Available online at www.ijpe-online.com

International Journal of Performability Engineering

vol. 14, no. 2, February 2018, pp. 349-356
DOI: 10.23940/ijpe.18.02.p16.349356

Inferring Gender of Micro-Blog Users based on Multi-Classifiers Fusion

Jinghua Zheng[a,*], Shize Guo[b], Liang Gao[b], Di Xue[c], Nan Zhao[d], Huimin Ma[a]

[a]*Hefei electronic engineering institute, Hefei, 230037, China*
[b]*Institute of North Electronic Equipment, Beijing, 100083, China*
[c]*Army Engineering University, Nanjing, 210007, China*
[d]*CAS Institute of Psychology, Beijing, 100101, China*

Abstract

Knowing user demographic traits offers a great potential for public information. Most researches have used local features to predict user demographic traits. Since this method did not make the most of user global features, the prediction performance was low. In this paper, our goal tries to use an ensemble learning method to improve the prediction performance through multi-classifiers fusion. Our work makes three important contributions. Firstly, we show how to predict Sina Micro-blog users' genders based on his/her text published on the social network. Secondly, we show that user's personality traits can also be used to infer gender. And last and thirdly, we propose multi-classifiers fusion to predict users' genders, and give the experimental results that validate our method by comparing it with a different local features dataset. Our experiment demonstrates that our method can improve the accuracy rate, the recall rate of prediction, and the F value.

Keywords: Sina micro-blog; gender prediction; multi-classifiers fusion; prediction accuracy

(Submitted on November 7, 2017; Revised on December 19, 2017; Accepted on January 26, 2018)

1. Introduction

In the Internet world, so many things must depend on the Internet. The rapid growth of social networks has produced an unprecedented amount of user-generated data, which provides an excellent opportunity for information mining. Thus, the data information becomes the most important and most crucial problem. Knowing users' views and demographic traits offers a great potential for personalizing web search results or related services such as query suggestion and query completion, as well as even determining search targets for the police. At present, many researchers have proven that information from social network can predict users' many attributes, such as gender, age, health, personality, etc. M. Kosinski et al. [4] testified that the information from social network can predict users' private traits and attributes such as sexual orientation, ethnicity, religion, gender, age, etc. Predicting a person's private traits from digital records of human behavior can not only be more real-time and precise, but also require less device utilization.

Most gender prediction researches are using social network data including connection relationship, status data, linguistic content and behaviour.

Inferring user demographics by using the connection relationship. The connection relationship includes connections between friends, group relationship, and node relationship, etc. A. Mislove et al. predict users' identity attributes. They predict users' identity attributes by using a community-finding algorithm [6]. J. He et al. created social network causal models with Bayesian network and testified that the connection relationship can be used to predict users' hidden attributes [2]. W. Xu et al. [16] predict users' identity attributes with different social relationships. For example, a user's age can be predicted by school connections, and friend connections can be used to predict a user's gender. E. Zheleva et al. [18] use both links and groups to predict a Facebook user's sensitive attributes, and the accuracy of gender prediction is 77.2%.

* Corresponding author.
E-mail address: zhengjh1001@163.com

Inferring user demographics by using the status data. The status data include the number of fans, the number of collection, the number of friends, browse website, etc. Michal K et al. [4] testified that the web browsing record can be used to predict a user's identity attribute including gender, age, occupation, political orientation, etc. With the ODP classification algorithm, B. Bin et al. [1] used a Facebook user's number of likes to infer his/her gender, age and religion, etc.

Inferring user demographics by using linguistic content. M. Pennacchiotti et al. used the linguistic content of user tweets, along with their other social features to predict the political orientation, ethnicity and the favourite business brands of Twitter users [8]. J. Otterbacher inferred the author gender of IMDB reviews based on stylistic and content features [7]. Q. Tang et al. obtained some gender-biased feature words from their texts, and showed the method of gender-biased feature words combined with gender-biased personal appellations and the features of a character's name has a good prediction performance [12].

Inferring user demographics by using user's behavior. I. Weber et al. [14] relied on user clicks on political blogs annotated with learning to assign a leaning score to queries. J. Ying et al. [17] showed that the user demographics can be predicted according to their mobile usage behavior, such as the number of text messages sent or received. C. Shen et al. [9] firstly used users' mouse dynamics including operation, keystroke and touch to predict users' gender, age, ethnicity, and etc.

