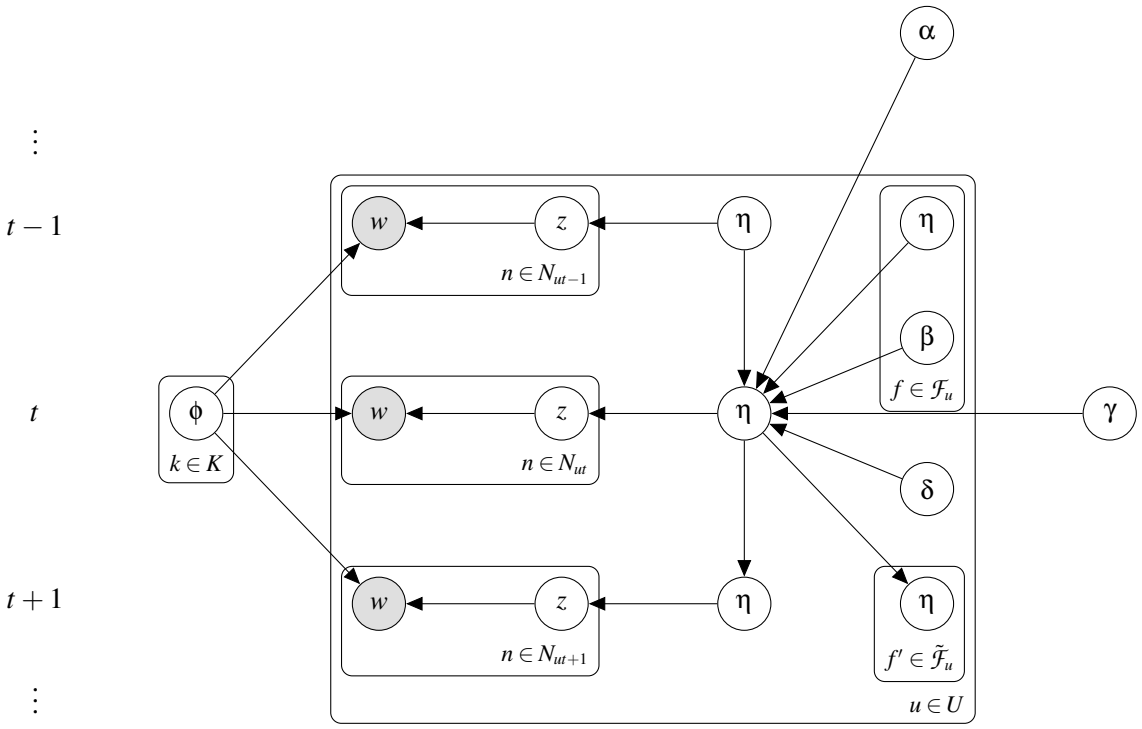

Figure 1 A plate diagram representing our proposed model

In our context, β_{ufk} represents the lagged social influence of friend-generated content on users' interest levels in various content topics. This influence varies depending on the specific topic and the network connection, drawing from previous studies that explore the moderating role of topics on consumer behavior (e.g., Liu et al. 2019). The parameters γ_{tk} and δ_{uk} are time-topic and user-topic fixed effects, respectively, which control for common shocks at a specific time and each user's intrinsic preference for a certain topic. These fixed effects help separates the impacts of homophily and correlated unobservables from social influence. The contribution of β_{ufk} and the two fixed effects to the model's identifiability is described below.

Previous research has highlighted key challenges in estimating social influence from behavioral data within social networks, as noted by Manski (1993) and Hartmann et al. (2008). The primary issues include endogenous group formation, correlated unobservables, and simultaneity. Endogenous group formation, or homophily, poses a problem because individuals with similar preferences tend to form social groups, potentially confounding the estimation of social influence. Various methods have been developed to address this, including the use of individual user-level fixed or random effects (Nair et al. 2010) and more complex heterogeneity specifications (Hartmann 2010).

Additionally, variables that simultaneously affect multiple users, such as advertising, seasonality, and trends, can introduce omitted variable bias if not properly accounted for in the model. These variables are examples of correlated unobservables. In our study, we mitigate biases from both correlated unobservables and homophily by incorporating user- and week-specific fixed effects. This approach enhances the accuracy of our estimation of social influence within social networks.

The simultaneity problem in social network research occurs when the behavior of a focal user and their network peers influence each other simultaneously. This mutual influence can create a significant correlation between explanatory variables and error terms, leading to biased coefficient estimates. To tackle this issue, methods such as instrumental variables (Nair et al. 2010) and equilibrium modeling (Hartmann 2010) have been widely adopted. In our study, we address the simultaneity problem by employing a lagged variable approach to describe friends' behaviors, similar to the method used in Ameri et al. (2019). This approach is based on the assumption that a user's behavior does not immediately reflect the influence of observing a friend's content. Instead, it posits that a significant accumulation of influence over time leads to changes in the user's level of interest in the content topic. By using lagged variables, we aim to more accurately capture the delayed impact of social influence on users' interests and behaviors.

3.2 Hierarchical Social Influence

In real social networks, a common phenomenon is the sparsity of social influence, raising the question of how to model this in our context. Previous studies, such as Karrer and Newman (2011), have considered node degree heterogeneity, where a few nodes have many connections while the majority have only a few. In the context of social influence, this suggests that most users are influenced by only a small number of friends, resulting in many network edges having negligible influence. This concept is reflected in the model by Trusov et al. (2010), which decomposes social influence among users into a continuous susceptibility parameter and a binary indicator. The binary indicator reflects whether a user distinctly influences their friends' behaviors, consistent with the sparsity assumption of social influence.

Furthermore, extensive research on social influence indicates that the strength of social influence can be moderated by various user- or network-level characteristics, such as the connectedness in a users' social network (Van den Bulte and Wuyts 2007, Iyengar et al. 2011) and the user's loyalty status (Viswanathan et al. 2017). These studies suggest that the extent of social influence one receives from other users in a network varies depending on both the individual's characteristics and those of others. However, given the sparse distribution of social influence within the network, this structure does not exist uniformly across all user pairs.

Building on these separate lines of research, we introduce a hierarchical mixture prior for the parameters of social influence to capture both the sparsity and the moderated effects. Specifically, we employ a Bayesian lasso prior (Park and Casella 2008) for representing its sparsity and a traditional linear regression model to account for the moderating factors. Our model assumes that only a few elements of social influence have significant non-zero values, and those that do are influenced by moderating effects from user- or network-level characteristics. Each element of social influence parameter, β_{ufk}, follows a hierarchical mixture prior.

$$\beta_{ufk} \sim \begin{cases} N(0, \sigma_{\beta}^2 \cdot \omega_{ufk}), & \text{if } \pi_{ufk} = 0 \\ N(x_{uf}^{\top} \rho_k, \sigma_{\beta}^2), & \text{if } \pi_{ufk} = 1 \end{cases}, \qquad \pi_{ufk} \sim Bernoulli(p_{\tau}), \qquad p_{\tau} \sim Beta(5,1)$$

$$\omega_{ufk} \sim Exp\left(\frac{\xi_k^2}{2}\right), \qquad \xi_k^2 \sim Gamma(1,1), \qquad \sigma_\beta^2 \sim IG(.001,.001), \qquad \rho_k \sim N(0, 10^2 \cdot I) \tag{4}$$

where π_{ufk} is a binary indicator of the presence of significant non-zero social influence, and if $\tau_{ufk} = 1$, then user f is expected to affect user u's interest in topic k, while there is no significant influence between them (i.e., assumed to follow the shrinkage prior of Bayesian lasso) otherwise. x_{uf} is a vector of explanatory variable containing user- or network-level characteristics. In this study, we observe the number of followees and followers and the proportion of the active week in the dataset of both user u and f. These variables are expected to be proxy variables for the connectivity in the users' network and the loyalty toward the social media platform, respectively. The settings of prior distribution for the remaining parameters including other parameters in Equation 3 are defined in EC.3.

4 Empirical Analysis

4.1 Data

We evaluate our proposed model using real data from Pinterest.com, an image-sharing social networking site where users post images from the internet or their own creations to share ideas and inspirations. On Pinterest, user-generated images are termed 'Pins,' and the act of creating content is referred to as 'Pinning.' Similar to other popular social media platforms like X (formerly Twitter) and Meta (formerly Facebook), Pinterest allows users to follow others. Users can view recent pins from those they follow on their feed page[1]. This setup suggests that attractive pins displayed on a user's feed may influence their interests in certain topics, indicating the presence of social influence among image content within Pinterest.

Figure 2 in our dataset illustrates the timelines of pins saved by two users. Initially, User A (followed by User B) showed interest in sweets and gifts, while User B was inclined towards gardening and interior design. However, following a period (weeks 70 to 80) where User A pinned sweets-related images, User B's interest in that topic seemed to increase, evidenced by her subsequent pinning of similar content, like cookie and cake recipes, after week 80. While the lack of web access log data in our study prevents us from confirming causality or whether User B actually viewed User A's pins, this lagged positive correlation in their pinned content suggests a potential positive social influence of User A on User B regarding sweets.

In this study, we collected[2] user-generated images and their follower network information from Pinterest. The target users were first selected in a snowball sampling fashion, and then images pinned by these users for 2 years (104 weeks) from December 1, 2017 were downloaded via the Pinterest API[3]. In our subsequent empirical analysis, images in first 100 weeks are used as in-sample data, and the hold-out sample comprises

[1] This feature of showing pins by the followees on the feed page has since been discontinued. Currently, pins, topics, and boards from followees are shown in the home feed tab (see, e.g., https://help.pinterest.com/en/article/following-and-followers, accessed January 8, 2024.).

[2] We collected data on December 1, 2019.

[3] https://developers.pinterest.com/

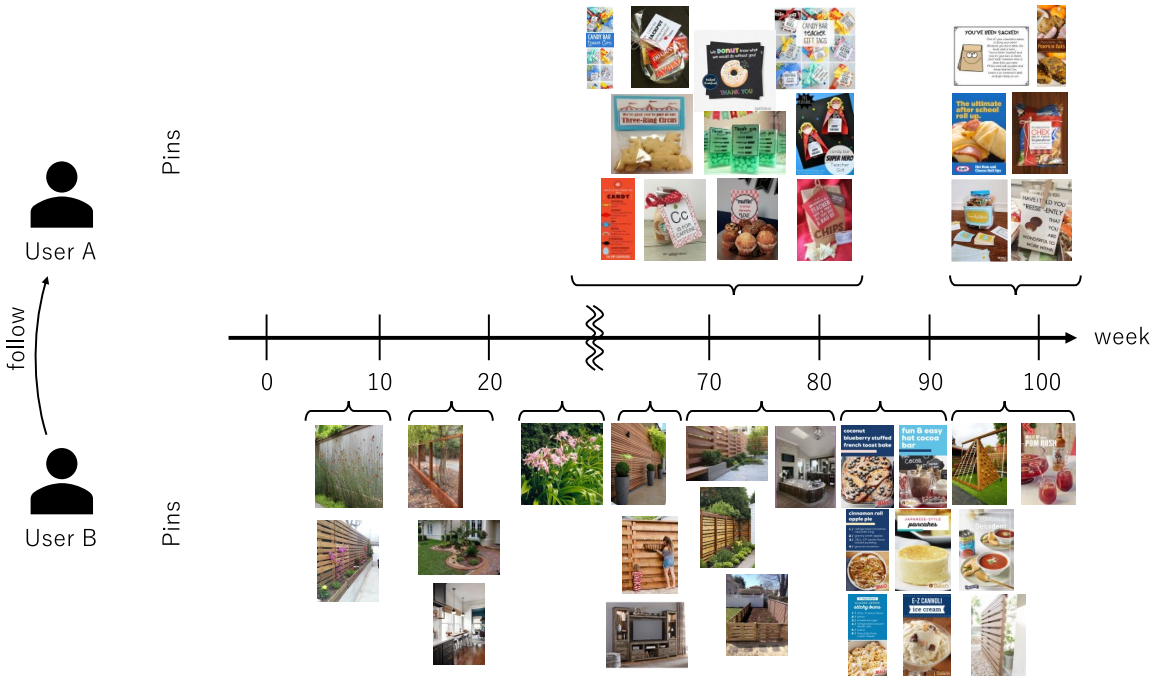

Figure 2 An illustration example of the timeline of pins saved by two users in our dataset

images in the last 4 weeks. To properly evaluate the performance of various models on this dataset, only users who have pinned in both in-sample and hold-out periods were selected. Note that we only keep users who followed at least one other user or were followed by at least one other user in the dataset. As a result, our final dataset contains 801 unique users, 125,691 pinned images, and 892 following relationships between users (i.e., the density of the network is 0.14%). In this study, we use this middle-size network data, but in EC.3, we investigate the computation time for model inference when applying it to larger networks through the simulation experiment.

To make the image data amenable for analysis using topic modeling, we used the object detection by Google Vision API[4] to extract embedded objects in images. Figure 3 shows such an example. Within the image, there are several objects, such as a bicycle, doors, and a picture frame. As a result, this image will be represented as a set of unordered textual objects, {bicycle, bicycle wheel, door, picture frame, ...}. This process is applied to all pinned images, and the size of all unique objects in the dataset is 9,019. This image data preprocessing is equivalent to word splitting when applying topic models to text data. We also note that some amount of information within the image might be lost because we ignore the proportion objects occupy an image if objects within an image are treated as a set of textual objects. To mitigate this, we simply use another API for calculating the coverage proportion of each object, or we can apply advanced image processing and representation techniques, such as convolutional neural networks (e.g., Liu et al. 2020). But

[4] https://cloud.google.com/vision/docs/object-localizer

Figure 3 An example of object detection by Google Vision

Table 3 Descriptive statistics of our dataset

	Mean	Median	SD	Minimum	Maximum
Number of pinned images in total	156.92	20	747.44	2	9,756
Number of pinned images per week	1.51	0	11.25	0	446
Number of followees	1.11	1	1.89	0	29
Number of followers	1.11	0	6.81	0	111
Number of objects in a image	8.83	10	2.22	1	12

it is beyond our study and we leave it for future work. Table 3 provides the descriptive statistics of the dataset.

