Social Community Analysis via Factor Graph Model

Zi Yang, Jie Tang, Juanzi Li, and Wenjun Yang
Department of Computer Science and Technology, Tsinghua University
{yangzi, tangjie, ljz}@keg.cs.tsinghua.edu.cn
Information Center, Planning and Engineering Institute Petrochina Corp. Ltd
ywj@petrochina.com.cn

Abstract

Community analysis attracts considerable interest from multiple discipline domains such as sociology, computer science, and physics. Previously, the macro-level community detection and micro-level pairwise influence between users were studied separately. In this paper, we try to give a systematical investigation of the two problems together. In particular, we formalize the social network in a factor graph model and employ a learning algorithm to estimate the pairwise social influence between nodes. Representative user finding and community structure discovery are then attacked based on the learned social influences. Our experimental results on a real dataset demonstrate the effectiveness and efficiency of the proposed methods.

Keyword: social network, community analysis, factor graph, social influence, representative user

1 Introduction

With the rapid development of web-based social applications and media such as Facebook, Twitter, and Flickr, community analysis in social networks has attracted considerable interest from many different domains. Community analysis in the complex social networks includes many challenging tasks such as social influence modeling, community structure discovery, and representative user finding [8].

Previously, quite a few methods [2] have been proposed for discovering community structures, also known as modularity property of a network [4], in the sense that nodes in the network are partitioned into groups such that there is a higher density of edges within groups than between them [10], but seldom study the other tasks, such as representative user finding and social influence modeling. Social influence is a complex and subtle force that governs the dynamics of all social networks. Learning the social influence can help understand the mechanisms by which communities emerge and change in the complex social networks. For example, a few influential users and their followers naturally form a community. However, previously, the different community analysis tasks, such as representative user finding and community structure discovery, were usually studied separately. In addition, most existing methods have focused on identifying the communities using heuristics [9]. For example Newman et al. [4] propose an algorithm based on greedy optimization of the quantity of *modularity* and Clauset et al. [1] extend the algorithm to scale it up to large-scale networks. However, the methods only consider the network structure and ignore the content information associated with each node. If one further wants to incorporate the

content information, he has to change the definition of modularity and accordingly change the heuristic rules for greedy optimization.

In this paper, we conduct a systematical investigation of the community analysis problem. In particular, we formally formulate the social network in a factor graph model. The model is general and flexible to adapt to different community analysis tasks. Two community analysis tasks, i.e., representative user finding and community structure discovery, are addressed using two instantiations of the graph model. The fundamental model and the learning algorithm for the instantiations are identical for different analysis tasks. The only difference lies in the definition of the factor functions for the two tasks. We evaluate the model on two different real datasets, i.e., a coauthor network and a social network from Digg.com. Experimental results demonstrate the effectiveness of the proposed method.

The rest of the paper is organized as follows: Section 2 and Section 3 formulate the two problems and explain the proposed approaches. Section 4 presents experimental results that validate our methodology. Finally, Section 5 concludes and discusses future work.

2 Representative User Finding

We first study the problem of representative user finding, which is motivated by the fact that many users' behaviors are influenced by another user [7, 10]. Typically, in a social network, some representative users (or influential users) may dominate the other users' behaviors in a community. For example, movie stars' behaviors may easily influence their fans. In this section, we present a factor graph model to formalize this problem and introduce how to solve the model using a loopy max-sum algorithm.

Problem Definition. The goal of representative user finding is to find a pairwise representativeness on each edge in the input social network, and estimate the most representative users for each user. Given a social network $G=(V,E)$, $V=\{v_i\}_i$ is the set of nodes (users), $E=\{e_{i,j}\}_{i.j}$ is a set of directional/undirectional edges representing reciprocal or parasocial relations between users, and $\alpha_{i,j}$ denotes the strength defined on each edge. We assume that each user v_i is associated with a probabilistic distribution on topics $\{\theta_{iz}\}_z$, the problem can be easily extended to find topic-level representative users. Formally, we introduce a set of variables $\{y_i^z\}_z$, each of whose components ranges from 1 to N representing the user that v_i mostly trusts (or relies on) w.r.t. a specific topic z.

