Detecting Community and Topic Co-Evolution in Social Networks

Juan Bi*[1], Zhiguang Qin[1], Jia Huang[2]
[1]School of Computer Science and Engineering, University of Electronic Science and Technology of China,
No.2006, Xiyuan Avenue, Chengdu, 611731, China
[2]China Science Publishing & Media Ltd (Science Press),
No.16 Sanse Road, Chengdu, China, 610061, China
Corresponding author, e-mail:bijuan66@gmail.com*, qinzg@uestc.edu.cn, huangjia@mail.sciencep.com

Abstract

In this paper we study how to discover the co-evolution of topics and communities over time in dynamic social networks. We present a topic model-based approach that automatically captures the dynamic features of communities and topics evolution. Our model can be viewed as an extension of the LDA model with the key addition that it can not only detect communities and topics simultaneously but also work in an online fashion. Instead of modeling communities and topics in statistical manner, the proposed model can simulate the user's interests drifting at different time epochs by taking into consideration the temporal information implied in the data, and observe how the community structure changes over time with the evolution of topics. Experiments on real-world data set have proved the ability of this model in discovering well-connected and topically meaningful communities and the co-evolution pattern of topics and communities.

Keywords: *community discovery, LDA, probabilistic generative model, social networks*

1. Introduction

The rapid development of online social networks has tremendously changed the way of people to communicate with each other. A lot of user-generated content is available on these online social networks. The rich source of text information can be exploited to extend the traditional social graph. Specifically, incorporating both linkage structure and text informant can provide a unique ability of detecting latent social structure among group of users. In this paper, we address the problem of automatically discovering latent communities of users from observed textual content and their relationships.

The study of community structure in networks is primarily based on the graph partitioning algorithm [1-6] and probabilistic model. The method presented in [1] is based on agglomerative algorithm where edges are removed from the network iteratively to split it into communities. These methods are purely based on graph partition algorithm, and they fail to account for other node attributes and communication content information.Meanwhile, the probabilistic generative models [7-10] have been gained significant attention in recent years. SSN-LDA [9] defines community as a distribution over the social link space. LDA-G [10] simply adapts the original LDA model for community discovery in a social graph, they merely consider the link structure in a graph. Several methods for analyzing the evolution of topics in large-scale corpora have been proposed [11-17]. These include the Dynamic Topic Model (DTM) [12], the Continuous Time Dynamic Topic Model (CTDM) [13] and Topic over Time (TOT) [14].

In this paper, we propose a probabilistic topic model to detect latent communities in a social network based on semantic information and the social relationships between users. In contrast to the previous works, the approach naturally allows the topic model to work in an online fashion. In such a way the user's interests drifting at different time epochs can be observed, and the evolution of topics, in turn, determine the changes of communities' structure and their topical features over time. In our work, we consider community and topic as different latent variables. The model cannot only discover communities and topics simultaneously, enable them to benefit each other, but also track the evolution of discovered communities and topics over time, which is useful in understanding the dynamic features of social networks.

Received October 17, 2013; Revised December 26, 2013; Accepted January 18, 2014

2. Proposed Method

Definition 1 (Topical community). A topical community is a group of users with more similar communication interests and stronger relationship strength between them within the group than between groups.

Definition 2 (time-stamped social graph). Let $\mathrm{G_t}$= (*U, E,X,W*) be the directed and weighted social graph at epoch *t*, where *U* is the user set in $\mathrm{G_t}$ and E is the link set where $e_{ij} \in E$ denotes a directed link from user*i* to user*j* which corresponds to a relationship between*i* and *j.X* is the set of weights where x_{ij} is the weight of the link e_{ij}. We denote the weight as the strength of relationship from user *i* to user *j.W* is the set of user-generated texts.

Definition 3 (Relationship strength). In our work, the relationship strength is the intensity of interactions such as mention, retweet between two connected users. We assume the stronger the relationship, the more number of interactions will take place between two users

Definition 4 (time-stamped documents). A collection of user-generated content are assumed to be divided into so called "epochs". The content generated by user *u* at the current epoch *t* is represented by$\mathrm{w_{t,u}} = \{\mathrm{w_{t,u,n}}\}_{\mathrm{n=1}}^{\mathrm{N_{t,u}}}$, *i.e.* the set of words in the content.We assume that the epoch *t* is a discrete variable, a time epoch can be a day, a month, or a year.