4.2 Model Comparison

In this section, we perform empirical evaluations to demonstrate the effectiveness of our model in terms of performance and ablation study.

Baselines: We compare our proposed model with three major baseline methods: latent Dirichlet allocation (LDA), dynamic topic model (DTM), and structured topic model (STM).

- LDA is a widely used benchmark approach to analyze text data. It is simple compared to the other two models because it does not assume either a state space model or some common structure behind the variations of the topic distributions. The major difference between LDA and our proposed model is in the prior of topic distributions in Equation 2. Specifically, LDA assumes the prior of topic distributions simply follows a Dirichlet distribution, $\theta_{ut} \sim Dirichlet(\theta_0)$ while ours follows a logistic-normal distribution with the mean parameter from a regression.
- DTM takes into account the dynamic evolution of the topic distribution as a state space model. The difference from ours is in the hierarchical structure in Equation 2 and 3, $\eta_{utk} \sim N(\alpha_{tk}, \delta^2)$, $\alpha_{tk} \sim N(\alpha_{t-1k}, a^2)$. That is, our proposed model extends DTM by considering influence from the surrounding network (i.e., social influence), including self-loops.

- STM considers the effects of common factors on the variations of the topic distributions by extending independent prior of a topic distribution in the LDA with a logistic-normal distribution and regression. In this study, common factors affecting the variations of the topic distribution represent images actually pinned by the user and followees of the user. Hence, the hierarchical structure in Equation 3 can be rewritten as $\lambda_{utk} = x_{ut-1}^{\top}\alpha_k + \sum_{f\in\mathcal{F}_u} x_{ft-1}^{\top}\beta_{ufk} + \gamma_{tk} + \delta_{uk}$, where x_{ut} represents vectorized images[5] pinned by user u at week t.

Deep Learning Model: In light of recent development of the machine learning community, it is important to compare the proposed topic model with deep learning models. Among the rapidly developing deep learning models, Dieng et al. (2020) propose embedded topic model (ETM) introducing the philosophy of embedded representation such as Word2vec (Mikolov et al. 2013) into topic models. In this model comparison, we apply ETM for the process of word generation, leaving the rest of our model as it is. In ETM model, because topic embedding vector and object embedding vector are introduced, the process of word generation in Equation 1 is changed into $w_{uti} \mid z_{uti} = k \sim Categorical(\text{softmax}(\rho^{\top}\alpha_k))$, where ρ and α_k are embedding vectors for object and topic k, respectively. We use the pre-trained embedding vector for ρ which was obtained by BERT, while the topic embedding vectors are estimated by MCMC as well as other parameters.

Model Ignoring Moderator on Social Influence: In addition, we also conduct ablation studies to understand the importance of moderating effects of user- and network-level characteristics on social influence. To do so, we intentionally created a variants of our proposed model assuming shrinkage prior for all pair of users and topics. In this model, the hierarchical structure in Equation 4 becomes $\beta_{ufk} \sim N(0, \sigma_{\beta}^2 \cdot \omega_{ufk})$, i.e., π_{ufk} is always zero which means the social influence parameter follows Bayesian lasso prior. We call this model BL in the following.

For performance evaluation, we use two standard metrics: perplexity and topic coherence. Perplexity is a well-known measure used in the evaluation of topic models to understand if the model predicts hold-out samples well. It is calculated based on the log density of the model for the hold-out data. The lower the value of perplexity, the better the predictive performance. Let H be the set of indices of users, time, and objects in the hold-out data, the definition of perplexity is given as follows.

$$Perplexity = \exp\left(-\frac{\sum_{(u,t,n)\in H}\log p(w_{uti})}{N^{out}}\right), \qquad p(w_{uti}) = \sum_{k=1}^{K}\hat{\theta}_{utk}\cdot\hat{\phi}_{kw_{uti}}, \tag{5}$$

where N^{out} denotes the total number of objects in the hold-out data. However, focusing only on the predictive performance of models neglects how well the model defines the topics along dimensions that are easy to

[5] The process of vectorization is as follows: first, we create an image-object matrix that stores the frequency of each object in the image. Next, we apply fast-greedy algorithm (Clauset et al. 2004) to the matrix to obtain clusters where each object is belonging into. Then, we summarize clustered objects into x_{ut} which is a vector representing the number of objects assigned to each cluster in the image content pinned by user u at week t.

understand for humans, which is usually measured by coherence. Thus, we use the topic coherence to show the interpretability of the obtained topics. The higher the coherence, the better the interpretability of topics. The coherence of the k^{th} topic obtained from the model is defined as follows.

$$Coherence(k; V^{(k)}) = \sum_{m=2}^{M} \sum_{n=1}^{m-1} \log \frac{C(v_m^{(k)}, v_n^{(k)}) + 1}{C(v_n^{(k)})}, \tag{6}$$

where $V^{(k)} = \{v_1^{(k)}, \ldots, v_M^{(k)}\}$ is the top M objects (i.e., $M = 10$ in this study) with the highest values in the estimated element distribution, $\hat{\phi}_k$. $C(v)$ denotes the content frequency, i.e., the number of content sets with at least one object v pinned by a given user at a particular time in the in-sample data, and $C(v, v')$ denotes the co-content frequency, i.e., the number of content sets containing both object v and v'. We use the average topic coherence across all topics to compare how well the model extracts interpretable topics.

We now report the performance for our proposed model and comparison models in terms of perplexity and topic coherence in Table 4 and 5. Note that perplexity is calculated using hold-out data while topic coherence is derived from in-sample data. From the tables, we have the following observations: (1) Our proposed model outperforms all comparison models for the best number of topics ($K = 14$) for the perplexity. Notably, our proposed model improves the perplexity by 66.1% and 29.1% over LDA and DTM with the optimal number of topics, respectively, which are extended by our model to take into account the influence from the first-order connections and self-loops in the network. (2) In the comparison based on coherence, which represents the interpretability of topics, the proposed model (as well as other comparison models) performed better than the deep learning model ETM. However, it is important to note that this result is based on the dataset used in this study and does not generally indicate that the proposed model is superior to ETM in terms of interpretability. The ETM used in this study is a single-layer neural network model, and it is noteworthy that it can be extended to a more expressive model by stacking multiple layers or incorporating convolutional models to handle images more appropriately.

Perplexity and coherence are metrics used to evaluate the predictive and interpretive performance of topic models, and they are essential for statistical comparison between models. On the other hand, they do not directly measure model characteristics that are applicable to practical marketing outcomes. Hence in addition to the model fit shown above, we compute the hit rate for the holdout data to further demonstrate the superiority of our model over baselines, specifically the prediction power. Since all benchmark models including our proposed one are kind of topic models, the predictive density $p(w_{uti})$ that the i^{th} object will be included in images pinned by user u in week t is given in Equation 5. The hit rate is determined by calculating the proportion of the top N predictions that match the actual images in the holdout data. Figure 4 presents hit rates for three different holdout periods: the first week, the first two weeks, and the entire first month. We vary N from 5 to 20 for these calculations. It's important to note that the hit rate is based on the 'best' model as determined by its perplexity measure, and we focus on models with a dynamic structure that

can predict the topic distribution per user (i.e., ETM, BL, and Proposed) to make the comparison clear[6]. The results show that the proposed model demonstrates competitive performance with the best models in every holdout periods. This outcome aligns with the perplexity findings, further reinforcing that our proposed model not only enhances predictive performance over conventional topic models but also offers valuable insights into the social influence exerted by each connection and across various topics.

[6] It is worth noting that other models, which are omitted here, have been confirmed by the authors to exhibit lower hit rates compared to these models.

Table 4 Performance comparison across models regarding the perplexity

	Number of topics								
model	2	3	4	5	6	7	8	9	10
LDA	1861.16	1670.03	1564.15	1463.01	1402.29	1366.99	1342.53	1306.27	1289.22
DTM	546.45	571.12	561.68	556.80	554.94	553.85	558.98	555.39	555.11
STM	485.92	444.08	434.17	409.58	405.91	404.47	409.36	403.62	405.48
ETM	740.92	625.49	509.51	408.48	402.65	431.56	497.86	631.67	498.81
BL	570.78	472.49	439.99	400.69	403.83	416.45	422.77	424.21	419.47
Proposed	558.78	452.32	432.86	410.34	509.50	389.18	426.14	431.06	400.67

	Number of topics									
model	11	12	13	14	15	16	17	18	19	20
LDA	1262.89	1248.53	1213.01	1223.13	1191.31	1178.69	1190.49	1158.19	1168.14	1142.07
DTM	554.85	556.09	555.60	556.97	550.52	549.95	552.43	550.33	555.95	551.79
STM	406.29	396.95	401.16	402.03	402.85	399.94	402.69	403.78	391.73	398.98
ETM	707.80	529.45	608.95	569.50	615.58	757.00	630.78	647.62	708.81	632.03
BL	444.61	400.45	430.19	394.65	396.79	411.74	389.09	486.78	441.10	421.46
Proposed	427.68	390.40	485.88	**387.57**	417.83	487.61	575.34	467.14	454.47	532.61

Table 5 Performance comparison across models regarding the topic coherence

	Number of topics								
model	2	3	4	5	6	7	8	9	10
LDA	-23.12	-28.30	-28.04	-32.39	-31.75	-38.74	-36.76	-35.01	-36.95
DTM	-23.12	-28.73	-29.19	-32.82	-37.45	-35.98	-39.51	-37.98	-38.30
STM	-23.12	-28.30	-28.04	-32.01	-35.57	-39.60	-39.19	-38.54	-36.61
ETM	-25.95	-29.14	-36.63	-31.84	-36.34	-39.99	-68.87	-73.83	-61.71
BL	-23.12	-28.30	-29.05	-31.98	-34.54	-30.35	-38.23	-35.14	-38.30
Proposed	**-23.12**	-28.30	-29.05	-31.98	-31.01	-31.71	-36.37	-37.32	-39.47

	Number of topics									
model	11	12	13	14	15	16	17	18	19	20
LDA	-36.45	-37.16	-36.75	-38.42	-40.07	-40.96	-39.60	-38.34	-42.42	-39.05
DTM	-40.49	-40.93	-39.04	-37.57	-39.66	-40.55	-44.23	-43.57	-42.07	-43.87
STM	-36.39	-39.18	-35.46	-38.27	-37.67	-39.97	-38.70	-41.17	-40.10	-40.75
ETM	-64.84	-64.52	-53.85	-69.74	-59.77	-70.66	-82.59	-81.66	-80.29	-81.03
BL	-36.57	-38.28	-35.28	-39.36	-38.99	-40.54	-37.80	-39.80	-41.15	-42.22
Proposed	-39.37	-36.55	-39.97	-39.58	-41.74	-41.13	-42.68	-38.82	-38.54	-40.09

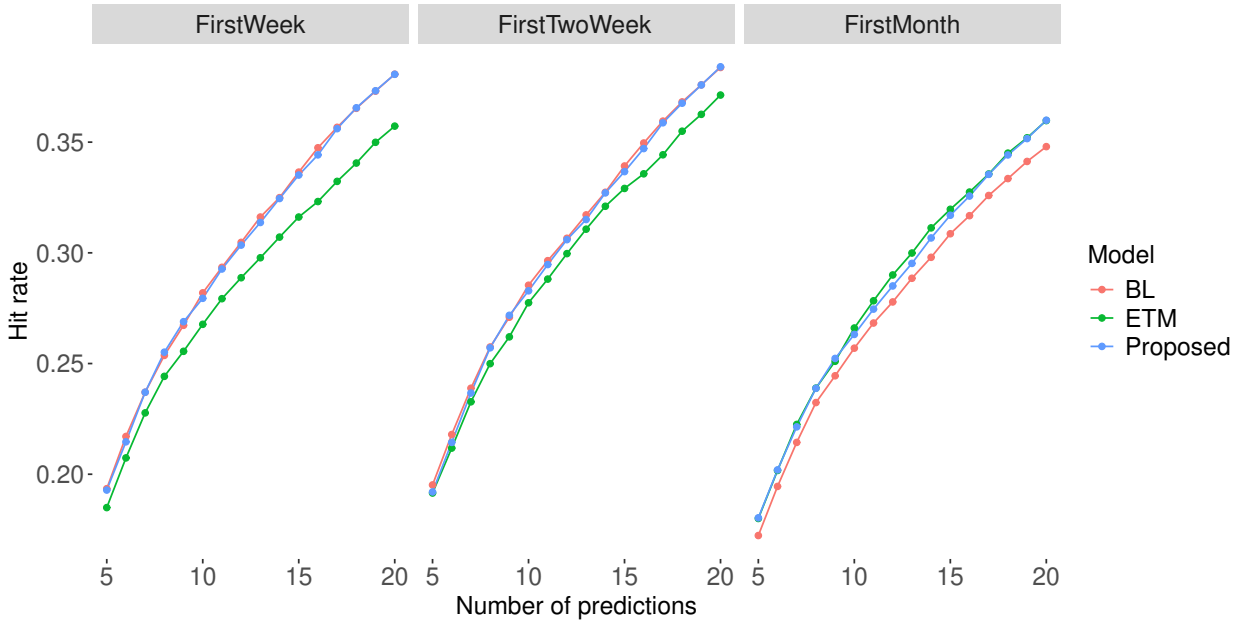

Figure 4 Performance comparison across models regarding the hit rate

4.3 Discussion

4.3.1 Interpretation of topics

In this section, we analyze the estimation results of our proposed model when applied to Pinterest data and discuss their potential implications. For all subsequent analyses, we have set the number of topics to 14, a decision informed by the perplexity measure discussed earlier. Table 6 presents the top 20 objects for each topic. The top objects are identified using the FREX score (Bischof and Airoldi 2012), which balances term frequency and exclusivity to that topic, with the weight parameter set to 0.5 to ensure equal influence from both factors. Based on these top objects, we have labeled each topic with terms such as 'Cake,' 'Animal,' 'Christmas,' and others, as indicated in the table head. Furthermore, these labelled topics exhibit a hierarchical structure, with related topics forming coherent clusters. We have summarized them into four main clusters: Food, Nature, Interior, and Fashion.

