International Journal of Database Theory and Application
Vol.9, No.6 (2016), pp.33-44
http://dx.doi.org/10.14257/ijdta.2016.9.6.04

Discovering Gangs of Criminals Using Data Fusion With Social Networks

Yang Jie[1] and Wang Huadong[2]

[1,2]*Modern Educational Technology Center, Zhengzhou University of Light Industry, Zhengzhou, Henan, China 450002*
yangjie1132@163.com

Abstract

Data mining technologies have been effectively applied to public security organizations and could help with the investigations. Among which, discovering and fighting against the gangs of criminals helps to the construction of peaceful and united environment. In this paper, we focus on the problem of discovering the gangs of criminals by integrating multiple data sources including the residential profiles from the public security bureau, the transfer data of banking accounts, the communication data from telecommunication operators, and the social interaction data from social networks. After employing a label propagation based method to discover communities in the criminal network, members within each group are ranked by their significance. Experiments exhibit the performance of proposed method.

Keywords*: Community detection, Social network analysis, Criminal gang, Data fusion*

1. Introduction

Data mining aims to discover hidden, unknown, interesting and useful knowledge and rules from massive data collection. Nowadays, data mining technologies have been effectively applied to public security organizations and could help with the investigations.

Unfortunately, in this materialistic society, with the increasing of crime rate, some criminals have already formed a gangsterdom force that conducts crimes in the manner of criminal groups. Accordingly, the public security and civil rights are continuously damaged. Therefore, discovering and fighting against the gangs of criminals helps to the construction of peaceful and united environment. For example, if criminal gangs are detected, further investigation focused on that specific gang might prevent future operations of the gang and even help catch the criminals. In this paper, we focus on the problem of discovering the gangs of criminals.

Specifically, we employ Social Network Analysis (SNA) fro criminal gangs discovery. In fact, SNA method has been successfully applied in modern sociology, psychology, economics and so on [1]. Criminals have structural characteristics just like other social groups if the motivations of group activities are ignored. The networks in criminal investigation field refer to organizations or gangs of criminals, where responsibilities are coordinated and distributed among criminals.

There are two major issues in detecting gangs of criminals. First, given known suspects, our goal is to discover any other members related to known criminals and the relationship between them. The second issue is to discover potential criminal network without any prior knowledge. Both problems can be solved by the fusion of multiple data sources. With the development of information technology, there are many data resources accumulated in a varied ways. For example, the residential profiles from the public security bureau, the transfer data of banking accounts, the communication data from telecommunication operators, and the social interaction data from social networks. In this

ISSN: 2005-4270 IJDTA

work, by integrating above four kinds of data sources with massive datasets, we explore the criminal gangs discovery problem.

Indeed, social media including social networking sites has been increasingly significant for aiding criminal activities investigation. According to statistics from the American Academy of matrimonial lawyers, 81% members believe that the increase of the number of marriage family cases is greatly related to social networks; and 66% divorce evidences are referenced from Facebook. Information on social networking sites is typically included as digital evidences for crime cases and events [2].

The remainder of this paper is organized as follows. Section 2 presents some related works. Proposed method of discovering potential gangs of criminals are discussed in Section 3. Then Section 4 gives the experiments and results, and Section 5 concludes the paper.

2. Related Work

The efforts of criminal data mining have been massively made ever since the 9-11 attack, especially the criminal gangs discovery. For example, Social Network Analysis (SNA) is introduced for crime investigation [3] and crime network analysis [4]. Klerks *et al*. [5] investigated the characteristics of criminal characteristics in terms of network size, density, correlation and centrality. Valdis *et al*. [6] examined the network surrounding the tragic events of September 11th, 2001 by collecting and analyzing the social relationship between members of the attack. Xu *et al*. [4] designed a CrimeNet Explorer system for knowledge discovery over criminal networks. Gao *et al*. [7] performed social network analysis on criminals with communication traces given the assumption of known a suspect within a gang. Qiao *et al*. [8] discovered key members of crime networks based on personality trait simulation over emails. Ma *et al*. [9] reviewed the applications and challenges of social network analysis in crime data mining. Li *et al*. [10] detected groups of financial criminals using PageRank [11] and Genetic Algorithm (GA).

However, most existing efforts listed above focus on theoretical or small-sized criminal network based on public dataset. Indeed, social networking sites have been proven to be a significant source of data collection for modern data mining applications [12], such as digital crime analysis [13]. Taylor *et al*. [14] examined the computer forensic process of obtaining digital evidence from social media. Holm *et al*. [15] explored the vulnerability of social network users to identity theft facilitated by the information they share on social networking sites. Broadhurst *et al*. [16] illustrated individual and group behavior of cyber crime. Moule *et al*. [17] examined patterns of Internet use in street crime. Pyrooz *et al*. [18] studied general online routine activities, online criminal and deviant behaviors, and gang-related online behaviors and processes.

In this paper, we try to integrate different data sources, including the residential profiles from the public security bureau, the transfer data of banking accounts, the communication data from telecommunication operators, and the social interaction data from social networks, as the evidence of group crime discovery. Specifically, based on the above data collection, a SNA based method is employed for criminal gang detection under two circumstances: (1) known some suspect which belongs to a criminal gang, and (2) without any information of a suspect being a criminal or not but only to discover the potential network between them.

3. Proposed Method

In crime investigation problem, we always should have some ground truth in hand, which is profiles of historical criminals. Those criminal records help to label people in the dataset. Therefore, in this paper, we assume history criminal records are known beforehand to identify future potential threats.

Recall that our objective is to discover groups of criminals with or without the knowledge of specific suspect. It can be split into two parts: first, discover all potential criminal groups; and second, given a specific suspect, determine his/her partners within a gang. Obviously, the latter task can be easily solved if we have all potential gangs. Therefore, we focus on the problem of discovering all potential criminal groups in this section.

