Prediction of Crowd-Powered Search Performance Based on Random Forest

Tao Wang[1,2], Weiming Zhang[1], Cheng Zhu[1], Kaiming Xiao[1], Zhong Liu[1], and Baoxin Xiu[1]
[1]Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology, Changsha, China, 410073
[2]The State Key Laboratory of Management and Control for Complex Systems, Chinese Academy of Sciences, Beijing 100190, China.
Email: {wangtao, wmzhang, zhucheng}@nudt.edu.cn

***Abstract*—Crowd-powered search is a form of crowdsourcing scheme which involves collaborations among voluntary Web users. Most popularly known episodes are succeed, while search tasks often failed in fact. In this research, we analyzed the factors which related to the performance of crowd-powered search though human flesh search (HFS) episodes, and predicted search performance based on these factors. We have analyzed 2.3 million microblogs about HFS which involved more than 1.3 million users over 2 years in Sina Weibo—the most popular social media site like twitter in China. Some useful features are found. Based on these features, we predict the performance of HFS episodes based on random forest method. The results of classification shown that our model performed good at differentiating these succeed and failed episodes automatically.**

***Index Terms*—crowdsourcing, crowd-powered search, human flesh search, social computing, random forest, online collaboration**

I. Introduction

Crowd-powered search [1], [2], one kind of crowdsourcing [3], is a new form of search and problem solving scheme that involves vast collaboration among voluntary Web users. Human flesh search (*HFS*)[4]-[6] is a Chinese original version of crowd-powered search[7], which is similar with the small world experiment[8]: the initiator starts a search of the identity of a person or the truth of an event through his or her connections or through posting the information online, if a participant does not know the answer, he or she would spread to his or her connections. The process continued until the results is found, or the participants lost interests. Human flesh search is an interesting phenomenon which first appeared in China and has triggered some sensational episodes, such as the South China Tiger event[9]. *HFS* also happened in other countries [10].

Crowdsourcing systems have shown the power of crowds, such as *ReCAPTCHA* [11], while it works not very well in every task, especially for complex tasks. In fact, it often failed even for simple tasks. In the small world experiment, Most letters did not send to the right person's hands [8]. Similar with this experiment, not all *HFS* succeeded. In fact, most of *HFS* episodes failed eventually. Which kind of *HFS* episodes would be succeed is an interesting issue. This paper attempts to find some useful features to identify the status of *HFS* episodes. We can make predictors with these features to predict one *HFS* episodes would be succeed or failed.

We define our task as a binary classification task. For each *HFS* episode, we predict it is 'succeed' or 'failed'. We make these predictions based on features extracted from each episode's profiles, which include statistics and collaboration networks topology attributes. The episode is failed if the search result did not appear until the end of the *HFS* episode.

The paper is organized as follows: In Section II, we introduced the data collection. Then we analyzed the factors which related to performance of *HFS* in Section III. These factors were used to make predictor in the next section to predict the status of *HFS*. At last, we conclude with discussion of future work.

II. Data Set

A. Profile of Data Set

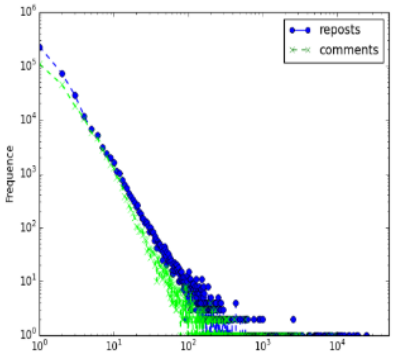
We have collected 530 *HFS* episodes from Jan. 2011 through Mar. 2013 from *Sina Weibo* [12], [13]. For all those episodes, we have collected the basic information including event profile, starting and ending time, population size of participants involved, final result, etc. For each participant, we collected his/her name, location, gender, followers' amount, registered time and so on. Furthermore, we have excluded the episodes with less than 100 participants. In the end, the dataset used in this study contains 307 *HFS* episodes with almost 2 million posts generated by nearly1.4 million users. The most influential episode is the search of a lost baby in a car which involved more than 200 thousand reposts.