Some topics in our study are less interpretable, likely due to the object detection algorithm's uniformly identification of all text in images as simply 'Text.' This inability to distinguish between different types of text content within images is a limitation of our study. Additionally, the model with 14 topics - selected based on the perplexity for this experiment - exhibited the worse interpretability than the best model selected based on coherence. Balancing the predictive performance of topic models with their interpretability is a well-known challenge. Therefore, enhancing the model's interpretability by implementing strategies such as setting certain data as seed words (Toubia et al. 2019) could offer a practical and valuable improvement. Despite these challenges, most topics are clearly identifiable by their labels, minimizing potential issues when using the estimated topics in subsequent analyses.

Table 6 Top 20 objects for each topic

	Food			Nature			
	Topic 1 Cake	Topic 5 Breakfast	Topic 10 Cuisine	Topic 2 Animal	Topic 8 Flower	Topic 11 Kids	Topic 14 Still Life
1	Dessert	Food	Ingredient	Cat	Flower	Font	Still life photography
2	Food	Cuisine	Dish	Carnivore	Plant	Text	Drink
3	Cuisine	Dish	Cuisine	Felidae	Cut flowers	Child	Still life
4	Dish	Ingredient	Produce	Small to medium-sized cats	Flowerpot	Adaptation	Tableware
5	Baked goods	Meal	Food	Whiskers	Floristry	Play	Alcoholic beverage
6	Ingredient	Recipe	Recipe	Mammal	Flowering plant	Line	Distilled beverage
7	Cake	Produce	Meat	Vertebrate	Floral design	Paper	Serveware
8	Buttercream	Dessert	Vegetable	Text	Bouquet	Toddler	Cup
9	Cream	Brunch	Comfort food	Kitten	Flower Arranging	Happy	Liqueur
10	Icing	Snack	Vegetarian food	Font	Herb	Paper product	Dishware
11	Chocolate	Comfort food	Staple food	Photo caption	Houseplant	Advertising	Cocktail
12	Recipe	Baking	Chicken meat	Tabby cat	Garden	Photography	Plate
13	Produce	Baked goods	Salad	European shorthair	Leaf	Photo caption	Stemware
14	Chocolate cake	Breakfast	Meal	Norwegian forest cat	Yard	Logo	Ceramic
15	Frozen dessert	Finger food	Fried food	Dog breed	Botany	Book cover	Alcohol
16	Cake decorating	Lunch	Soup	Canidae	Petal	Graphic design	Classic cocktail
17	Carrot cake	Superfood	Finger food	Domestic short-haired cat	Grass	Brand	Champagne stemware
18	Chocolate brownie	Party	Side dish	Dog	Wildflower	Document	Wine glass
19	Cheesecake	Cookies and crackers	Curry	Snout	Spring	Poster	Bowl
20	Sponge cake	Bake sale	Leaf vegetable	Domestic long-haired cat	Shrub	Games	Porcelain

	Interior			Fashion			
	Topic 3 Christmas	Topic 6 Furniture	Topic 9 Interior Design	Topic 4 Accessory	Topic 7 Fashion	Topic 12 Makeup	Topic 13 Nail Art
1	Pattern	Furniture	Property	Jewellery	Clothing	Hair	Natural landscape
2	Christmas decoration	Table	Interior design	Gemstone	Fashion	Hairstyle	Sky
3	Textile	Room	Room	Fashion accessory	Dress	Chin	Nature
4	Crochet	Interior design	Floor	Body jewelry	Outerwear	Face	Nail care
5	Design	Product	Building	Silver	Shoulder	Eyebrow	Manicure
6	Christmas tree	Shelf	Furniture	Ring	Footwear	Beauty	Nail polish
7	Christmas ornament	Wall	Ceiling	Metal	Sleeve	Skin	Nail
8	Christmas	Chair	House	Diamond	Fashion model	Lip	Art
9	Ornament	Living room	Tile	Turquoise	Neck	Forehead	Finger
10	Wool	Couch	Wall	Engagement ring	Street fashion	Nose	Sea
11	Needlework	Bed	Home	Necklace	Waist	Long hair	Hand
12	Art	Bedding	Cabinetry	Pendant	Jeans	Blond	Painting
13	Craft	Shelving	Kitchen	Earrings	Leg	Brown hair	Illustration
14	Knitting	Bedroom	Countertop	Crystal	Shoe	Hair coloring	City
15	Plant	Coffee table	Architecture	Jewelry making	Beige	Black hair	Drawing
16	Tree	Wood	Bathroom	Bracelet	Gown	Cheek	Service
17	Fashion accessory	Bed sheet	Living room	Chain	Cocktail dress	Muscle	Cosmetics
18	Thread	Desk	Real estate	Emerald	Joint	Head	Water
19	Pumpkin	studio couch	Wood flooring	Gold	Denim	Layered hair	Landmark
20	Embroidery	Rectangle	Flooring	Wedding ring	Jacket	Photography	Atmospheric phenomenon

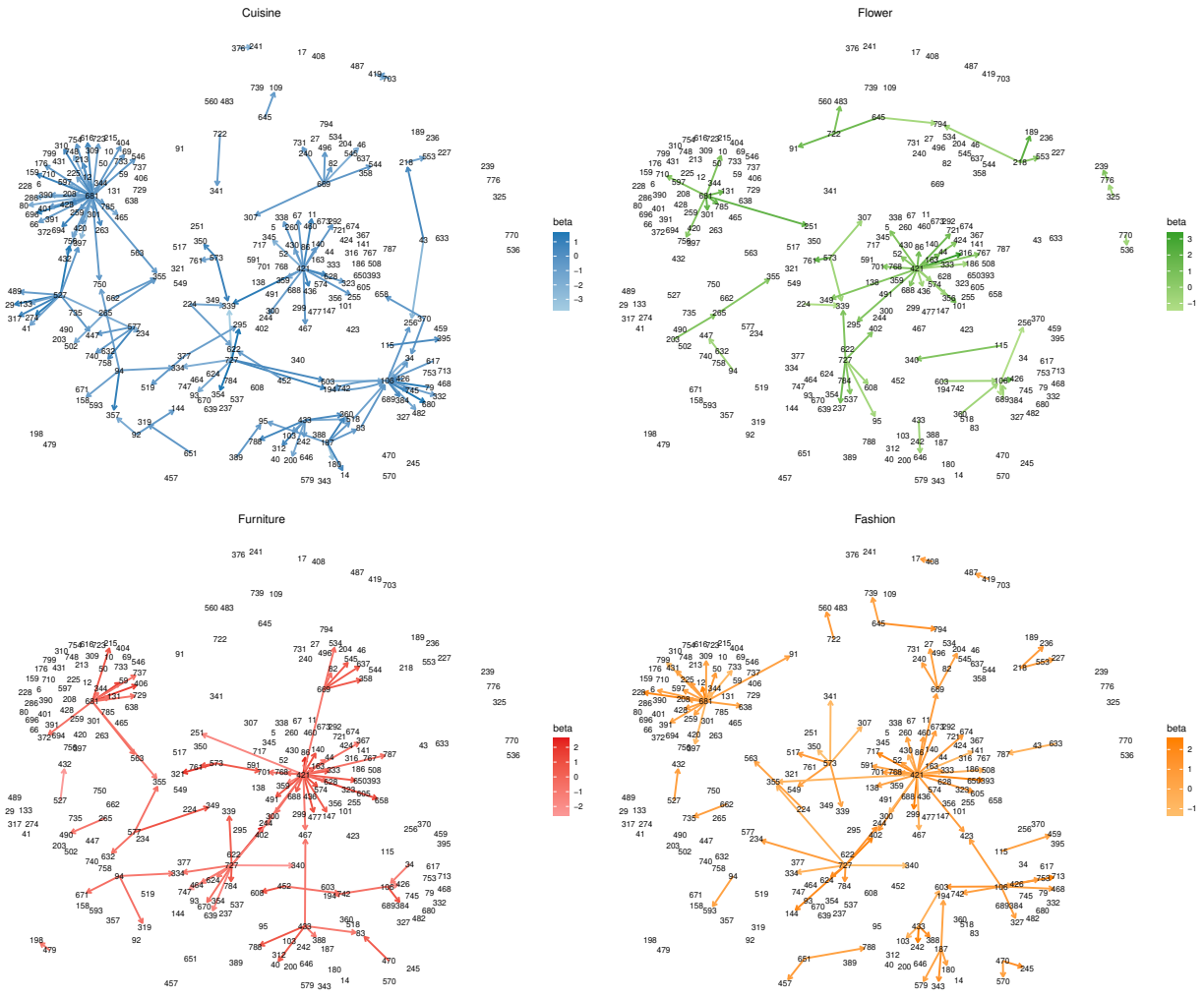

Figure 5 Visualization of Pinterest user-user following network characterized by the social influence under four different topics

4.3.2 Networks with social influences

Next, we visualize social networks using the estimated social influence scores, $\hat{\beta}_{ufk}$. Figure 5 illustrates the networks among Pinterest users, highlighting the top 10% of these scores for each topic. Doe to space limitations, we focus representative topics for each topic cluster: 'Cuisine', 'Flower', 'Furniture', and 'Fashion', corresponding to the clusters of Food, Nature, Interior, and Fashion, respectively. User IDs are denoted by numbers, with arrows representing the directional influence from one user to another. The color intensity of each arrow corresponds to the degree of influence (i.e., the value of estimates), as detailed in Equation 3.

Our model defines social influence as the impact of users' topic proportion in their posted content on the topic proportion of another user who follows them. Therefore, a positive influence implies that the user likely to post content with topics similar to those of followees. The networks illustrated in Figure 5 reveal that such social influences vary not only across user pairs but also across topics. For instance, user 421, an influencer with 107 followers in our dataset and over 4.5K in reality, significantly raised followers' interest

in ‘Fashion’ compared to other topics because she has many strong influence on her followers for that topic. Such insights are highly practical for businesses. Firms can use our model to identify effective seed users for specific topics in campaigns. The level of interest shown by followees can help gauge the demand for such topics/content/products helping forecasting and demand management. Moreover, the influences of the same user on multiple topics were also estimated, such as users 86 and 688 in ‘Flower’, and ‘Furniture’. This suggests the possibility of topic-spillover. Therefore, selecting user 421 as the seed user in a campaign could also heighten users’ interests in topics beyond the campaign’s primary focus.

However, the network in the figure also reveals many negative social influence. For the interpretation of negative influence, consider a scenario where user B follows user A, and user A has negative social influence on user B regarding the cuisine topic. This indicates that when user A posts some content related to cuisine, user B’s posts on the same topic decreases relative to other topics. This phenomenon could imply that user B feels unfavorable sentiments toward user A or her created content. However, a more compelling interpretation is that it represents a strategic behavior aimed at maintaining uniqueness or differentiating oneself from other users within the community. Several prior studies have observed such behaviors on social media where users gradually diverge from the content posted by their connected peers in the network (e.g., Jiang et al. 2024, Zeng and Wei 2013). Similarly, in this study, negative social influence is understood to reflect such strategic behaviors on social media platforms. Furthermore, the proposed model successfully captures social influence, which can be either positive or negative, at both the user-pair and topic-specific levels.

This study proposes a method for estimating social influence on a topic-by-topic basis for the content generated by users. In contrast, traditional seed marketing campaigns often assign ‘static labels’ to influencers based on the typical topics of their posts. For instance, an influencer who frequently posts about makeup and cosmetic products to a young audience might be labeled with tags such as ‘makeup,’ ‘beauty,’ and ’youth’. However, the topics an influencer posts about and the topics on which they exert significant influence do not necessarily align. Our proposed model accounts for the impact of an influencer’s posts on the interests of the surrounding users, allowing for a more accurate estimation of topic-specific influence. We believe that by considering these estimated influences, campaigns can be designed to be more effective, potentially enhancing the overall impact of traditional seed marketing strategies[7].

4.3.3 User importance index

Recognizing the significance and practicality of selecting seed users, and their identification through our model, we proceed to quantify this aspect. For this purpose, we introduce a user importance index based

[7] This paragraph incorporates valuable feedback from anonymous reviewers. We sincerely appreciate their constructive comments, which significantly improve this work.

on each topic's social influence and network connection information. Two basic network metrics for assessing node importance are degree centrality and strength[8]. Degree centrality indicates the number of direct connections (i.e., neighbors) a user has within the network. While it can be defined as both indegree and outdegree in a directed network, we focus on indegree in this study, i.e., the number of followers a user has, to gauge a user's influence on their network.

Strength is the average weight of the edges connected to a node, indicating the extent of influence a node has over its followers. In a directed network, strength can be calculated for each direction; here, we consider in-strength. To integrate both degree and strength in measuring user importance for seed selection, we introduce an importance index for user u in topic k, drawing upon prior literature (Opsahl et al. 2010).

$$\text{importance}_{uk} = \begin{cases} d_u^{(1-\rho)} \cdot s_{uk}^{\rho} & s_{uk} \geq 0 \text{ or } \rho = 0 \\ -d_u^{(1-\rho)} \cdot |s_{uk}|^{\rho} & \text{otherwise,} \end{cases} \tag{7}$$

where d_u is an indegree of user u, and $s_{uk} = \sum_{f \in \tilde{\mathcal{F}}_u} \beta_{fuk}$ is an in-strength of user u for topic k. ρ is a tuning parameter between 0 and 1 and can be predefined according to data and applications. For example, when $\rho = 0$, the importance index considers only degree, and when $\rho = 1$, it only has strength. The value of ρ plays a critical role in balancing degree and strength when measuring user importance. For instance, in campaigns where the key performance index is the number of page views of diffused content, degree should be weighted more heavily. This approach aligns with the concept that well-connected users can rapidly disseminate content in a word-of-mouth campaign (Hinz et al. 2011). However, research like Watts and Dodds (2007) indicates that targeting well-connected individuals does not always significantly impact influence cascades. Conversely, focusing on the average influences means that while high impressions of the content might not be immediately achievable, the content could still be gradually disseminated to areas where topic-specific influence is strong and steady. Moreover, leveraging topic-spillover influence can be an effective strategy for spreading viral content. It allows for capturing attention beyond the primary topic of the campaign, potentially with minimal additional investment.