Learning with Factor Graph. In the representative user finding problem, two kinds of information should be considered: 1) users are influenced by their friends; 2) users behaviors are finally determined by their own characteristics. Our main idea is to leverage the factor graph model [3] to solve this problem, in which the observation data are cohesive on both local attributes and relationships.

Now we define the proposed model, which is an alternative TFG (Topical Factor Graph) model [7]. In [7], users' similarities are considered to quantify social influence between users, while in our model, the pair-wise relation delivers their representative degrees towards their common friends. The model has the following components: a set of observed variables $\{v_i\}_{i=1}^N$ and a set of hidden vectors $\{\mathbf{y}_i\}_{i=1}^N$. For example, Figure 1 shows a simple example of how the factor graph is built upon a social network. The observed data consist of four nodes $\{v_1,\ldots,v_4\}$, and the edges between the nodes indicate the four social relationships in the social network (see Figure 1a).

The four nodes correspond to hidden vectors $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_4\}$ (see Figure 1b).

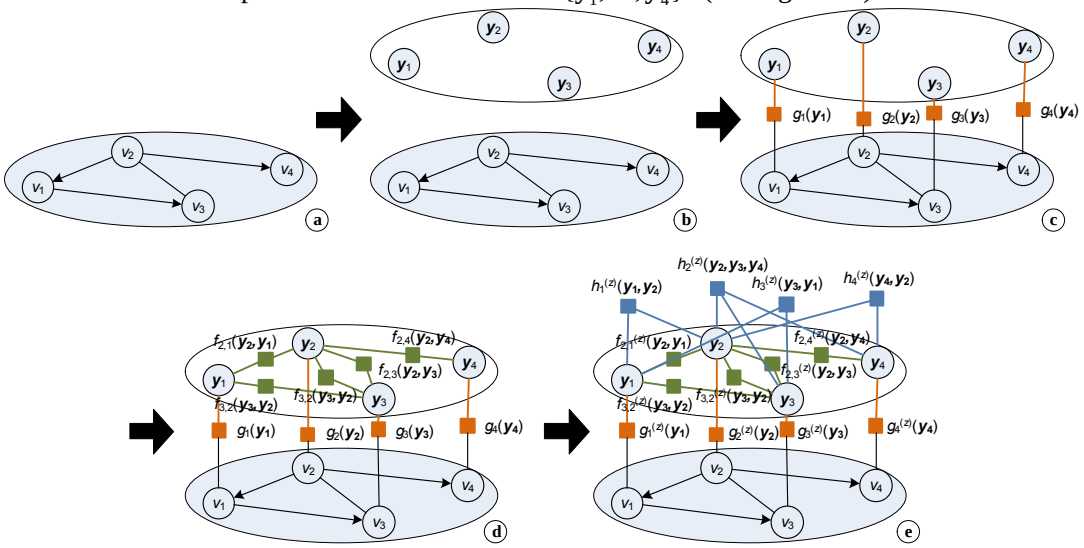


Figure 1: Example of a factor graph for representative user finding

There are three kinds of feature functions: node feature function, edge feature function, and regional feature function. Basically, users in a network would have preferences to follow other users behaviors or to act according to their own wills on a specific topic z. The former is characterized by the edge feature function and the latter is characterized by the node feature function. In this problem, the feature functions are formally defined as follows.

Node Feature Function. $g_i^z(\mathbf{y_i})$ is a function defined on node v_i specific to topic z, describing how likely the user v_i chooses a different user v_j (i.e., $j \neq i$) or himself as a representative. It can be represented by factors shown in Figure 1c, and formally defined as follows. If the user choose his neighbor as his representative, then the node feature function is proportional to the topical similarity and the interaction strength between the two users by considering θ_{jz} and $\alpha_{i,j}$. If he chooses self-representative, then his impact and trustiness should be significant (e.g., at least some other users also choose him as a representative user), and thus the node feature function sums up the topical similarity and the interaction strength between the user and those pointing to him.