2.1. Model

The graphic model representation of our model which we present in this paper is illustrated in Figure 1. In this model, the mixture components, *i.e.* communities and topics, are shared explicitly across all time epochs, but the mixing weights of each component evolve over time, for example, some topics may become more popular while others may become outdated.

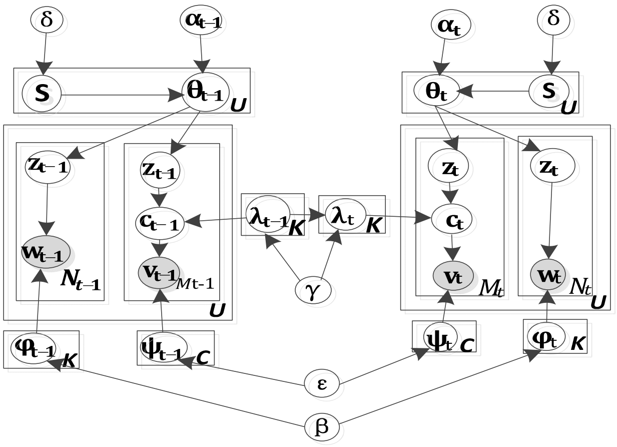


Figure 1. Graphical Model Representation of the Proposed Dynamic Model

At time epoch *t*, the proposed model consists of two parts. First, we model the interests of each user in the corpus. Specifically, we represent each user as a multinomial distribution over topics θ_t, thus each word written by the user is generated from one topic selected from the distribution. In order to model the evolution of topic, we assume that the current topic at epoch*t* can be generated in two ways, either depending on the topic distribution of the previous time epochs or being not influenced by historical information but current status. In the model, we use a parameter *s* to control the influence situation. The *s* is generated from a Bernoulli distribution whose parameter is δ. When *s* = 0, a new topic distribution will be sampled from symmetric Dirichlet distribution *α*. When *s* = 1, it means user's current interest is determined by his previous status. In this case, we assume topic smoothly changes from time *t-1* to *t*. A topic with a higher mixture weight at the current epoch is more likely to have a higher weight in the next epoch. Motivated by dHDP [15], the priors of a topic *z* at epoch *t* can be constructed as follows:

$$\alpha_{t,z} = s_t \sum_{\tau=1}^{t-1} exp\{-\eta(t-\tau)\} n_{\tau,z} \times (1-s_t)\alpha$$

Where *η* is a smooth parameter, $n_{t,z}$ is the number of words assigned to topic *z* at time epoch *t*. The second stage of the generative process is derived the community membership for user depends on this user's topics. Hence, we select a community assignment for a specific user from the topic-community distribution *λ* and finally each link (interaction) of this user generated from the community-specific distribution.

Thus, the generative process for time epoch *t* of the proposed model is given as follows:

a) For each community $c_t \in$ *C*, draw a multinomial distribution $\vec{\psi}_{c_t}$~ Dirichlet *($\vec{\gamma}$)*
b) For each topic $z_t \in$ *K*, draw a multinomial distribution $\vec{\varphi}_{z_t}$~ Dirchlet($\vec{\beta}$)
c) For each topic $z_t \in$ *K*, draw influence probabilityδ_t ~ Beta(π)
d) For each user *i* ∈ *U*:
   (1) Draw an influence indicator $s_{t,i}$~ Bernoulli (δ_t)
   (2) Draw a multinomial topic distribution$\vec{\theta}_{t,i}$~Dirichlet ($\vec{\alpha}_t$)
   (3) For each topic$z_t \in$ *K*, draw community distribution$\vec{\lambda}_{t,z}$~ Dirichlet($\vec{\varepsilon}\vec{\lambda}_{t-1,z}$)
   (4) For each word $w_{t,i,j} \in N_{t,i}$ associated with user *i*:
      a. Draw a topic $z_{t,i,j}$~ Multinomial($\vec{\theta}_{t,i}$)
      b. Draw a word $w_{t,i,j}$~ Multinomial($\vec{\varphi}_{z_{t,i,j}}$)
   (5) For each link$v_{t,i,j} \in M_{t,i}$ for user *i*:
      a. Draw a topic $z_{t,i,j}$~ Multinomial($\vec{\theta}_{t,i}$)
      b. Draw a community $c_{t,i,j}$~Multinomial($\vec{\lambda}_{t,z_{t,i,j}}$)
      c. Draw a link $v_{t,i,j}$~Multinomial($\vec{\psi}_{c_{t,i,j}}$)