Until now, most researchers have used machine learning to predict a person's identity attribute. M. Kosinski et al. used linear regression and logistic regression algorithm to predict a user's gender and age, and Area Under Curve(AUC) to evaluate the performance of their model, the AUC being 0.87 and 0.7 respectively [4]. B. Bin et al. used the number of likes on Facebook to predict a user's gender and age, and the AUC is 0.84 and 0.77 respectively [1]. With Log-linear models, S. Volkova et al. predict a user's gender, age and political affiliation based on the Twitter profiles and postings of her friends [13]; the accuracy of user's gender prediction being 0.802.

In recent years, most researches have used classification/regression machine learning to train different dataset, i.e., they used local features to predict a user's demographic traits without using the user's global features. The main challenge lies in the fact that only a very limited amount of data is available to allow training models to predict the demographics based on social network data. It is so hard to get enough data related for training that the prediction performance is not satisfactory.

In this paper, our dataset includes three parts. Firstly, we collected 10 Micro-blog features of 1604 Sina Micro-blog users. Secondly, we got 3.7 million texts from these Sina Micro-blog users' postings to extract 102 psychology features based TextMind System. Thirdly, we extract the users' Big-Five personality information to predict the attribute based on our prior work of personality prediction of Sina Micro-Blog users. We use multi-classifiers decision fusion method to improve the prediction performance.

Hence, the paper makes three important contributions:

1) We show how to predict a user's demographics traits based on his/her text published on the network, such as Sina Micro-blog.

2) We show that a user's personality traits can also be used to infer gender.

3) We propose multi-classifiers fusion to predict users' gender, and give the experimental results to validate our method by comparing it with different local features dataset.

We trained the three type datasets independently, and then trained them simultaneously. We utilize a logistic regression algorithm to predict Sina Micro-Blog users' genders and ages. Our experiment proves that the dataset including psychological and personality data can improve the predictive performance of precision rate, recall rate and F-measure.

The rest of this paper is organized as follows: In Section 2, we will introduce data and method. In Section 3, we will explicate our dataset and experiment, and give the experiment results followed by a discussion about the results. Finally, we will give a conclusion in Section 4.

2. Data and Method

2.1. Data

Our experiment will use the dataset of Sina Micro-blog Users, including Micro-blog feature data as well as micro-blog text. A user's feature vector consisted of a set of N features computed over his Micro-Blog content. We collected 1752 Sina Micro-blog users' micro-blog data. Some features required pre-processing a subset of labeled users (e.g., transforming the

non-numerical data into numerical data). So, we got 12 micro-blog features, 102 psychology features and 5 personality features. And finally, we got 1106 data samples of Sina Micro-blog users that are active.

Micro-blog features

Name (the nickname of the Micro-blog user), age (517 18+ users and 589 25+ users), city, followers_count (the number of the user's fans), friends_count (the number of the user's friends), statuses_count (the number of the user's poster), favourites_count (the number of collection), bi_followers_count (the number of users following to each other), description (Sina Micro-Blog profile), verification (whether the user is verified).

Psychology features (102 features)

The first person singular/plural pronoun (the frequency on the use of first personal pronouns in the text), the second person singular/plural pronoun (the frequency on the use of second personal pronouns in the text), the third person singular/plural pronoun (the frequency on the use of third personal pronouns in the text), sentiment word (the number of sentiment word in the text, such as happy), social word (the number of social word in the text, such as study), positive emotional word (the number of positive emotional word in the text, such as love.), negative emotional word (the number of negative emotional word in the text, such as disgusting), cognitive word (the number of cognitive word in the text, such as imagine), physiology word (the number of physiology word in the text, such as awkward), etc.

We extracted the 102 text-features related to a person's psychology with TextMind (http://ccpl.psych.ac.cn/textmind/), which is a Chinese language psychological analysis system developed by the Computational Cyber Psychology Lab, Institute of Psychology, Chinese Academy of Sciences. The dictionary, text and punctuation used in TextMind are optimized to Simpilified Chinese, and the categories are compatible to LIWC.

Big-five personality traits are as follows (5 features)

The Big-Five personality traits, also known as the Five Factor Model (FFM), is a model based on common language descriptors of personality and is the most frequently used among personality-related research. The five factors have been defined as neuroticism, agreeableness, extraversion, conscientiousness and openness, and are often represented by the initials (N, A, E, C, O). It is easy to calculate in the computer. Each dimension describes a person's personality from a different aspect.