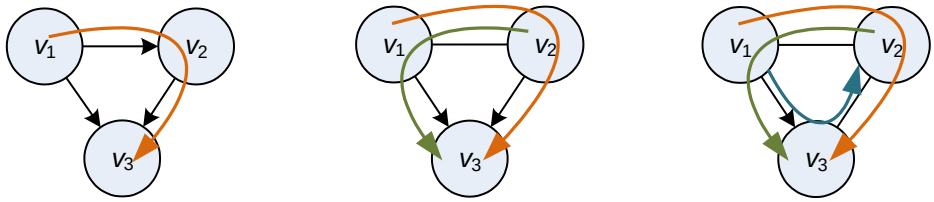
Edge Feature Function. $f_{i,j}^z(\mathbf{y_i}, \mathbf{y_j})$ is a feature function defined on the edge $e_{i,j}$ of the input network specific to topic z to capture dependencies between friend pairs v_i and v_j, and bias the user v_i to be influenced by the user that also influences one of his/her friends v_j, following the theory of social homophily. Formally, for each edge $e_{i,j}$, if two neighboring users v_i and v_i choose the same representative ($y_i^z = y_j^z$), the edge feature function takes a larger value than if two neighboring users choose different representatives, where the value is denoted as bias coefficient. Note that if the edge $e_{i,j}$ is directional, then the suggestion is valid only along the direction of the edge, i.e., v_i may suggest representatives for v_j, but the converse is not. In Figure 1d, 5 dependencies of homophily are defined for 3 directional edges and a bi-directional edge respectively.

Regional Feature Function. $h_k^z(\mathbf{y}_{\mathbf{I(i)} \cup \{\mathbf{i}\}})$ is a feature function defined on the set of neighboring nodes of v_i and itself w.r.t. topic z to avoid "leaders without followers" in the learned model. More specifically, if a representative node v_k is the representative of himself on topic z, then it

must be a representative of at least another node v_i on the same topic z. If the predicted $\{\mathbf{y}_i\}_{i=1}^N$ result in an invalid configuration, then the regional feature function will take 0 as an punishment for v_k's behavior, otherwise it takes 1 to approval it. In Figure 1e, we subsequently append 4 factors representing 4 regional factors on 4 nodes and their neighbors.

In summary, we can define an objective function by considering all the feature functions based on the factor graph theory [3]. Solving such a factor graph with cycles is often intractable. We use a loopy max-sum algorithm to solve the maximization problem, where message variables p_{ijk}^z and c_{ijk}^z are defined on nodes, edges or triangles (three neighboring users) in a social network. One advantage of max-sum algorithm is that all message variables usually have nice explanations as real social interaction processes among users. For example, p_{ijk}^z represent how likely user v_i persuades v_j to take v_k as his representative on topic z, and c_{ijk}^z represents how likely user v_j compliances the suggestion from v_i that he considers v_k as his representative on topic z.

Note that if an edge $e_{i,j}$ is directed, then the suggestion messages are defined along the direction. For example, in Figure 2(a), three connected users v_1, v_2 and v_3 form a triangle structure, but only $p_{1,2,3}^z$ is valid for $p_{i,j,k}^z$, and the arc can be simply understood as v_i recommend to v_k the user v_j. In Figure 2(b), two edges are directed, and $p_{1,2,3}^z$ (the orange arc), and $p_{2,1,3}^z$ (the green arc) are valid messages for $p_{i,j,k}^z$. In Figure 2(c), where only v_1 and v_3 are connected by a directed edge, $p_{1,3,2}^z$ (the blue arc), $p_{1,2,3}^z$, and $p_{2,1,3}^z$ are valid messages for $p_{i,j,k}^z$. That is, influences can only be propagated along the specific direction if it is given.



(a) Triangle with three directed edges

(b) Triangle with two directed edges

(c) Triangle with one directed edge

Figure 2: Example of triangles with directed edges

Based on results of the proposed factor graph model, we can calculate the *representative degree* of each node, which is proportional to the optimized probability $p(y_i^z = j)$, and a sigmoid function converts each value into the ranges of [0,1]. And accordingly *self-representative degree* of person v_i is proportional to the probability $p(y_i^z = i)$.

3 Community Discovery

Many different community analysis tasks can be also solved using the proposed factor graph model. Here we introduce how to solve the community discovery problem using the factor graph. The goal of community discovery is to partition the social network into groups such that there is a higher density of edges within groups than between them [10]. The problem is different, but closely relevant to, the representative user finding problem. For example, a representative user and his/her followers naturally form a community. Based on this intuition, we again formalize the

community discovery problem in the factor graph model.