The distribution of the reposts amount is as Table I. The data collection involved identifying *HFS* episodes manually (via browsing and searching on the Web), and searching news media for second-hand reporting and comments about *HFS* episodes both manually and automatically. Fig. 1 shown some profile of our data set. Participants' reposts and comments distribution for all episodes is shown in Fig. 1(a), comments are almost less

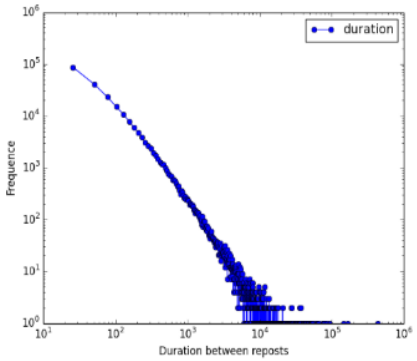
Manuscript received February 19, 2015; revised April 28, 2015.

doi: 10.12720/jiii.3.4.293-298

than reposts in our dataset. The distribution of interval (in seconds) between two participants are shown in Fig. 1(b). One participant often joined *HFS* after another in several seconds, while some episodes may also last several days when next person joined it.



(a) Reposts and Comments



(b) Interval

Figure 1. The profile of dataset.

B. Episodes Status

For studying the factors about episodes status, the first step is to identify which episodes are successful. We identified its status by annotators manually. To avoid biased judgment from one person, we use cross-checking among multiple annotators, as what they have done in many other works. Three individual experts identified successful *HFS* episodes, and then we consider one episodes to be succeed or failed if two or more individuals agreed on the decision. 47 successful episodes and 261 failure episodes are identified at last. The profile of retweets distribution is as Table I.

TABLE I. DISTRIBUTION OF THE COUNT OF RETWEETS

	Failed		**Succeed**	
Range	*Episodes Count*	*Average Forward Count*	*Episodes Count*	*Average Forward Count*
100 and 500	128	217	15	250
500 and 1000	42	729	8	786
1000 and 5000	52	2196	14	2208
5000 and 10000	12	7462	2	5962
10000 and 50000	22	22251	3	17892
50000 and 150000	4	71380	4	80533
150000 and 250000	1	160104	1	235073
Total	261	4587	47	14122

III. THE IMPACT FACTORS OF CROWD-POWERED SEARCH

We analyzed the factors which related to the crowd-powered search tasks performance from two perspectives in this section. Statistics factors and collaboration networks' factors are analyzed by corresponding metrics.

A. Statistics Factors

Fig. 2 shows the statistic features of *HFS* episodes, which include duration, average duration, followers, average followers, average mentioned frequency and frequency of users who participated more than once in the same episode. In these sub-figures, the x-axis represents the value of features, the y-axis represents the accumulated succeed rate corresponding to the feature. For avoiding bias situation, we statistics the success rate in different situations. For one, two, three, four succeed episode, we compute its success rate, respectively. The four lines in the sub-figures represents the statuses under the four situations.

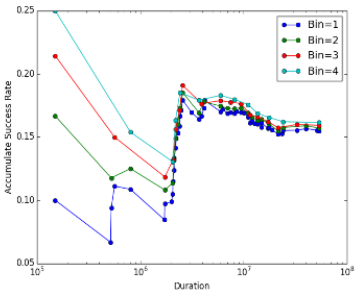
Duration of the *HFS* episode reflect how long the *HFS* episode lasted. As shown in Fig. 2(a) and Fig. 2(b), in our data set, the *HFS* episode lasted from less than one hour to more than one year, the average time duration is nearly half a month. Some persons want to find the target of *HFS* episodes after even one year, which resulting in a long average duration at the end of the stage. Surprisingly, not the longer the episode lasts, the greater of success rate, while episodes that last longer often ended in failure. Collectively, the average interval of each participants lasted in one episode is inverse proportion to the succeed rate, even though it is not true at the beginning of the average interval, and some outliers in middle stage.

The spreading scope is critical to connect the initiator and the target in *HFS*. Two communication channels, followers and mention mechanism, are considered in this work. Followers can receive information automatically from their master, and the participants can invite others person who may be help by mention channel. Followers played extremely critical role in the event propagation as the spread mechanisms of *Weibo*. Followers will receive a reminder when their masters send one *Weibo*. If they are not interested in and know nothing about the truth, they may just be onlookers, while they would participate if they are interested in and retweets the *HFS* event, this behavior will continue to spread to their followers, and so on.

