Uncertainty Community Detection in Social Networks

Hong-wei Liu, Li Chen, Hui Zhu, Tao Lu , Fei Liang
Guangdong University of Technology, Guangdong Province, 510520, China
Email: hwliu@gdut.edu.cn, 173404946@qq.com, lutao0211@126.com, 253002415@qq.com, 549940317@qq.com

***Abstract*—Community detection in social networks has a great uncertainty and it has commercial value for business. Many of these communities are the target markets to business. Research found that some factors played important roles for social community mining, such as user profile, information of out-degree and in-degree. This paper takes full advantage of user information to model Hybrid Bayesian networks which has strong ability to deal with uncertain event. We begin mining data from users' profile in social networks and process these data into binary logic data and discrete ones. Combining these two kinds of data, we build Hybrid Bayesian networks, then use graph theory to simplify the process of calculation. Finally, we find that this Hybrid Bayesian networks can provide more accurate and intelligent community detection and it can be applied to different target users and target network communities. Then firms can provide better services for target market.**

***Index Terms*—Bayesian Network, Graph theory, Community detection, Social network**

I. INTRODUCTION

The last decade has witnessed the development of internet and information technology. Various online applications and social networks spring up rapidly [1]. All of these social networks provide new methods for communication and collaboration platforms. People will, consciously or unconsciously, cluster into communities of different scales because of the similar interests, user profiles and topics. In other words, the social network can be divided into various communities. Then it becomes a hotspot how to excavate these social communities better.

In fact, there exist a lot of methods of social community mining. The initial theories include graph Partitioning [2], k-means clustering [2], hierarchical clustering algorithm [3] ,etc, which have evolved into some more operable methods, such as GN algorithm [4], and W-H algorithm. They divide communities based on the degree of closeness and strength of interaction between nodes. However, as we know, users in social networks have many attributes, like different profiles [5], different labels etc. But these methods seldom take advantage of user's profiles and cannot provide personal community mining. There are a lot of factors that affect community mining. So community detection has a certain degree of uncertainty. For this reason, we use Bayesian Network which has strong capabilities to process uncertain events to explore social communities.

Figure1. Users cluster into different communities in social network.

II. THE STRUCTURE OF A SOCIAL NETWORK

In figure1, we can see each node connects with another and cluster into groups of different scales. The structure of a social network can be viewed as a network topology. We classify nodes in a social network into two types, one is target node, and the other is degree node. The number of nodes adjacent to a target node is called its degree. In this paper, degree node means nodes that adjacent to a target node and it consists of in-degree node and out-degree node. In-degree node means followee [6], and out-degree node means follower.

The set of target node is defined as:

$$V = \{v_i\}, \quad i = 1, \dots, n \tag{1}$$

and the set of degree node as:

$$f(v_i) = \{v_i^k\}, \quad \mathrm{k} = 1, \dots, n \tag{2}$$

v_i^k means degree node k which is connected with v_i.

Manuscript received April 31, 2013; revised May 31, 2013

The work was supported by grants (No. 70971027) from the Nationall Natural Science Foundation of China.

This work also was supported by Key Research Institute of Humanities and Social Scienc.

doi:10.4304/jsw.9.4.1045-1049

Through analyzing the attributes of v_i from its profile, its static attribute can be derived:

$$C_g(v_i) = \{1,0\}, \quad g = 1, \dots, n \tag{3}$$

If $v_i$conforms to some static characteristics, $C_g(v_i) = 1$, otherwise $C_g(v_i) = 0$. $C_g(v_i)$ means different kinds of static attributes we extracted, for example, $C_1(v_i)$ represents user's gender and $C(v_i)$ represents user's information of its label.

Similarly, degree node also has its static attribute:

$$C'_g(v_i^k) = \{1,0\}, \quad g = 1, \dots, n \tag{4}$$

If v_i^kconforms to some static characteristics, $C'_g(v_i^k) = 1$, otherwise $C'_g(v_i^k) = 0$. The static attribute of degree node is use to defining dynamic attribute which will be described as followes.