Table 7 lists the top 10 important users, along with their importance indices for specific topics at given values of the tuning parameter ρ. It is noteworthy that when the index is based solely on indegree (i.e., $\rho = 0$), the importance value remains consistent across all topics. However, as ρ increases, the importance of certain users may shift, reflecting their social influence on specific topics. For instance, user 421, who has many followers and a strong influence on the topic of 'Fashion', maintains a high importance level when ρ is close to 0. However, as ρ approaches 1, her importance declines, suggesting that she has both positive and negative influences on her followers, affecting their interests in this topic. On the other hand,

[8] There are many other sociometric measures of node importance in social networks, such as closeness centrality and betweenness centrality, but since calculating them requires knowledge of the entire network, these indices of importance may not be desirable from a practical standpoint.

Table 7 Top10 important users and their importance indices

(a) $\rho = 0$

Cuisine		Flower		Furniture		Fashion	
user	importance	user	importance	user	importance	user	importance
681	111.00	681	111.00	681	111.00	681	111.00
421	107.00	421	107.00	421	107.00	421	107.00
645	42.00	645	42.00	645	42.00	645	42.00
370	41.00	370	41.00	370	41.00	370	41.00
727	37.00	727	37.00	727	37.00	727	37.00
651	34.00	651	34.00	651	34.00	651	34.00
187	33.00	187	33.00	187	33.00	187	33.00
106	29.00	106	29.00	106	29.00	106	29.00
218	24.00	218	24.00	218	24.00	218	24.00
433	24.00	433	24.00	433	24.00	433	24.00

(b) $\rho = 0.5$

Cuisine		Flower		Furniture		Fashion	
user	importance	user	importance	user	importance	user	importance
79	1.21	681	3.76	573	1.73	681	4.49
573	1.16	421	3.70	426	0.53	421	3.62
426	0.67	370	1.83	211	0.46	645	2.83
482	0.51	218	1.70	470	0.44	370	2.62
503	0.48	645	1.60	745	0.32	727	2.57
689	0.42	651	1.44	194	0.31	187	2.49
211	0.40	669	1.37	256	0.28	651	2.32
470	0.39	187	1.36	79	0.20	106	2.25
203	0.31	27	1.24	503	0.12	722	1.86
327	0.27	261	1.19	680	0.12	433	1.81

(c) $\rho = 1$

Cuisine		Flower		Furniture		Fashion	
user	importance	user	importance	user	importance	user	importance
79	1.46	203	0.64	426	0.28	470	0.30
426	0.45	776	0.22	573	0.15	426	0.29
482	0.26	34	0.16	745	0.10	203	0.22
503	0.23	211	0.15	194	0.10	722	0.20
689	0.18	421	0.13	256	0.08	645	0.19
203	0.10	681	0.13	211	0.05	187	0.19
327	0.07	218	0.12	79	0.04	681	0.18
573	0.07	376	0.10	470	0.02	727	0.18
211	0.04	518	0.09	503	0.02	106	0.17
332	0.02	27	0.09	680	0.01	79	0.17

user 681, who has the largest follower count in the dataset, exhibits a consistently high influence on the topic of 'Flower', regardless of the value of ρ. Additionally, user 79, despite having a smaller follower base, demonstrating strong influence on both 'Cuisine' and 'Furniture' topics, with stable importance even as ρ increases. Such users, who maintain high importance across multiple topics, are likely to facilitate topic-spillover. For practitioners executing seeded marketing campaigns (SMCs), it is crucial to identify the optimal seed users based on the user importance index relevant to the campaign topic and budget. However, determining the appropriate tuning parameter remains a key challenge. The next section will explore the campaign's effectiveness under various values of the tuning parameter through simulations, evaluating how the network-wide level of interest changes when selected seed users increase their topic-related posts.

Table 8 Estimates of hierarchical coefficients

	Cuisine		Flower		Furniture		Fashion	
	Est	95% HPD	Est	95% HPD	Est	95% HPD	Est	95% HPD
Intercept	-0.12	(-0.69, 0.51)	0.11	(-0.49, 0.67)	0.05	(-0.54, 0.66)	-0.23	(-0.80, 0.41)
#Follows (sender)	-0.20	(-0.70, 0.30)	-0.01	(-0.51, 0.54)	-0.16	(-0.68, 0.43)	-0.42	(-0.93, 0.02)
#Followers (sender)	0.00	(-0.56, 0.54)	-0.01	(-0.55, 0.59)	-0.05	(-0.67, 0.51)	-0.20	(-0.70, 0.29)
#Follows (receiver)	0.01	(-0.55, 0.58)	-0.06	(-0.60, 0.50)	0.10	(-0.46, 0.67)	-0.02	(-0.70, 0.68)
#Followers (receiver)	-0.25	(-0.94, 0.32)	**0.45**	(0.17, 0.74)	**0.45**	(0.12, 0.80)	**-0.37**	(-0.75, -0.03)
Activity (sender)	-0.05	(-0.65, 0.58)	-0.02	(-0.65, 0.53)	-0.05	(-0.61, 0.62)	-0.03	(-0.63, 0.60)
Activity (receiver)	-0.05	(-0.70, 0.56)	0.02	(-0.62, 0.63)	0.01	(-0.63, 0.58)	-0.01	(-0.62, 0.58)

4.3.4 Hierarchical coefficients

Finally, we examine the estimated regression coefficients for variables moderating social influence if they are not reduced to zero. Table 8 presents the estimated coefficients along with their 95% HPD intervals. Notably, the number of followers on the recipient side of the user pair is significantly and positively estimated for the 'Flower' and 'Furniture' topics. This suggests that users with a larger follower base are more likely to generate interest among surrounding users when they post about these topics. Previous studies have also indicated that the behavior of well-connected users can influence the behavior of others and therefore contribute to the success of SMCs (Hinz et al. 2011). The findings of this study are consistent with these earlier studies, which suggest that well-connected users tend to have strong influence. However, this insight does not apply across all topics; for instance, in topics like 'Cuisine,' the number of followers appears to have little impact on the strength of social influence, while in topics like 'Fashion,' a moderating effect that may weaken social influence is observed.

While the number of followers on the recipient side was shown to have a significant moderating effect on social influence across many topics, other variables, particularly the activity level introduced as a proxy for platform loyalty, did not exhibit significant effects. Due to data limitations, this study could not capture variables that adequately represent loyalty, but incorporating various features, such as the number of years since platform registration, engagement of previous posts, and interactions with other users, could enhance the expressive ability of hierarchical model. Additionally, the proposed model assumed that moderating effects operate only when social influence is not reduced to zero. When the number of such non-zero social influence is not so much, the estimation of coefficients of moderating effects with weak signals may have large variances. Indeed, when aggregating the estimated indicators representing the presence of significant social influence, π_{ufk} in Equation 4, it was found that the structure of shrinkage prior was chosen for 80-90% of user pairs rather than the structure of hierarchical regression. While this aligns with our expectations, as it suggests the sparsity of social influence, it may also give rise to the limited number of significantly estimated coefficients. Refining the structure of the hierarchical model could potentially capture moderating effects with weak signals, but this extensions beyond the scope of the current study and remains a subject for future research.

4.4 Simulation for Seeded Marketing Campaigns

In this section, we visualize the effectiveness of SMCs depending on how to select seed users, determined by the user importance index proposed in the previous section, through numerical simulations. The scenarios for the SMCs simulation are as follows:

1. Select the top n users as seed users based on the user importance index (n is fixed at 10 in this simulation). The method for determining the value of tuning parameter will be discussed below.
2. The seed users, receiving special requests from companies managing the campaign, will increase their post related to the topic by 10% within the campaign period. Increasing by 10% means that $\theta_{ut+1k} = \theta_{utk} + 0.1$, so the posts on topics k' $(k' \neq k)$ will be proportionally reduced[9], resulting in $\theta_{ut+1k'} = \theta_{utk'} - \frac{0.1}{K-1}$.
3. Topic distributions and posts for non-seed users will be simulated according to the proposed model.
4. In each period, we measure the change in interest in topic k across the entire network which are defined as the average value of the topic distribution across all users from the start of the campaign period.

In the scenario described above, it is assumed that the company managing the campaign selects seed users based on the user importance index. However, the emphasis on either the users' network position or their social influence on specific topics varies depending on the settings of tuning parameter, ρ in Equation 7. Therefore, the simulation also examines how the campaign's effectiveness changes with different tuning parameter values. As mentioned in the previous section, when $\rho = 0$, the user importance is calculated solely based on the network, accordingly leading to the selection of users with a large number of followers. This setting allows consideration of the potential reach of seed users but does not account for the certainty with regard to increased interest among surrounding users. Conversely, when $\rho = 1$, the user importance is determined entirely by the social influence on the specific topic, disregarding the number of followers the user has in the network. While this setting allows for consideration of the certainty of influence, it does not account for the user's potential reach. The setting of $0 < \rho < 1$ takes a balance between these two aspects, considering both potential reach and certainty of influence. The simulation evaluates the campaign's effectiveness using these three tuning parameter settings and a method that randomly selects seed users without considering the importance index, resulting in a total of four approaches. For the random approach, we use the average value of 20 simulations to mitigate biases due to randomness.

Figure 6 illustrates the effects of SMCs through simulation, as well as the differences resulting from various methods of selecting seed users. The vertical axis represents the percentage increase in users' interests for that topic across the entire network compared to the start of the campaign in each period. The colors in the graph indicate the method of seed user selection: red (Random) represents random selection, green (Network) represents to the consideration of network position ($\rho = 0$), blue (Topic) represents the consideration of social influence by topic ($\rho = 1$), and purple (Balance) shows the results of the most effective setting when ρ was varied from 0.1 to 0.9.

[9] If $\theta_{ut+1k'} < 0$, reset $\theta_{ut+1k'} = 1e-10$ and re-normalized.

The results indicate that the network-based index performed most effectively for topics such as cuisine and furniture, while the balance-based index competed with or outperformed network-based index for topics like flowers and fashion. The superior performance of the network index in certain topics aligns with prior research asserting that selecting well-connected or high-status users as seeds maximizes the impact of SMCs (e.g., Hinz et al. 2011). However, studies on SMCs' effectiveness also emphasize that, beyond network-based status, the alignment between the content or brand's topic and the interests of influencers and their followers can significantly influence outcomes (Leung et al. 2022). The finding that balance, which accounts for both influencers' potential reach and topic-specific influence, outperformed the network approach in certain topics resonates with this emerging trend. Furthermore, recent studies highlight that focusing exclusively on high-status users within networks may not guarantee SMCs success. Instead, leveraging not only the extensive reach of high-status users but also the strong relationships that low-status users maintain with their followers has become increasingly recognized as critical (Gu et al. 2024, Wies et al. 2023). While much of the existing research provides guidelines for selecting seed users based solely on network metrics, this study underscores the importance of incorporating users' topic-specific influence into the decision-making process. The proposed model offers an effective tool for such analysis, suggesting that considering both network-level status and topic-specific influence can enhance SMCs performance. In this regard, the simulation results presented in this section emphasize the potential to enhance existing SMCs by considering both the status at the network-level and social influence on a topic-by-topic basis derived from the proposed framework.

Thus, we suggest that integrating these factors can lead to more effective and targeted SMCs strategies, potentially refining the way influencer campaigns are designed and executed. Specifically, instead of relying solely on easily accessible network-based metrics such as the number of followers when selecting seed users, it is crucial to create a small user pool dataset comprising the seed candidates and their surrounding users. By applying the proposed model to this dataset, we can calculate user importance while accounting for topic-specific influence. In the above simulation, Network, Topic, and Balance were distinctly separated to highlight the characteristics of each seed user selection method. However, since these methods are represented by a single tuning parameter, ρ, we can encompass all three methods by exploring the SMCs effectiveness while varying ρ from 0 to 1 to select the optimal seed users. This approach allows for the selection of seed users who can most effectively elevate interest levels within the community, based on the topics of the content or brand being promoted.

