A Co-evolution Model of Network Formation and Content Generation on Social Reading Platform

Mirai Igarashi[1] and Nobuhiko Terui[2]

1. Faculty of Business Science, Tsukuba University

2. Department of Business Economics, School of Management, Tokyo University of Science

Abstract

Understanding how individuals form social relationships and generate contents on social media is fundamental for both academic and practitioners. In the context of the co-evolution modeling, the formation of social networks and the behaviors of individuals are jointly modeled as they influence each other. However, too little attention has been paid to qualitative aspect of the behavior such as sentiment and topic of contents. This study proposes a Bayesian co-evolution model incorporating dynamic network model and topic model to describe the interdependent processes of network formation and content generation. The proposed model is empirically applied to the data in Japanese story-telling platform.

1. Introduction

To understand what kinds of structures lie when consumers become friends with other users in social media and produce content of text and images, we need to unveil the interaction between the network formation and the content production. Users who post some contents that are attractive or similar to their own tastes are more likely to be followed than those who do not. This is an example of how the content production of those users have influenced connections between users. Conversely, users may post similar content to other attractive content posted by friends, which is an example of how being connected to other users on the network change the content production behavior.

In the literature, the statistical models describing the co-evolution of network and behavior have been developed mainly by Snijders and his coauthors (e.g., Snijders et al., 2007; Steglich et al., 2010). Bhattacharya et al. (2019), which extends the model of Snijders et al. (2007) to the online environment, found that users tend to connect with others that have similar posting behavior, but after doing so, these users tend to diverge in their posting behavior. In the field of marketing, Ameri et al. (2022) applied the co-evolution model to identify the key drivers of the social media users' on-site behavior and discussed the most effective strategy to increase users' activity level.

However, these existing studies only consider numerical information, such as the presence or absence of behaviors and the number of times users do that behavior. On the other hand, it is more important to consider qualitative aspects of the behavior, such as what kind of content users produce. For example, existing studies have shown that users with similar behaviors are more likely to connect with each other, but no matter how active both of them are (i.e., they are similar in terms of the number of times of behavior), the probability of connection should also depend on the qualitative similarity such as topics of content they post or view. The effect of network formation on the qualitative aspects of behavior and how network formation is affected by the content of behavior remains unclear. This study fills these research gaps.

The purpose of this study is to construct a co-evolution model of network formation and content production behavior that takes into account both the effect of posting content on social media on network formation with other users and the effect of connected users on own content production behavior as a result of network formation. As for the network formation behavior, we apply a binary choice model to describe whether or not a user follows other known users on the platform, and for content production behavior, we describe the generation process of text content by a topic model assuming a hierarchical structure considering the influence of connected users.

2. Model

2.1 Network Formation

We observe the relationship between user i and j as a binary data y_{ijt}, that is, $y_{ijt} = 1$ if user i is following user j at time t and $y_{ijt} = 0$ otherwise. In this study, we assume a directed graph, thus in general, y_{ijt} and y_{jit} are not necessarily equal. This binary relationship follows a binary probit model.

$$y_{ijt} = \begin{cases} 1 & \text{if } y^*_{ijt} > 0 \\ 0 & \text{otherwise} \end{cases}, \qquad y^*_{ijt} = x_{ijt}^\top \beta_i^{(y)} + \epsilon_{ijt}, \qquad \epsilon_{ijt} \sim N(0,1)$$

x_{ijt} is a vector of covariates that are classified in network characteristics, such as number of degrees and transitivity, and qualitative similarities between those users defined by cosine similarity of topic distribution, which will be introduced in the next part of our model.

2.2 Content Production

Our model assumes a situation where social media users generate text contents (e.g., tweets, blogs, and novel stories). We observe text content m as data split into the smallest units (i.e., a set of words) in a fashion of Bag-of-Words (BoW) and denote it as $w_m = \{w_{m1}, \dots, w_{mN_m}\}^\top$. An element w_{mn} is generated by following a conventional topic model (Blei et al., 2003)

$$u_{mn} \sim Categorical(\theta_m), \qquad w_{mn} \mid u_{mn} = k \sim Categorical(\phi_k)$$

$\theta_m = \{\theta_{m1}, \dots, \theta_{mK}\}^\top$ is a topic distribution that represents a topic proportion within content m, and each dimension, θ_{mk}, that is nonnegative and on a simplex vector is realized through a softmax transformation of a normally distributed natural parameter, η_{mk}.

$$\theta_{mk} = \frac{\exp(\eta_{mk})}{\sum_{k'} \exp(\eta_{mk'})}, \qquad \eta_{mk} \sim N\left(\alpha_{t_m k} + \gamma_{i_m k} + z_{mk}^\top \beta_{i_m}^{(\eta)}, \sigma^2\right)$$

where t_m represents a posted time point of content m and i_m represents a user who generate content m, and $\alpha_{t_m k}$ and $\gamma_{i_m k}$ represent random effect of a specific time and user, respectively. z_{mk} is a vector of covariates for content m and topic k, and this is defined as a sum of topic distributions corresponding to contents posted or viewed by connected users with user i_m.