The graphical model representation is shown in Figure 1 where the gray circles correspond to observed variables of textual information and link information respectively. Others denote the latent variables and parameters. This generative model represents content information as a mixture of topics and link information as a mixture of communities. At epoch *t*, wefirst generate topic for each word for a specific user *u* from multinomial $\theta_{t,u}$, and the topic for each word generated by this user represents the interests of him. Then we generate the community assignment for this user depending on the user and the topics which the user is really interested in. $\lambda_{t,z}$represents the community distribution for topic *z*. The communitymembership of a user is derived from the topic's community mixture. In other words, users who share series of topics with each other should be members of the same community. The link information was assumed to be random mixture over communities, and each link of a user was finally generated from the community-specific distribution.

At first epoch *t* = 1, the topic distribution$\theta_{1,u}$ is drawn from a Dirichlet prior *α*, and the topic-community distribution $\lambda_{1,z}$ is drawn from a Dirichlet priorε, where *α* and *ε* are initialized to symmetric constant, as done in original LDA modeling.

Formally, let *Z* and *C* be the set of latent topics and latent communities respectively, *W* be the set of words in the corpus, *V* be the set of interactions that were observed on the social graph. The joint probability on the texts, links and the latent variables at epoch *t* is given by:

$$P\left(W_{t,}V_{t,}Z_{t,}C_{t,}S_t|\alpha,\beta,\varepsilon,\gamma,\alpha_{t,}\delta_t\right)$$

$$= P\left(W_t|Z_t;\varphi_t\right)P\left(V_t|C_t;\psi_t\right)P(Z_t|\theta_t)P(C_t|Z_t,\Lambda_t)P(\Lambda_t|\Lambda_{t-1},\varepsilon) \times P\left(\varphi_t|\beta\right)P\left(\psi_t|\gamma\right)P(\theta_t|\alpha_t,S_t = i)$$

$$= \int\prod_{i=1}^{U}\prod_{n=1}^{N_{t,i}} p(w_{t,i,n}|\vec{\varphi}_{z_{t,i,n}})\prod_{z=1}^{K} p(\vec{\varphi}_{z_t}|\vec{\beta})\, d\varphi \times \int\prod_{i=1}^{U}\prod_{n=1}^{M_{t,i}} p(v_{t,i,n}|\vec{\psi}_{c_{t,i,n}})\prod_{c=1}^{C} p(\vec{\psi}_{c_t}|\vec{\gamma})\, d\psi$$

$$\times \int\prod_{i=1}^{U}\prod_{n=1}^{M_{t,i}} p\left(z_{t,i,n}\middle|\vec{\theta}_{t,i}\right) p\left(c_{t,i,n}\middle|\vec{\lambda}_{i,z_{z_{t,i,n}}}\right)\prod_{i=1}^{U}\prod_{z=1}^{K} p\left(\vec{\lambda}_{t,z}\middle|\vec{\lambda}_{t-1,z},\vec{\varepsilon}\right) d\lambda$$

$$\times \int\prod_{i=1}^{U}(\prod_{n=1}^{N_{t,i}} p\left(z_{t,i,n}\middle|\vec{\theta}_{t,i}\right) p\left(\vec{\theta}_{t,i}\middle|\vec{\alpha_t},s_{t,i}\right))\ d\theta \times \prod_{i=1}^{U} p\left(s_{t,i}|\delta_t\right) p(\delta_t|\vec{\pi})d\delta$$