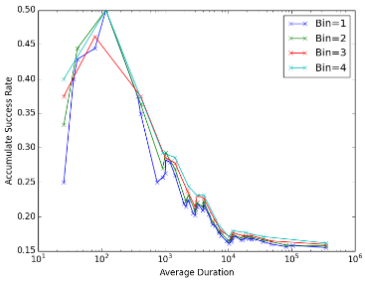
As shown in Fig. 2(c) and Fig. 2(d), the success rate is proportional to absolute amount of one episode's followers at large scale episodes, while it does not work at small scale. To our surprise, there is negative correlation between amount of average followers and its success rate in a *HFS* episode. Lower average followers amount one participant owned, higher success rate in the episode, even it does not fit all episodes very well.

Mention others is another spread way on *Weibo*, one can contact with person who does not follow her/him through mention her/him. As shown in Fig. 2(e), the number of average mention of each participant is

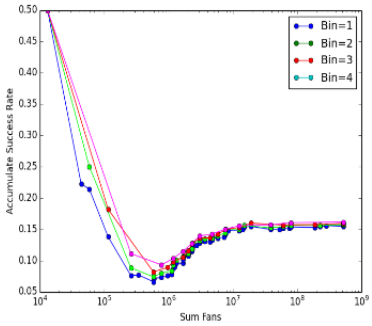
proportional to success rate. But this effect fluctuate at the middle stage.



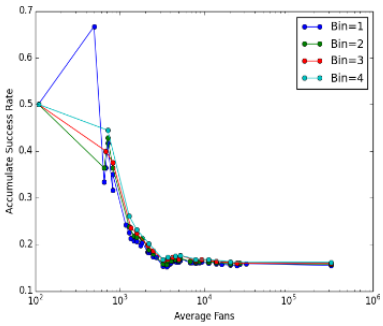
(a) # of Duration vs. Success Rate



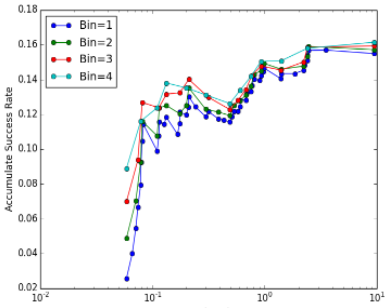
(b) # of Avg. Duration vs. Success Rate



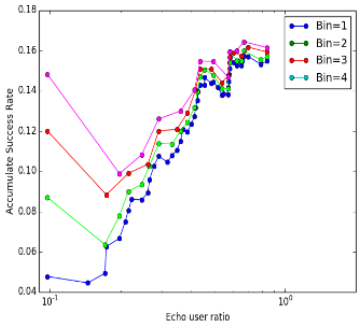
(c) # of Followers vs. Success Rate



(d) # of Avg. Followers vs. Success Rate



(e) # of Avg. @ Times vs. Success Rate



(f) # of Repeated Participants vs. Success Rate

Figure 2. The statistics factors vs. success rate.

There are always some participants involved in the search process more than once in a *HFS* episode. They may report for some search progress, or ask the results of someone else's, or make some inferences, etc. These people often constitute key part of *HFS* search groups, who is also the key to finish the search process. Fig. 2(f) shows the repetition rate is proportional to the success rate. The higher repetition rate, the greater the chances of success, although there are also some exceptions.

B. Collaboration Networks' Factors

HFS tasks often finished by crowds collaborated with each other. We constructed collaboration networks to describe the collaboration in each *HFS* episode. In a *HFS* collaboration network, each node is corresponding to a unique participant. If participant A retweeted participant B's tweets, or mentioned B, there is a link from A to B. *Figure 3* shown the networks' features vs. success rate. In these sub-figures, the x-axis represents the value of networks' features, the y-axis represents the accumulated succeed rate corresponding to the feature. Same as Fig. 2, the four lines in the sub-figures represents the statuses under the four situations.

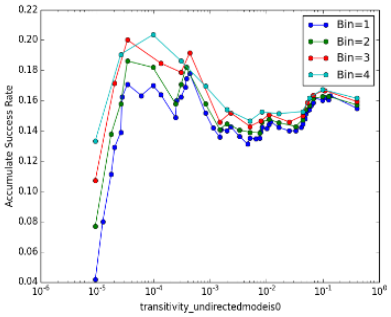
Transitivity[14] measures the probability that two neighbors of a node are connected. The average transitivity vs. successful rate of *HFS* collaboration networks is shown as Fig. 3(a). In our dataset, the average transitivity of the *HFS* collaboration network is range from less than 0.0001 to more than 0.1.The success rate fluctuate with average transitivity increase. The transitivity of collaboration network between 0.01 and 0.1 has more chance to succeed, even though there are some outliers.