There are, however, several cautions in interpreting this simulation analysis. The outcome metric used in this study focuses on the extent to which seed users can elevate the interest levels of surrounding users, inferred from their posted content. But SMCs involve additional metrics that warrant attention, such as engagement (e.g., likes and comments) and increased sales driven by the campaign. Estimating how seed user contributions impact these metrics would require extending the proposed model, a task constrained in

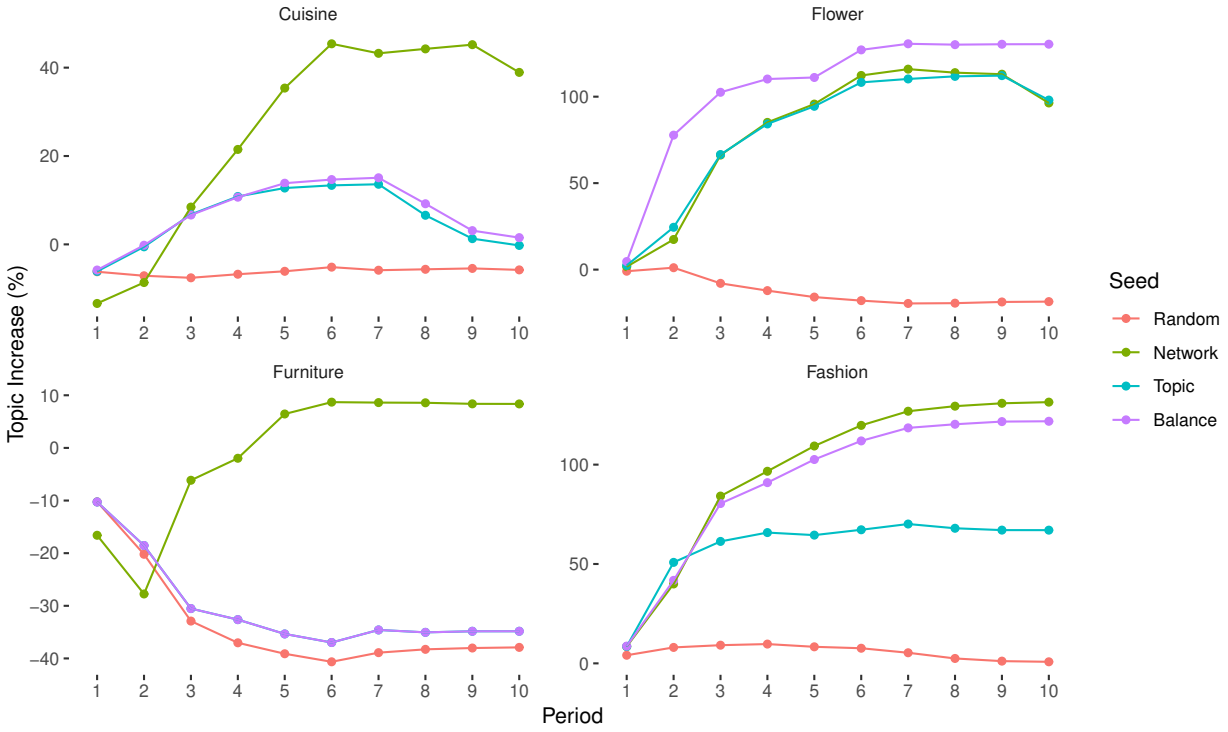

Figure 6 Results of SMCs simulation studies under different seed user selection strategies[10]

this study by data limitations. Additionally, while the interest levels of surrounding users are inferred from their posting behavior, this approach may underestimate actual changes in interest. Typically, users only post content when their interest in a topic surpasses a certain threshold. Consequently, interest level changes that fall short of prompting posting behavior may go unobserved, even if such changes are substantial. On the other hand, data on viewed content and reactions to it generally involve lower psychological thresholds compared to posting behavior and thus are likely to capture more nuanced changes in interest levels. Extending the proposed model to incorporate such data into an analytical framework presents a compelling avenue for future research.

5 Managerial Implications and Conclusion

In this paper, we developed a methodology to estimate social influence among social media users generating unstructured content. Our approach extends the conventional topic model by incorporating lagged influence on user interests within a hierarchical topic distribution structure. While previous studies have focused on numerical aspects of consumer behavior, our approach advances the field by addressing the influence on unstructured content generation, such as text and images. This methodology identifies key users for specific topics, enhancing the efficiency of seeded marketing campaigns in two ways: by ensuring consistent dissemination through alignment of influential topics with the company's content, and by attracting attention

[10] Note: In the bottom-left, the "Topic" line overlaps with the "Balance" line, making it visually indistinguishable.

to additional topics with minimal investment by targeting users who have influence across multiple areas, thus facilitating topic-spillover.

Our empirical analysis of real-world image-sharing social media data demonstrates that our model outperforms conventional topic models in both predictive accuracy and topic interpretability. We visualized the model's results to interpret topics and identify important users based on network features and influence level. Firms can leverage these key users to optimize their marketing campaigns, aligning with user importance for each content topic. Numerical simulations for campaigns indicated the potential to enhance the existing approaches by considering social influence on the relevant topic as well as the user's network-level status. Although our analysis focused on image data, the model's applicability to other unstructured data like text (Büschken and Allenby 2016) and purchase data (Jacobs et al. 2016) is promising, given its adaptability within the topic modeling framework.

When applied to the context of products like fast fashion apparel, this methodology can provide firms with valuable insights for using influencers as seeds for marketing and estimating potential demand. The improved predictive accuracy through identification of influencer seeds and their followers suggests that manufacturers can use these models to better predict market trends and consumer demand, allowing for more efficient production planning and inventory management. Improved topic interpretability allows manufacturers to gain clearer insights into specific areas of interest within their consumer base. The ability to analyze unstructured data, such as images and text, enables manufacturers to engage more deeply with consumer feedback, leading to rapid iterations and improvements that align more closely with consumer expectations and needs. This, in turn, can inform research and development efforts, driving innovations that resonate with consumer interests.

This study also acknowledges several limitations. First, the scalability of the proposed model is constrained, which limits the size of the network that the morel can handle to a medium scale by today's standards. Since the proposed model assumes heterogeneous social influence between each pair of users, which leads to estimation of a large influence network, the computational cost increases exponentially as the number of user grows. According to this limitation, recent advancements in machine learning, particularly in low-rank matrix approximations, offer promising avenues for reducing computational demands. By leveraging these techniques, it may be possible to approximate topic-specific influence between users efficiently while maintaining accurate estimates of the topic composition in user-generated content. Although such approximations may introduce estimation errors, balancing this trade-off with scalability is critical for incorporating the proposed method into decision-making processes in practical applications.

Second, the proposed model, grounded in topic modeling, does not fully capture the rich information embedded within images, While the model adopts an approach that detects objects within images and treats them as object sets for analysis, it overlooks crucial aspects such as object positioning, the extent of visual coverage, image resolution, aesthetic quality, and the relationships between objects. Such challenges are

pervasive in studies dealing with unstructured data such as text. In recent years, there has been growing interest in text-based research that incorporates previously overlooked information, such as aesthetic quality and writing style, and a similar shift is necessary for image data. Moving beyond simple aggregation or applying topic models is essential to unlock the richer analytical potential of image content.

Looking ahead, our model hold promise for valuable extensions. One possibility is to explore the influence of external stimuli and marketing variables on social influence by assuming a hierarchical structure for the coefficients. Although not covered in this paper due to data limitations, research suggests that social influences can be moderated by factors such as consumer status and network position (Susarla et al. 2012), in addition to those considered in this study. Additionally, incorporating time-varying social influence parameters, similar to time-varying vector autoregressive models (Primiceri 2005), could enable real-time analysis of consumer interactions and the evolving influence of users. These potential extensions present exciting opportunities for future research.

References

Ameri, Mina, Elisabeth Honka, Ying Xie. 2019. Word of Mouth, Observed Adoptions, and Anime-Watching Decisions: The Role of the Personal vs. the Community Network. *Marketing Science*, 38 (4), 567-583.

Archak, Nikolay, Anindya Ghose, Panagiotis G. Ipeirotis. 2011. Deriving the Pricing Power of Product Features by Mining Consumer Reviews. *Management Science*, 57 (8), 1485-1509.

Bao, Tong, Tung Lung Steven Chang. 2014. Finding disseminators via electronic word of mouth message for effective marketing communications. *Decision Support Systems*, 67 21-29.

Bass, Frank M. 1969. A New Product Growth for Model Consumer Durables. *Management Science*, 15 (5), 215-227.

Bhattacharya, Anirban, Debdeep Pati, Natesh S. Pillai, David B. Dunson. 2015. Dirichlet–Laplace Priors for Optimal Shrinkage. *Journal of the American Statistical Association*, 110 (512), 1479-1490.

Bischof, Jonathan, Edoardo M. Airoldi. 2012. Summarizing topical content with word frequency and exclusivity. *International Conference on Machine Learning*. ICML_International Conferene on Machine Learning.

Blei, David M., John D. Lafferty. 2005. Correlated topic models. *Advances in Neural Information Processing Systems*. NeurIPS_Neural Information Processing Systems, 147-154.

Blei, David M., John D. Lafferty. 2006. Dynamic topic models. *ACM International Conference Proceeding Series*, vol. 148. ICML_International Conferene on Machine Learning, 113-120. doi:10.1145/1143844.1143859.

Blei, David M., Andrew Y. Ng, Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3 (4-5), 993-1022.

Bollinger, Bryan, Kenneth Gillingham. 2012. Peer effects in the diffusion of solar photovoltaic panels. *Marketing Science*, 31 (6), 900-912.

Büschken, Joachim, Greg M. Allenby. 2016. Sentence-based text analysis for customer reviews. *Marketing Science*, 35 (6), 953-975.

Carlson, Keith, Praveen K. Kopalle, Allen Riddell, Daniel Rockmore, Prasad Vana. 2023. Complementing human effort in online reviews: A deep learning approach to automatic content generation and review synthesis. *International Journal of Research in Marketing*, 40 (1), 54-74.

Carvalho, Carlos M., Nicholas G. Polson, James G. Scott. 2010. The horseshoe estimator for sparse signals. *Biometrika*, 97 (2), 465-480.

Cater, C. K., R. Kohn. 1994. On Gibbs sampling for state space models. *Biometrika*, 81 (3), 541-553.

Chae, Inyoung, Andrew T. Stephen, Yakov Bart, Dai Yao. 2017. Spillover effects in seeded word-of-mouth marketing campaigns. *Marketing Science*, 36 (1), 89-104.

Chen, Xi, Ralf Van der Lans, Tuan Q. Phan. 2017. Uncovering the Importance of Relationship Characteristics in Social Networks: Implications for Seeding Strategies. *Journal of Marketing Research*, 54 (2), 187-201.

Chintagunta, Pradeep K., Shyam Gopinath, Sriram Venkataraman. 2010. The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets. *Marketing Science*, 29 (5), 944-957.

Clauset, Aaron, M. E. J. Newman, Cristopher Moore. 2004. Finding community structure in very large networks. *Physical Review E*, 70 (6), 066111.

Dieng, Adji B., Francisco J.R. Ruiz, David M. Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8 439-453.

Dost, Florian, Ulrike Phieler, Michael Haenlein, Barak Libai. 2019. Seeding as Part of the Marketing Mix: Word-of-Mouth Program Interactions for Fast-Moving Consumer Goods. *Journal of Marketing*, 83 (2), 62-81.

Glynn, Chris, Surya T. Tokdar, Brian Howard, David L. Banks. 2019. Bayesian Analysis of Dynamic Linear Topic Models. *Bayesian Analysis*, 14 (1), 53-80.

Godes, David, Dina Mayzlin. 2004. Using online conversations to study word-of-mouth communication. *Marketing Science*, 23 (4),.

Godes, David, Dina Mayzlin. 2009. Firm-created word-of-mouth communication: Evidence from a field test. *Marketing Science*, 28 (4), 721-739.

Gong, Shiyang, Juanjuan Zhang, Ping Zhao, Xuping Jiang. 2017. Tweeting as a marketing tool: A field experiment in the TV industry. *Journal of Marketing Research*, 54 (6), 833-850.

Griffiths, T. L., M. Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101 (Supplement 1), 5228-5235.

Gu, Xian, Xiaoxi Zhang, P.K. Kannan. 2024. Influencer Mix Strategies in Livestream Commerce: Impact on Product Sales. *Journal of Marketing*, 88 (4), 64-83.

Hartmann, Wesley R. 2010. Demand estimation with social interactions and the implications for targeted marketing. *Marketing Science*, 29 (4), 585-601.

Hartmann, Wesley R., Puneet Manchanda, Harikesh Nair, Matthew Bothner, Peter Dodds, David Godes, Kartik Hosanagar, Catherine Tucker. 2008. Modeling social interactions: Identification, empirical methods and policy

implications. *Marketing Letters*, 19 (3-4), 287-304.

Hinz, Oliver, Bernd Skiera, Christian Barrot, Jan U. Becker. 2011. Seeding strategies for viral marketing: An empirical comparison. *Journal of Marketing*, 75 (6), 55-71.

Iyengar, Raghuram, Christophe Van den Bulte, Thomas W. Valente. 2011. Opinion leadership and social contagion in new product diffusion. *Marketing Science*, 30 (2), 195-212.

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

Jansen, Nora, Oliver Hinz. 2022. Inferring opinion leadership from digital footprints. *Journal of Business Research*, 139 (October 2021), 1123-1137.

Jiang, Ling, Xingyu Chen, Sentao Miao, Cong Shi. 2024. Play it safe or leave the comfort zone? Optimal content strategies for social media influencers on streaming video platforms. *Decision Support Systems*, 179 114148.

Karrer, Brian, M. E.J. Newman. 2011. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83 (1), 1-10.

Krijestorac, Haris, Rajiv Garg, Vijay Mahajan. 2020. Cross-Platform Spillover Effects in Consumption of Viral Content: A Quasi-Experimental Analysis Using Synthetic Controls. *Information Systems Research*, 31 (2), 449-472.

Kwark, Young, Gene Moo Lee, Paul A. Pavlou, Liangfei Qiu. 2021. On the Spillover Effects of Online Product Reviews on Purchases: Evidence from Clickstream Data. *Information Systems Research*, 32 (3), 895-913.

Lanz, Andreas, Jacob Goldenberg, Daniel Shapira, Florian Stahl. 2019. Climb or Jump: Status-Based Seeding in User-Generated Content Networks. *Journal of Marketing Research*, 56 (3), 361-378.

Lee, Thomas Y., Eric T. Bradlow. 2011. Automated marketing research using online customer reviews. *Journal of Marketing Research*, 48 (5), 881-894.

Leung, Fine F., Flora F. Gu, Yiwei Li, Jonathan Z. Zhang, Robert W. Palmatier. 2022. Influencer Marketing Effectiveness. *Journal of Marketing*, 86 (6), 93-115.

Liu, Liu, Daria Dzyabura, Natalie Mizik. 2020. Visual Listening In: Extracting Brand Image Portrayed on Social Media. *Marketing Science*, 39 (4), 669-686.

Liu, Xiao, Dokyun Lee, Kannan Srinivasan. 2019. Large-Scale Cross-Category Analysis of Consumer Review Content on Sales Conversion Leveraging Deep Learning. *Journal of Marketing Research*, 56 (6), 918-943.