2.2. Parameter Estimation

We adopt the collapsed Gibbs sampling, a stochastic approach for approximate inference in high-dimensional models. We need to derive $p(c_i = c|Z_t, C_{t,-i,}V_t, W_t, \theta, \varphi, \lambda, \psi)$ and $p(z_i = z|Z_{t,-i,}C_t, V_t, W_t, \theta, \varphi, \lambda, \psi)$, the conditional distribution of a community and topic based on all other variables. In particular, the conditional distribution of the topic assignment (when $s_t = 1$) is given while the other case (when $s_t = 0$) is omitted due to the space limited.

$$p(z_i = z|C_t, Z_{t,-i,}V_t, W_t, S_t\theta, \varphi, \lambda, \psi) = p(c_i = c, z_i = z, w_i = w, s_t = 1)|Z_{t,-i}, C_{t,-i}, W_{t,-i}, \theta, \varphi, \lambda, \psi)$$

$$= \frac{n_{-i,(t,z)}^{(w)} + \beta}{n_{-i,(t,z)}^{(.)} + T\beta} \times \frac{n_{-i,(t,z)}^{(c)} + \varepsilon\lambda_{t-1,z}}{n_{-i,(t,z)}^{(.)} + C\varepsilon} \times \frac{n_{-i,(t,u)}^{(z)} + \alpha_{t,z}}{\sum_{z=1}^{K}(n_{-i,(t,u)}^{(z)} + \alpha_{t,z})} \times \frac{N_{-i,(t,z)}^{(1)} + \eta}{N_{t,z}^{(1)} + 2\eta}$$

Where $n_{-i,(t,z)}^{(w)}$ is the number of times of word*w*assigned to topic *z* at epoch *t*, excluding the current word *i*. $n_{-i,(t,z)}^{(.)}$ is the total number of words assigned to topic *z* at epoch *t* excluding current word *i*. Similarly, $n_{-i,(t,z)}^{(c)}$ is the number of times of community *c* sampled from topic *z* at epoch *t*, not including the current community. $n_{-i,(t,u)}^{(z)}$ is number of words generated by user *u* at epoch *t* assigned to community topic *z,* not including the current one. The last term measures the probability of having the influence indicator variable *s* equal to 1. Further, the conditional distribution of a community assignment is given by:

$$p(c_i = c|Z_t, C_{t,-i,}V_t, W_t, \theta, \varphi, \lambda, \psi) = p(c_i = c, v_i = v|C_{t,-i}, V_{t,-i}, \theta, \varphi, \lambda, \psi) = \frac{n_{-i,(t,c)}^{(v)} + \gamma}{n_{-i,(t,c)}^{(.)} + E\gamma} \times \frac{n_{-i,(t,z)}^{(c)} + \varepsilon\lambda_{t-1,z}}{n_{-i,(t,z)}^{(.)} + C\varepsilon}$$

Where $n_{-i,(t,c)}^{(v)}$ is the number of times of user *v* assigned to community *c* at epoch *t*, not including the current user. $n_{-i,(t,c)}^{(.)}$ is the total number of users assigned to community *c* at epoch *t*, not including the current one.

Finally, the multinomial parameters $\theta_{(t,u),z}$, $\lambda_{(t,z),c}$, $\varphi_{(t,z),w}$,$\psi_{(t,c),v}$ are obtained as follows:

$$\theta_{(t,u),z} = p\{z|u\} = \frac{n_{t,u}^{(z)} + \alpha_{t,z}}{\sum_{z=1}^{K}(n_{t,u}^{(z)} + \alpha_{t,z})} \quad \varphi_{(t,z),w} = p\{w|z\} = \frac{n_{t,z}^{(w)} + \beta}{n_{t,z}^{(.)} + T\beta}$$

$$\lambda_{(t,z),c} = p\{c|z\} = \frac{n_{t,z}^{(c)} + \varepsilon\lambda_{t-1,z}}{n_{t,z}^{(.)} + C\varepsilon} \quad \psi_{(t,c),v} = p\{v|c\} = \frac{n_{t,c}^{(v)} + \gamma}{n_{t,c}^{(.)} + E\gamma}$$