2.3 Hierarchical Structure

In the binomial probit model for network formation and the topic model for content production, we assume user heterogeneity on the coefficients of the covariates, and control distributions of the heterogeneous parameters by a hierarchical regression with user attributes, i.e.,

$$\beta_i = \left\{\beta_i^{(y)\top}, \beta_i^{(\eta)\top}\right\}^\top, \qquad \beta_i \sim MVN(\delta g_i, \Sigma_\beta)$$

where δ is a coefficient matrix of user i's attribute, g_i, and Σ_β is a covariance matrix.

Then, the joint likelihood of our model is as follows:

$$p(Y, W \mid X, Z, G, \alpha, \beta, \gamma, \delta, \Sigma_\beta) = \prod_{i=1}^{n} p(\beta_i \mid g_i, \delta, \Sigma_\beta) \left\{ \prod_{j \in J_i} \prod_{t \in T_{ij}} p(y_{ijt} \mid y_{ijt}^*) p(y_{ijt}^* \mid x_{ijt}, \beta_i) \right\}$$

$$\times \prod_{m=1}^{M} p(\eta_m \mid z_m, \alpha_{t_m}, \beta_{i_m}, \gamma_{i_m}) \left\{ \prod_{n=1}^{N_m} p(w_{mn} \mid u_{mn}, \phi) p(u_{mn} \mid \eta_m) \right\}$$

We adopt Markov Chain Monte Carlo (MCMC) method to generate samples from the posterior distributions for parameters of out model.

3 Empirical Analysis

3.1 Data

We evaluate our model using Japanese story-telling platform dataset. The platform allows users to freely post and read stories, and works in a variety of genres, including science fiction and love romance, are updated day after day. In addition, it has social networking features that allow users to follow their favorite works and authors to check daily updates and easily interact with them. Therefore, it can be expected that a co-evolutionary structure is inherent in the platform, where users form network and write stories according to the works posted and read by other users who are connected in the network.

As for model estimation, the target users were selected in a snowball sampling fashion, and then the corresponding access logs, which are (i) recognizing and (ii) following users and (iii) posting and (iv) following works. The works data consists of story introduction written by authors and their tags for 53 weeks from January 1, 2016. Totally, our dataset includes 1,066 users. Table 1 provides the descriptive statistics of the dataset.

Table 1 Descriptive statistics of dataset

		Number of log per user				
Acess Log Data	Total	Mean	Median	SD	Max	Min
Recognition (user)	40,961	76.8	40	91.1	563	0
Follow (user)	15,295	16.9	3	36	466	0
Post (work)	3,700	3.47	2	7.04	139	0
Follow (work)	9,098	8.53	2	17	169	0

			Number of words per work				
Works Data	# words	# vocablary	Mean	Median	SD	Max	Min
Story Introduction	84,565	4,085	22.86	15	33.26	972	0
Tag	9,728	249	2.63	2	2.05	8	0

3.2 Estimation Result

Figure 1 and 2 show histograms of the distribution with respect to the estimated heterogeneous coefficients for the models of network formation and content production. First, as for the network formation model, Figure 1 shows that most of users have positively estimated effect of out-degree on the following probability, while the most effect of in-degree and transitivity are negative. This is a natural interpretation for both: if a user follows many

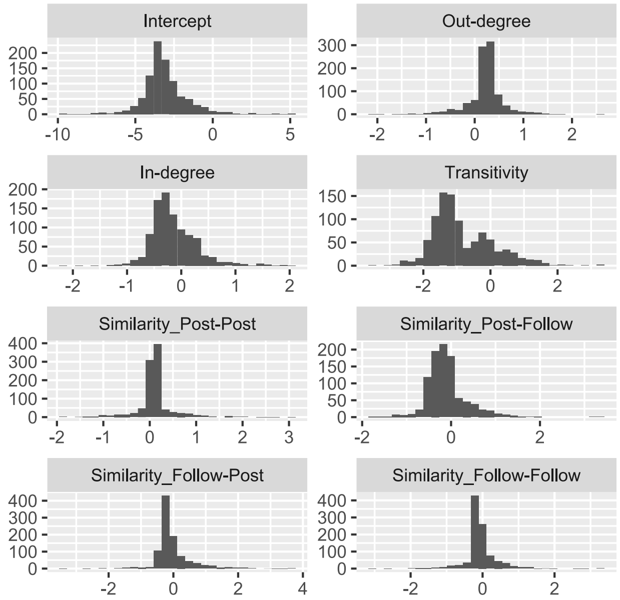

Figure 1 Histograms of heterogeneous coefficients (network formation model)