Liu, Yong. 2006. Word of mouth for movies: Its dynamics and impact on box office revenue. *Journal of Marketing*, 70 (3), 74-89.

Lu, Shuya, Jianan Wu, Shih Lun (Allen) Tseng. 2018. How Online Reviews Become Helpful: A Dynamic Perspective. *Journal of Interactive Marketing*, 44 17-28.

Manski, Charles F. 1993. Identification of Endogenous Social Effects: The Reflection Problem. *The Review of Economic Studies*, 60 (3), 531.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems*. NeurIPS_Neural Information Processing Systems, 3111-3119. doi:10.18653/v1/d16-1146.

Moe, Wendy W., Michael Trusov, Robert H. Smith. 2011. The value of social dynamics in online product ratings forums. *Journal of Marketing Research*, 48 (3), 444-456.

Moldovan, Sarit, Eitan Muller, Yossi Richter, Elad Yom-Tov. 2017. Opinion leadership in small groups. *International Journal of Research in Marketing*, 34 (2), 536-552.

Moon, Sangkil, Wagner A. Kamakura. 2017. A picture is worth a thousand words: Translating product reviews into a product positioning map. *International Journal of Research in Marketing*, 34 (1), 265-285.

Nair, Harikesh S., Puneet Manchanda, Tulikaa Bhatia. 2010. Asymmetric Social Interactions in Physician Prescription Behavior: The Role of Opinion Leaders. *Journal of Marketing Research*, 47 (5), 883-895.

Opsahl, Tore, Filip Agneessens, John Skvoretz. 2010. Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, 32 (3), 245-251.

Park, Eunho, Rishika Rishika, Ramkumar Janakiraman, Mark B. Houston, Byungjoon Yoo. 2018. Social dollars in online communities: The effect of product, user, and network characteristics. *Journal of Marketing*, 82 (1), 93-114.

Park, Trevor, George Casella. 2008. The Bayesian Lasso. *Journal of the American Statistical Association*, 103 (482), 681-686.

Polson, Nicholas G., James G. Scott, Jesse Windle. 2013. Bayesian inference for logistic models using Pólya-Gamma latent variables. *Journal of the American Statistical Association*, 108 (504), 1339-1349.

Primiceri, Giorgio E. 2005. Time Varying Structural Vector Autoregressions and Monetary Policy. *The Review of Economic Studies*, 72 (3), 821-852.

Rishika, Rishika, Jui Ramaprasad. 2019. The Effects of Asymmetric Social Ties, Structural Embeddedness, and Tie Strength on Online Content Contribution Behavior. *Management Science*, 65 (7), 3398-3422.

Roberts, Margaret E., Brandon M Stewart, Edoardo M. Airoldi. 2016. A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, 1459 (March), 988-1003.

Rossi, Peter E., Greg M. Allenby, Robert McCulloch. 2005. *Bayesian Statistics and Marketing*, vol. 1. John Wiley & Sons, Ltd, Chichester, UK.

Sanchez, Joaquin, Carmen Abril, Michael Haenlein. 2020. Competitive spillover elasticities of electronic word of mouth: An application to the soft drink industry. *Journal of the Academy of Marketing Science*, 48 (2), 270-287.

Sarkka, Simo. 2013. *Bayesian Filtering and Smoothing*. Cambridge University Press, Cambridge.

Schulze, Christian, Lisa Schöler, Bernd Skiera. 2014. Not all fun and games: Viral marketing for utilitarian products. *Journal of Marketing*, 78 (1), 1-19.

Sonnier, Garrett P., Leigh Mcalister, Oliver J. Rutz. 2011. A dynamic model of the effect of online communications on firm sales. *Marketing Science*, 30 (4), 702-716.

Stephen, Andrew T., Donald R. Lehmann. 2016. How word-of-mouth transmission encouragement affects consumers' transmission decisions, receiver selection, and diffusion speed. *International Journal of Research in Marketing*, 33 (4), 755-766.

Susarla, Anjana, Jeong Ha Oh, Yong Tan. 2012. Social networks and the diffusion of user-generated content: Evidence from youtube. *Information Systems Research*, 23 (1), 23-41.

Tirunillai, Seshadri, Gerard J. Tellis. 2014. Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation. *Journal of Marketing Research*, 51 (4), 463-479.

Toubia, Olivier, Garud Iyengar, Renée Bunnell, Alain Lemaire. 2019. Extracting Features of Entertainment Products: A Guided Latent Dirichlet Allocation Approach Informed by the Psychology of Media Consumption. *Journal of Marketing Research*, 56 (1), 18-36.

Trusov, Michael, Anand V. Bodapati, Randolph E. Bucklin. 2010. Determining Influential Users in Internet Social Networks. *Journal of Marketing Research*, 47 (4), 643-658.

Van den Bulte, Christophe, Stefan Wuyts. 2007. *Social Networks and Marketing*. Marketing Science Institute, Cambridge, MA.

Viswanathan, Vijay, F. Javier Sese, Manfred Krafft. 2017. Social influence in the adoption of a B2B loyalty program: The role of elite status members. *International Journal of Research in Marketing*, 34 (4), 901-918.

Wang, Jing, Anocha Aribarg, Yves F. Atchadé. 2013. Modeling choice interdependence in a social network. *Marketing Science*, 32 (6), 977-997.

Wang, Xin (Shane), Jun Hyun (Joseph) Ryoo, Neil Bendle, Praveen K. Kopalle. 2021. The role of machine learning analytics and metrics in retailing research. *Journal of Retailing*, 97 (4), 658-675.

Watts, Duncan J, Peter Sheridan Dodds. 2007. Influentials, Networks, and Public Opinion Formation. *Journal of Consumer Research*, 34 (4), 441-458.

Wies, Simone, Alexander Bleier, Alexander Edeling. 2023. Finding Goldilocks Influencers: How Follower Count Drives Social Media Engagement. *Journal of Marketing*, 87 (3), 383-405.

Wu, Chunhua, Hai Che, Tat Y. Chan, Xianghua Lu. 2015. The economic value of online reviews. *Marketing Science*, 34 (5), 739-754.

Zeng, Xiaohua, Liyuan Wei. 2013. Social Ties and User Content Generation: Evidence from Flickr. *Information Systems Research*, 24 (1), 71-87.

Zhu, Feng, Xiaoquan Zhang. 2010. Impact of online consumer reviews on Sales: The moderating role of product and consumer characteristics. *Journal of Marketing*, 74 (2), 133-148.

E-Companion for Identifying Influential Users by Topic in Unstructured User-generated Content

EC.1 Model Inference Procedure

According to the model specification in Section 3, the joint likelihood of our proposed model becomes:

$$\begin{aligned} & p(w,z,\eta,\phi,\alpha,\beta,\gamma,\delta) \\ &= \prod_{u=1}^{U}\prod_{t=1}^{T}\left\{\prod_{i=1}^{N_{ut}} p(w_{uti}\mid z_{uti},\phi)p(z_{uti}\mid \eta_{ut})\right\}\prod_{k=1}^{K} p(\eta_{utk}\mid \eta_{\cdot t-1k},\alpha_k,\beta_{u\cdot k},\gamma_{tk},\delta_{uk})\times \\ & \quad p(\phi,\eta_{\cdot 0},\alpha,\beta,\gamma,\delta) \end{aligned} \tag{EC.1}$$

where $p(\phi,\eta_{\cdot 0},\alpha,\beta,\gamma,\delta)$ represents the joint prior distribution of the model unknowns, which will be specified in Section EC.2.

We now derive the procedure for model inference. Our approach utilizes the Markov Chain Monte Carlo (MCMC) method to generate samples from the posterior distributions for a predetermined number of K topics. A notable challenge in this process is the lack of a closed-form expression for the posterior distribution, primarily due to the logistic normal prior, as defined in Equation 2. To overcome this, we employ the technique proposed by Polson et al. (2013), which represents the categorical likelihood as a mixture of Gaussians with respect to a Pólya-Gamma distribution. As a result, the likelihood of the topic distribution for corresponding topic assignments is defined as follows:

$$\begin{aligned} p(z_{ut}\mid \eta_{ut}) &\propto \left(\frac{\exp(\eta_{ut1})}{\sum_{k'}\exp(\eta_{utk'})}\right)^{N_{ut1}}\cdots\left(\frac{\exp(\eta_{utK})}{\sum_{k'}\exp(\eta_{utk'})}\right)^{N_{utK}} \\ &\propto \exp\left(\kappa_{utk}\psi_{utk}-\frac{\zeta_{utk}}{2}\psi_{utk}^2\right), \end{aligned} \tag{EC.2}$$

where $\psi_{utk}=\eta_{utk}-\log\sum_{k'\neq K}\exp(\eta_{utk'})$, $\kappa_{utk}=N_{utk}-\frac{N_{ut}}{2}$, and N_{utk} is the number of elements assigned to topic k in the content generated by user u at time t. ζ_{utk} is an auxiliary variable following the Pólya-Gamma distribution $\zeta_{utk}\sim PG(N_{ut},0)$. As a result, we can derive the posterior distribution in a closed form.

To make inference on the dynamic processes that govern the topic distribution, we apply the forward filtering and backward sampling algorithm (Cater and Kohn 1994). Thus, we can rewrite the model representation of the hierarchical regression structure of the topic distribution in Equation 3 for derivation of the posterior distribution. Let $\eta_{tk}=(\eta_{1tk},\ldots,\eta_{Utk})^\top$ be a stacked form of η_{utk} over all users.

$$\eta_{tk}\sim N(B_k\eta_{t-1k}+C_{tk},I), \tag{EC.3}$$

$$\text{where } [B_k]_{uu'}=\begin{cases}\alpha_k & \text{if } u=u'\\ \beta_{uu'k} & \text{if } u\neq u', u'\in\mathcal{F}_u\\ 0 & \text{otherwise}\end{cases}, \qquad C_{tk}=\gamma_{tk}+\begin{pmatrix}\delta_{1k}\\ \vdots\\ \delta_{Uk}\end{pmatrix}$$

A coefficient matrix B_k is a $U \times U$ square matrix that can be interpreted as a social network weighted by autoregressive coefficients for self-loops (diagonal) and social influences on network edges (non-diagonal). This reformulation allows us to derive the filtering distribution and the smoothing distribution as follows.

$$
\begin{aligned}
p(\eta_{tk} \mid z_{1:t}, \cdots) &\propto N(\mu_{tk}, \Sigma_{tk}) \qquad \text{(EC.4)}\\
\Sigma_{tk} &= \left(S_{tk}^{-1} + diag(\zeta_{tk})\right)^{-1}, \quad S_{tk} = I + B_k \Sigma_{t-1k} B_k^\top \\
\mu_{tk} &= \Sigma_{tk}\left(\kappa_{tk} + \zeta_{tk} \cdot \log \sum_{k' \neq k} \exp(\eta_{tk'}) + S_{tk}^{-1}(B_k \mu_{t-1k} + C_{tk})\right) \\
p(\eta_{tk} \mid z_{1:T}, \cdots) &\propto N(\tilde{\mu}_{tk}, \tilde{\Sigma}_{tk}) \qquad \text{(EC.5)}\\
\tilde{\mu}_{tk} &= \mu_{tk} + G_{tk}\left(\tilde{\mu}_{t+1k} - B_k \mu_{tk} - C_{t+1k}\right), \quad G_{tk} = \Sigma_{tk} B_k^\top \left(I + B_k \Sigma_{tk} B_k^\top\right)^{-1} \\
\tilde{\Sigma}_{tk} &= \tilde{S}_{tk} + G_{tk} \tilde{\Sigma}_{t+1k} G_{tk}^\top, \quad \tilde{S}_{tk} = \Sigma_{tk} - G_{tk}(I + B_k \Sigma_{tk} B_k^\top) G_{tk}
\end{aligned}
$$

More details of this derivation and posterior distributions of the remaining parameters are in EC.2. Furthermore, we conduct numerical experiments on synthetic data to validate the inference procedure regarding the performance of recovering model parameters. These results are presented in EC.3.

