Extracting Evolutionary Communities in Community Question Answering

Zhongfeng Zhang and Qiudan Li
The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China. E-mail: {zhongfeng.zhang,qiudan.li}@ia.ac.cn

Daniel Zeng
The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China. Department of Management Information Systems, University of Arizona, Tucson, Arizona. E-mail: zeng@email.arizona.edu

Heng Gao
The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China. E-mail: heng.gao@ia.ac.cn

With the rapid growth of Web 2.0, community question answering (CQA) has become a prevalent information seeking channel, in which users form interactive communities by posting questions and providing answers. Communities may evolve over time, because of changes in users' interests, activities, and new users joining the network. To better understand user interactions in CQA communities, it is necessary to analyze the community structures and track community evolution over time. Existing work in CQA focuses on question searching or content quality detection, and the important problems of community extraction and evolutionary pattern detection have not been studied. In this article, we propose a probabilistic community model (PCM) to extract overlapping community structures and capture their evolution patterns in CQA. The empirical results show that our algorithm appears to improve the community extraction quality. We show empirically, using the iPhone data set, that interesting community evolution patterns can be discovered, with each evolution pattern reflecting the variation of users' interests over time. Our analysis suggests that individual users could benefit to gain comprehensive information from tracking the transition of products. We also show that the communities provide a decision-making basis for business.

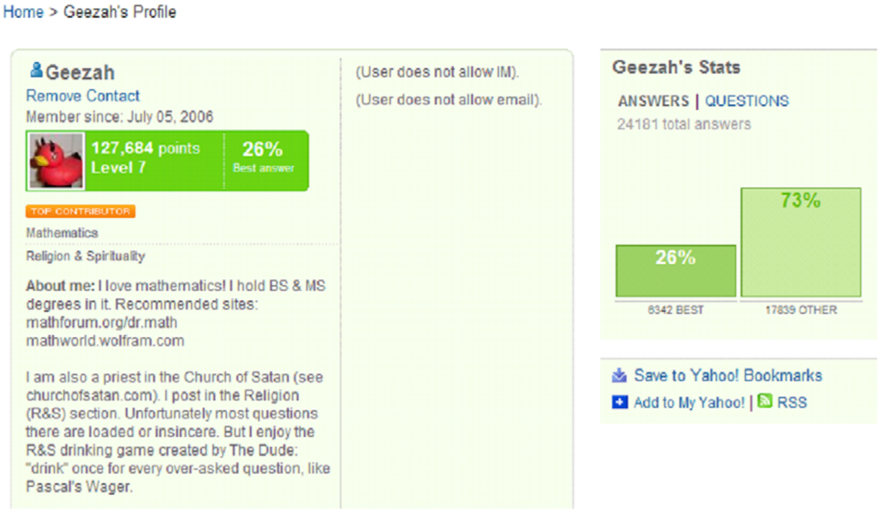
Introduction

The rapidly increasing popularity of community question answering (CQA) has made it an alternative information-seeking channel to search engines. Major search engines have released their own CQA services, such as Yahoo! Answers, Baidu Zhidao, and Naver. In a CQA service, an asker first posts a question for discussion, and then other users choose to answer questions they are interested in and familiar with. Users could also comment or vote on the quality of questions and answers. Through this process, users form interactive networks. In such networks, densely connected individuals form communities by interacting with each other. Meanwhile, communities may evolve over time, because of changes in users' interests, activities, and new users joining the network. Thus, discovering the community structures and community evolution over time is necessary to better understand user interactions in CQA. However, previous work in CQA has mainly focused on providing better service for individual users, for example, detecting the high quality content and authority users (Bian et al., 2009), searching for similar questions (Cao, Duan, Lin, Yu, & Hon, 2008), and recommending proper answers to the newly submitted question (Bian, Liu, Agichtein, & Zha, 2008), etc. Extracting user communities and mining their evolution over time are topics that have not been systematically studied in CQA.

Decision-making departments may track the communities' reaction to a pandemic or other social event for an informed policy. Communities could also act as a feedback channel to an announced policy, thus the departments could track the communities' comments and adjust their policies accordingly. An individual user may track other people's reaction to a social event and gain information about it.

Received January 19, 2011; revised April 6, 2013; accepted April 22, 2013

© 2014 ASIS&T • Published online 29 January 2014 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/asi.23003



Geezah

Member since: July 05, 2006

127,684 points
Level 7

26%
Best answer

Mathematics

Religion & Spirituality

About me: I love mathematics! I hold BS & MS degrees in it. Recommended sites:
mathforum.org/dr.math
mathworld.wolfram.com

I am also a priest in the Church of Satan (see churchofsatan.com). I post in the Religion (R&S) section. Unfortunately most questions there are loaded or insincere. But I enjoy the R&S drinking game created by The Dude: "drink" once for every over-asked question, like Pascal's Wager.

(User does not allow IM).

(User does not allow email).

Geezah's Stats

24181 total answers

FIG. 1. An example of a user profile. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

In this article, we focus on discovering communities from CQA interactive networks and analyze community evolution. Several community detection techniques have been studied in other forms of social networks, including paper citation networks and the blogosphere. Most of them are based on the assumption that an individual in a network only belongs to one community at one time. However, in many cases, a user may be engaged in several communities simultaneously. For example, in Figure 1, the user is a top contributor both in mathematics and religion & spirituality. Taking this characteristic of networks into consideration, an overlapping community structure would better capture the nature of networks. The soft clustering of communities is studied in this article to identify the overlapping community structures in social networks.

Recently, some studies have been conducted that analyze dynamic communities in social networks, and this works well in undirected networks. However, a CQA network is typically a directed network, and they simply ignore the edge directions for directed networks. It is clear that taking information on edge directions into consideration could allow us make more accurate analysis of community evolution. In CQA, the actors in the heterogeneous network involve askers, questions, answers, and answerers. We propose a novel computation model that extracts community structures with a probabilistic community model (PCM) and captures the community evolution patterns over time in directed networks. The PCM model benefits soft clustering of communities at each time step.

We empirically evaluate the performance of our method on several static networks and a dynamic network from Yahoo! Answers. Experimental results on static networks show that our probabilistic community model could improve the community extraction quality. We show that CQA communities could serve as a product feedback channel for both businesses and consumers. A consumer could rely on the communities' opinions when making an informed decision whether to buy a product. The firms could better understand users' comments on their products, and take specific measures to improve their services.

The rest of this article is organized as follows: In Related Work, we discuss relevant studies in the literature. The detailed procedure of our algorithm is presented in Community Evolutionary Discovery. In Parameter Estimation and Extension, we discuss parameter estimation for our model and introduce extensions to handle some practical issues. We empirically evaluate our algorithm in Empirical Evaluation. The Conclusion sums up our study and discusses future research directions.

Related Work

Because our work is related with CQA and community evolution, we first review previous research in CQA. Then, we focus on existing community detection and evolution tracking techniques in the literature.

Community Question Answering

The popularity of Yahoo! Answers and other online CQA systems has attracted much attention for people to mine their commercial benefit. CQA offers opportunities for some businesses and professionals to gain exposure, improve branding, and acquire website traffic. Webmasters have been trying to attract people to their websites by posting high quality answers on topics related to their websites.

The success of CQA systems has also aroused much research interest. Cao et al. (2008) studied the problem of recommending related questions to the queried question. Given a question as query, Bian et al. (2008) proposed GBrank to rank related question-answer pairs. Agichtein, Castillo, Donato, Gionis, and Mishne (2008) studied the problem of detecting high-quality content from both questions and answers. A mutual reinforcement semisupervised

DOI: 10.1002/asi

learning framework was proposed (Bian, Liu, Zhou, Agichtein, & Zha, 2009), to detect high quality questions and answers, and identify authority users simultaneously.

In a CQA site, users form interactive communities, and the communities may evolve over time. Consequently, community extraction and evolution tracking are also important issues in CQA. We suggest that the interest in CQA communities could serve as an alternative feedback channel for social events, helping decision-making departments develop informed policy and individuals gain comprehensive information.

The rapid adoption of Web 2.0 technologies and platforms, including CQA, has made available an expanding and enriched set of channels for information seeking and sharing. It not only reveals the hidden topic structures, but also reflects users' changing interests over time. The application of this type of approach is potentially broad. For instance, in business situations, CQA might be used as an alternative information and feedback channel to learn about people's interests and concerns, and help track people's reactions to both products and services. Such knowledge can lead to better informed decisions and more effective business practices (Zhang, Li, & Zeng, 2010). Our analysis suggests that individual users could gain comprehensive information from tracking the transition of products. We also show that the communities provide a decision-making basis for business.

Community Detection and Evolution

Community evolution detection involves three major tasks: community extraction, community evolution tracking, and community quality measurement. In this section, we will first review the applications of community analysis, and then we will focus on existing works for each task.

The Application of Community Analysis

Trust is a fundamental factor for a community to be attractive to its members, and is essentially relevant to the success of online communities. Geng, Whinston, and Zhang (2004) studied the issue of trust in an electronic community from a dynamic process perspective, and developed an evolutionary model to find factors for a community to be healthy and attractive. Leimeister, Ebner, and Krcmar (2005) described how to design and implement trust-enabling functionalities in a visual community for patients, and introduced a model to achieve supporting trust in a community. Pavlou and Gefen (2004) analyzed the impact of institution-based factors on trust building in online marketplaces. User community changes the way people communicate and affects social interaction (Zeng, Wang, & Carley, 2007; Wang, Carley, Zeng, & Mao, 2010).