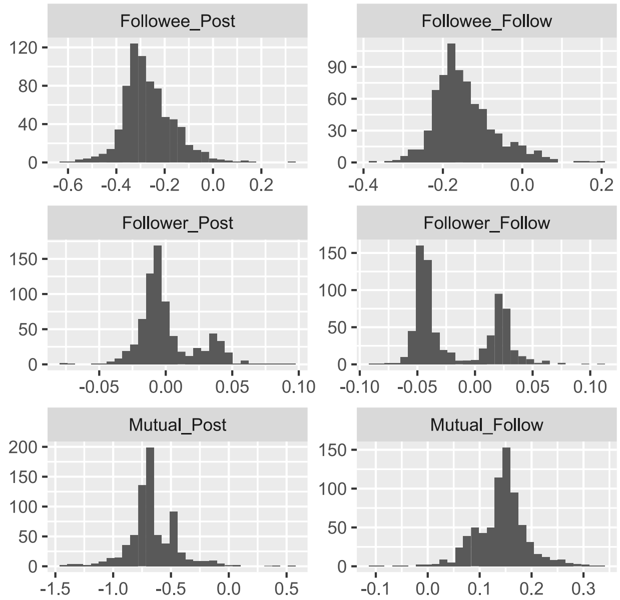

Figure 2 Histograms of heterogeneous coefficients (content production model)

others (i.e., with high out-degree), her following probability is likely to increase because the hurdle to follow new users is relatively low, while if a user has many followers (i.e., with high in-degree), her following probability tends to decrease because her works will be read by many users in the platform, responding to such as ranking page and top of search result, and then not a few users will not like that works.

Next, as for the content production model, Figure 2 show that most of coefficients are negatively estimated with respect to the effect of the topics of works that followees post and follow on topics of own works. In other words, when followees post or follow works about a certain topic, users following them tend to avoid that topic in their own works, which is consistent with the result of previous study (e.g., Bhattacharya et al. 2019) showing that users who connected in the network gradually diverge in their posting behavior over time to create a unique presence on the platform. However, this is a novel finding because our model clarified this tendency in the qualitative aspect of the behavior, i.e., the topic of the content, while the previous studies only referred to the quantitative aspect of the behavior. The coefficients for followers are smaller than those for followees, and this suggests that users who are one way follow with a weak relationship with you do not have a strong effect on your behavior. On the other hand, the effect for mutual follower are larger than those for followees, and it reflects the strength of relationship between the two users.

4 Conclusion

In this study, we proposed a methodology that estimates both the effect of behavior in the network on new network formation and the effect of other users connected in the network on the own behavior, i.e., co-evolution structure in the way of accommodating qualitative contents. Previous studies have developed co-evolution model considering only numerical information, but the qualitative aspects of the behavior, particularly how the topics of content change by forming a new network relationship, or the effect of similarities regarding the topics of content on network formation, remained unclear. This study contributes to the statistical modeling literature of co-evolutional structure between network formation and content production. Furthermore, the empirical analysis using a real-world dataset of Japanese story-telling platform has uncovered the significant bi-directional effect of network formation and content production by considering qualitative aspects of contents produced by users.

Reference

- Ameri, M., Honka, E., & Xie, Y. (2022). From strangers to friends: Tie formations and online activities in an evolving social network. *Journal of Marketing Research*.
- Bhattacharya, P., Phan, T. Q., Bai, X., & Airoldi, E. M. (2019). A coevolution model of network structure and user behavior: The case of content generation in online social networks. *Information Systems Research*, 30(1), 117-132.
- Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3(4-5):993–1022.
- Snijders T, Steglich C, Schweinberger M (2007). Modeling the coevolution of networks and behavior. van Montfort K, Oud H, Satorra A, eds. *Longitudinal Models in the Behavioral and Related Sciences* (Lawrence Erlbaum, Mahwah, NJ), 41–71.
- Steglich, C., Snijders, T. A., & Pearson, M. (2010). Dynamic networks and behavior: separating selection from influence. *Sociological methodology*, 40(1), 329-393.