The density vs. successful rate of *HFS* collaboration networks is shown as Fig. 3(b). The density of the *HFS* collaboration network is range from less than 0.0001 to more than 0.1, most of them are around 0.001. We observe significant success rate decrease with the density increasing. This means that the sparser of *HFS* collaboration network is, the closer to succeed.

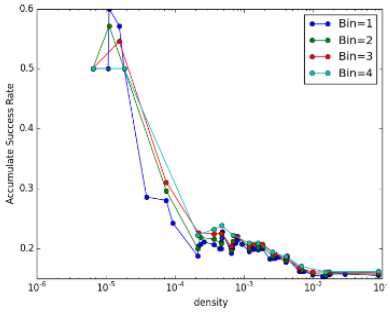
In the collaboration networks, the cluster may represents the discussion or collaboration groups. It is critical for *HFS* episodes to find the targets. In this work, we detected the clusters based on *G-N* algorithm[15]. The two nodes connected each other are identified as strong connection, while just one connected the other is

identified as weak connection. The number of clusters can reflect the discussion efforts in the *HFS* episodes. The number of strong connected clusters and weak connected clusters vs. successful rate of *HFS* collaboration networks are shown in Fig. 3(c), and Fig. 3(d), respectively. They have the same trend as a whole, more clusters, closer to succeed, while the increase trend is more obvious with weak connected clusters. We also note that a few outliers exist at the beginning. They have something not in common as well. Compared with strong connected and weak connected clusters, we can find that the there is no causality between the number of weak and strong connected clusters. Some collaboration networks have more weak connected clusters, while they have fewer strong connected clusters, and vice versa.

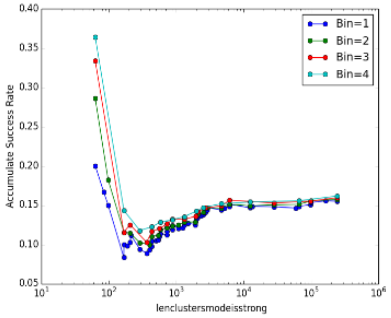
The average degree vs. successful rate of *HFS* collaboration networks is shown in Fig. 3(e). The average degree range from 2 to 11, which means one participant only have two neighbors in some episodes on average, while it may have more than 10 neighbors on average in some other episodes. We noticed that, the *HFS* episode with the higher average degree, the closer to succeed, which means the participant have more neighbors on average in the collaboration networks, the episodes have more chance to succeed.



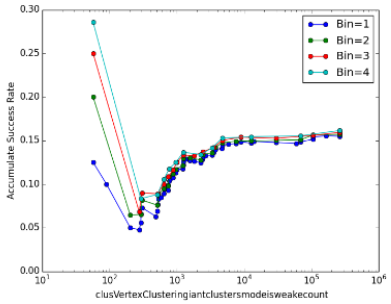
(a) Transitivity vs. Success Rate



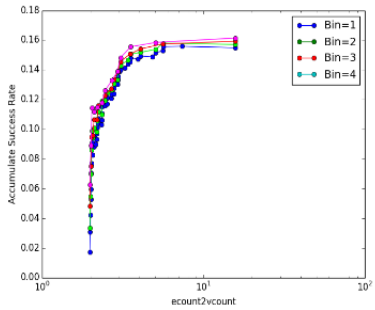
(b) Density vs. Success Rate



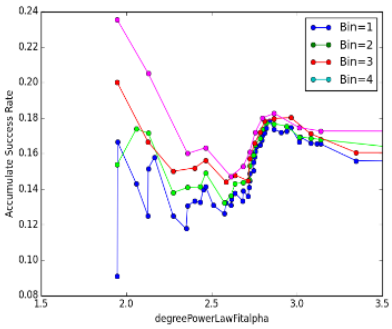
(c) # of strong connected clusters vs. Success Rate



(d) # of weak connected clusters vs. Success Rate



(e) Avg. Degree vs. Success Rate



(f) Power of Degree Distribution vs. Success Rate

Figure 3. The collaboration networks' factors vs. success rate.

In our dataset, the degree of collaboration networks of episodes are almost followed power-law distribution. The power-law exponents of degree distribution vs. successful rate of *HFS* collaboration networks is shown in Fig. 3(f). It is obviously that most of the power-law exponents are less than 3.0 but more than 2.5. They are portioned with success rate at this scope, while it fluctuates beyond this scope, no matter it is less than 2.5 or more than 3.0.