An important role of online communities is to provide a channel for users to communicate, interact, and share knowledge with each other. Ma and Agarwal (2007) described an identity-based view to understand the relationship between information technology (IT)-based features in online communities and online knowledge contribution. The authors offered a new perspective on the mechanisms through which IT features facilitate knowledge sharing. Gu, Konana, Rajagopalan, and Chen (2007) studied the trade-off between information quality and quantity, and explored how users with different characteristics value virtual communities. Bieber, Engelbart, and Furuta (2002) introduced an architecture for a community knowledge evolution system, which served as an ever-evolving repository of the community's knowledge.

Several researchers have studied the impact of the online community's behavior on e-commerce, such as new products and services planning, product sales and auction fraud prevention. Wagner and Majchrzak (2006) studied wikis and the wiki way, which can enable the creation of successful customer-centric business. The authors examined the six characteristics that affect customer engagement. Forman, Ghose, and Wiesenfeld (2008) studied the impact of reviewer identity disclosure on consumers when making purchase decisions and evaluating the helpfulness of reviews. They also studied the role of a reviewer's geographical location in product sales. Li and Hitt (2008) examined how the idiosyncratic preferences of early buyers may affect long-term consumer purchase behavior as well as the social welfare benefits of review systems. Chua, Wareham, and Robey (2007) studied the role of community in managing auction fraud and showed that the community is an important factor for the perpetration and prevention of crime. Discovering user communities may assist the setup of efficient recommender systems for targeted marketing, improving the quality of social information retrieval, among others (Wang, Zeng, Hendler, Zhang, Feng, Gao, Wang, & Lai, 2010).

To explain why people stay or leave a community in the long run, Fang and Neufeld (2009) offered a longitudinal investigation of one open source software (OSS) community, and showed that situated learning and identity construction behaviors were positively linked to sustained participation.

To improve our understanding of the evolution of online communities, Trier (2008) presented a method for event-based dynamic network visualization and analysis. The Commetrix software was introduced to describe the community formation process and network lifecycles.

Previous research has shown the importance of community analysis from different perspectives, such as trust, knowledge sharing, the impact of community on e-commerce, and the dynamic visualization of communities, etc. Nevertheless, the problem of detecting densely connected communities and their evolution patterns has not been well addressed. This article focuses on this problem and the application of the discovered evolution patterns. The importance of detecting densely connected communities and their evolution patterns is reflected in the following areas: First, the densely connected communities reveal people who form a close relationship with each other during their participation in common topics, it will help understand people's static concern about topics during a short time interval; second, the communities' evolution patterns reflect people's interest change trajectory, with which people will gain insight into the development and change about events concerning them.

Community Extraction

Community analysis has been an active research area in recent years. Approaches such as degree-based and matrix perturbation-based have been proposed to identify cohesive subgroups from social networks (Wasserman & Faust, 1994). Relying on people's information in a database, Sycara and Zeng (1994) developed an intelligent secretary agent system to help arrange efficient meetings among people who shared similar intertests. In web community analysis, there has also been extensive research on community extraction. Flake, Lawrence, and Giles (2000) proposed algorithms based on maximum flow and minimum cut framework for web community identification. Priebe, Conroy, Marchette, and Park (2005) introduced the theory of scan statistics on graph partitions, and used the theory for anomaly detection from the Enron e-mail graph. White and Smyth (2005) presented several algorithms that combined the modularity Q function and spectral clustering algorithm. Complex network theory was applied to analyze open-source software systems and structural properties of social interaction in collaborative tagging systems, respectively (Zheng, Zeng, Li, & Wang, 2008; Zeng & Li, 2008).

Within the these works, one basic assumption is that an entity in a network only belongs to one community, which always results in hard community memberships. Unfortunately, this assumption does not always make sense in real-world applications, in which a user may be interested in several communities. An overlapping or soft clustering of these communities could better capture the nature of the network. Several researchers have conducted work on this task. Yu, Yu, and Tresp (2005) proposed SNMF (symmetric nonnegative matrix factorization) algorithm for soft clustering on graphs in which graphs were clustered in a probabilistic way. Wei, Wang, Ma, and Zhou (2008) first performed a hard clustering on the community with the spectral partition, and then extended the seed sets with lazy random walk technique. Gregory (2007) presented a hierarchical clustering algorithm by extending the Girvan and Newman (GN) algorithm based on the betweenness centrality measure. Most recently, Shen, Cheng, and Guo (2009) proposed to construct a maximal clique network from the original network, and adopt the existing modularity optimization methods for identifying the overlapping community structure. Zhang, Li, Zeng, and Gao (2013) proposed a unified framework for user community detection in social network services,which integrates the user friendship networks and user-generated contents. Peng, Zeng, Zhao, and Wang (2010) proposed a unified user profiling scheme which makes good use of all types of co-occurrence information in the tagging data.

In this article, inspired by the idea of modularity function Q and latent Dirichlet allocation (LDA) topic model in document analysis, we propose a soft clustering-based generative probabilistic model for community structure detection. Zhang, Qui, Giles, Foley, and Yen (2007) have proposed to use LDA for community structure discovery. Our work differs from theirs in the following ways: (a) The network in our work is directed and dynamic; (b) the evolution of the network is studied in our article; (c) our work also discusses the interest of the community.

The aforementioned soft clustering algorithms focus on finding different community groups, in which members possess a certain probabilistic distribution. Compared to the hard clustering algorithms, the soft clustering algorithms find more meaningful communities in real-world applications. As to the critiques of the overlapping communities, Newman and Girvan (2004) proposed a modular function Q first to detect the quality of hard clustering algorithms, which was further generalized by Lin, Chi, Zhu, Sundaram, and Tseng (2009), so that it can measure the quality of the overlapping communities. The higher modular function Q means the closer relationship of members belonging to the same community.

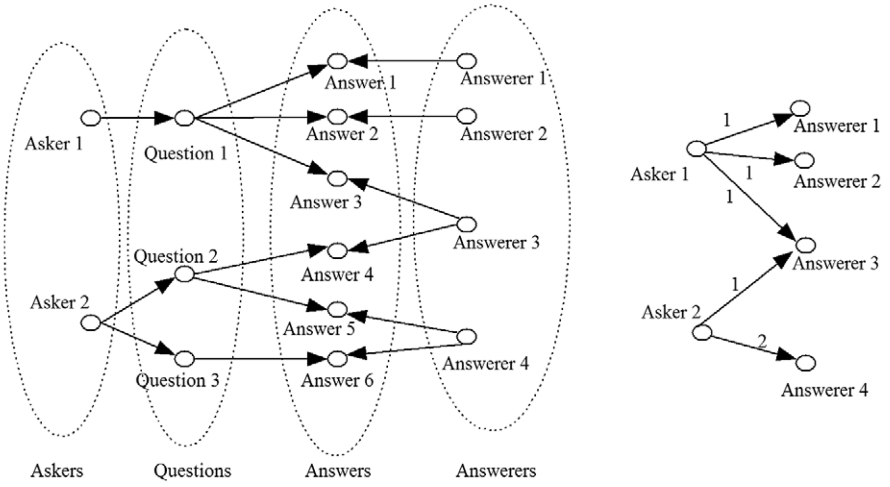
Community Evolution Tracking

There exists several works on analyzing the evolution of communities in dynamic networks. Sun, Faloutos, Papadimitriou, and Yu (2007) presented a parameter-free scheme, GraphScope, to mine graph evolutions over time. Asur, Parthasarathy, and Ucar (2007) proposed an event-based approach to capture the evolution behaviors of individuals and communities. Chi, Song, Zhou, Hino, and Tseng (2007) proposed an evolutionary spectral clustering framework, which incorporated temporal smoothness to measure the cluster quality. In Chi et al.'s framework, a good clustering result was considered to fit the current data while simultaneously not deviating dramatically from historical data. Lin, Chi, Zhu, Sundaram, and Tseng (2008; 2009) presented a unified framework, FacetNet, to discover communities and analyze their evolutions simultaneously. Yang, Chi, Zhu, Gong, and Jin (2009) introduced a probabilistic generative model, which unified community detection and evolution analysis. Yang et al.'s model employed the Bayesian treatment to estimate the posterior distributions of unknown parameters, which was believed to give robust estimation of community memberships.

These studies work well when applied to undirected networks. However, in CQA, the user interactive networks are usually directed networks. To detect community evolutions in directed networks, previous studies simply ignore the edge directions. It is clear that in discarding the directions of edges, a great deal of information about the networks' structure is lost. Taking this information into consideration may allow us to make more accurate analysis of community evolution. In this article, we propose a unified algorithm based on a generative probabilistic model for this problem, which is able to handle directed networks.

Community Quality Evaluation

To evaluate the strength of community structures, many weight functions or metrics including min-cut (Ding, He,



(a) Relationship of users, questions and answers

(b) User Network

FIG. 2. Construction of a user network.

Zha, Gu, & Sinon, 2001), intensity ratio, and edge ratio have been proposed. These metrics focused on simply counting edges, and did not take the statistical information of communities into consideration. Wang, Philips, Schreiber, and Wilkinson (2008) introduced the Poisson discrepancy for measuring the quality of graph clusters, which was derived from spatial scan statistics on point sets. Modularity Q proposed by Newman and Girvan (2004) has been widely adopted for evaluating the quality of community structures. In Newman & Girvan (2004), they further extended the Q function from undirected networks to directed networks. Lin, Chi, Zhu, Sundaram, and Tseng (2008) further generalized the modularity function to handle soft memberships. Because of the nice performance and convincing results of community quality detection with the application of the modularity Q in previous community work, in this article, the modularity Q will be used for measuring the quality of discovered communities. We will explain the computation method for modularity Q in Community Number Selection in detail. The higher Q value means the better community quality, whereas the community quality reflects the quality of mined community structure. Good community quality means that people belonging to one community are densely connected.