We assume:

$$S_g(v_i) = \frac{\sum_{i=1,k=1}^{n}(C'_g(v_i^k)=1)}{f(v_i)} \tag{5}$$

$$g = 1, \dots, n$$

When $S_g(v_i)$reaches a certain threshold, we define dynamic attribute as:

$$B_g = \begin{cases} b_{g1}, & \text{if} S_g(v_i) > \sigma_1 \\ b_{g2}, & \text{if} \sigma_2\% > S_g(v_i) > \sigma_3\% \\ & \dots \\ b_{gm}, & \text{if} S_g(v_i) < \sigma_s\% \end{cases} \tag{6}$$

$$g = 1, \dots, n$$

We use Graph theory to simplify the process of calculation of $S_g(v_i)$. Suppose $D \in \{0,1\}^{\mathrm{n} \times \mathrm{n}}$ is the adjacency matrix [7] of target node. $d_{\mathrm{ik}} \in \{0,1\}$. If $d_{\mathrm{ik}} = 1$, it means that v_i connects with v_k.

Assume static matrix of target node and degree node as:

$$u_i \begin{cases} 1, \text{if } C_g(v_i) = 1 \ \wedge C'_g(v_i^k) = 1 \\ 0, \text{if } C_g(v_i) = 0 \ \wedge C'_g(v_i^k) = 0 \end{cases} \tag{7}$$

Then

$$U = [u_i] = [u_1 \quad u_2 \quad \dots u_n]^T \tag{8}$$

$$S_g(v_i) = \frac{d_{ik}\, U}{\|d_{ik}\|^2} \tag{9}$$

III. Building Bayesian Network

A. Definition of parameter in Bayesian Network

Bayesian network (BN) is made up of conditional probability and topological structure [8]. The relationship between variables is depends on the reality [9].Thus the connection between variables is the structure of BN while the conditional probability of variable is parameter of BN [10]. BN has a strong self-study ability which means it can learn structure and parameter from data.

We build a BN model so to detect whether a target node belongs to a target community through the attribute of target node and degree node. According to the theory of BN, we assume target parameter T:

$$T = \begin{cases} \text{true, if } v_i \text{belongs to target community} \\ \text{false, otherwise} \end{cases} \tag{10}$$

Then assume C_g and L_g as a parameter:

$$C_g = \begin{cases} \text{true,} & \text{if } C_g(v_i) = 1 \\ \text{false,} & \text{if } \ C_g(v_i) = 0 \end{cases} \tag{11}$$

$$g = 1, \dots, n$$

$$L_g = \begin{cases} \text{maybe} \\ \text{notbe} \end{cases} \tag{12}$$

If L_g=maybe, means that v_i probably belongs to the target community. If L_g=notbe, $v_i$does not belong to target community.

According to the BN theory:

$$\mathrm{P}(B_g|L_g) = \frac{\mathrm{P}(L_g|B_g)\mathrm{P}(B_g)}{\mathrm{P}(L_g)} \tag{13}$$

B. Learning Parameter in Bayesian Network

There are two kinds of methods to learn parameter in BN. One is Maximum likelihood estimate [11], the other is Bayesian estimation method [12]. Because the lack of prior probability, and if there are no examples with missing values in the training set, we assume parameter independence, we use maximum likelihood estimates.

TABLE Ⅰ.
DECISION TABLE

B_1	S_1	$P(L_1 = maybe\|b_{1m})$	$P(L_1 = notbe\|b_1$
b_{11}	$\geq 50\%$	75%	25%
b_{12}	$35\% \leq S_1 < 50\%$	60%	40%
b_{13}	$20\% \leq S_1 < 35\%$	50%	50%
b_{14}	$5\% \leq S_1 < 20\%$	15%	85%
b_{15}	$S_1 < 5\%$	10%	90%

IV. Experimental Studies

A. Defining Attributes of Node and Selecting Parameter

We took microblog which is one of the most popular social networks as an object of the research. First, we assume Z community as the target community. Second, through data mining, we collected structured and unstructured data.

According to equation (10), we assume parameter as:

$$T = \begin{cases} true, & \text{if } v_i \text{belongs to Z} \\ false, & \text{otherwise} \end{cases}$$

To simplify, we assume one static attribute of v_i:

$$C_1(v_i) = \{1,0\}, \quad g = 1, \dots, n$$

If v_i's profile contains the fields that are associated with target community Z, then $C_1(v_i)$=1.For example, the profile of target user v_{123} contains community Z's name, then $C_1(v_{123})$=1.