IV. Crowd-Powered Search Performance Prediction

For predicting the performance of *HFS*, we simply treated this as a binary classification problem. We follow a classification framework where the goal is to predict whether a candidate episode *e* is a succeed episode or failed episode. To build a classifier *c*

$$c : e \rightarrow \{succeed\ episodes, failed\ episodes\}$$

As our dataset is very small and bias. There are many outliers may destroy the performance of classifier. We tested several classification algorithms, such as logistic regression, naïve bayes, support vector machine *(SVM)*

and tree-based algorithms etc. At last, we choose *random forest* as candidate machine learning method to predict the status of *HFS* episodes, as it performed more robust for outliers than other algorithms in our dataset.

Random forest algorithm is proposed by *Breiman and Cutler* in 2001. It runs by constructing multiple decision trees while training the class that is the mode of the classes output by individual trees[16]. It has improved performance over single decision trees.

Given that the train and test sets are sometimes highly imbalanced, we favor evaluating these models using the F1 scores etc. To measure the effectiveness of our classifier, we compute precision, recall, and F-measure under different percent of training dataset.

A. Features

Most of features are analyzed based on the findings we reported in the previous parts. We mainly illustrate the space of features from two aspects: statistics view and collaboration networks view which arise from our findings. Except for the basic features we analyzed above, we expand the features space from participant's collaboration relationships etc. We note that not all participants contribute clues for *HFS* tasks. Therefore, we do further filtering for the participants in order to delete nodes and links which has little contribution as much as possible. As of this, we defined two types of collaboration networks, one is the whole collaboration network, and the other is the core part of the whole collaboration networks which excluded the participants with little contribution. The participants with little contribution are identified as the participant neither mention others, nor be mentioned or be retweeted by others, which means the indegree of the nodes is zero in the whole collaboration networks. In other words, the participants in the core networks have retweeting or mentioning activities, or being mentioned by other participants. As of this, each feature have two dimensions.

B. Results

TABLE II. CLASSIFICATION PERFORMANCE.

#% of Training set	Precision	Recall	F1
60%	73.96	26.66	39.19
70%	70.95	30.21	42.38
80%	70.22	34.13	45.93
90%	68.44	38.85	49.56
95%	71.30	37.46	49.11

Using the classifiers setup described above. We run the classifier for 100 times for each experiment. Their precision ranges from 68.4% to 74% under different percent of training dataset. In particular, 60% training data set performed best at precision, while the recall is proportion to the percent of training dataset. It ranged from 26% to 38.85%. The details of prediction performance is listed as Table II.

This positive result shows that our classification model is promising to predict new *HFS* episodes in the future. Our framework works effectively, though the results is not very good. It is a beneficial attempt, which verified that the statuses of *HFS* can be predicted from its profile etc. Similar with *HFS*, some other crowdsourcing projects may also be predicted as they developed. Even though other crowdsourcing episodes may show different styles, their behaviors and observable features may share the same patterns. This can help scientists to get more understanding on the power of crowds, and help business man todiscover more promising projects.

V. CONCLUSION AND DISCUSSION

In this paper, we analyzed the factors about *HFS* succeed or not, including the structure of the spreading scope and spread or propagation quality. Based on those observations, we presented a classification model for predicting the *HFS* episodes status. The results shown that our model is effective for classification.

There are still a lot of efforts to make next. Firstly, more episodes are needed for training the machine learning algorithms. Secondly, more features can be extracted, such as the difficulty of the episode itself, further dividing *HFS* into different types, the behavior pattern under various stages, and so on, are worth further study in future. Thirdly, the preprocessing is needed for the features in this work as they are not independent, some of them rely on each other.

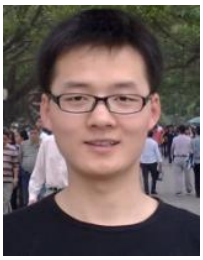
ACKNOWLEDGMENT

This work is partly supported by Hunan Provincial Innovation Foundation for Postgraduate, the National Natural Science Foundation of China though grants No. 61273322, 71471176, and No. 71232006.