Community Evolutionary Discovery

In a CQA community, users form an interactive network by posting questions and providing answers. Figure 2(a) shows the interaction process connecting these elements. An edge from an asker to a question indicates that the user asked the question; an edge from an answerer to an answer means that the answer was provided by this user. The edge from a question to an answer indicates that the answer is associated with the question. We further summarize the relationships between askers and answerers in a directed network, as shown in Figure 2(b). The weight of the edge between asker 2 and answerer 4 indicates that answerer 4 has answered two questions posted by asker 2.

We represent the user interaction network at time t as a temporal graph Gt(V_{ask},V_{answer},ξ_t,W_t), in which $v_i \in V_{ask}$ represents an asker, $v_j \in V_{answer}$ represents an answerer, edge $e_{t;ij} \in \xi_t$ represents the interactions between asker$v_i$and answererv_j, $\omega_{t;ij} \in W_t$ denotes the weight of edge $e_{t;ij}$. Assuming there are m askers and n answerers in G_t, $W_t \in R^{m*n}$ is the adjacency matrix for G_t. Over time, the user interaction history can be represented by a sequence of temporal graphs < G_1,. . .,G_t,. . . >. In the following sections, we may also elide the subscript t whenever it does not cause confusion.

A community is formed by connecting the askers and answerers with similar interests. The translation of a community from one time step to another is described as the "evolution" of the community. Given a user interaction network, the problem of community evolutionary discovery is defined as:

Definition 1. (community evolutionary discovery) The challenge for community evolutionary discovery is to extract temporal communities in each time step, and track the evolution procedure of these communities.

The following sections depict the detailed design of our community evolution discovery algorithm. We first present the formulation of our model in detail. Then, we describe how to extract community structures and their evolutions.

Probabilistic Community Model (PCM)

In an actor-network, actors may have a diverse set of interest (Monteiro, 2000). This may cause a user to engage in multiple communities. Thus, the soft clustering of communities is studied in this section.

Inspired by the LDA model (Blei, Ng, & Jordan, 2003; Steyvers & Griffiths, 2007) in document analysis, we make basic assumptions about the user interaction network in CQA as follows:

1. When an answerer answers a question posed by an asker, he shows similar interests with the asker. Thus, the interest space of an asker can be represented by his associated answerers.
2. Each user can be modeled as a multinomial distribution over communities and a community is a probabilistic distribution over users grouped into it.

Based on the these assumptions, we derive our PCM to detect the soft community structures in CQA.

Answerer Generation Procedure

We use$p(c)$ for the distribution over community C of a particular asker, $p(v_j|C)$is the probability distribution over answerers given a community C. Each answerer v_j associated with an asker is generated by the following procedure: First, a community is sampled from the community distribution; then, an answerer is chosen from the community-answerer distribution. Specifically, the procedure can be formulated as:

Definition 2. The distribution of answerers connected with an asker can be calculated as: $p(v_j) = \sum_{k=1}^{T} p(v_j \mid c = k) p(c = k)$ where $p(c = k)$ is the probability that the kth community was sampled for v_j, and $p(v_j|c = k)$ is the probability of v_j belonging to community k.

The multinomial distribution over answerers for community k is represented by $\phi^{(k)} = p(v_j|c = k)$. The multinomial distribution over askers for community k is represented by $\theta^{(v_i)} = p(c)$. Parameters ϕ and θ indicate the importance of the answerers to a community and the importance of the askers to a community respectively.

Dirichlet Distribution Assumption

Because the conjugate prior for multinomial distribution is a Dirichlet distribution, it is natural to choose Dirichlet distribution as the prior distributions of ϕ and θ.

Definition 3. Given the hyperparameter α as the conjugate prior of θ, the probability density of θ is defined as: $p(\theta \mid \alpha) = \frac{\Gamma\left(\sum_{k=1}^{T} \alpha_k\right)}{\prod_{k=1}^{T} \Gamma(\alpha_k)} \prod_{k=1}^{T} \theta_k^{\alpha_k - 1}$, where $\Gamma(x)$is the gamma function.

Similarly, by placing Dirichlet prior β on ϕ, we get $p(\phi \mid \beta) = \frac{\Gamma\left(\sum_{j=1}^{m} \beta_j\right)}{\prod_{j=1}^{m} \Gamma(\beta_j)} \prod_{j=1}^{m} \phi_j^{\beta_j - 1}$.

Each hyperparameter α_k can be interpreted as the number of times community k is sampled for an asker, before having observed any actual answerers for this asker. Whereas each hyperparameter β_j can be interpreted as the number of times answerer j is sampled from a community before any actual answerers from the network are observed.

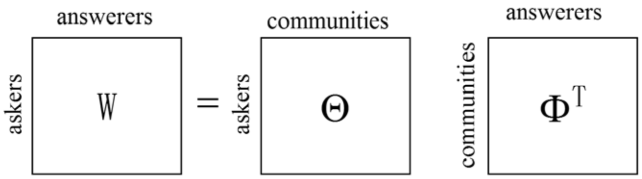


FIG. 3 Matrix factorization of the community model.

Matrix Factorization Interpretation

In our PCM model, the user interaction network is split into two parts: the asker-community relationship and the community-answerer relationship. This procedure can be interpreted as the matrix factorization illustrated in Figure 3.

ϕ is a sparse answerer-community matrix with dimension $n \times T$ and θ is a sparse asker-community matrix with dimension $m \times T$. $\theta_{ik} = p_{k \to i}$ indicates the probability that an interaction is observed in community k involving asker i and $\phi_{jk} = p_{k \to j}$ indicates the probability that an interaction is observed in community k involving answerer j. Obviously, $\sum_j \phi_{jk} = 1$, $\sum_i \theta_{ik} = 1$.

Community Topic Generation

The topic of a community is extracted to define the common interest of the community. It is typically represented as a probabilistic distribution over keywords.

Given p_{ik} being the probability of asker/answer i belonging to community C_k, w_{ik} being the weight of user i in C_k ($w_{ik} = \phi_{ik}$ or $w_{ik} = \theta_{ik}$). The community topic is generated as follows.

Algorithm 1, community topic generation:

```
1. For each user $i$ in $C_k$,
2. For each word $t_i$ in user $i$'s contents, with $T_{t,i}$ being the weight of $t$ in the content. In this paper, $T_{t,i} = \mathrm{TFIDF}_{t,i}$
    a. $t$ is sampled with probability of $p_{ik}$;
    b. the weight of $t$ in community topic is updated as: $Z'_{t,k} = Z_{t,k} + w_{ik} * T_{t,i}$
3. The topical distribution for $C_k$ over the word list is by normalizing over all $Z_{t,k}$, i.e. $p_{t,k} = \frac{Z_{t,k}}{\sum_j Z_{j,k}}$
```

The Overall Model

Based on this analysis, the generative procedure of PCM, which describes how answerers associated with askers are generated, can be summarized as follows:

1. For each temporal community C, a multinomial distribution ϕ_c is sampled from a Dirichlet prior β;
2. For each asker v_i, a multinomial distribution θ_{v_i} is sampled from a Dirichlet prior α;
3. For each answerer v_j connected with asker v_i:
    a. A community C_i is sampled from multinomial distribution θ_{v_i};$(p(C_i|\alpha))$
    b. An answerer v_j is sampled from multinomial distribution ϕ_c;$(p(v_j|c_i,\beta))$
4. Generate the community topic for Ci using Algorithm 1.

DOI: 10.1002/asi

Community Extraction and Their Evolution Detection

The problem of community evolutionary discovery is to extract communities and then track their evolution, wherein the task of community extraction involves identifying community structure and determining community membership.

Given a user interaction network G, the community structure C of network G can be represented as: $\mathrm{C} = (\theta,\phi,\Lambda)$, where Λ is a $T \times T$ nonnegative diagonal matrix, $\Lambda_k = p_k$ indicates the probability of generating community k, and $\sum_{k=1}^{T} \Lambda_k = 1$.

The community membership of askers and answerers can be defined as $WP = D^{-1}\phi\Lambda$ and $DP = B^{-1}\theta\Lambda$, respectively, where D and B are diagonal matrices, D_{ii} $D_{ii} = \sum_k \phi_{ik}\Lambda_k$, $B_{ii} = \sum_k \theta_{ik}\Lambda_k$. $(WP)_{jk}$ indicates the probability of answerer j belonging to community k, and $(DP)_{ik}$ indicates the probability of asker i being clustered into community k.

To track community evolutions, we compute the evolution probability of communities along a timeline. The historic community structure also contains valuable information related to current community structure. The quality of discovered community structure could be improved by considering recent historic communities. To fulfill this purpose, we propose to perform temporal smoothing on community structures. Our approach models the shift between community structures as a Markov chain. Specifically, instead of randomly initializing the community structure, we fold-in the network at time step t to the model learned at time step *t-1*. With temporal smoothing, the kth community at time step t is smoothly evolved from the kth community at time step *t-1*.

Parameter Estimation and Extension

In this section, we employ the Gibbs sampling method for parameter estimation in our PCM model. We also introduce some extensions of our method to handle undirected networks and automatically determine the number of communities in a network.

Parameter Estimation

In the PCM model described in the previous section, the main variables of interest are the asker-community distribution θ, and answerer-community distribution Φ. Estimation methods such as EM can be adopted to obtain direct estimates of θ and Φ. However, EM suffers from problems of local maxima and high computational burden.

