TOHOKU UNIVERSITY

A Topic Model Using Text Information on Social Media for Social Network Analysis

Mirai Igarashi[1], Nobuhiko Terui[2]

[1]Graduate School of Economics and Management, Tohoku University; mirai.igarashi.s7@dc.tohoku.ac.jp

[2]Graduate School of Economics and Management, Tohoku University; terui@econ.tohoku.ac.jp

1. INTRODUCTION

- **Overview**

In modern social media development, we should properly analyze social network using not only network information but also text information. In this research, we propose topic model which takes these information into account, then construct weighted network each topic by model parameters, finally show that influential users on the network can be detected according to some centrality measures.

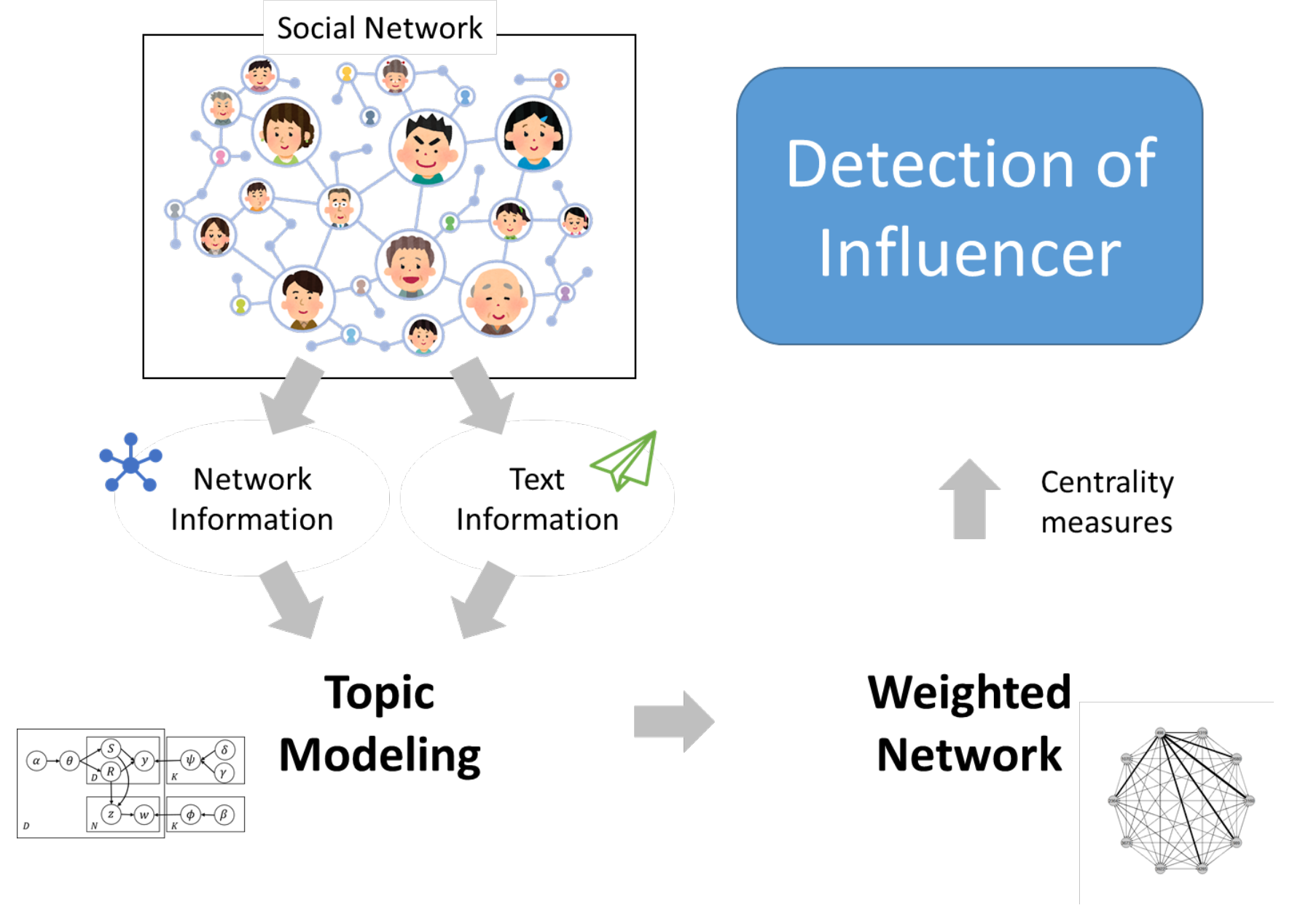


Figure 1: Overview of this research

- **Previous Research**

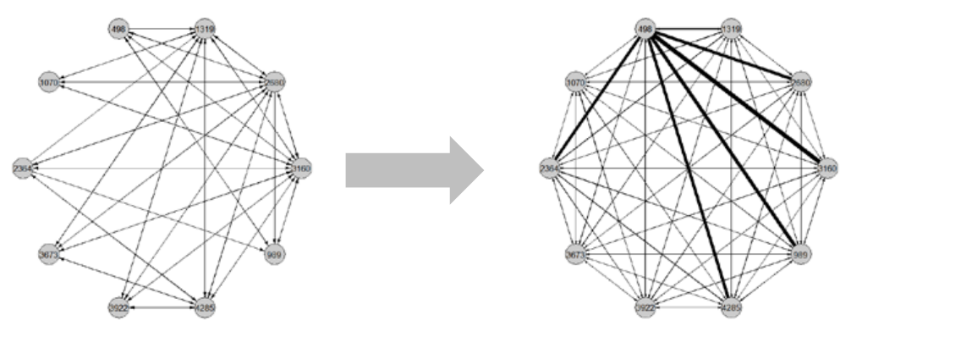
In the field of sociology or marketing, researchers argue that social relations have strength and weakness, and propose the weighted network models.

Issues that unsolved in previous research are handling of non-link, link with dissimilar person, and use of text information on social media.

- **Handling of Non-link**

Many researchers (e.g. Chen et al. 2017) give zero weight for all non-links on weighted network. But, such links should have potential or future influence, so it is necessary to construct weighted network model which gives different values for all links (link and non-link).

MMSB (proposed by Airoldi et al. 2008), which suppose latent topics for observed link relationships, is one of the solution, we also construct weighted network based on the concept of MMSB.



- **Link with Dissimilar Person**

In sociology or marketing, on the basis of homophile, it is thought that similar people tend to link together, dissimilar people unlikely to link. In modern social media, however, where people have social connections with various people all over the world, it is easy to imagine that we are connected to dissimilar people without common hobbies or circumstances and influenced from them. We propose the model that do not give low weight for all links with dissimilar person in accordance with the link probability among topics.

- **Use of Text Information on Social Media**

In previous research, weights of network are estimated by not only the link relationships but also the characteristics of each node. We consider that text information posted on social media can be used as what represents the characteristics of nodes, because users with various hobbies and circumstance link with people in accordance with what they are interested in or what topics they post. Our proposed topic model take two kinds of information into account, network and text information.