Problem Definition the objective of *community discovery* is to find a community y_i from $\{1,\ldots,C\}$ for each person v_i, which represents the community that v_i belongs to, such that the preservation of structure is maximized, or the modularity Q of the community is maximized. Modularity is a description on the structure of the community, which indicates the significance of the community structure. The higher value of modularity corresponds to a better division of a network [1]. Intuitively tightly associated pairs in the social network are more probable to be in close proximity in the discovered community structure. To model the community discovery problem using factor graph, we introduce a virtual node u_c for each community c.

Learning with Factor Graph. The chief concern in this problem is the modularity, a representation of the modularity Q is defined as the sum of the edge modularity $q_{i,j}$ on $e_{i,j}$: $\exp(\alpha_{i,j} - k_i k_j / 2m)$, where k_i is the sum of weights of out-edges of node v_i, i.e., $k_i = \sum_{j \in O(i)} \alpha_{i,j}$ and m is the sum of weights of all edges in the network, i.e., $m = \sum_{i,j} \alpha_{i,j}$. Therefore, we respectively define node features and edge features to capture this intuition, and then a factor graph model is constructed based on the factors. Figure 3 shows an example of a factor graph. Similar to the factor graph model for the representative user finding problem, the observed data consist of four nodes $\{v_1,\ldots,v_4\}$, which have corresponding hidden vectors $Y = \{y_1,\ldots,y_4\}$. The difference lies in that we introduce $C = 2$ latent variables for communities, and accordingly define different feature functions for communities.

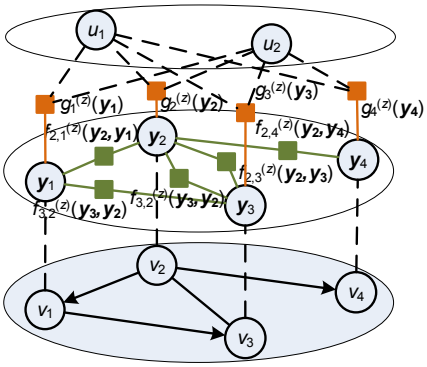


Figure 3: Example of a factor graph for community discovery

There are two kinds of feature functions: node feature function and edge feature function.

Edge Feature Function. $f_{i,j}(y_i, y_j)$ is a feature function defined on the edge $e_{i,j}$ of the input network, representing how likely two connected nodes choose the same community node for each edge, where the bias coefficient of the objective function in the representative user finding is not constant, but dependent on the strength of the edge and the out-degree of the persons, which is equivalent to the definition of modularity $q_{i,j}$ on edges. Formally, if both neighboring nodes v_i and v_j choose the same community ($y_i = y_j$), the factor takes the value of its local modularity; otherwise, it takes the value 1.

Node Feature Function. $g_i(y_i)$ is a feature function defined on node v_i, describing how likely a node v_i takes u_c as his/her social community. Although the product of all the edge feature factors $f_{i,j}$ is sufficient to guarantee the structure of community achieves the maximum significance, and a message passing algorithm could be applied to solve the objective function, but since all the communities are identical to each other, the algorithm cannot make a decision for

each individual. Here we simply achieve this by considering that the communities are represented by the tendency of his/her neighbors' choosing communities in majority.

Again, we can define an objective function by combining all the feature functions and solve it using a message passing algorithm. Specifically, two types of variables p_{ijc} and c_{ijc} are defined on edges in a social network, and explanations given to each type of variables are analogous to those for representative user discovery. Here, p_{ijc} implies how likely user v_i suggests v_j to take u_c as his community and c_{ijc} implies how likely user v_j compliances the suggestion from v_i that he considers u_c as his community. We can also similarly define the representative degree of a person v_i to a community c based on the learned variables p_{ijc} and c_{ijc}.

4 Experiments

In this section, we present the evaluation of the results, and demonstrate the model effectiveness using a case study. The proposed unified framework based on the factor graph model is implemented and publicly available for downloading as well as other related data sets and materials[1].