In this article, Gibbs sampling (Steyvers & Griffiths, 2007) is employed. Instead of estimating parameters θ and Φ directly, Gibbs sampling evaluates the posterior distributions of community k. The probability of assigning the current answerer to each temporal community, conditioned on the community assignment of all the other answerers, can be calculated as (Steyvers & Griffiths, 2007):

$$p(c = k \mid c_{-k}, v_j, v_i, \cdot) \propto \frac{C^{nT}_{v_j,k} + \beta}{\sum_{v=1}^{n} C^{nT}_{vk} + n\beta} \frac{C^{mT}_{v_i,k} + \alpha}{\sum_{t=1}^{T} C^{mT}_{v_i,t} + T\alpha}$$

where C^{nT} and C^{mT} are matrixes of counts with dimensions $n \times T$ and $m \times T$, respectively; $C^{nT}_{v_j,k}$is the number of times that answerer v_j is assigned to community k, not including the current instance; $C^{mT}_{v_i,k}$is the number of times that community k is assigned to askerv_i, not including the current instance; c_{-k} represents the community assignments of all other answerers.

After a number of iterations, we can obtain the parameters estimation as:

$$\phi^{(k)}_{v_j} = \frac{C^{nT}_{v_j,k} + \beta}{\sum_{v=1}^{n} C^{nT}_{vk} + n\beta} \tag{1}$$

$$\theta^{(v_i)}_{k} = \frac{C^{mT}_{v_i,k} + \alpha}{\sum_{t=1}^{T} C^{mT}_{v_i,t} + T\alpha} \tag{2}$$

Extension to Undirected Networks

It is obvious that our PCM model could also be used for community detection in other forms of directed networks. In a directed network, adjacent matrix W usually is a rectangle matrix.

An undirected network can be represented by an undirected graph. In this case, the adjacent matrix W becomes a square and symmetric matrix. Intuitively, when performed on an undirected network, the resulting probabilistic matrix θ should be equal to Φ. Unfortunately, it is not the case. As Gibbs sampling performs an approximation on θ and Φ by iteratively sampling, there may be a slight difference between the resulting θ and Φ.

We may also conclude this from the estimation Formulas (1) and (2).When W is a symmetric matrix, C^{nT} and C^{mT} would be equal. However, as generally $\alpha \neq \beta$, $\mathrm{n} \neq \mathrm{T}$, $\theta^{(v_i)}_{k}$ and $\phi^{(k)}_{v_j}$are seldom equal.

Because our model is basically derived from the concept of LDA in document analysis, results of our model are comparable to LDA. We argue that Φ is more appropriate as the final community clustering result, and $WP = DP = D^{-1}\phi\Lambda$. In our experiments, we chose Φ as the community partition result and evaluated the performance of PCM model on undirected networks.

Community Number Selection

So far, we have assumed the community number T is set beforehand. In many cases, people may have difficulties in estimating the community number. In this section, we try to answer these questions: how to evaluate the quality of our community partition results and how to determine the number of communities at each time step.

The modularity function Q proposed in (Newman & Girvan, 2004) has been widely used to measure the quality of community structures.

DOI: 10.1002/asi

Definition 4. (Modularity Q) Given the community partitionP_T of a network, the modularity function Q is defined as (Newman & Girvan, 2004):

$$Q(P_T) = \sum_{k=1}^{T}\left[\frac{A(V_k, V_k)}{A(V,V)} - \left(\frac{A(V_k, V)}{A(V,V)}\right)^2\right] \qquad (3),$$

where $A(V_p, V_q) = \sum_{\mu \in V_p, \upsilon \in V_q} \omega_{\mu\upsilon}$, $\omega_{\mu v}$ is the weight of edge $e_{\mu v}$.

A high value of Q generally indicates a good community structure.

This formulation was initially defined to measure community structures on undirected networks. Leicht and Newman (2008) further extended it to measure community structures in directed networks. Lin et al. (2009) extended Q to handle soft memberships, and defined soft modularity Qs. Unfortunately, Qs was intended for undirected networks, and required the weight matrix W to be square.

To measure the community quality in our case, in which the weight matrix W may be rectangular (the row m may not be equal to column n), we extend the concept of modularity to handle soft memberships in directed networks.

Definition 5. (soft modularity) Given the weight matrix W of a directed network, and the community membership matrixes WP, DP, we define the soft modularity in directed networks as:

$$Q_{ds} = Tr\left[WP^T \cdot P \cdot DP\right] - E^T \cdot P \cdot DP \cdot DP^T \cdot P^T \cdot E \quad (4),$$

where E is a vector whose elements are all ones, and Tr[A] is the trace of squared matrix A. $p_{ij} \in P$ is the probability of interactions between asker i and answerer j, and P can be obtained by normalizing W.

Theorem 1. The soft modularity Q_{ds} defined in Equation (4) has the same probabilistic interpretation with Q defined in Equation (3), but is generalized to soft community membership in directed networks.

Proof. In the standard Q formula, the first term $\frac{A(V_k, V_k)}{A(V,V)}$ measures the probability of edge that has both vertexes in the same community k. As proven in (Lin et al., 2009), in soft clustering cases, this empirical probability can be rewritten as: $\Sigma_{\mathrm{ij}} p_{ij} p_{(k|i)} p_{(k|j)}$. $\frac{A(V_k, V)}{A(V,V)}$ measures the probability of edge that has at least one vertex in the community k. In soft clustering cases, it can be rewritten as $\Sigma_{\mathrm{i}} p_{(k|i)} \Sigma_{\mathrm{j}} p_{ij}$. Note that $p_{(k|i)} = (B^{-1}\theta\Lambda)_{ik} = (DP)_{ik}$, $p_{(k|j)} = (D^{-1}\phi\Lambda)_{jk} = (WP)_{jk}$. By summing these two terms over all k ($k = 1,. . .,T$), we get the extended soft modularity Q_{ds} as Equation (4).

When graph G is an undirected graph, in which W is a symmetric matrix, it is straightforward to see that $WP = DP$, and Q_{ds} is equal to Q_s defined in (Lin et al., 2009). Further, when WP and DP become 0/1 indicator matrixes, which is the case in a hard communication partition, Q_{ds} is equal to Q. Thus, Q_{ds} defined in our paper is a general case of standard Q formula. □

TABLE 1. Detailed information of the data sets.

Data set name	# of Vertices	# of Edges	Features
Football Games	115	616	undirected
Political Blogs	1,490	19,091	directed,
Neural Networks	297	2,359	directed, weighted

To detect the best community structure at each time step t, we run our model to search for the best candidate community number T, which would gain the highest value Q_{ds}.

Empirical Evaluation

In this section, we report the empirical evaluation of the proposed algorithm, aimed at answering the following research questions: (a) Can the proposed PCM model improve the detection performance of community structure? (b) Can we discover meaningful community evolution patterns in CQA? (c) Will the communities discovered in CQA provide valuable feedbacks for decision making departments and individuals?

To answer the first question, we select several static networks with different features and diverse sizes. As the previous techniques such as maximum flow/minimum cut and scan statistics on graph partitions, which are reviewed in Community Extraction, are only applicable for hard clustering research, we compare the performance of our model with the two mature techniques SNMF and spectral clustering, which were successfully used for soft community detection in (Lin et al., 2009; Chi et al., 2007), respectively. These two mature techniques are classical and representative, and their wide adoption and deep theoretical basis in the community mining field influenced us to choose them as baseline techniques. To answer questions (b) and (c), we collected the iPhone data set with more than 16,000 question threads from Yahoo! Answers. We analyzed the evolution of community interest over time on these data sets.

Evaluation of Community Extraction Model

We performed several empirical studies on real-world data sets, to evaluate the performance of our PCM. Table 1 summarizes the data sets[1] used in our experiments. For all the experiments, parameters are set to be: $\alpha = 0.001{*}N$, $\beta = 0.01$, where N is the number of edges in the network.

We compared our PCM model with SNMF (Yu et al., 2005), which also performs soft clustering on graphs. We also compared our method with the spectral clustering method, which is a typically a hard community partition

[1] These data sets are downloaded from http://www-personal.umich.edu/%7Emejn/netdata/

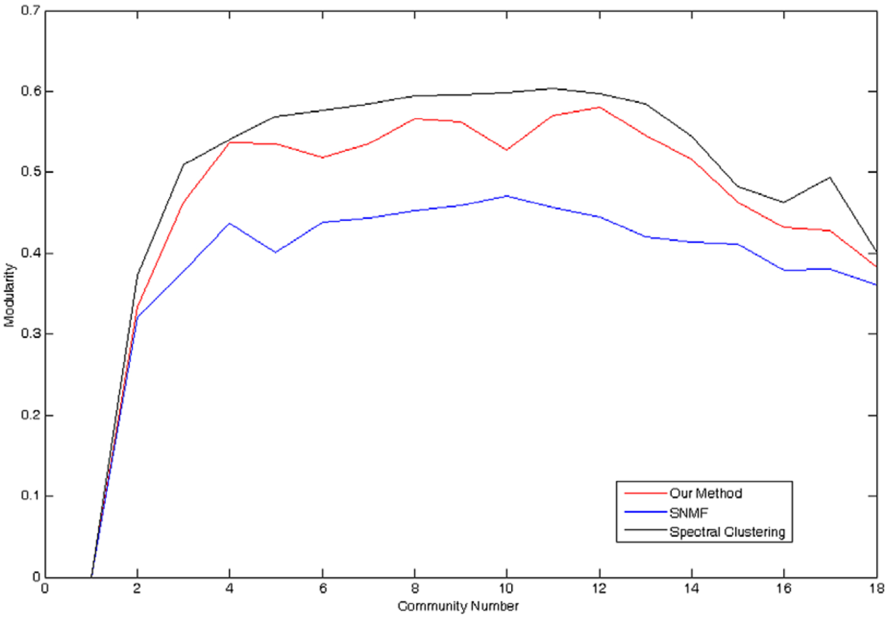


FIG. 4. Modularity on football games. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

TABLE 2. Best clustering results on football games.