2. MODEL

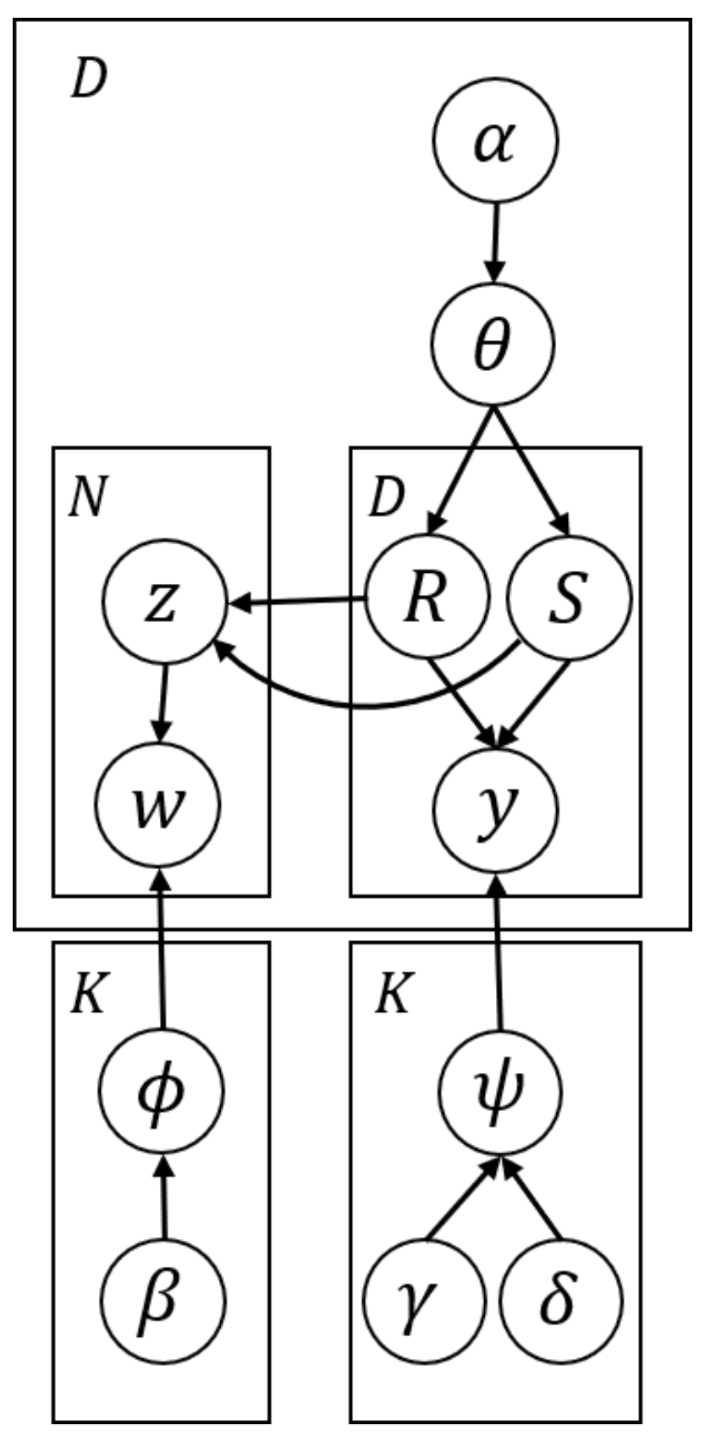


Figure 2: Proposed model

- **Proposed Topic Model**

Proposed topic model take two kinds of information into account, network and text information. Our model is based on the MMSB (Airoldi et al. 2008) concept for the network, the LDA (Blei et al. 2003) concept for the text, and the CTM (Blei and Jordan 2003) concept for the combination between two kinds of topics.

Users on the network have the unique topic distribution θ_d, link topics $S_{dd'}, R_{dd'}$ are assigned to the link according to θ_d. Word topic z_{dn} is decided in accordance with the ratio of link topics.

$$S_{dd'} \sim Categorical(\theta_d),$$
$$R_{dd'} \sim Categorical(\theta_{d'}),$$
$$z_{dn} \sim Categorical\left(\frac{N_{d1}}{2(D-1)}, \dots, \frac{N_{dK}}{2(D-1)}\right)$$

The link $y_{dd'}$ is generated according to binomial distribution with link topics and link probability $\psi^{(d\prime)}$, the word w_{dn} is generated according to categorical distribution with word topic and word distribution ϕ.

$$y_{dd'} \sim Binomial\left(\psi^{(d')}_{S_{dd'}R_{dd'}}\right),$$
$$w_{dn} \sim Categorical\left(\phi_{z_{dn}}\right)$$

- **Constructing Weighted Network**

Our model estimate the probability that the link is generated. We regard the probability as the strength of relationship between users, that is, the higher the probability between user $d \rightarrow d'$, the bigger the influence of information sent by d' to d is.