Neuroticism (sensitive/nervous vs. secure/confident). Neuroticism refers to the degree of emotional stability and impulse control. This dimension is sometimes called emotional stability. High neuroticism is an anxious, hostile, fragile, sensitive personality.

Agreeableness (friendly/compassionate vs. challenging/detached.) Agreeableness is a measure of one's trusting and helpful nature, and whether a person is generally well-tempered. High agreeableness is often seen as naïve, submissive, altruistic and modest.

Extraversion (outgoing/energetic vs. solitary/reserved). Extraversion refers to energy, positive emotions, assertiveness, talkativeness, sociability personality, and tendency to seek stimulation in the company of others. High extraversion is often perceived as attention-seeking, and domineering.

Conscientiousness (efficient/organized vs. easy-going/careless). Conscientiousness has a tendency to be organized and dependable, show self-discipline, act dutifully, aim for achievement, and prefer planned rather than improvised behavior. High conscientiousness is often perceived as stubbornness and obsession.

Openness (inventive/curious vs. consistent/cautious). Openness reflects the degree of intellectual curiosity, creativity and a preference for novelty and variety a person has. Openness is often seen as appreciation for art, emotion, adventure, unusual ideas, curiosity and varieties of experience. High openness can be perceived as unpredictability or lack of focus.

In this paper, we treat the user's Big-five personality features as gender prediction variables based on our previous research [10].

Gender

In this case, the labels were self-evident: male and female. 444 male and 662 female labeled users were collected.

We got the Pearson correlation coefficient between feature and gender using SPSS. Table 1 gives us the Pearson correlation coefficient, where * shows the correlation on 0.05 and ** on 0.01.

Table 1. The Pearson correlation coefficient between feature and gender

Feature	gender	feature	gender	feature	gender
First person	-.077*	Love words	-.082*	Words more than six letters	.104**
Second person plural	-.072*	Colon	.116**	Micro-Blog emotions	-.101**
Sad words	-.067*	semicolon	.122**	Auxiliary words	-.115**
Physiological words	-.082*	Exclamation mark	-.072*	URL	.117**
Body words	-.083*	Dash	.083*	Nickname	-.086**
Sex words	-.129**	Quotes	.091**	Friends count	-.106**
Feeding words	-.072*	Abbreviated symbol	-.067*	Favourites count	.071*
Working words	.136**	Brackets	-.072*	description	.209**
Career	.100**	Arabic numerals	.307**	conscientiousness	.180**
Leisure	.075*	Money	.088**	Neuroticism	-.144**

** the correlation is marked on 0.01, * the correlation is marked on 0.05.

2.2. Machine Learning Framework

Recently, most of the identification prediction based on social network had used the traditional classification or regression algorithm in machine learning by different data features [1,2,4,5,6,9,13,16,18]. In this paper, we consider multiple local features and use multi-classifiers decision fusion method to predict Sina Micro-blog users' genders. The multi-classifiers fusion makes full use of global features from different feature points and improves performance indicators [3,11,15].

On the basis of three classifiers trained with users' micro-blog features, psychology features and personality traits, multiplication rule is employed to combine the three classifiers to make the prediction with classification knowledge from all the data samples. It is hard to resolve just by relying on a single classifier to improve the performance of a single classifier. Therefore, multi-classifiers fusion methods have been widely used. By using complementary information from different classifiers, we can get a better result. E. H. Ibrahim et al. showed that the classification accuracy by using multi-classifiers fusion improved in three datasets out of four [3]. H. Wei et al. proposed a face verification model based on confusing local Bayesian classifier and eliminated the limitation of using global face feature for face verification [15]. M. Takruri et al. proposed an automated non-invasive multi-classifier system for skin cancer detection and showed that this method got a better result than standalone Skin Lesion classification [11].

2.2.1. Multi-classifiers Fusion

In this paper, we infer Sina Micro-blog users' genders on different datasets by using multi-classifiers fusion, as is shown in Figure 1. The logistic regression model is used as the standalone classification. Meanwhile, in order to thoroughly characterize the contributions of these dataset to attribute inference accuracy, we tested (1) using different subsets of a user's attribute features, (2) predicting them by using simply summed on the different dataset and (3) using ensemble learning algorithm to predict the user's gender.