	Our method	Spectral clustering	SNMF
No. of Clusters	12	11	10
Highest Qds	0.58	0.60	0.47

TABLE 3. *p*-value for pairwise t-test.

	PCM (undir) vs SNMF (undir)
p-Values	<0.001

method. The extended soft modularity Q_{ds} is computed to measure the quality of extracted community structures.

Network of Football Games

We start with the network of the NCAA college football. It is a game scheduled during the regular season Fall 2000, between Division IA colleges. Obviously, this data set is a typically undirected network. In total, there are 115 teams, with 616 games played over the course of the season. These teams are divided into 12 conferences, with each conference containing about eight to 13 teams. Generally, games are more frequent between members of the same conference than between teams of different conferences.

Figure 4 shows the modularity Q_{ds} varying with the number of communities. Table 2 shows the value of the highest modularity and the corresponding number of communities for each algorithm. It can be concluded that our method is able to correctly detect the 12 conferences, whereas spectral cluster and SNMF detect 11 and 10 conferences, respectively. Our method could gain a higher soft modularity value for each community partition than SNMF.

We verified our comparison result with the help of a pairwise *t*-test. As shown in Table 3, it is concievable that our method could gain a higher soft modularity value for each community partition than SNMF.

Network of Political Blogs

We then studied the performance of our algorithm on the network of political blogs, which is a directed network of hyperlinks between weblogs on US politics. In total, there are 1,490 blogs, with 19,091 links. Blogs with hyperlinks between them tend to have similar political leanings, which could be left (liberal) or right (conservative). This data set was first proposed by Adamic and Glance (2005). According to the authors, the blogs were collected 2 months preceding the US presidential election of 2004. Only liberal and conservative blog communities were collected. They did not gather blogs from URLs of independents or moderate blogs, which were far fewer in number. The study in Adamic and Glance (2005) also showed that 91% of the links originating within either the conservative or liberal communities stay within that community. Thus, there are exactly two groups in this data set, and liberals and conservatives link primarily within their separate communities.

We first performed our method on the directed networks that were collected. As SNMF and spectral clustering are

DOI: 10.1002/asi

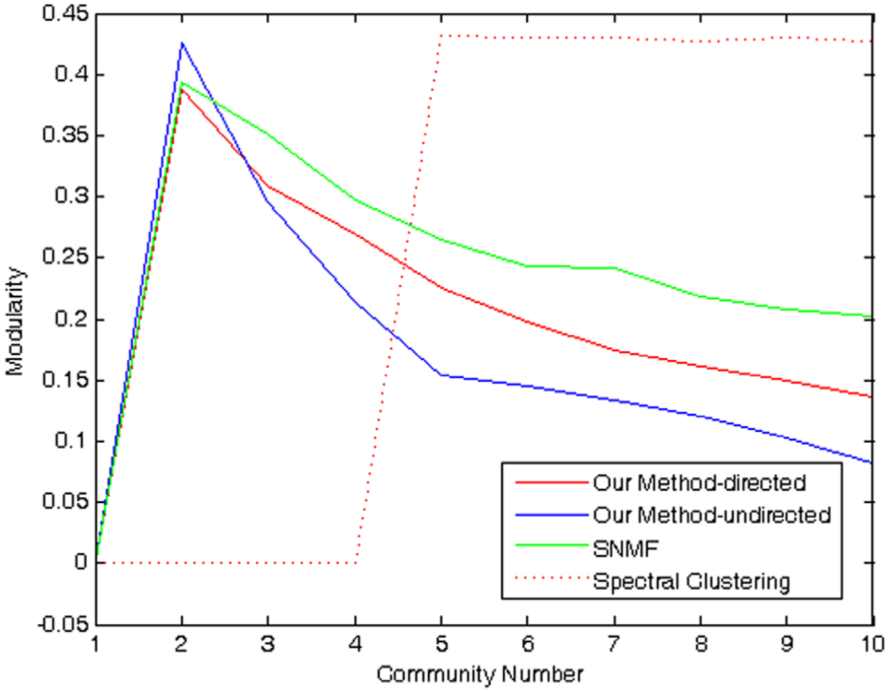


FIG. 5. Modularity on political blogs. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

TABLE 4. Best clustering result on political blogs.

	Our method (directed)	Our method (undirected)	Spectral clustering	SNMF
No. of Clusters	2	2	5	2
Highest Qds	0.39	0.43	0.43	0.39

TABLE 5. *p*-value for pairwise t-test.

	PCM (dir) vs PCM (undir)	PCM (dir) vs spectral (undir)	PCM (dir) vs SNMF (undir)	PCM (undir) vs spectral (undir)	PCM (undir) vs SNMF (undir)
p-Values	0.001	<0.001	<0.001	<0.001	0.001

suitable for undirected networks only, we then ignored the direction of blog links, and converted the blog network into an undirected network.

Figure 5 shows the modularity values varying with the community number. The clustering results with the highest Q_{ds} for each algorithm are listed in Table 4. It can be seen that our method could properly detect the two political leanings on both directed and undirected networks. For undirected networks, SNMF could also detect the community structure accurately, but with a lower modularity value. The spectral clustering method tended to split the networks into smaller groups, and get "confused" when clustering the networks into two communities. In this section, we also conducted the pairwise *t*-test. As shown in Table 5, the results verified what we have just stated.

Neural Network

Finally, we tested our algorithm on neural networks. In contrast to the political blog data set, in which the hyperlinks between blogs are Boolean values, the neural network is a directed, weighted graph. There are 297 vertices and 2,359 weighted edges in total. This data set can be applied to directed networks, it can also be used to discover undirected network structure under the condition of leaving out the edge directions. Thus this data set is applicable for mining different kinds of community networks. The experimental setups are similar to the previous section. First, we test our method on the directed, weighted network. Then, an undirected, weighted network is constructed by ignoring the directions of edges.

Figure 6 shows the modularity values varying with the number of communities on the neural network. The clustering results with highest modularity value are listed in Table 6. We can observe from Figure 6 that for undirected weighted networks, our method could gain the highest modularity value 0.42, when the community number is set to 3. SNMF gains its highest modularity value 0.41 when the community number is set to 5. The spectral clustering algorithm, which performs hard community partition, achieves its highest modularity value 0.37, when the number of community is set to 5. When considering the direction of edges, our method gains the highest

DOI: 10.1002/asi

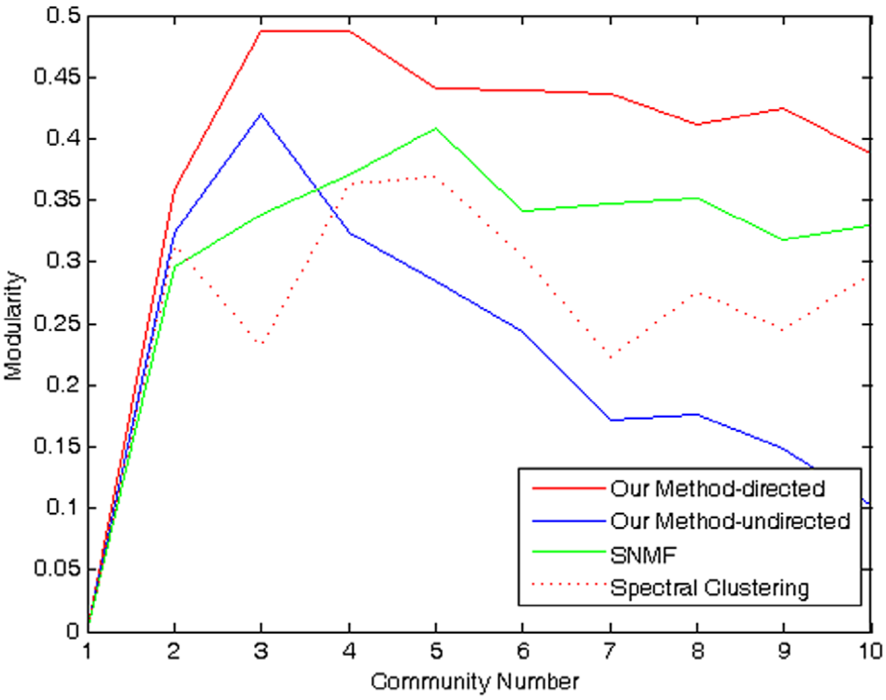


FIG. 6. Modularity on neural networks. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

TABLE 6. Best clustering result on neural networks.

	Our method (directed)	Our method (undirected)	Spectral clustering	SNMF
No. of Clusters	4	3	5	5
Highest Qds	0.49	0.42	0.37	0.41

TABLE 7. *p*-value for pairwise *t*-test.

	PCM (dir) vs PCM (undir)	PCM (dir) vs spectral (undir)	PCM (dir) vs SNMF (undir)	PCM (undir) vs spectral (undir)	PCM (undir) vs SNMF (undir)
p-Values	<0.001	<0.001	<0.001	0.001	0.001

modularity $Q_{ds} = 0.49$ with $T = 4$. From Figure 6, we also note that our method gains higher modularity in most partition settings. To make our experiment result in Table 6 more convincing, we conducted the pairwise *t*-test. As shown in Table 7, the results verified what we have stated above.