EC.2 Details of the Posterior Distributions

In this appendix, we show the detailed derivation process of the filtering distribution and smoothing distribution (the derivation is based on Sarkka 2013), which were omitted in the text. As described in Section EC.1, we adopt the forward filtering and backward sampling algorithm to sample from the posterior distribution of the topic distribution. Note that in the following, we omit the notation of parameters without our focus in this section for simplicity. First, let the filtering distribution at time $t-1$ be $p(\eta_{t-1k} \mid z_{1:t-1}) = N(\mu_{t-1k}, \Sigma_{t-1k})$, and then the joint distribution of η_{t-1k} and η_{tk} given data up to $t-1$ is defined as follows.

$$
\begin{aligned}
p(\eta_{t-1k}, \eta_{tk} \mid z_{1:t-1}) &= p(\eta_{tk} \mid \eta_{t-1k}) p(\eta_{t-1k} \mid z_{1:t-1}) \\
&= N(\eta_{tk}; B_k \eta_{t-1k} + C_{tk}, I) N(\eta_{t-1k}; \mu_{t-1k}, \Sigma_{t-1k}) \\
&= N(m_1, S_1), \qquad \text{(EC.6)}\\
\text{where } m_1 = \begin{pmatrix} \mu_{t-1k} \\ B_k \mu_{t-1k} + C_{tk} \end{pmatrix}, &\qquad S_1 = \begin{pmatrix} \Sigma_{t-1k} & \Sigma_{t-1k} B_k^\top \\ B_k \Sigma_{t-1k} & I + B_k \Sigma_{t-1k} B_k^\top \end{pmatrix}
\end{aligned}
$$

The last line is obtained by using Lemma A.1 of Sarkka (2013). By marginalizing the joint distribution with respect to η_{t-1k}, we obtain the following conditional distribution.

$$
\begin{aligned}
p(\eta_{tk} \mid z_{1:t-1}) &= N(m_2, S_2) \qquad \text{(EC.7)}\\
\text{where } m_2 = B_k \mu_{t-1k} + C_{tk}, &\qquad S_2 = I + B_k \Sigma_{t-1k} B_k^\top
\end{aligned}
$$

The likelihood of η_{tk} with respect to z_t can be obtained by using (EC.2).

$$
\begin{aligned}
p(z_t \mid \eta_{tk}) &\propto \exp\left(\psi_{tk}^\top \kappa_{tk} - \frac{1}{2}\psi_{tk}^\top diag(\zeta_{tk})\psi_{tk}\right) \qquad \text{(EC.8)}\\
\text{where } \psi_{tk} &= \{\psi_{1tk}, \ldots, \psi_{Utk}\}^\top = \eta_{tk} - \log \sum_{k' \neq k} \exp(\eta_{tk'})\\
\kappa_{tk} &= \{\kappa_{1tk}, \ldots, \kappa_{Utk}\}^\top, \qquad \zeta_{tk} = \{\zeta_{1tk}, \ldots, \zeta_{Utk}\}^\top
\end{aligned}
$$

Therefore, the posterior distribution of η_{tk} when observing the data up to t (i.e., filtering distribution) is given as follows.

$$
\begin{aligned}
p(\eta_{tk} \mid z_t, z_{1:t-1}) &\propto p(z_t \mid \eta_{tk}) p(\eta_{tk} \mid z_{1:t-1})\\
&\propto \exp\left(\psi_{tk}^\top \kappa_{tk} - \frac{1}{2}\psi_{tk}^\top diag(\zeta_{tk})\psi_{tk}\right) N(m_2, S_s)\\
&= N(\mu_{tk}, \Sigma_{tk}) \qquad \text{(EC.9)}\\
\text{where } \Sigma_{tk} &= \left(S_2^{-1} + diag(\zeta_{tk})\right)^{-1},\\
\mu_{tk} &= \Sigma_{tk}\left(\kappa_{tk} + \zeta_{tk} \cdot \log \sum_{k' \neq k} \exp(\eta_{tk'}) + S_2^{-1} m_2\right)
\end{aligned}
$$

Next, as with the above, the joint distribution of η_{tk} and η_{t+1k} given $z_{1:t}$ is as follows.

$$
\begin{aligned}
p(\eta_{tk}, \eta_{t+1k} \mid z_{1:t}) &= N(\tilde{m}_1, \tilde{S}_1) \qquad \text{(EC.10)}\\
\text{where } \tilde{m}_1 &= \begin{pmatrix} \mu_{tk} \\ B_k \mu_{tk} + C_{t+1k} \end{pmatrix}, \qquad \tilde{S}_1 = \begin{pmatrix} \Sigma_{tk} & \Sigma_{tk} B_k^\top \\ B_k \Sigma_{tk} & I + B_k \Sigma_{tk} B_k^\top \end{pmatrix}
\end{aligned}
$$

Since the joint distribution is Gaussian, the conditional distribution is easily obtained as follows.

$$
\begin{aligned}
p(\eta_{tk} \mid \eta_{t+1k}, z_{1:t}) &= N(\tilde{m}_2, \tilde{S}_2) \qquad \text{(EC.11)}\\
\text{where } \tilde{m}_2 &= \mu_{tk} + G_{tk}(\eta_{t+1k} - B_k \mu_{tk} - C_{t+1k})\\
G_{tk} &= B_k \Sigma_{tk} (I + B_k \Sigma_{tk} B_k^\top)^{-1}\\
\tilde{S}_2 &= \Sigma_{tk} - G_{tk}(I + B_k \Sigma_{tk} B_k^\top) G_{tk}^\top
\end{aligned}
$$

Let the smoothing distribution at $t+1$ be $p(\eta_{t+1k} \mid z_{1:t}) = N(\tilde{\mu}_{t+1k}, \tilde{\Sigma}_{t+1k})$, and since $p(\eta_{tk} \mid \eta_{t+1k}, z_{1:T}) = p(\eta_{tk} \mid \eta_{t+1k}, z_{1:t})$ from the model specification, we can obtain the joint distribution when observing the whole data as follows.

$$
\begin{aligned}
p(\eta_{tk}, \eta_{t+1k} \mid z_{1:T}) &= p(\eta_{tk} \mid \eta_{t+1k}, z_{1:t}) p(\eta_{t+1k} \mid z_{1:T})\\
&= N(\tilde{m}_2, \tilde{S}_2) N(\tilde{\mu}_{t+1k}, \tilde{\Sigma}_{t+1k})\\
&= N(\tilde{m}_3, \tilde{S}_3) \qquad \text{(EC.12)}\\
\text{where } \tilde{m}_3 &= \begin{pmatrix} \tilde{\mu}_{t+1k} \\ \mu_{tk} + G_{tk}(\tilde{\mu}_{t+1k} - B_k \mu_{tk} - C_{t+1k}) \end{pmatrix}\\
\tilde{S}_2 &= \begin{pmatrix} \tilde{\Sigma}_{t+1k} & \tilde{\Sigma}_{t+1k} G_{tk}^\top \\ G_{tk} \tilde{\Sigma}_{t+1k} & \tilde{S}_2 + G_{tk} \tilde{\Sigma}_{t+1k} G_{tk}^\top \end{pmatrix}
\end{aligned}
$$

Therefore, we can obtain the posterior distribution of η_{tk} when observing the whole data (i.e., smoothing distribution) by marginalizing the joint distribution with respect to η_{t+1k}.

$$p(\eta_{tk} \mid z_{1:T}) = N(\tilde{\mu}_{tk}, \tilde{\Sigma}_{tk}) \tag{EC.13}$$
$$\text{where } \tilde{\mu}_{tk} = \mu_{tk} + G_{tk}(\tilde{\mu}_{t+1k} - B_k\mu_{tk} - C_{t+1k}), \qquad \tilde{\Sigma}_{tk} = \tilde{S}_2 + G_{tk}\tilde{\Sigma}_{t+1k}G_{tk}^{\top}$$

In the MCMC iteration, we calculate the filtering distribution forwards, and then let us regard $\tilde{\mu}_{Tk} = \mu_{Tk}, \tilde{\Sigma}_{Tk} = \Sigma_{Tk}$ to sample from the smoothing distribution backwards.

If η_{ut} and z_{ut} are given, ζ_{utk} can be also sampled from the following Pólya-Gamma distribution (Polson et al. 2013).

$$p(\zeta_{utk} \mid \eta_{ut}, z_{ut}) \propto PG\left(N_{ut}, \eta_{utk} - \log \sum_{k' \neq k} \exp(\eta_{utk'})\right) \tag{EC.14}$$

Since we can easily derive the posterior distributions of the remaining parameters as with the conventional Bayesian estimation of the normal linear regression model (Rossi et al. 2005) and topic models (Griffiths and Steyvers 2004), only the obtained distributions are displayed in the following.

$$p(z_{uti} = k \mid \cdots) \propto \exp(\eta_{utk}) \times \frac{N_{kv\setminus uti} + \phi_0}{N_{k\setminus uti} + \phi_0 \cdot V}, \qquad \text{where} \quad \phi_k \sim Dirichlet(\phi_0) \tag{EC.15}$$

$$p(\alpha_k \mid \cdots) \propto N(\mu, \sigma^2), \qquad \text{where} \quad \sigma^2 = \left(\sum_{u=1}^{U}\sum_{t=1}^{T} \eta_{ut-1k}^2 + \frac{1}{\sigma_{\alpha 0}^2}\right)^{-1} \tag{EC.16}$$
$$\mu = \sigma^2 \left(\sum_{u=1}^{U}\sum_{t=1}^{T} \eta_{ut-1k}\left(\eta_{utk} - \sum_{f \in \mathcal{F}_u} \beta_{ufk} \cdot \eta_{ft-1k} - \gamma_{tk} - \delta_{uk}\right)\right)$$

$$p(\beta_{ufk} \mid \pi_{ufk} = 1, \ldots) \propto N(\mu, \sigma^2), \qquad \text{where} \quad \sigma^2 = \left(\sum_{t=1}^{T} \eta_{ft-1k}^2 + \frac{1}{\sigma_{\beta}^2}\right)^{-1} \tag{EC.17}$$
$$\mu = \sigma^2 \left(\sum_{t=1}^{T} \eta_{ft-1k}\left(\eta_{utk} - \alpha_k \cdot \eta_{ut-1k} - \sum_{f' \in \mathcal{F}_u} \beta_{uf'k} \cdot \eta_{f't-1k} - \gamma_{tk} - \delta_{uk}\right) + \frac{x_{uf}^{\top}\rho_k}{\sigma_{\beta}^2}\right)$$

$$p(\beta_{ufk} \mid \pi_{ufk} = 0, \ldots) \propto N(\mu, \sigma^2), \qquad \text{where} \quad \sigma^2 = \left(\sum_{t=1}^{T} \eta_{ft-1k}^2 + \frac{1}{\sigma_{\beta}^2 \cdot \omega_{ufk}}\right)^{-1} \tag{EC.18}$$
$$\mu = \sigma^2 \left(\sum_{t=1}^{T} \eta_{ft-1k}\left(\eta_{utk} - \alpha_k \cdot \eta_{ut-1k} - \sum_{f' \in \mathcal{F}_u} \beta_{uf'k} \cdot \eta_{f't-1k} - \gamma_{tk} - \delta_{uk}\right)\right)$$

$$p(\gamma_{tk} \mid \cdots) \propto N(\mu, \sigma^2), \qquad \text{where} \quad \sigma^2 = \left(U + \frac{1}{\sigma_{\gamma 0}^2}\right)^{-1} \tag{EC.19}$$
$$\mu = \sigma^2 \left(\sum_{u=1}^{U} \eta_{utk} - \alpha_k \cdot \eta_{ut-1k} - \sum_{f \in \mathcal{F}_u} \beta_{ufk} \cdot \eta_{ft-1k} - \delta_{uk}\right)$$

$$p(\delta_{uk} \mid \cdots) \propto N(\mu, \sigma^2), \qquad \text{where} \quad \sigma^2 = \left(T + \frac{1}{\sigma_{\delta 0}^2}\right)^{-1} \tag{EC.20}$$
$$\mu = \sigma^2 \left(\sum_{t=1}^{T} \eta_{utk} - \alpha_k \cdot \eta_{ut-1k} - \sum_{f \in \mathcal{F}_u} \beta_{ufk} \cdot \eta_{ft-1k} - \gamma_{tk}\right)$$

EC.3 Performance of Parameter Recovery

In this section, we conduct parameter recovery experiments using synthetic data to validate the performance of the proposed model and the estimation procedure. As introduced in Section 3, since the proposed model defines social influence as the lagged correlation between users' latent topic distributions, we should demonstrate the reliability of the estimated social influence on imaginary variables through a numerical experiment.

To evaluate the performance of parameter recovery, now we suppose several scenarios. The number of users (U) and the number of times (T) are set to be 100 or 200, and the number of topics (K) is set to be 5, 10, or 20. Another scenario is the sparsity proportion of social influence, specifically, in the case of $s\%$ sparsity, only randomly chosen $s\%$ of all edges in the generated network are given non-zero value of β, while the remaining $1-s\%$ of edges do not have any influence. Following the setting of each scenario, a random network and the values of parameters in Equation 3 are initialized, and then the topic distribution is set according to the hierarchical structure. The element distribution is also randomly set with $\phi_k \sim Dirichlet(\phi_{0k}),\ \phi_{0k} = \{\phi_{0k1}, \ldots, \phi_{0kV}\}^\top$, where ϕ_{0kv} corresponding 50 unique objects for each topic is ten times the others, specifically, in the case of $k = 1$, $\phi_{0k1} = \phi_{0k50} = 10$, while $\phi_{0k51} = \phi_{0kV} = 1$. Thus, the size of vocabulary is $V = 50 \times K$. Given the generated topic distribution and element distribution, the data w is generated according to the generative process in Section 3.1. Using the generated data, the model is estimated by MCMC described in the previous section.

Figure EC.1 and EC.2 show the values of root mean square error (RMSE) and correlation coefficient between the true values and the estimated values for each scenario and parameter. Although the parameters of the hierarchical regression model in Equation 3 are not recovered well, the topic distribution and the element distribution, which are the parameters of interest, are correctly estimated with about 0.8 of the correlation coefficient in most scenarios. For each scenario, the accuracy of the estimation tends to be worse as the number of data (U, T) and the number of topics (K), that is, the number of parameters to be estimated, increase. Since the parameters of interest in this study are topic distribution, element distribution, and social influence as can be seen from the discussion in Section 4.3, improving the estimation accuracy of the remaining parameters (α, γ, δ) is out of scope.