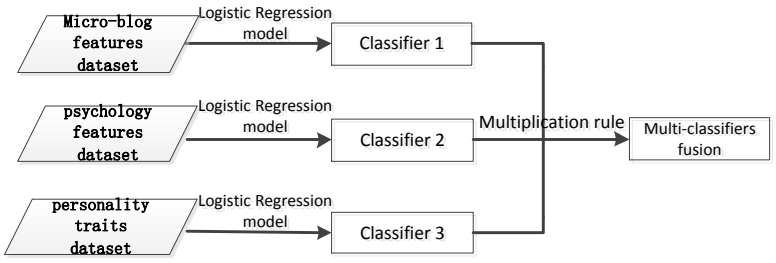

Figure 1. The machine learning framework

For z is the classified model, $\theta = \{\theta_1, \cdots, \theta_c\}$ is the labels, and K classifies are represented as $f_i (i = 1, \cdots, K)$. The data sample $x^{(k)}$, $k = 1, 2, \cdots, K$, is feature vector of the *k*th classifier. The output of K classifies is calculated by $p(x^{(k)} \in \theta_i \mid x^{(k)})$, where p(.) is the probability that $x^{(k)}$ belongs to θ_i.

Whether model z is assigned exactly into which label is determined at its posterior probability, as shown in Equation 1. That is:

$$p(\theta_j \mid x^{(1)}, \cdots, x^{(K)}) = \frac{p(x^{(1)}, \cdots, x^{(K)} \mid \theta_j) p(\theta_j)}{p(x^{(1)}, \cdots, x^{(K)})} \tag{1}$$

Assign $z \to \theta_i$ if $p(\theta_i \mid x^{(1)}, \cdots, x^{(K)}) = \max_{j=1}^{c} p(\theta_j \mid x^{(1)}, \cdots, x^{(K)})$

Where $p(x^{(1)}, \cdots, x^{(K)})$ is the joint probability density of K feature vectors.

In this paper, we use logistical regression to train three classifiers and we choose the multiplication rule to compromise the three classifiers, as shown in Equation 2.

Assign $z \to \theta_i$ if

$$\theta_i = \arg\max p(\theta_i) \prod_{k=1}^{K} p(\theta_i \mid x_k) \tag{2}$$

2.2.2. Logistic regression

Logistic regression (LR) is an interpretable model and has good generalization ability. It has been widely used in the field of deep learning. In this paper, LR algorithm is applied to make the base classifier.

For every task belonging to the space X×Y, its training dataset is denoted as

$$\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$$

where x_i is a dimension feature vector, and y_i is the true value for this point.

For binary classification, Figure 1 below shows LR model, as shown in Equation 3.

$$\ln\left(\frac{p(y=1 \mid x)}{p(y=0 \mid x)}\right) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n = w^T x \tag{3}$$

Where w is the fitting parameter.

$$\sum_{j=0}^{1} p(y=j \mid x) = 1$$

So

$$p(y=1) = \frac{1}{1 + e^{-w^T x}} \tag{4}$$

Equation 4 is a sigmoid function, which results in (0, 1). The value of w is the algorithm's learning goal that is commonly obtained by maximum likelihood estimation. The Log likelihood function is Equation 5.

$$L(w) = \sum_{j=0}^{1} \sum_{i=1}^{N_j} \ln p(x_i^{(j)} \mid y=j; w) \tag{5}$$

Gradient ascent algorithm is used for optimizing the LR model. So, we got the iterative Equation 6.

$$w := w + \alpha \nabla_w f(w) \tag{6}$$

Where α is called as step, $\nabla_w f(w)$ is the gradient of function $f(w)$.

3. Experiments and Comparision

3.1. Data Acquisition

Our experiment will use the dataset of Sina Micro-Blog Users including three kinds of features, such as 12 Micro-blog features, 102 psychology features (extracted by the TextMind system) and 5 Big-Five personality traits [14].

Finally, we got a total of 1106 users who are active users, in which there are 444 males and 662 females, and 517 18+ users and 589 25+ users from 19 provinces and cities around the nation. Therefore, the practical dataset is infective.

3.2. Result

Prediction of users' genders with their public information on Micro-Blog is the goal of our experiment. We chose LR algorithm as the classification prediction model, and we use 10-fold cross validation to test the performance of the prediction mode.