Through a series of experiments just described, we can conclude that our PCM improves the quality of extracted community structures. In the following sections, we adopt our PCM algorithm for community extraction in CQA communities.

TABLE 8. Detailed information concerning the smartphone data set.

Data set	# of questions	Time Span	# of askers	# of answerers
Smartphone	49,022	Jan, 2011- Dec, 2012	39,504	55,293

Community Evolution in CQA

To analyze the community evolution patterns in CQA, we collected a smartphone data set from Yahoo! Answers, by crawling all the resolved questions about iPhone, Android, and Windows Mobile Phone in 2012. We also collected about 8,000 questions before January 2012 as the cold start for the community evolution. For each question thread, we extracted the following fields: the asker, the question context, the question time stamp, and the associated answers. Each associated answer to a question thread is composed of the answer content, the answerer who posted the answer and the answer time stamp. Table 8 summarizes the key statistics for the smartphone data set. The interactive network is constructed by making a unique user in the data set as a vertex, and an edge <*i*,*j*> indicates that the question posed by the asker *i* was answered by answerer *j*. We weight each edge <*i*,*j*> in the network by the number of asker *i* 's questions that answerer *j* has answered.

The reason that we use this data set is based on the following considerations: First, Yahoo! Answers has been the largest knowledge sharing community on the web, and the use of smartphones is increasing rapidly. Thus,

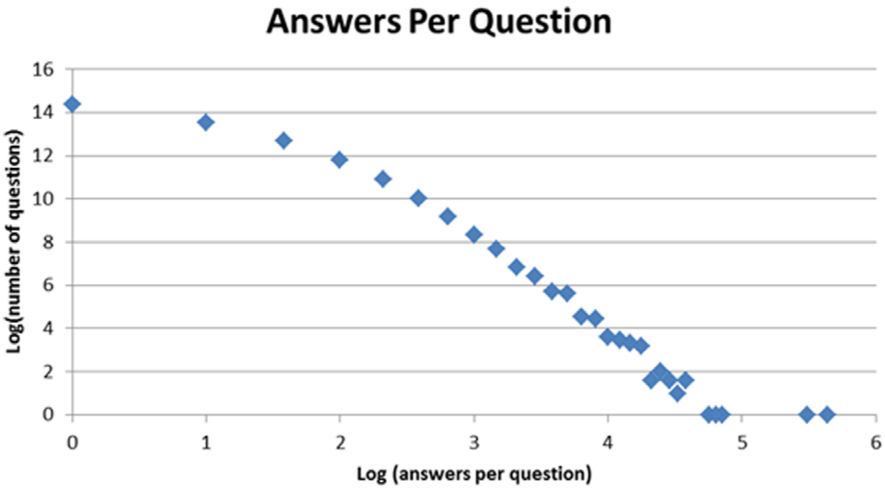


FIG. 7. Distribution of answers per question. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

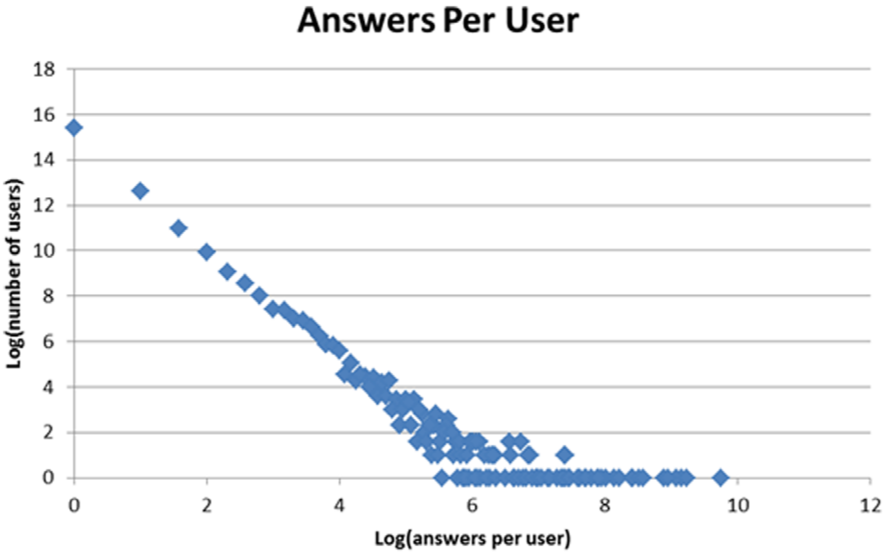


FIG. 8. Distribution of answers per user. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

collecting smartphone data set from Yahoo! Answers provides us a sufficient data source to analyze user behaviors. Second, everyone could post their information needs and share their knowledge on Yahoo! Answers, causing the quality of the content to vary drastically. Yahoo! Answers encourages users to report abuse of the system to avoid inappropriate questions and answers, and filter out malicious users. Question threads from the answered questions are generally considered to be appropriate for knowledge sharing purposes by the community users. Using the resolved questions from Yahoo! Answers ensures the quality of our data set, and provides us a more reasonable conclusion.

Figure 7 shows the number of received answers per question as a log-log distribution plot. We can see that a large number of questions get very few answers, whereas a small portion of questions get a large number of answers. The possible explanation for this phenomenon are: (a) the questions are only open for a few days; (b) without a proper recommendation policy, many questions do not receive access to answers by proper users; (c) many questions are meant to trigger discussion, encourage the users to express their opinion, thus receiving a large number of answers. Figure 8 shows the log-log distribution of answers posted per user. Again, we can see that a few users are very active whereas many users have limited activity in sharing answers. From Figure 7 and Figure 8, it can be seen that the user activity in CQA follows a power-law distribution, which has been discovered in other forms of online social networks. Similar with other real-world networks, the CQA networks exhibit a fat-tailed distribution, implying that a minority of users are far more gregarious and popular than others. This property can be seen as evidence of a universal

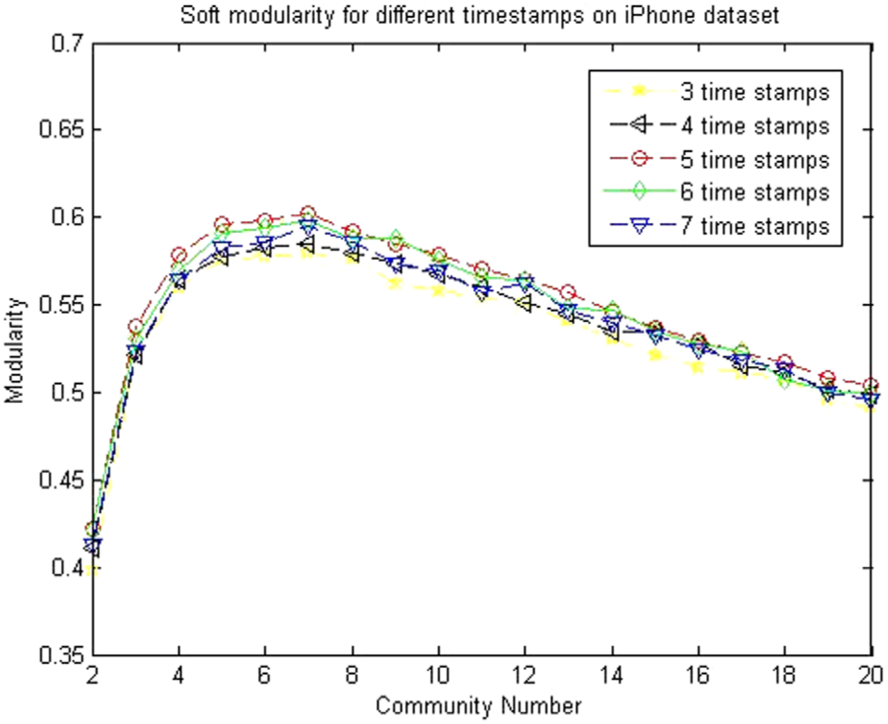


FIG. 9. Soft modularity for each timestep on smartphone data set. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

organizing principle that governs the network as it evolves over time.

To get the best time stamp number, we have conducted an experiment using different time stamps. The modularity value Q was introduced to measure the impact of different time stamps on the community structure. As we have mentioned in Community Quality Evaluation, the higher value Q means the better community structure. The choice of time interval is trivial, which should be neither too big nor too small. We empirically evaluated the time stamps ranging from 3 to 7 on Q value. To compare their performance directly, we computed the mean value of Q for each of these time stamps. The experimental results are shown in Figure 9, from which we can easily obtain the best time stamp number 5. It shows a higher value Q when compared to other numbers, the smoother curve of this number also confirmed our choice for accepting it reasonably.

We calculate the soft modularity value for each time step on different settings of community numbers. Figure 10 plots the soft modularity values under different community numbers for each time step. Ii (i=1,. . ., 5) represents the ith timestep. It can be noted that the soft modularity achieves relatively stable and high value for each time step when the community number $T = 10$. Thus, in our later study, the community number is fixed at ten during each time stamp.

Table 9 shows the p-value for the pairwise t-test on modularity when the community number $T = 10$. We perform PCM modeling on both directed networks and undirected networks (obtained by omitting the edge direction in the user network). Newman's spectral optimization methods are selected as a baseline, which combines spectral partition and a modularity matrix for community structure detection in both undirected networks (Newman, 2006) and directed networks (Leicht & Newman, 2008). In the spectral models, a user is assigned to one community at a time. The PCM model significantly outperforms spectral modeling in both directed and undirected networks. This verifies that the overlapping community structure exists in the process of electronic word-of-mouth (eWOM) in CQA. The PCM model performs better in directed networks than in undirected networks. This suggests that keeping direction information would allow us to make more accurate determination of the communities. The following studies are conducted on the directed network with the PCM model.