In this research, the strength of relationship between users on topic k is defined by product of the similarity of topic distributions and link probability, as follows.

$$x^{(k)}_{dd'} = \sum_{k=1}^{K} \theta_{dk}\theta_{d'k'}\psi^{(d')}_{kk'}$$

- **Detecting Influencers from Network**

Borgatti (2005) argues that two centrality measures, degree centrality and eigenvector centrality, are suite for grasping the influence on network. Given the weighted network on topic k, these two measures are defined as follows.

$$deg^{(k)}_d = \sum_{d'=1}^{D} x^{(k)}_{dd\prime}$$
$$ev^{(k)}_d = v^{(k)}_d, \qquad x^{(k)}v^{(k)} = \lambda^{x^{(k)}}_{max} v^{(k)}$$

3. ANALYSIS

- **Data**

We use Twitter data collected for this research. Target users are those who follow Nintendo of America's Twitter account (1st-step) and those who follow 1st-step users (2nd-step), in total 5,028 users. Text data is gathered from their tweets, and the number of vocabulary is 19,805.

Model parameters are estimated by collapsed Gibbs sampling. The number of topics with smallest perplexity is selected from 5, 10, 15, 20, and 25, 15 is selected.

- **Link Prediction**

Given estimated parameters from training data (90%), the link predictions of test data (10%) are followed by this equation.

$$p\left(y_{dd'} = 1 | \theta_d, \theta_{d'}, \psi^{(d')}\right) = \theta_d \psi^{(d')} \theta^T_{d'}$$

Figure 3 represents ROC curves (and AUC) of training and test data, and Table2 represents confusion matrixes when cut-off which classifies link and non-link is set as the value with maximized MCC. These results show that our proposed model is fitted well to the network situation on Twitter.