In our experiments, single classification models and Naïve Bayes prediction model on all dataset methods are selected to compare with the multi-classifiers fusion method mentioned in this paper. Our experimental results are shown in Figure 2 from precision rate, recall rate and the F measure. Experimental results demonstrate that the proposed approach yields a nice performance to gender classification, and the multi- classifiers fusion method outperforms the individual classifier trained with only user single messages and the simple classifier on all messages.

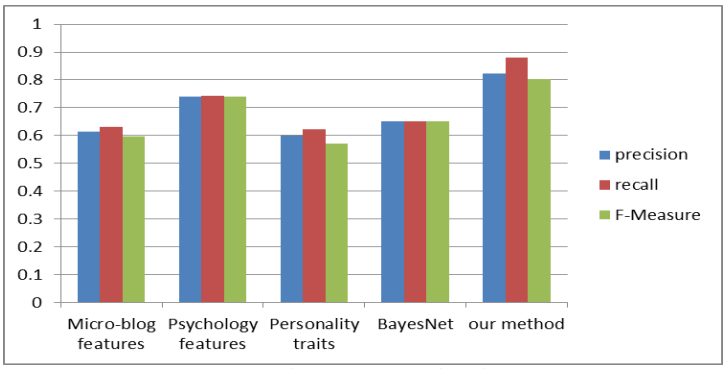

Figure 2. The performance result of the five methods

Figure 2 shows that the results of the three former methods are the same, and the prediction performance by single personality traits is the worst. That is to say, there is not full influence to classifier for a user's personality traits. Then, the prediction performance based on Bayes Net by training all of the three data samples is very low. It cannot improve the prediction performance by different data features' superposition. A single classifier cannot make the best of the global feature, so the result is relatively poor. But, the multi-classifiers fusion method can make full use of the complementary information from different classifiers to improve the prediction performance greatly. In order to ensure the stability and the performance of the prediction, it is necessary to make the multi-classifiers fusion so that the neighbourhood feature of all feature points is used to the fullest.

But, there are also some false classifications: 1) Some features in the test samples did not appear in the training samples, so the classifier cannot study these features. 2) The length of a user's text is so short that many features are sparse. Thus, it is useless to make the prediction model.

4. Conclusions

In this paper, we analysed Sina Micro-blog users' gender prediction and created the prediction model based on the multi-classifiers fusion method. An experiment was conducted to demonstrate the good performance of our models.

With the rapid development of Internet technology, the data structure is more complex and the calculation has more to compute. It is relatively simple using static state data, but the accuracy is very low. So, we use the multiplication rule to fuse the multi classifiers and improve the prediction performance. Figure 2 shows that our model got a better result and better generalization performance.

There is a great application value to predict Micro-blog users' genders based off of their public information. Our experiments use multiplication classifier, and each classifier has a great impact on the overall prediction results. But, it will affect the overall prediction performance due to a classifier's error. In the future, we will do further work on multi-classifiers model and feature extraction. We will also look for better models and features on a larger scale to improve the accuracy of prediction models.

Acknowledgements

The author would like to thank the anonymous reviewers for their valuable comments. This work is supported by the National Natural Science Foundation of China (Grant Number: 61602491). We would also like to thank the Institute of Psychology, Chinese Academy of Sciences for their generous support of our research.