Community Extraction and Evolution Patterns

Table 10 lists the topical keywords for six community evolutionary patterns. Because iPhone, Android, and phone etc. are common words in our data set, such words are also added as stop words. That is, these words will not be included in the topical keyword list. With these keywords, we could analyze the community evolution patterns in detail:

- **Thread a:** The community thread labeled with a seems to be about comparison among different mobile platforms,

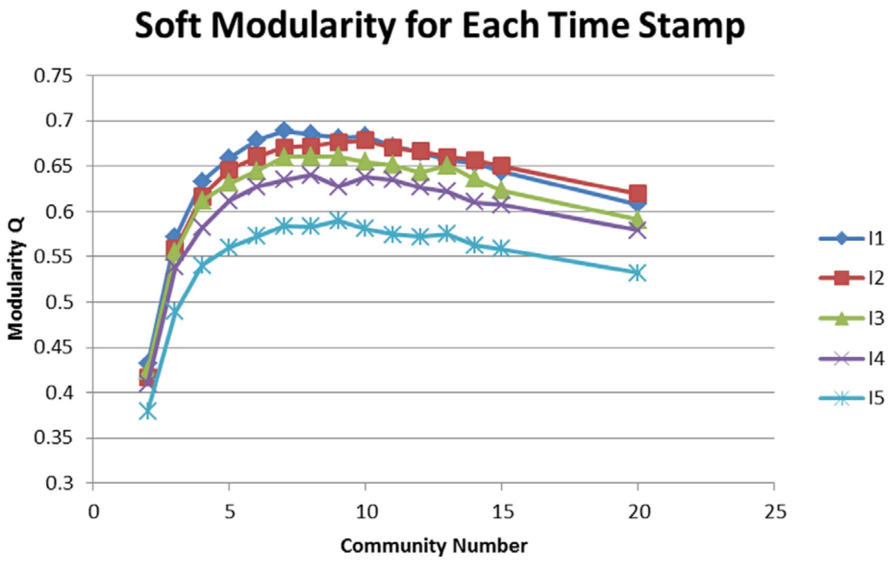


FIG. 10. Soft modularity for each time stamp on smartphone data. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

TABLE 9. *p*-value for pairwise *t*-test.

	PCM (dir) vs PCM (undir)	PCM (dir) vs spectral optimization(dir)	PCM (dir) vs spectral optimization (undir)	PCM (undir) vs spectral optimization (dir)	PCM (undir) vs spectral optimization (undir)
p-Values	0.001	<0.001	<0.001	<0.001	0.001

especially IOS and Android. The Windows phone from Microsoft and Lumia from Nokia, etc. are also compared with these two major mobile OS. Many people tend to ask for advice when they are about to change mobile devices. With the release of Nokia Lumia 820, which runs on Windows Phone 8, in September 2012, it is generally compared with Android and iPhone during I3. Influenced by the release of iPad mini and Nexus 4, people have made malicious comparisons among iPad, Kindle, Nexus, etc. during I4 and I5.

- **Thread b:** Thread b is about jail breaking of iPhone and data plan contact with AT&T. To get an iPhone, users have to sign a contract with AT&T. Jail breaking of iPhone for access to other network service providers such as Verizon seems to be a permanent topic.
- **Thread c**: Besides communication functions, the major use of smartphones may be for entertainment, especially playing games, listening to music, and watching videos. Communities in thread c mainly focus on entertainment on smartphones. Users in this community tend to share their game experiences, favorite music, and videos.
- **Thread d:** The rapid development of mobile app markets has been a major drivers of smartphone adoption. Communities in thread d may be formed by a group of app developers. Various questions involved in application development and app markets are discussed. Apps about games have attracted major concern, which is the focus of community topics during I4 and I5.
- **Thread e**: Communities in thread e may be about the daily use of different smartphones. Questions in these communities cover a great variety of usage problems. During I1, the file synchronization problem occupies the main topic, which evolves into the app downloading problem during I2. During I3, questions about networking are the primary concern of the community. Then, business apps such as emailing and lifestyle apps are discussed during I4.
- **Thread f:** Communities in thread f mainly focus on hardware comparisons among different mobile devices. Questions about keyboard, camera, screen, and battery etc are discussed in thread f.

The development of Web 2.0 has provided users and firms multiple channels to get information about products, including the products' official news, product reviews, product blogs, CQA, etc. The official news would generally focus on the merits and uniqueness of the products, without contributions from actual users. By contrast, user-generated content from product reviews, blogs and CQA are usually shared by consumers. Thus, feedback from these channels is more objective and could better reflect the consumers' experiences in using a product. Different from product reviews and blogs, in which a user posts his/her experience and other users choose to like or dislike his opinion, CQA supports user interaction by posting questions and providing answers and thus could better reflect the requirements of the majority of users.

DOI: 10.1002/asi

TABLE 10. Top keywords for several community evaluation patterns of smartphone data set. The labels are our own interpretations of these communities.

	Thread a	Thread b	Thread c	Thread d	Thread e	Thread f
I1	phone	iphon	android	android	iphon	ipad
	iphon	phone	phone	app	itun	tablet
	window	plan	app	phone	comput	video
	connect	data	download	free	music	camera
	internet	wait	market	game	ipod	galaxi
	make	appl	free	account	transfer	laptop
	chang	verizon	game	make	sync	inch
	call	unlock	screen	develop	contact	displai
	wireless	contract	samsung	download	song	mini
	us	jailbreak	googl	googl	click	keyboard
Label	platform comparison	data plan & jailbreak	entertaining–games	app developing	file syncing	hardwares
I2	iphon	iphon	iphon	android	android	phone
	window	unlock	comput	phone	phone	window
	microsoft	phone	appl	app	download	android
	comput	jailbreak	music	download	app	iphon
	facebook	appl	itun	game	free	appl
	laptop	itun	phone	free	iphon	nokia
	phone	plan	video	mobil	googl	galaxi
	make	card	click	market	music	batteri
	charger	contract	call	googl	market	samsung
	white	verizon	connect	samsung	click	camera
Label	platform changing	jailbreak & data plan	entertaining–music & video	app store & developing	downloading	hardwares
I3	phone	iphon	iphon	phone	android	phone
	nokia	phone	phone	iphon	facebook	pictur
	android	unlock	video	android	ipad	galaxi
	button	servic	case	appl	wifi	screen
	window	data	download	make	chang	appl
	lumia	card	look	where	new	note
	app	look	music	monei	answer	camera
	hold	free	comput	price	number	come
	power	plan	window	peopl	call	android
	amaz	text	android	game	link	googl
Label	platform comparison	text & data plan	entertainment–music	app developing	networking	hardwares
I4	ipad	iphon	android	iphon	mail	iphon
	mini	phone	phone	phone	chang	phone
	ipod	appl	samsung	android	window	android
	appl	android	galaxi	app	yahoo	appl
	nexu	galaxi	mail	download	password	app
	kindl	data	game	devic	sound	connect
	tablet	plan	yahoo	music	account	batteri
	touch	unlock	screen	comput	address	comput
	fire	contract	nokia	game	hous	us
	screen	us	account	samsung	contact	music
Label	tablets: ipad, kindle, nexus	data plan &contract	entertainment:games	music & games	office applications	hardware
I5	iphon	iphon	iphon	ipad	phone	phone
	phone	phone	phone	iphon	android	android
	ipad	ipad	android	phone	iphon	iphon
	android	android	galaxi	mini	ipad	galaxi
	mini	appl	app	android	app	nokia
	look	app	samsung	appl	galaxi	app
	galaxi	make	make	game	look	screen
	make	ipod	game	app	samsung	look
	ipod	unlock	download	screen	tablet	camera
	tablet	music	music	tablet	download	samsung
Label	ipad, kindle, galaxy nexus	jailbreak	entertainment–game & music	game apps	app download	hardware comparison

This article presents a study of evolutionary communities serving as a eWOM purpose in CQA. With the empirical analysis of the smartphone data set, it can be concluded that communities from CQA can reflect variations in users' interests during different periods of a product's lifecycle. From a commercial perspective, these applications involve analyzing market reactions, gathering market information, and providing personalized information for the targeted population. These applications are beneficial for both business and individuals.

Conclusion

Detecting communities and tracking their evolution in dynamic user networks has been an active research area in the literature. In this article, we propose a generative probabilistic model for detecting community evolution in CQA. Modularity Q is extended for measuring the overlapping community quality in directed networks. Extensive experimental studies demonstrate that our method appears to improve community extraction quality, and is capable of discovering meaningful community evolution patterns in CQA. The experimental results on the smartphone data set suggest that CQA communities could serve as an effective eWOM mechanism, and an alternative channel for product/service feedback.

The research reported in this article represents one of the first studies on community evolution discovery in CQA. Many issues remain for further study. First, though we propose to use extended soft modularity for selecting the proper community partition number T, it is usually computationally expensive to search T in a wide range of candidates. As part of our future research, we are developing better schemes to automatically determine the number of communities. Second, our work only focuses on communities in CQA. The growth of Web 2.0 has triggered multiple channels for users to share information, including microblogging, blogs, forums, etc. A framework integrating these different channels could provide a more comprehensive insight into social events.