Then, transform $C_1(v_i)$, to parameters in BN according to equation (11):

$$C_1 = \begin{cases} \text{true}, & \text{if } C_1(v_i) = 1 \\ \text{false}, & \text{if } \ C_1(v_i) = 0 \end{cases}$$

Similarly, static attribute of out-degree node:

$$C'_1\left(v_i^k\right) = \{1,0\}, \quad g = 1, \dots, n$$

Static attribute of in-degree node:

$$C'_2\left(v_i^k\right) = \{1,0\}, \quad g = 1, \dots, n$$

Besides, in microblog, users can follow a group called microgroup. So we view Microgroup as a kind of an out-degree node. Static attribute of Microgroup:

$$C'_3\left(v_i^k\right) = \{1,0\}, \quad g = 1, \dots, n$$

Calculate $S_1(v_i)$, $S_2(v_i)$and $S_3(v_i)$ that corresponding to out-degree node, in-degree node and Microgroup.

$$S_1(v_i) = \frac{\sum_{i=1,k=1}^{n}\left(C'_1\left(v_i^k\right) = 1\right)}{f(v_i)}$$

$$S_2(v_i) = \frac{\sum_{i=1,k=1}^{n}\left(C'_2\left(v_i^k\right) = 1\right)}{f(v_i)}$$

$$S_3(v_i) = \frac{\sum_{i=1,k=1}^{n}\left(C'_3\left(v_i^k\right) = 1\right)}{f(v_i)}$$

In the following section we talk about the calculation of remaining parameters , and structure of BN.

B. *Parameter Learning in Hybrid Bayesian Networks*

L_g is parent node of B_g so that we can assume its initial prior probability value as 0.5. The value of L_g will change with conditional probability of B_g. That is why the initial prior probability value will not affect the final result.

We take B_1 as an example. It is difficult to assess $\mathrm{P}(b_{1m}|L_1 = \text{maybe})$ and $\mathrm{P}(b_{1m}|L_1 = \text{notbe})$. Sowe assess $\mathrm{P}(L_1 = \text{maybe}|b_{1m})$ and $\mathrm{P}(L_1 = \text{notbe}|b_{1m})$ first, then figure out $\mathrm{P}(b_{1m}|L_1 = \text{maybe})$ and $\mathrm{P}(b_{1m}|L_1 = \text{notbe})$. TABLE Ⅰ is the decision table of B_1.

In TABLE Ⅰ, for example$\mathrm{P}(L_1 = \text{maybe}|b_{1m})$=75% means when $S_1(v_i)$ reach over 50%, we consider the probability of $L_1 = \text{maybe}$ is 75%.

B_1 has five variables that means m=5. Then assume $\mathrm{P}(b_{1m})$=0.2. According to Bayes formula, we can acquire $\mathrm{P}(b_{1m}|L_1 = \text{maybe})$:

$$\mathrm{P}(b_{1m}|L_1 = \text{maybe}) = \frac{\mathrm{P}(L_1 = maybe|b_{1m})\mathrm{P}(b_{1m})}{\mathrm{P}(L_1)}$$

Similarly, we can calculate the conditional probability of B_2, B_3.As can be seen in Table Ⅱ (a), Table Ⅱ(b)and Table Ⅱ (c).

TABLE Ⅱ **(a).**
PROBABILITY OF B_1

L_1	maybe	notbe
b_{11}	0.357	0.086
b_{12}	0.286	0.138
b_{13}	0.238	0.172
b_{14}	0.071	0.293
b_{15}	0.048	0.311

TABLE Ⅱ **(b).**
PROBABILITY OF B_2

L_2	maybe	notbe
b_{21}	0.357	0.086
b_{22}	0.286	0.138
b_{23}	0.238	0.172
b_{24}	0.071	0.293
b_{25}	0.048	0.311

TABLE Ⅱ **(c)** .
PROBABILITY OF B_3

L_3	maybe	notbe
b_{31}	0.468	0.179
b_{32}	0.344	0.321
b_{33}	0.188	0.5

As mentioned above, in lack of prior probability of T we use maximum likelihood estimates.

After learning parameters in GeNIe which is a kind of software for BN, we acquired prior probability of C_1 and T which shown in TABLE III and TABLE IV (a)(b).