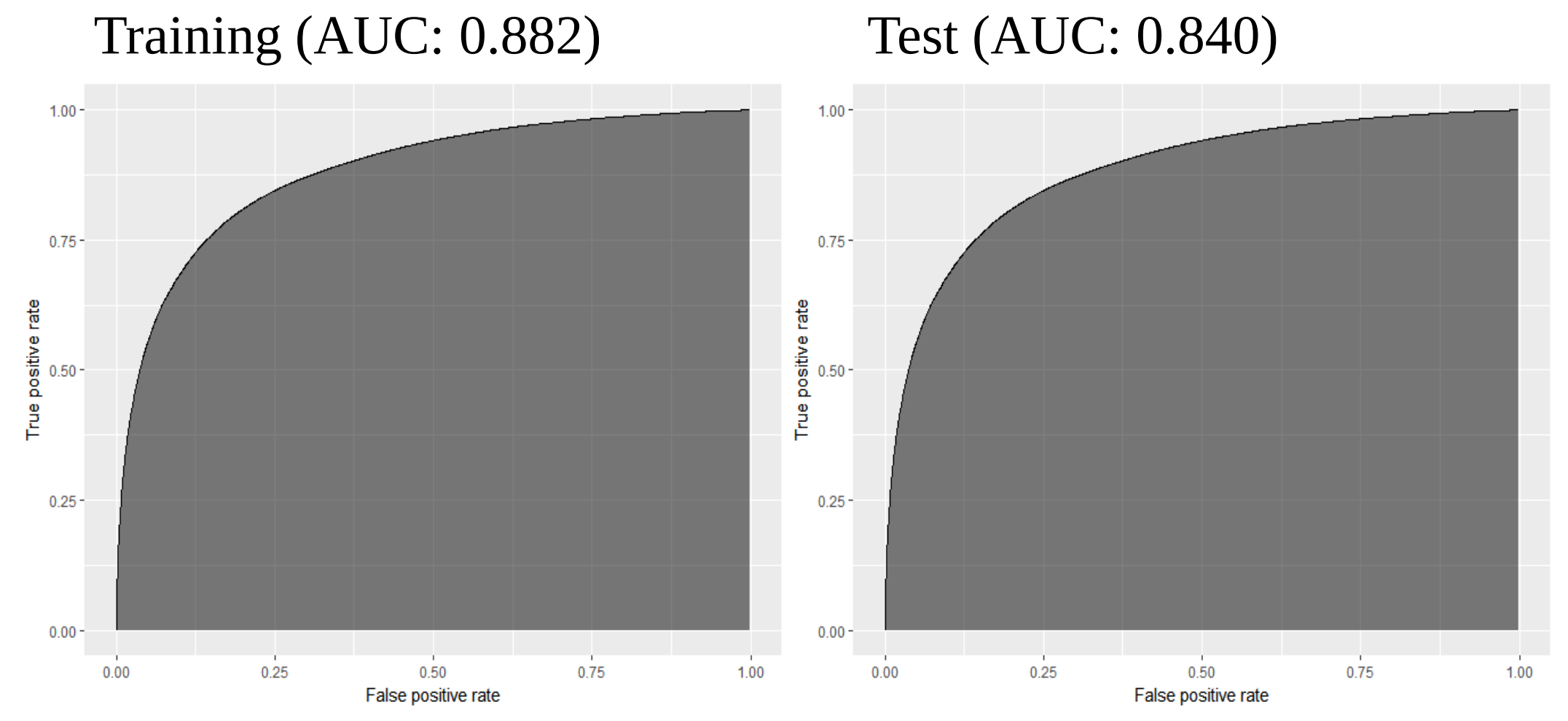


Figure 3: ROC curves

Table 2: Confusion Matrices

Training		Predict	
		Link	Non-Link
Data	Link	20,467	47,965
	Non-Link	250,825	22,427,415

Cut-off: 0.0467
Accuracy: 0.987
F1: 0.120
MCC: 0.145

Test		Predict	
		Link	Non-Link
Data	Link	1,725	5,659
	Non-Link	20,371	22,427,415

Cut-off: 0.0553
Accuracy: 0.990
F1: 0.117
MCC: 0.131

- **Diffusion Simulation**

We simulate the comparison of the performance of influencers (degree influencers and eigenvector influencers) detected from weighted network with binary network.

The simulation results (Topic 4, 6, 9) are showed in figure 4, in eigenvector influencer's simulation, those who detected from weighted network have more influence. This result means that in the situation, a user who followed by influential users is further regard as influencer, we should consider weighted network.