Acknowledgment

This research is supported by NSFC grants (No. 61172106, No. 71025001, No. 91024030, and No. 70890084), the MOH Grant (No. 2012ZX10004801 and No. 2013ZX10004218), and BJNSF(No. 4112062).

References

Adamic, L.A., & Glance, N. (2005). The political blogosphere and the 2004 US Election. In Proceedings of the Word Wide Web, Workshop on the Weblogging Ecosystem, New York, USA (pp. 36–43).

Agichtein, E., Castillo, C., Donato, D., Gionis, A., & Mishne, G. (2008). Finding high quality content in social media. In Proceedings of the International Conference on Web Search and Web Data Mining, New York, USA (pp. 183–194).

Asur, S., Parthasarathy, S., & Ucar, D. (2007). An event-based framework for characterizing the evolutionary behavior of interaction graphs. In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, California, USA (pp. 913–921).

Bian, J., Liu, Y., Agichtein, E., & Zha, H. (2008). Finding the right facts in the crowd: factoid question answering over social media. The 17th International World Wide Web Conference, New York, USA (pp. 467–476).

Bian, J., Liu, Y., Zhou, D., Aigchtein, E., & Zha, H. (2009). Learning to recognize reliable users and content in social media with coupled mutual reinforcement. In Proceedings of the 18th International World Wide Web Conference, Madrid, Spain (pp. 51–60).

Bieber, M., Engelbart, D., & Furuta, R. (2002). Toward virtual community knowledge evolution. Journal of Management Information Systems, 18(4), 11–35.

Blei, D., Ng, A., & Jordan, M. (2003). Latent dirichlet allocation. Journal of Machine Learning Research, V(3), 993–1022.

Cao, Y., Duan, H., Lin, C., Yu, Y., & Hon, H.W. (2008). Recommending questions using the MDL-based tree cut model. The 17th International World Wide Web Conference, New York, USA (pp. 81–90).

Chi, Y., Song, X., Zhou, D., Hino, K., & Tseng, B.L. (2007). Evolutionary spectral clustering by incorporating temporal smoothness. In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2007, San Jose, California, USA (pp. 153–162).

Chua, C.E.H., Wareham, J., & Robey, D. (2007). The role of online trading communities in managing internet auction fraud. MIS Quarterly, 31(4), 759–781.

Ding, C.H.Q., He, X., Zha, H., Gu, M., & Sinon, H.D. (2001). A min-max cut algorithm for graph partitioning and data clustering. In Proceedings 2001 IEEE International Conference on Data Mining, San Jose, California, USA (pp. 107–114).

Fang, Y., & Neufeld, D. (2009). Understanding sustained participation in open source software projects. Journal of Management Information Systems, 25(4), 9–50.

Flake, G.W., Lawrence, S., & Giles, C. (2000). Efficient identification of web communities. In Proceedings of 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston, MA, USA (pp. 150–160).

Forman, C., Ghose, A., & Wiesenfeld, B. (2008). Examining the relationship between reviews and sales: the role of reviewer identity disclosure in electronic markets. Information Systems Research, 19(3), 291–313.

Geng, X.J., Whinston, A.B., & Zhang, H. (2004). Health of electronic communities: an evolutionary game approach. Journal of Management Information Systems, 21(3), 83–110.

Gregory, S. (2007). An algorithm to find overlapping community structure in networks. In Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases, Warsaw, Poland (pp. 91–102).

Gu, B., Konana, P., Rajagopalan, B., & Chen, H.W.M. (2007). Competition among virtual communities and user valuation: the case of investing-related communities. Information Systems Research, 18(1), 68–85.

Leimeister, J.M., Ebner, W., & Krcmar, H. (2005). Design, implementation, and evaluation of trust-supporting components in virtual communities for patients. Journal of Management Information Systems, 21(4), 101–135.

Leicht, E.A., & Newman, M.E.J. (2008). Community structure in directed networks. Phys. Physical Review Letters, 100(11).

Li, X., & Hitt, L.M. (2008). Self-selection and information role of online product reviews. Information Systems Research, 19(4), 456–474.

Lin, Y., Chi, Y., Zhu, S., Sundaram, H., & Tseng, B.L. (2008). FaceNet: a framework for analyzing communities and their evolutions in dynamics networks. In Proceedings of the 18th International World Wide Web Conference, New York, USA. (pp. 685–694).

Lin, Y., Chi, Y., Zhu, S., Sundaram, H., & Tseng, B.L. (2009). Analyzing communities and their evolutions in dynamic social networks. ACM Transactions on Knowledge Discovery from Data (TKDD), special issue on Social Computing, Behavioral Modeling, and Prediction, 3(2).

DOI: 10.1002/asi

Ma, M., & Agarwal, R. (2007). Through a glass darkly: information technology design, identity verification, and knowledge contribution in online communities. Information Systems Research, 18(1), 42–67.

Monteiro, E. (2000). Actor-network theory and information infrastructure. In From Control to Drift. The Dynamics of Corporate Information Infrastructure, Oxford University Press (pp. 71–83).

Newman, M.E.J., & Girvan, M. (2004). Finding and evaluating community structure in networks. Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics, 69(2).

Newman, M.E.J. (2006). Finding community structure in networks using the eigenvectors of matrices. Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics, 74(3).

Pavlou, P.A., & Gefen, D. (2004). Building effective online marketplaces with institution-based trust. Information Systems Research, 15(1), 37–59.

Peng, J., Zeng, D., Zhao, H., & Wang, F. (2010). Collaborative filtering in social tagging systems based on joint item-tag recommendations. Proceedings of the 19th ACM International Conference on Information and Knowledge Management, pp. 809–818.

Priebe, C.E., Conroy, J.M., Marchette, D.J., & Park, Y. (2005). Scan statistics on Enron graphs. Computational and Mathematical Organization Theory, 11(3), 229–247.

Shen, H., Cheng, X., & Guo, J. (2009). Quantifying and identifying the overlapping community structure in networks. Journal of Statistical Mechanics: Theory and Experiment, V.

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. In *Handbook of Latent Semantic Analysis*.

Sun, J., Faloutos, C., Papadimitriou, S., & Yu, P.S. (2007). Graphscope: parameter-free mining of large time-evolving graphs. In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, California, USA (pp. 687–696).

Sycara, K.P., & Zeng, D. (1994). Towards an intelligent electronic secretary. International Conference on Information and Knowledge Management, Intelligent Information Agents Workshop.

Trier, M. (2008). Research note—towards dynamic visualization for understanding evolution of digital communication networks. Information Systems Research, 19(3), 335–350.

Wagner, C., & Majchrzak, A. (2006). Enabling customer-centricity using wikis and the wiki way. Journal of Management Information Systems, 23(3), 17–43.

Wang, F.Y., Carley, K.M., Zeng, D., & Mao, W. (2010). Social computing: From social informatics to social intelligence. IEEE Intelligent Systems, 22(2), 45–53.

Wang, B., Phillips, J.M., Schreiber, R., & Wilkinson, D.M. (2008). Spatial scan statistics for graph clustering. In Proceedings of the SIAM International Conference on Data Mining, Georgia, USA (pp. 727–738).

Wang, F.Y., Zeng, D., Hendler, J.A., Zhang, Q., Feng, Z., Gao, Y., Wang, H., & Lai, G. (2010). A study of the human flesh search engine: crowd-powered expansion of online knowledge. Computer, 43(8), 45–53.

Wasserman, S., & Faust, K. (1994). Social network analysis: methods and applications. Cambridge University Press.

Wei, F., Wang, C., Ma, L., & Zhou, A. (2008). Detecting overlapping community structures in networks with global partition and local expansion. In Proceedings of WWW Research and Development, 10th Asia-Pacific Web Conference, Shenyang, China (pp. 43–55).

White, S., & Smyth, P. (2005). A spectral clustering approach to finding communities in graph. In Proceedings of the Fourth SIAM International Conference on Data Mining, Newport Beach, California, USA (pp. 274–285).

Yang, T., Chi, Y., Zhu, S., Gong, Y., & Jin, R. (2009). A Bayesian approach toward finding communities and their evolutions in dynamic social networks. In Proceedings of the SIAM International Conference on Data Mining, Sparks, Nevada, USA (pp. 990–1001).

Yu, K., Yu, S., & Tresp, V. (2005). Soft clustering on graphs. Advances in Neural Information Processing Systems, Vancouver, British Columbia.

Zeng, D., & Li, H. (2008). How useful are tags?—an empirical analysis of collaborative tagging for Web page recommendation, Intelligence and Security Informatics. Lecture Notes in Computer Science, 5075, 320–330.

Zeng, D., Wang, F.Y., & Carley, K.M. (2007). Social computing. IEEE Intelligent Systems, 22(5) 20–22.

Zhang, H., Qiu, B., Giles, C.L., Foley, H.C., & Yen, J. (2007). An LDA-based community structure discovery approach for large-scale social networks. In Proceedings of 56th Session of the International Statistical Institute, Lisboa, Portugal. (pp. 200–207).

Zhang, Z., Li, Q.D., & Zeng, D. (2010). Evolutionary Community Discovery from Dynamic Multi-relational CQA Networks. Workshop IWCSN 2010 for 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, pp. 83–86.

Zhang, Z., Li, Q.D., Zeng, D., & Gao, H. (2013). User community discovery from multi-relational networks. Decision Support Systems, 54(2), 870–879.

Zheng, X., Zeng, D., Li, H., & Wang, F. (2008). Analyzing open-source software systems as complex networks. Physica A: Statistical Mechanics and its Applications, 387(24) 6190–6200.