TABLE III.
PRIOR PROBABILITY OF C_1

true	0.304
false	0.696

TABLE IV. **(a)**

C_1	false							
L_1	maybe				notbe			
L_2	maybe		notbe		maybe		notbe	
L_3	maybe	notbe	maybe	notbe	maybe	notbe	maybe	notbe
true	0.99	0.51	0.40	0.53	0.81	0.89	0.65	0.001
false	0.01	0.49	0.60	0.47	0.19	0.11	0.35	0.999

PRIOR PROBABILITY OF T WHENC_1 = true

TABLE IV. **(b)**
PRIOR PROBABILITY OF T WHENC_1 = false

C_1	true							
L_1	maybe				notbe			
L_2	maybe		notbe		maybe		notbe	
L_3	maybe	notbe	maybe	notbe	maybe	notbe	maybe	notbe
true	0.99	0.97	0.52	0.68	0.41	0.43	0.82	0.89
false	0.01	0.03	0.48	0.32	0.59	0.57	0.18	0.11

C. *Calculating $S_g(v_i)$ based on Graph Theory*

Figure 2 indicats a simple relationship between six users in social networks. The black nodes signify$C'_g\left(v_i^k\right) = 1$ or $C_g(v_i) = 1$. The white ones mean $C'_g\left(v_i^k\right) = 0$ or $C_g(v_i) = 0$.

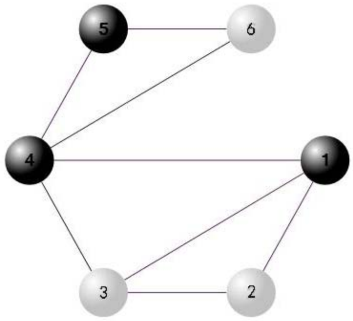


Figure2. Relationship of users in social networks

Suppose node 6 is target user which means we want to calculate $C_g(6)$.Then its adjacency matrix is:

$$\begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix} = \text{D} = \begin{bmatrix} d_1 & d_2 & \dots d_6 \end{bmatrix}^T$$

According to equation (7), static matrix of target node and degree nodes is:

$$U = [u_i] = [1 \quad 0 \quad 0 \quad 1 \quad 1 \quad 0]^T$$

Then

$$S_g(6) = \frac{d_6 U}{\|d_6\|^2}$$

In

$$= \frac{1}{2}[0 \quad 0 \quad 0 \quad 1 \quad 1 \quad 0][1 \quad 0 \quad 0 \quad 1 \quad 1 \quad 0]^T = 1$$

the same way, we can also calculate other target nodes as followes:

$$S_g(v_i) = \begin{bmatrix} \frac{1}{3} & \frac{1}{2} & \frac{2}{3} & \frac{1}{2} & \frac{1}{2} & 1 \end{bmatrix}^T$$

D. *Reliability Test of Hybrid Bayesian Networks*

In order to test the reliability of Hybrid Bayesian, we input the data of training set into this model. For example, there is a user whose number is 59321. His profile contains some fields associated with target community Z, meaning $C_1(59321) = 1$. And its $S_1(59321) = 70\%$, $S_2(59321) = 88\%$, $S_3(59321) = 20\%$. From this percentage, we knew this user's dynamic attribute belongs to $b_{11}, b_{21,} b_{33}$.Then this parameter is into Hybrid Bayesian networks shown in Figure3.

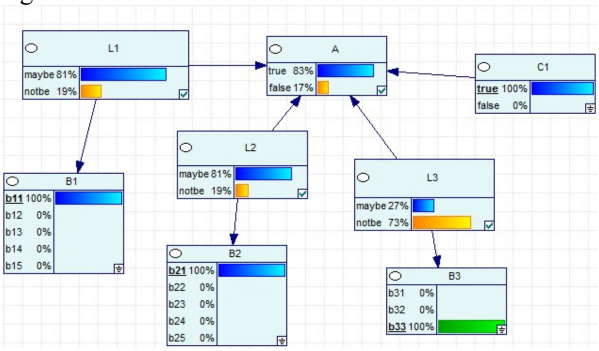


Figure3. The test result of user no.59321

The model in Figure 3 proves that the probability of reliability of the one who is identified as belonging to community Z in reality is 83%.

Figure4 shows the result of probability of user reliability through teston Hybrid Bayesian networks. If a user's mark is 1, that means we can affirm this user is part of community Z. The "Result" means $P(A = true \mid C_1, B_1, B_2, B_3)$.It can be seen from Figure 4 that the simulation result is close to reality.Then we can classify the threshold to meet different requests, for example, when the percentages in "Result" is over 55%, users which satisfies this condition can be considered to belong to community Z.