3. Results and Analysis

3.1. Dataset Description

Here, we present the data collected from Twitter. Since our goal is to explore the relationship between user' interests and their interactions in the social network, we need to collect information about users, content and link structure. The content in Twitter refers to tweets. And we connect two users only if an interaction took place between them via mention actions (*@user name*) or retweet actions (*RT*), each link weighted by counting the number of times these actions have taken place between the two users. All the data is collected via Twitter API from July 1, 2012 to October 31, 2012. We applied pre-processing to tweets content by removing non-English tweets, punctuations and stop words. We also excluded a small number of short tweets, in which less than ten words remained in the bag of words after the stop-words had been removed. Finally, our collection contains 3054 users, 183675 links and 137633 distinct words. For simplicity, the unit epoch was set to one month, so there were 4 epochs (i.e. July, August, September and October).

3.2. Experiment Results

Our model is evaluated in three problem domains: the evolution of topics, the evolution of communities, and the dynamic relationship between topic and community. *Z* and *C*, the number of topics and the number of communities respectively are fixed and shared across all time epochs. We set the number of communities *C* at 10 and topics *Z* at 20.

3.2.1. Topic Trend Analysis

To analyze the evolution of the topics over time, e.g. whether they are emerging or declining, we calculate the topic popularity along four time epochs. The more users who communicated on a topic, the more popular the topic is. Because each user is interested in each topic with a different degree, the popularity of topic *z* at epoch *t* is formally defined as:

$$popularity(z,t) = \frac{1}{|U|}\sum_{u\in U}\theta_{t,u}[z]$$

Where $\theta_{t,u}$ is the topic distribution for user *u* at epoch *t*, which indicates the level of participation of each user in each topic.

In Figure 2, we present the mixing proportion of topics at each epoch. Each topic is represented by a stripe. The width of a stripe corresponds to the popularity of the corresponding topic over time. The wider the stripe, the more popular the topic is. From Figure 2, we find that the popularity of most topics in each epoch varied smoothly. Since the mixture proportion for each topic may be influenced by the history topics information, the users' topic (interest) may not change too much between adjacent epochs; but after a long time, interests may drift.

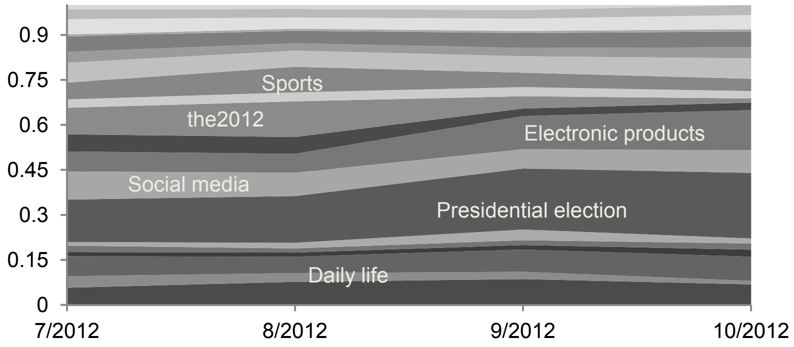


Figure 2. Overview of the *E*volution of *T*opics

3.2.2. Community Evolution Analysis

In our model, a community assignment of a user is dependent on the user and his inherent interest (latent topics). Therefore, the community structure and its topic distribution also change according to the topic evolution over time. In order to reveal the hidden evolution patterns of communities, we selected two communities for the comparative analysis.

For each community, we can get the topic distribution at each epoch. The topical interest of each community can be described by the most occurring topics in it. Those topics correspond to the most representative topics for the selected community. Formally, given a selected community c, the set of most important topics $\mathrm{Repr}_{\mathrm{z,c}}$ can be computed as:

$$Repr_{z,c} = n\, \underset{z\in[1:K]}{arg\, max}\left\{\sum_{t=1}^{4}\lambda_{t,z,c}\right\}$$

Where *n* argmax denotes the function returning the n topics with the highest values. On this way, we can have an overview of topics in the community.