We extract a coauthor network from an academic search system Arnetminer.org (http://arnetminer.org) [6]. The data set consists of 1,050,021 authors and 3,154,643 coauthor relations. For representative user finding, topic distributions of authors are discovered using a statistical topic modeling approach, Author-Conference-Topic (ACT) model [6]. A sample of such topics is listed on Arnetminer.[2] We will show the result of the representative authors found on seven topics in our experiments. Moreover, for each topic, we extract a subset of the original network with nodes having the highest distribution on this topic.

Results. Table 1 lists 10 most representative persons on 7 different topics discovered by the proposed factor graph model on the coauthor network, and Table 2 lists 5 most representative persons of each community discovered by community discovery method for these different topics. Here we set the number of communities as 5.

We see that the persons who are mostly representative to the community are not necessarily the most "self-representative persons". And there are some persons who are active in 2 or more communities, e.g., Prof. Jiawei Han and Dr. Xiaohua Hu in "Data Mining" network, Prof. Philip S. Yu in "Database System" network, etc.

Table 1: 10 most self-representative persons on 7 different topics for coauthor network

TYPE	REPRESENTATIVE PERSONS
Data Mining	Jiawei Han, Jian Pei, Philip S. Yu, Ke Wang, Qiang Yang, Heikki Mannila, Wei Wang, Eamonn J. Keogh, Martin Ester, Ada Wai-Chee Fu
Database System	Gerhard Weikum, Michael Stonebraker, Michael J. Franklin, Divesh Srivastava, Jennifer Widom, Michael J. Carey, Richard T. Snodgrass, Beng Chin Ooi, Joseph M. Hellerstein, Philip A. Bernstein
Information Retrieval	Mounia Lalmas, Nicholas J. Belkin, Ophir Frieder, Alan F. Smeaton, Mark Sanderson, Stephen E. Robertson, James P. Callan, Chris Buckley, Norbert Fuhr, Amanda Spink
Web Service	Jen-Yao Chung, Wil M. P. van der Aalst, Patrick C. K. Hung, Jun-Jang Jeng, Manfred Rei-

[1] http://arnetminer.org/soinf
[2] http://www.arnetminer.org/topicBrowser.do

	chert, Ying Li, Jian Wu, Schahram Dustdar, Jian Yang, Claude Godart
Bayesian Networks	Didier Dubois, Henri Prade, Philippe Smets, Serafin Moral, David Heckerman, Salem Benferhat, Lluis Godo, Daphne Koller, Luis M. de Campos, Finn Verner Jensen
Semantic Web	Steffen Staab, Stefan Decker, Dieter Fensel, Ian Horrocks, Enrico Motta, York Sure, Bijan Parsia, Carole A. Goble, Deborah L. McGuinness, Jeff Z. Pan
Machine Learning	Bernhard Scholkopf, Michael I. Jordan, Robert E. Schapire, Manfred K. Warmuth, Alex J. Smola, Yoram Singer, John Shawe-Taylor, Satinder P. Singh, Peter L. Bartlett, Yoav Freund