V. Conclusions

Both products and marketing strategy should be of pertinence. Therefore, in big data era, whether E-commerce companies can accurately find out the target market has become the key to success. Plenty of communities exist in social networks. These communities must include the target markets of companies. So our model is to help enterprises to find out their target communities in social networks. We built a Hybrid Bayesian network with the data from users' profiles. Through structured data, we could obtain Binary logic data and Discrete ones and employ graph theory to optimize the model. This model can provide personal community detection while making use of users' profiles and features of target community. Therefore we could provide personal community detection for each user in social network. Mining community becomes more flexible resulting from taking advantage of conditional probability. This research only provides a basic model and its parameters can be modified depending on different characteristics of different target communities.

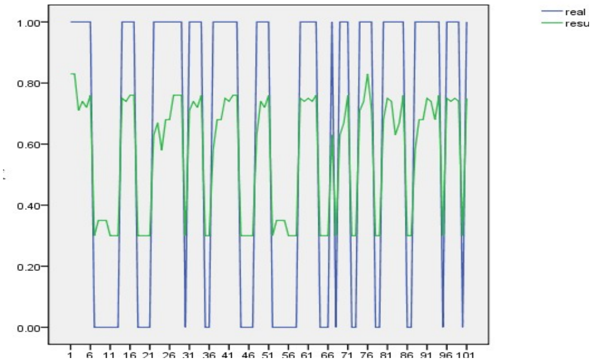


Figure 4. The result of probability of users' reliability

ACKNOWLEDGMENT

The work was supported by grants (No. 70971027) from the National Natural Science Foundation of China. This work also was supported by Key Research Institute of Humanities and Social Sciences at Universities, No. 12ZS0112.

REFERENCES

[1] J.F.F. Mendes, S.N. Dorogovtsev, *Evolution of Networks*. Advances in Physics, vol. 51, issue 4,2002.

[2] M.E.J. Newmana. *Detecting community structure in networks*. The European physical journal B.2004, pp.321-330.

[3] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison Wesley, 2005, pp. 36.

[4] M.E.J. Newmana. *Fast algorithm for detecting community structure in network*. Physical Review E.2004, pp. 69.

[5] R. Xiang, J. Neville, and M. Rogati. Modeling relationship strength in online social networks. In *WWW'10: Proceedings of the 19th international conference on World wide web*, USA, 2010.

[6] Lei Tang , Huan Liu .*Community Detection and Mining in Social Media*[M] .Morgan & Claypool, 2010, pp.18.

[7] S.Wasserman and K.Faust. *SocialNetwork Analysis: Methods and Applications*. Cambridge University Press, 1994, pp. 3.

[8] Heckerman D,Geiger D,Chickering D. *Learning Bayesian Networks: the combination of knowledge and statistical data*. Machine Learning, 1995, pp.197-243.

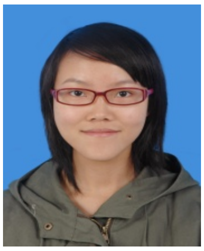
[9] R. Chen and K. Sivakumar. *A new algorithm for learning parameters of a Bayesian network from distributed data*. In Proceedings of the 2002 IEEE International Conference on Data Mining, 2002, pp.585–588,

[10] F. V. Jensen. K. G. Olesen. and S. k. Andersen. *An algebra of Bayesian belief universes for knowledge-based systems*. Network. 1990, pp. 637-660.

[11] Daniel Grossman, Pedro Domingos. *Learning Bayesian Network Classifiers by Maximizing Conditional Likelihood*. ICML '04 Proceedings of the twenty-first international conference on Machine learning, 2004, pp. 46.

[12] George H. John, Pat Langley. *Estimating Continuous Distribution in Bayesian Classifiers*. UAI'95 Proceedings of the Eleventh conference on Uncertainty in artificial intelligence,2011, pp.338-345.

Hong-wei Liu received a B.S in Sun Yat-Sen University in 1983 and received M.E. in South China University of Technology. He is now the professor in Guangdong University of technology of management. His research interests include Network Management and Algorithms of privacy in e-business.



Li Chen was born in 1989. She received a B.S. in 2008 from Guangdong University of technology. Currently she is a master student in Guangdong University of technology. Her major is Management Science. She now focuses on privacy in e-business and data mining in social network.

Tao Lu received a B.S. in 2007 from Hainan Normal University. Currently he is a master student in Guangdong University of technology. He now focuses on Network Management and Privacy in e-business.

Hui Zhu was born in 1988. She received a B.S. in 2007 from Guangdong University of technology. Currently she is a master student in Guangdong University of technology .She now focuses on Social Networks and Empirical research on privacy,

Fei Liang was born in 1988. He is a master student in Guangdong University of technology. He now focuses on Network Management and Algorithms of privacy in e-business.