In Table 1 we give top five topics (n = 5) and their corresponding key words for each community. For example, the dominant topic in *community 1* is *topic7* (recall that *topic7* is about presidential election) and the topmost topic in the *community 4* is *topic13* (sports).

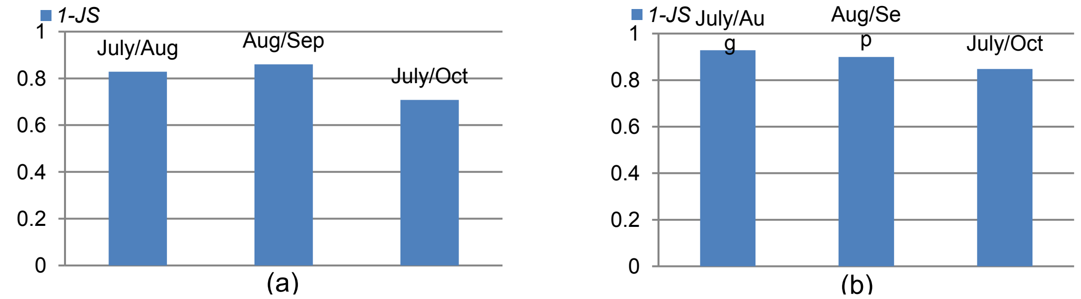
Table 1. Top Five Topics in Community 1 and Community 4

Community 4									
Topic 13	0.1882	Topic 18	0.1081	Topic 11	0.0674	Topic 7	0.0604	Topic10	0.0412
nfl	0.1891	fishing	0.0345	Olympics2012	0.0561	vote	0.05618	teaching	0.0248
football	0.0530	Friday	0.0237	BBC	0.0275	presidential	0.02753	education	0.0214
player	0.0173	weekend	0.01774	Soccer	0.0117	cnn	0.01173	academic	0.0122
shot	0.0096	night	0.00843	beer	0.0082	election	0.00823	math	0.0105
espn	0.0935	great	0.00504	succeed	0.0076	campaign	0.00766	examine	0.07810
Community 1									
Topic7	0.2318	Topic 9	0.0981	Topic 8	0.0708	Topic 11	0.06782	Topic13	0.0532
president	0.08305	iPhone	0.05432	Twitter	0.02883	London	0.0223	game	0.0225
Romney	0.07292	Apple	0.05396	Social	0.02573	Olympics	0.0117	espn	0.0117
election	0.02528	iPhone5	0.04782	Facebook	0.02225	USA	0.00845	shot	0.0085
speech	0.02065	launch	0.03440	Youtube	0.01634	Live	0.00762	winning	0.0076
presidential	0.01455	mobile	0.02238	Google	0.01239	wining	0.00703	dead	0.0574

To have a clear insight into the evolution in each community over time, we furtherleveraged the *JS* (Jensen-Shannon) divergence to measure the similarity between communities generated at different epochs. *JS (p || q)* represents the dissimilarity between two probability distributions *p* and *q*, which is defined as:

$$\square\square(\square,\square) = \frac{1}{2}[\square\square(\square,\square) + \square\square(\square,\square)]$$

Where *KL (p, q)* is the Kullback-Leibler divergence and m = 0.5(p + q). To compute the JS-divergence between two communities, we represent each community as a vector of probabilities over topics $\{\square_{\square,\square}\}_{\square=1}^{\square}$and a vector of probabilities over links$\{\square_{\square,\square}\}_{\square=1}^{\square}$. The topic similarity and link structure similarity of *community 1* between different time epochs were calculated and displayed in Figure 3(a) and (b), respectively. All of the results exhibit ahigh similarity between two contiguous time periods. Especially, link structure of *community 1* does not change significantly for the entire time epochs. This may be because there are strong evolution dependencies between these epochs. By constructing the priors as a weighted combination of the history information, the distribution of each component at epoch *t* is influenced by its past distribution. Consequently, community structure and topic between adjacent epochs may stay the same or change smoothly. However, the similarity between epoch *1* and epoch *4* is relatively smaller than others, which means the community has changed after a long time.