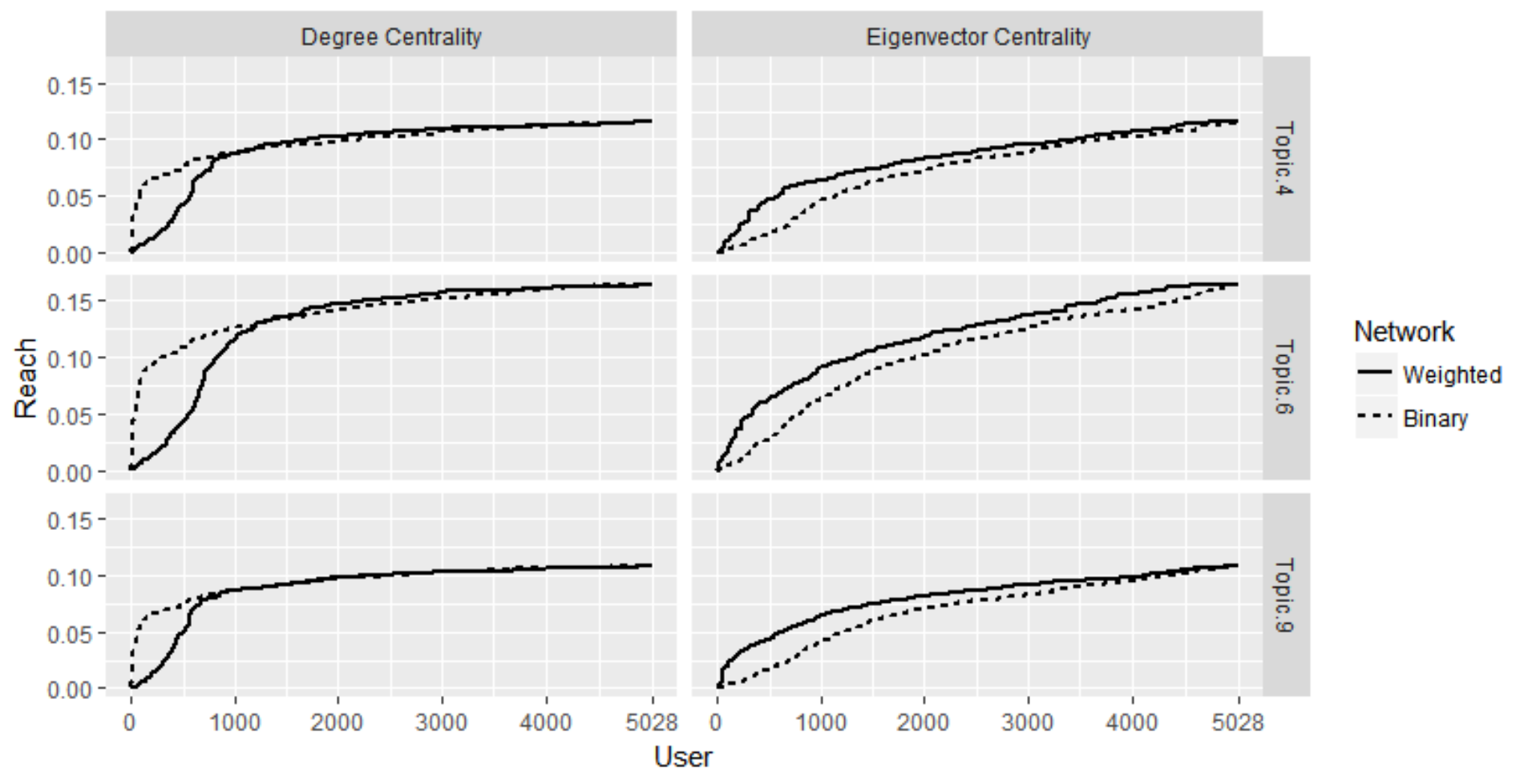


Figure 4: Simulation result

4. NEXT STEP

- **Application for Diffusion Model**

Structure of the network affects the diffusion of products and information, many researchers tackle to the modeling of diffusion process which regard weighted network as main variables. Proposed weighted network also can be applied this area and should be compared to previous models. To do so, we need to collect data tracking the diffusion process and estimate the influence of proposed networks of each topic on the diffusion from real data.

5. BIBLIOGRAPHY

1. Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008), "Mixed Membership Stochastic Blockmodels," *Journal of Machine Learning Research*, 9(Sep), 1981-2014.
2. Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003), "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, 3(Jan), 993-1022.
3. Blei, D. M. and Jordan, M. I. (2003),"Modeling Annotated Data," In *Proceedings of the 26th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, 127-134.
4. Borgatti, S. P. (2005), "Centrality and Network Flow," *Social Networks*, 23(3), 191-201.
5. Chen, X., van der Lans, R., and Phan, T. Q. (2017), "Uncovering the Importance of Relationship characteristics in Social Networks: Implications for Seeding Strategies," *Journal of Marketing Research*, 54(2), 187-201.