Table 2: Five most representative persons of 5 communities on 4 different topics

TYPE	REPRESENTATIVE PERSONS
Data Mining	Christos Faloutsos Elaine P. M. de Sousa, Spiros Papadimitriou, Floriana Esposito, Longbing Cao
	Hongjun Lu, Wei Wang, Jiawei Han, Jeffrey Xu Yu, Xifeng Yan
	Philip S. Yu, Ke Wang, Jaideep Srivastava, Chengqi Zhang, Michail Vlachos
	Jiawei Han, Dimitrios Gunopulos, Jian Pei, Xifeng Yan, Xiaohua Hu
	Zheng Chen, Qiang Yang, Hua-Jun Zeng, Xiaohua Hu, Chengqi Zhang
Database System	Anthony K. H. Tung, Gao Cong, Limsoon Wong, Heng Tao Shen, Mong-Li Lee
	Christian S. Jensen, Kyu-Young Whang, Richard T. Snodgrass, Dieter Pfoser, Thomas Schwarz
	Philip S. Yu, Hector Garcia-Molina, S. Sudarshan, Philip Bohannon, Jennifer Widom
	C. Mohan, Beng Chin Ooi, Bruce G. Lindsay, Donald D. Chamberlin, Philip S. Yu
	Michael Stonebraker, Guy M. Lohman, Joseph M. Hellerstein, Kenneth A. Ross, Yuqing Wu
Information Retrieval	Michael A. Shepherd, Carolyn R. Watters, Edward A. Fox, Wensi Xi, Bernard J. Jansen
	Chris Buckley, James P. Callan, Djoerd Hiemstra, Wessel Kraaij, Amit Singhal
	Diane Kelly, C. J. van Rijsbergen, Mounia Lalmas, Nicholas J. Belkin, W. Bruce Croft
	Justin Zobel, Kareem Darwish, Nivio Ziviani, Martin Franz, Ian Soboroff
	Weiguo Fan, Edward A. Fox, Ming Luo, Joemon M. Jose, Micheline Hancock-Beaulieu
Web Service	Marlon Dumas, Liangzhao Zeng, Boualem Benatallah, Francisco Curbera, Wil M. P. van der Aalst
	Claude Godart, Olivier Perrin, Vincenzo D'Andrea, Florian Rosenberg, Nirmal Mukhi
	Wil M. P. van der Aalst, Remco M. Dijkman, Boualem Benatallah, David Edmond, Haifei Li
	Francisco Curbera, Athman Bouguettaya, Fabio Casati, Boualem Benatallah, Schahram Dustdar
	Haifei Li, Rama Akkiraju, Ying Huang, Frank Leymann, Jia Zhang

For some other people, they may be willing to consider other persons as their representatives as much as (or even more than) to consider themselves. Each person can choose another person or themselves as their representatives according to different values of bias coefficient. A higher value will broaden the view of each person, and force them to rely on another person more than themselves. In our experiments, we find that with the value set as 1 (no bias), 64.1% of the persons choose themselves as the most representative node, while with the bias set as 2, only 8.0% of persons remain to choose self-representative.

Case Study. We present a case study on the coauthor network with respect to topic "data mining". We discover the most self-represented person is Prof. Jiawei Han with self-representative degree 0.00158 (Cf. Figure 4). We plot six most self-represented persons (in red) together with persons who are mostly influenced by them (in black). The scores in red denote the

self-representative degrees and the score on each edge denotes the pairwise representative degree.

We find that the six self-represented persons are not easily to be influenced by the others, and specifically their representative degrees on others are mostly less than a half of their self-representative degrees. We can also see that some influential users may be also strongly influenced by the others. For example, Prof. Jiawei Han has a high representative degree (0.0064) on Dr. Jian Pei. While some influential users might be very independent, for example, Prof. Heikki Mannila is not associated with others.

Thus we analyze the probability distribution of these top self-representative persons on 5 communities [5], see Figure 5. All the persons contribute to two or more communities, e.g., Prof. Jiawei Han mostly contributes to the second and the fourth community, and Prof. Philip S. Yu devotes to the third community as well, however Prof. Heikki Mannila almost equally contributes to all the communities other than the third community. Nevertheless, we find that different persons have different preferences on communities. As most of the persons exhibit a relatively strong association with the third community, the association between this community and Prof. Heikki Mannila appears weaker than others, and in turn he tends to cooperate with many other persons from the first and the fifth communities that have weaker connections with other self-representative persons, e.g., Prof. Ke Wang, Dr. Jian Pei and Prof. Jiawei Han.

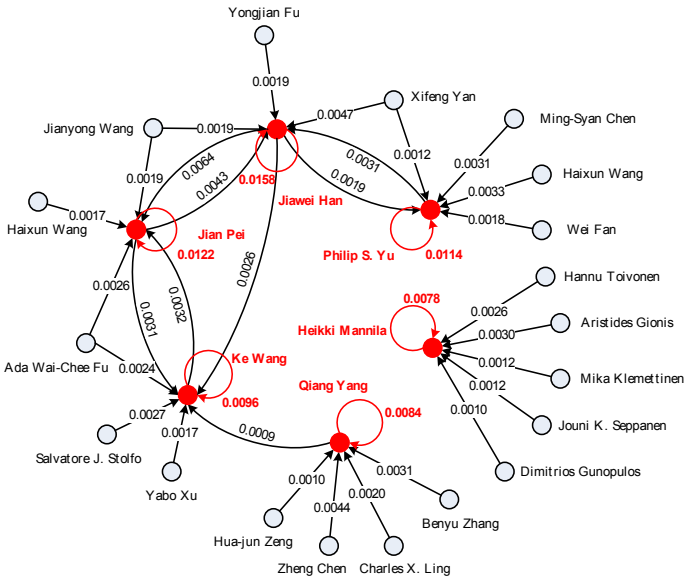


Figure 4: Representative persons and their most represented persons discovered by our method