(a) (b)

Figure 3. Similarity Comparisons of Community 1 in (a) Topics, (b) Link structure over all epochs

To better understand the evolution behaviors as above illustrated, the evolutionary process of topics profile for a specific *community 1* over all epochs is presented in Figure 4. Topical peaks for a community indicate the dominant topics for that particular community. We can see that *topic7* is very prominent in *community 1* across all epochs. This has been expected because topic7 is related to the "US presidential election in 2012", which is the most dominant and widely discussed topic in the selected dataset. However, topics can rise and fall in prominence. It is not necessary for every topic distribution to stay the same at different evolutionary epochs. *Topic11* ("the2012 Olympics") is clearly identified in *community 1* by our model, and our model correctly shows its rise and fall in prominence during the four epochs. These analysis results demonstrate that our model can not only model the temporal evolution of topics over time based on historical information, but also capture the emerging topics during the evolutionary process, which can be done by sampling the influence indicator s.

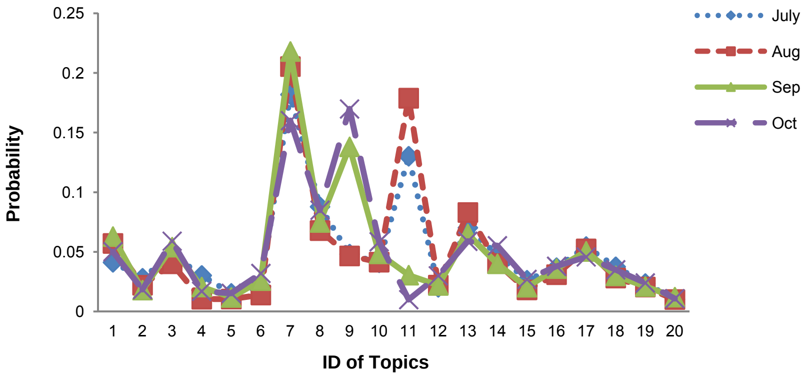


Figure 4. Topic Evolution of Community 1 Over All Epochs

3.2.3. Perplexity Analysis

Perplexity is a common criterion for evaluating the quality of clustering. It measures the predictive performance and the ability of a model to generalize to unseen data. The higher the predictive performance is, the lower the perplexity will be, and hence, better generalization performance can be achieved. We compute the perplexity of observing both link structure and words.

We divided the data into training set *D* and test set$\widetilde{\square}$randomly. Let *N* be the size of training set and *M* the size of test set. Formally, the perplexity of a test set at epoch t given the training set is:

$$Perplexity(t) = exp\left(-\frac{log\,P\left(\widetilde{D}^t\middle|Model,D\right)}{|\widetilde{D}^t|}\right) = exp\,(-\frac{\sum_{i=N+1}^{N+|\widetilde{D}^t|} logP(d_i{}^t|Model,D)}{|\widetilde{D}^t|})$$

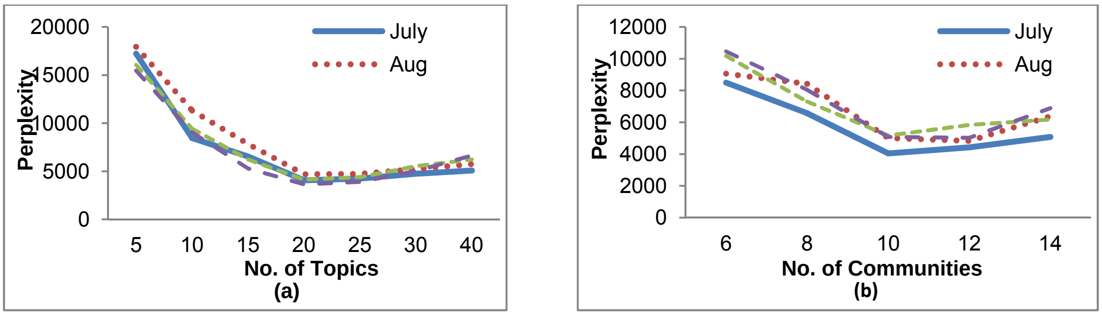


Figure 5. Perplexity Value for (a) No. of topics, (b) No. of communities

We examined the perplexity value for each epoch on different setting of topic number and community number. Figure 5(a) plots the perplexities against the number of topics, the number of communities was set to 10 for this experiment. In both epochs, the perplexities can get their minimum at around 20 topics. Figure 5(b) plots the perplexities against the number of communities. For both epochs, the perplexities get their minimum at around 10 communities.