REFERENCES

[1] Q. Zhang, F. Y. Wang, D. Zeng, and T. Wang, "Understanding crowd-powered search groups: A social network perspective," *PloS One*, vol. 7, no. 6, pp. e39749, 2012.
[2] Q. Zhang, Z. Feng, F. Y. Wang, and D. Zeng, "Modeling cyber-enabled crowd-powered search," presented at The Second Chinese Conference on Social Computing, Beijing, December 18-19, 2010.
[3] A. Tarrell, N. Tahmasbi, D. Kocsis, A. Tripathi, *et al*., "Crowdsourcing: A snapshot of published research," presented at the Nineteenth Americas Conference on Information System, Chicago, Illinois, August 15-17, 2013.
[4] F. Y. Wang, D. Zeng, J. A. Hendler, Q. Zhang, *et al*., "A study of the human flesh search engine: Crowd-powered expansion of online knowledge," *Computer*, vol. 43, no. 8, pp. 45-53, 2010.
[5] C. H. Chao and Y. H. Tao, "Human flesh search: A supplemental review," *Cyberpsychology Behav. Soc. Netw.*, vol. 15, no. 7, pp. 350-356, 2012.
[6] T. Wang, Q. Zhang, J. Fu, X. Wang, and S. Zheng, "What is the uniqueness of growth pattern in human flesh search organization?" *Intelligence and Security Informatics*, vol. 8039, pp. 75-81, 2013.
[7] F. Y. Wang, D. Zeng, Q. Zhang, J. A. Hendler, and J. Cao, "The Chinese 'Human Flesh' Web: The first decade and beyond," *Chin. Sci. Bull.*, vol. 59, pp. 3352-3361, September 2014.
[8] S. Milgram, "The small world problem," *Psychol. Today*, vol. 2, no. 1, pp. 60-67, 1967.
[9] C. Holden, "Rare-tiger photo flap makes fur fly in China," *Science*, vol. 318, no. 5852, pp. 893, 2007.
[10] C. Shirky, *Here Comes Everybody: The Power of Organizing Without Organizations*, Penguin, 2008.
[11] L. Von Ahn, B. Maurer, C. McMillen, D. Abraham, and M. Blum, "Recaptcha: Human-based character recognition via web security measures," *Science*, vol. 321, no. 5895, pp. 1465-1468, 2008.

[12] S. Chen, H. Zhang, M. Lin, and S. Lv, "Comparision of microblogging service between *Sina Weibo* and Twitter," in *Proc. International Conference on Computer Science and Network Technology*, 2011, vol. 4, pp. 2259-2263.
[13] L. Yu, S. Asur, and B. A. Huberman, "What trends in Chinese social media," arXiv:1107.3522 [cs.CY], 2011.
[14] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world'networks," *Nature*, vol. 393, no. 6684, pp. 440-442, 1998.
[15] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Phys. Rev. E 69, 026113*, 2004.
[16] L. Breiman, "Random forest," *Mach. Learn.*, vol. 45, no. 1, pp. 5-32, 2001.



Tao Wang is a PhD candidate in the College of Information Systems and Management at the National University of Defense Technology, Changsha, China. He is a visiting student in The State Key Laboratory of Management and Control for Complex Systems in Chinese Academy of Sciences. His major interests include social computing, social network analysis and cyber-enabled social movement organization.

Weiming Zhang received the M.Sc. and Ph.D. degrees from National University of Defense Technology (NUDT), Changsha, China, in 1992and 2001, respectively. Currently, he is a professor of NUDT. His research interests are in system of systems, system engineering.

Cheng Zhu received the M.Sc. and Ph.D. degrees from National University of Defense Technology (NUDT), Changsha, China, in 2000and 2004, respectively. Currently, he is a professor of NUDT. His research interests are in information management and decision-making support technology.

Kaiming Xiao is a postgraduate student in the College of Information Systems and Management at the National University of Defense Technology, Changsha, China. His major interests include complex network analysis, network optimization and game theory for networks.

Zhong Liu received the M.Sc. and Ph.D. degrees from National University of Defense Technology (NUDT), Changsha, China, in 1997 and 2000, respectively. Currently, he is a professor of NUDT. His research interests are in information management and decision-making support technology.

Baoxin Xiu was born in 1977. He received the B.S. degree in Applied Mathematics, and the Ph.D. degree in Management Science and Engineering from the National University of Defense Technology (NUDT). Now he is an associate professor in Information Systems Engineering Laboratory at NUDT. His research interests include granular computing, computational and mathematical organization theory.