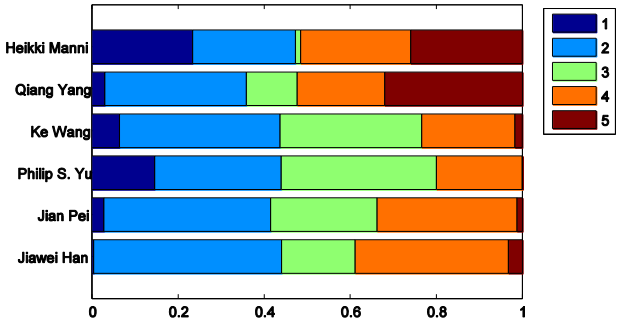


Figure 5: Community distribution analysis for self-representative persons

5 Conclusion and Future Work

In this paper, we study the problem of social community analysis. We formally define the problem of social community analysis and propose a unified factor graph model. The model is very general and flexible. Based on different definitions of the factor functions, we can adapt the factor model to deal with the representative user finding and the community discovery subproblems. Experimental results on two different genres of data sets demonstrate that the proposed approach can effectively find the representative users from social networks. Analysis also shows some interesting results.

The general problem of social community analysis presents an interesting research direction for social network analysis. There are many potential future directions of this work. One interesting issue is to further consider the other sub tasks, e.g., authoritative users finding, in the factor graph model. Another interesting problem is to incorporate some supervised information into the sampling process. For example, for community discovery, two users are restricted to be clustered into one group. It is also interesting to consider the temporal information for community analysis by leveraging a dynamic factor graph, and systematical comparison with existing method for community detection will be conducted in the future work. We are also studying how to integrate the topic detection task into the current factor graph model. Finally, to scale up to massive data, the iterative sum-product algorithm can be designed with a distributed learning algorithm.

6 *ACKNOWLEDGMENTS

The work is supported by the Natural Science Foundation of China (No. 60703059, No. 60973102, No. 61073073), National Key Foundation Research (No. 60933013), National High-tech R&D Program (No. 2009AA01Z138).

References

[1] A. Clauset, M. E. Newman, and C. Moore. Finding community structure in very large networks. *Phys. Rev. E*, 70, 2004.

[2] R. Guimera and L. A. Nunes Amaral. Functional cartography of complex metabolic networks. *Nature*, 433(7028):895–900, February 2005.

[3] F. Kschischang, S. Member, B. J. Frey, and H. andrea Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47:498–519, 2001.

[4] M. E. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69, 2004.

[5] S. Staab, P. Domingos, P. Mika, J. Golbeck, L. Ding, T. Finin, A. Joshi, A. Nowak, and R. R. Vallacher. Social Networks Applied. *IEEE Intelligent Systems*, 20:80-93, 2005.

[6] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. ArnetMiner: Extraction and Mining of Academic Social Networks. In *Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'08)*, pages 990-998, 2008.

[7] J. Tang, J. Sun, C. Wang, and Z. Yang. Social influence analysis in large-scale networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'09)*, pages 807–816, 2009.

[8] F. Wang, K. M. Carley, D. Zeng, and W. Mao. Social Computing: From Social Informatics to Social Intelligence. IEEE Intelligent Systems, 22:79-83, 2007.
[9] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press, 1994.
[10] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

Author's Bio

Zi Yang is a master student at Department of Computer Science and Technology of Tsinghua University. His research interests include machine learning and social network mining.
Jie Tang is an associate professor at Department of Computer Science and Technology of Tsinghua University. His research interests include machine learning, social network mining, and semantic web.
Juanzi Li is a professor at Department of Computer Science and Technology of Tsinghua University. Her research interest includes semantic web and knowledge discovery.
Wenjun Yang is the department manager, of Sales Information Technology Research Department in information center, Petrochina Planning and Engineer Institute. His research interest includes planning, requirement analyzing.