Next, to validate the prior distribution for social influence in the proposed model, several models with different prior settings are estimated for each scenario. In the field of Bayesian statistics, in addition to the Bayesian lasso prior assumed in the proposed model, the horseshoe prior (Carvalho et al. 2010) and the Dirichlet-Laplace prior (Bhattacharya et al. 2015) have been used as shrinkage priors. The definitions of the horseshoe prior and the Dirichlet-Laplace prior are

$$\beta_{ufk} \mid \omega_{ufk}, \tau_k \sim N(0, \omega_{ufk}^2 \cdot \tau_k^2), \qquad \omega_{ufk}^2 \sim C^+(0,1), \qquad \tau_k^2 \sim C^+(0,1) \tag{EC.21}$$

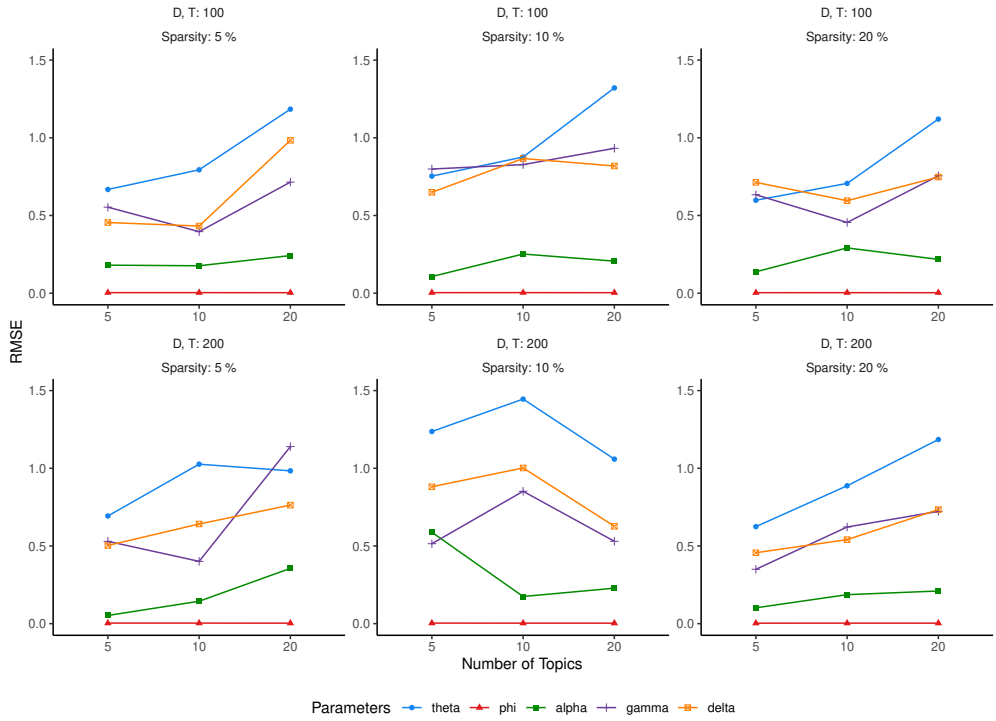

Figure EC.1 RMSE

and

$$\beta_{ufk} \mid \xi_{ufk}, \omega_{ufk}, \tau_k \sim N(0, \xi_{ufk} \cdot \omega_{ufk}^2 \cdot \tau_k^2) \tag{EC.22}$$
$$\xi_{ufk} \sim Exp\left(\frac{1}{2}\right), \qquad \omega_k \sim Dirichlet\left(\frac{1}{\sum_{u=1}^{U} |\mathcal{F}_u|}\right), \qquad \tau_k \sim Gamma\left(1, \frac{1}{2}\right),$$

respectively. In this simulation, we compare the performance of the Bayesian lasso prior with three different prior distributions, including these plus weakly informative prior ($\beta_{ufk} \sim N(0, 10^2)$), which does not assume sparsity. Figure EC.3 shows the values of RMSE between true and estimates. All models can accurately recover the true values, among which Bayesian lasso is superior to others in most of the scenarios. Moreover, even when the scale of the model (U, T, K) is increased, the RMSEs of social influence do not get so worse as the other parameters. Figure EC.4 shows the F-measure which is calculated by regarding users as influential when the estimated β is 0.5 or higher in absolute value. The F-measures are high in all scenarios, among which Bayesian lasso outperforms the others, and it indicates that the proposed model can provide reliable estimates of social influences among users.

EC.4 Simulation Experiments for Large Network Data

In this section, we discuss the computational cost of applying the proposed model to larger-scale networks, rather than the medium-scale network dataset used in the empirical analysis. Figure EC.5 illustrates the computation time required per iteration for estimating the proposed model as the number of users. As

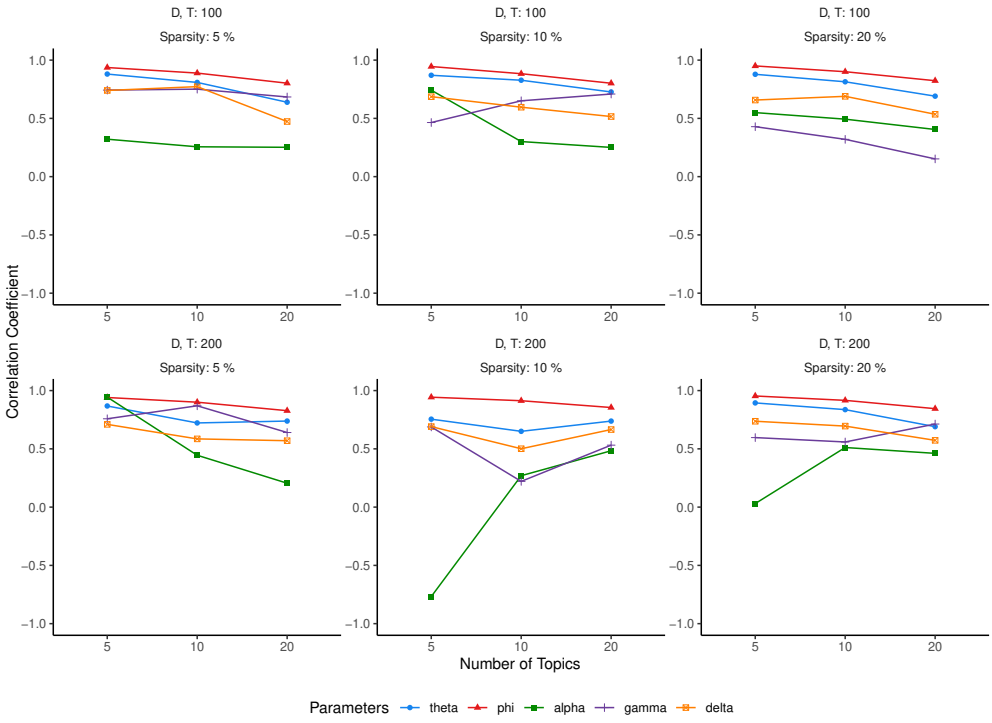

Figure EC.2 Correlation coefficient

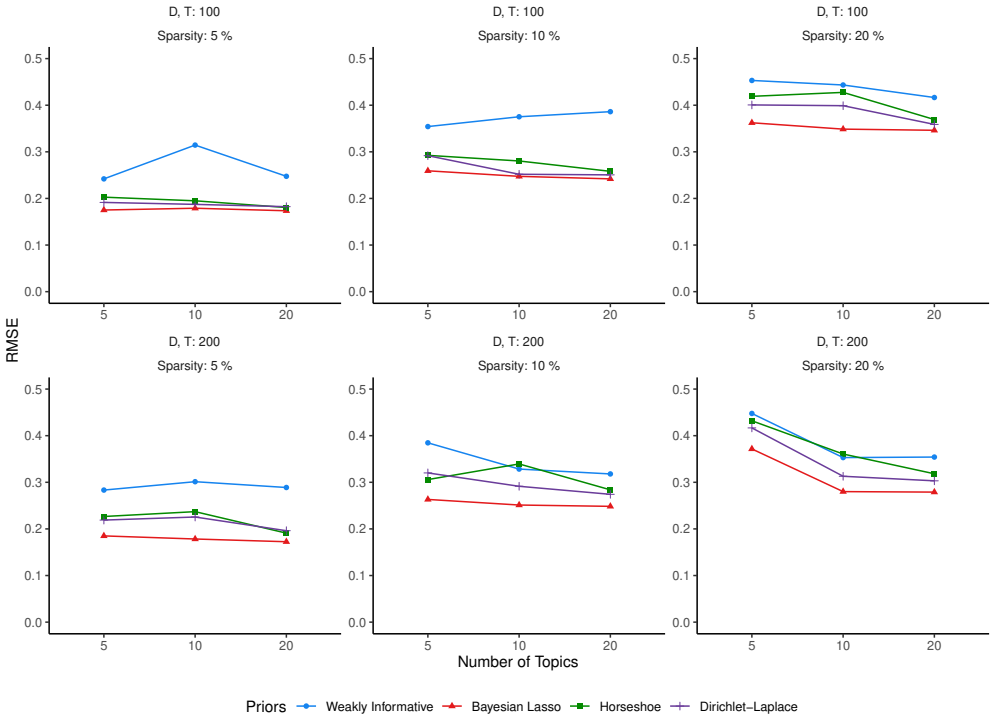

Figure EC.3 RMSE

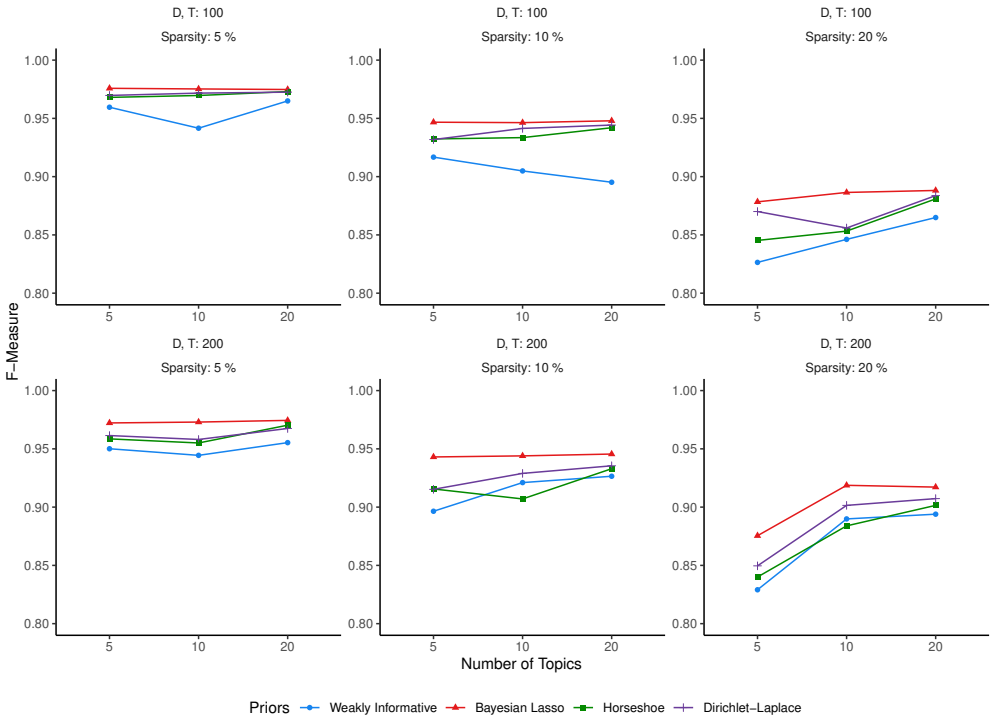

Figure EC.4 F-measure

shown, the computation time increases exponentially with the number of users. This exponential growth is likely due to the bottleneck caused by the matrix multiplication and inversion in Equation EC.10 and EC.14, which could pose a significant challenge when applying the proposed model to large-scale networks. Although this study has not identified a solution to this high computational cost, recent advancements in machine learning, particularly in low-rank matrix approximations, offer promising avenues for reducing computational demands. By leveraging these techniques, it may be possible to estimate user-specific topic distributions with realistic computational costs by approximating the large-scale social influence network, while still accurately estimating the social influence between user pairs. While addressing this challenge is beyond the scope of the current study and remains a topic for future research, it is an essential issue given the high demand in practice for effectively designing SMCs within large-scale networks involving thousands or even millions of users.

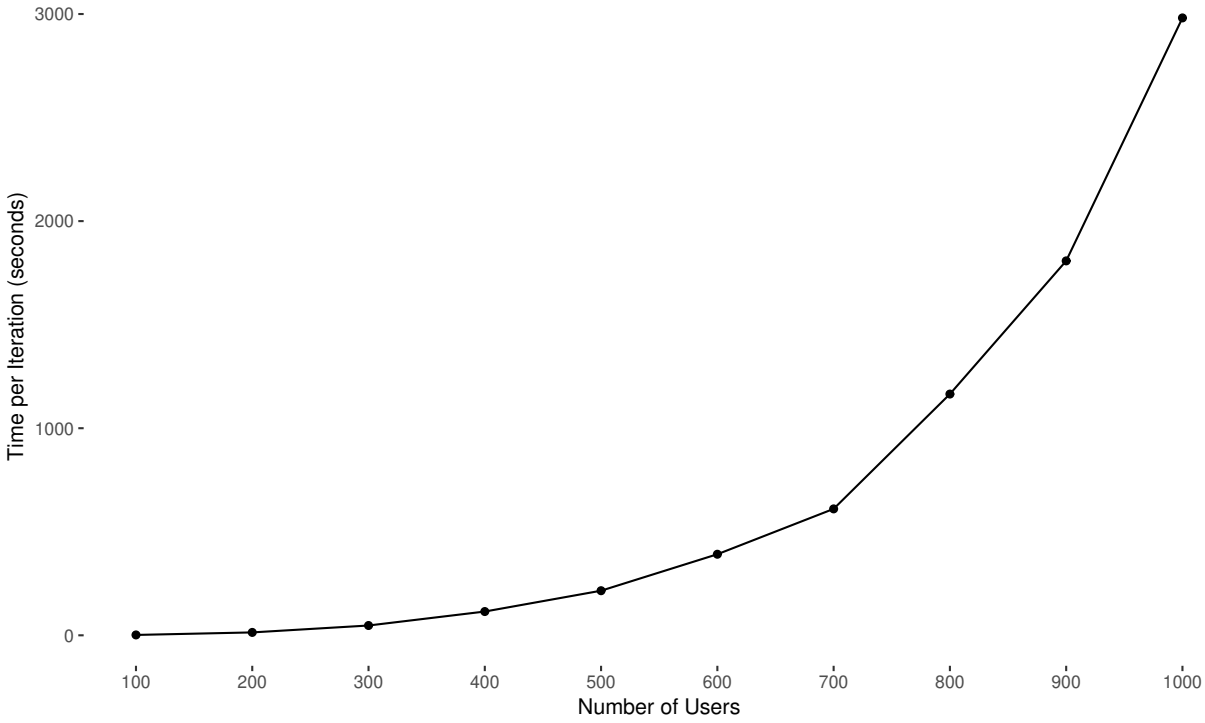

Figure EC.5 Computation time of the proposed model