4. Conclusion

With the advent of online social networking,various modes of communication enable users not only to create relationships with others but also to share interests by generating texts. In this paper, we present a unified probabilistic generative model that not only detect communities and topics in social networks simultaneously, but also capture the dynamic features of communities and topics evolution. This model extends prior works on community discovery by incorporating both the temporal information of relationships and the textural content generated by users. Community is detected dependent on not only the explicit links between individuals but also the topics they communicate about. The model is able to identify important consistent topics, as well as capture the emerging topics which are intensively covered only in a certain time period. The experiment results are demonstrated that the model have the capability to detect well-connected and topically meaningful communities and the co-evolution of communities and topics.

References

[1] Girvan, Michelle, Mark EJ Newman. *Community structure in social and biological networks*. Proceedings of the National Academy of Sciences. 2002; 99(12): 7821-7826.

[2] FortunatoS. Community detection in graphs. *Physics Reports*. 2010; 486(3): 75-174.

[3] Newman, Mark E, Michelle G. Finding and evaluating community structure in networks. *Physical review E*. 2004; 69(20): 26-31.

[4] Newman, Mark EJ. Fast algorithm for detecting community structure in networks. *Physical review E*, 2004; 69(6): 066133.

[5] Mingwei Leng, Jinjin Wang, Pengfei Wang, Xiaoyun Chen. Hierarchical Agglomeration Community Detection Algorithm via Community Similarity Measures. *TELKOMNIKA Indonesian Journal of Electrical Engineering.* 2012; 10(6): 1510-1518.

[6] Jian Li, Huiwen Deng. Community Structure Detection Algorithm Based on the Node Belonging Degree. *TELKOMNIKA Indonesian Journal of Electrical Engineering*. 2013; 11(12): 7649-7654.

[7] Blei David M, Andrew Y Ng, Michael I Jordan. Latent dirichlet allocation.*the Journal of machine Learning research 3*. 2003; 993-1022.

[8] Zhou, Ding, et al. *Probabilistic models for discovering e-communities.* Proceedings of the 15th international conference on *World Wide Web*. ACM. 2006; 173-182.

[9] Zhang H, et al. *An LDA-based community structure discovery approach for large-scale social networks*. IEEE Conference on Intelligence and Security Informatics. 2007; 200-207.

[10] Henderson, Keith, Tina Eliassi-Rad. *Applying latent dirichlet allocation to group discovery in large graphs.* Proceedings of the 2009 ACM symposium on Applied Computing. ACM. 2009; 1456–1461.

[11] Huang, Hsun-Hui, Horng-Chang Yang. *Semantic Clustering-Based Community Detection in an Evolving Social Network.* Genetic and Evolutionary Computing (ICGEC). Sixth International Conference on. IEEE. 2012; 91-94.

[12] Blei David M, John D Lafferty. *Dynamic topic models*. Proceedings of the 23rd international conference on Machine learning. ACM. 2006; 113-120.

[13] Wang, Chong, David Blei, David Heckerman. *Continuous time dynamic topic models.* Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence (UAI). 2008.

[14] Wang, Xuerui, Andrew McCallum. *Topics over time: a non-Markov continuous-time model of topical trends*. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM. 2006; 424-433.

[15] Pruteanu-Malinici, Iulian, et al. *Dynamic hierarchical Dirichlet process for modeling topics in timestamped documents. IEEE Trans*. PAMI 2010; 32(6): 996-1011.

[16] Iwata, Tomoharu, et al. *Online multiscale dynamic topic models*. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM. 2010; 663-672.

[17] Tang, Xuning, and Christopher C. Yang. *TUT: a statistical model for detecting trends, topics and user interests in social media.* Proceedings of the 21st ACM international conference on Information and knowledge management. ACM. 2012; 972-981.