References

1. B. Bin, and S. Milad, "Inferring Demographic of Search Users: Social Data Meets Search Queries," in *International World Wide Web Conference Committee*, pp. 13-17, Rio de Janeiro, Brazil, May 2013
2. J. He, W. Chu, and Z. Liu, "Inferring Privacy Information from Social Networks," in *IEEE International Conference on Intelligence & Security Informatics*, pp. 154-165, vol.3975 of Lecture Notes in Computer Science, Berlin, 2006
3. E. H. Ibrahim, A. I. Hashad, E. Shawky, and A. Maher, "Robust Breast Cancer Diagnosis on Four Different Datasets Using Multi-classifiers Fusion," *International Journal of Engineering Research & Technology(IJERT)*, vol. 4, no. 3, pp. 114-118, March 2015
4. M. Kosinski, D. Stillwell, and T. Graepel, "Private traits and Attributes are Predictable from Digital Records of Human Behavior," *Proceedings of the National Academy of Sciences*, vol. 110, no. 15, pp. 5802-5805, February 2013
5. J. Lindamood, R. Heatherly, M. Kantarcioglu, and B. Thuraisingham. "Inferring Private Information Using Social Network Data," in *Proceeding of International Conference on World Wide Web*, pp. 1145-1146, Madrid, Spain, 2009
6. A. Mislove, B. Viswanath, and K. P. Gummadi, "You are Know: Inferring User Profiles in Online Social Networks," in *Proceedings of the 3rd International Conference on Web Search and Web Data Mining*, *WSDM*, pp. 251-260, New York, USA, February 2010
7. J. Otterbacher, "Inferring Gender of Movie Reviewers: Exploiting Writing Style, Content and Metadata," in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM'10*, pp. 369-378, Toronto, On, 2010
8. M. Pennacchiotti, and A. Popescu, "Democrats, Republicans and Starbucks Afficionados: User Classification in Twitter," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *KDD'11,* pp. 430-438, San Diego, CA, 2011
9. C. Shen, Z. Cai, X. Guan, Y. Du, and R.A. Maxion, "User Authentication through Mouse Dynamics," *IEEE Transaction on Information forensics and Security*, vol.8, no. 1, pp. 16-30, January 2013
10. H. Shu, J. Zheng, D. Xu, and N. Zhao, "Predicting Big-Five Personality for Micro-blog Based on Robust Multi-task Learning," in *The 3rd International Conference of Pioneering Computer Scientists,Engineers and Educators*, pp. 486-499, Changsha, China, September 2018
11. M. Takruri, M. W. Rashad, and H. Attia, "Multi-classifier Decision Fusion for Enhancing Melanoma Recognition Accuracy," in *The 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA)*, pp. Ras Al Khaimah, United Arab Emirates, November 2016
12. Q. Tang, H. Lin, "Research on Gender Recognition for Character in Text," *Journal of Chinese Information Processing*, vol 24, no. 2, pp. 46-51, 2010
13. S. Volkova, Y. Bachrach, M. Armstrong, and V. Sharma, "Inferring Latent User Properties from Texts Published in Social Media," in *The 29th Aaai Conference on Artificial Intelligence* (www.aaai.org), pp. 4296-4297, Austin Texas, USA, January 2015
14. I. Weber, V. Garimella, and E. Borra, "Political Search Trends," in *Processdings of the 35th International ACM SIGIR Conference on Research and Development in Information retrieval*, SIGIR'12, pp. 1012-1012, Portland, OR, 2012
15. H. Wei, S. He, K. Lu, "Face Verification by Confusing Local Bayesian Classifier," *Opto-Electronic Engineering*, vol. 43, no. 3, pp. 80-87, March 2016
16. W. Xu, and X. Zhou, "Inferring Privacy Information via Social Relations," in *Proceeding of IEEE 24th International conference on Data Engineering workshop*, pp. 525-530, Cancun, Mexico, April 2008
17. J. Ying, Y. Chang, C. Huang, and V. Tseng, "Demographic Prediction Based on Users Mobile Behaviors," in *Mobile Data Challenge 2012 Workshop*, Newcastle, UK, June 2012
18. E. Zheleva, and L. Getoor, "To Join or not to Join: the Illusion of Privacy in Social Networks with Mixed Public and Private User Profiles," in *Proceedings of the 19th International Conference on World Wide Web*, pp. 531-540, Madrid, Spain, April 2009

Jinghua Zheng graduated from Electronic Engineering Institute, for the degree of Master and Ph. D. Now she is a lecture of National University of Defense Technology, Hefei, China. Her current research interests include information security, machine learning, and intelligent information processing.

Shize Guo professor, PhD supervision, receives a special government subsidy of the State Council. He was selected for the national key talents project. His current research interests include information security, cyberspace Security.

Liang Gao graduated from National University of Defense Technology, for the degree of Master and Ph. D, Changsha, China. Now, he is an engineer of the Institute of North Electronic Equipment, Beijing. His research fields concern information security and privacy, social network analysis and mining.

Di Xue received her M. S. and B. S. from PLA University of Science and Technology in 2015 and 2012. Now she is a Ph.D. student of Army engineering university. Her research fields concern information security and privacy, social network analysis and mining.

Nan Zhao is an associate research fellow of the Institute of Psychology, CAS. His research fields concern feature analysis based on social network dataset.

Huimin Ma received her M. S. from North China Electric Power University. She is a lecture of National University of Defense Technology, Hefei, China. Her current research fields concern information security.