The basic idea is to first construct a structural graph of persons, and then apply SNA based method to detect associated clusters. The overall process can be illustrated in Figure 1, and the details are explained as follows.

Input: the set of persons associated with residential profiles, banking transfer data, phone data and Twitter data.

Output: groups of potential criminals.

Step 1: Data preparation, including data cleaning and mapping between residential profiles, banking transfer data, phone data and Twitter data. This is data preprocessing step. After obtaining data collection from different sources, original dataset should be cleaned and integrated for further analysis, among which data mapping is significant for data fusion. In our case, data mapping refers to mapping identification number, banking account number, phone number and Twitter account into one single person entity. In this way, we combine four kinds of data sources together to represent an entity.

Step 2: Construct a social network graph, where each node represents a person entity, and the edge represents connections in banking transfer, phone communication and Twitter interactions. Since we employ a graph-based method for analysis, constructing a graph with connections between entities is the second step. Note that the edges include connections generated by different data sources.

Step 3: Label some nodes within above graph based on history crime records as criminals. Given the history crime records from local public security bureau, we can label the entities in the social graph as known criminals, and others are unknown. Those labeled nodes provide supervised information for further analysis.

Step 4: Apply SNA based algorithms to detect groups of potential criminals, along with the most significant potential criminals within each group. In this step, real data analysis is performed, including the gang detection and the key members within each gang group. Note that the groups of criminals are discovered based on the fact of known criminals and the connections between them, which might involve temporary innocent suspects, and therefore could detect potential criminal organizations and prevent future criminal operations.

Step 5: Decision making and possible measures for potential criminal gangs. After identifying potential criminal gangs, some decisions and measures are made by concerned parties. Note that models or algorithms give the possible criminal groups without complete confirmation about the guilty of the groups or individuals, and the decisions are up to relevant decision makers.

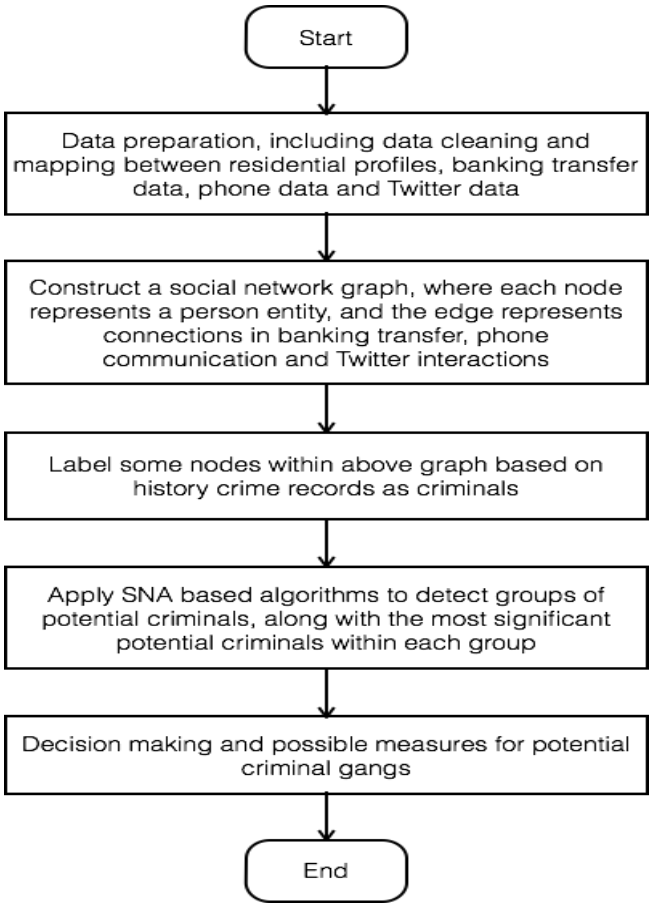

Figure 1. Workflow of Criminal Gangs Detection with Data Fusion

3.1. Constructing Social Network Graph

After data preparation and mapping in the previous steps, we have the dataset consisting records with the identification fields as follows:

$$(ID, PID, PN, BID, TID), \quad (1)$$

Where ID is the entity ID denoting each single record, PID is the identification number of each person, PN is the phone number of that person, BID is the banking account number, and TID is the Twitter account number. Note that even though in this study we use Twitter, a popular social networking site as an example, the social media data source could be any social networking sites indeed. ID is the primary key, and (PID, PN, BID, TID) are the composite keys.

Basically, the nodes within the graph are represented by ID, and the edges are constructed by the relationships among (PID, PN, BID, TID). Note that all ID fields are anonymous in this study. For example, given two entities with ID 001 and 002, there exist edges between them when any of the following conditions are satisfied: (1) 001's PN has called or texted 002's PN, or on the contrary; (2) 001's BID has made a transfer request to 002's BID or on the contrary; or (3) 001's TID has followed or sent tweets to 002's TID or on the contrary. The edge is directed based on the initiator and recipient of the activities, and the weights is simply calculated as summation of occurrences combined all activities.

In this way, we get a graph $G=(V,E)$, where node $v \in V$ is represented by ID, and edge $e_{ij} \in E$ between nodes v_i, v_j is the connections between entities drawn from social interactions from different sources. Notate adjacent matrix A as:

$$A=\begin{pmatrix} - & a_{12} & \cdots & a_{1n} \\ a_{21} & - & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & - \end{pmatrix}, \tag{2}$$

Where a_{ij} is the weight of edge between v_i, v_j, and the diagonal elements are meaningless. If there exists no connections between them, a_{ij} is set to 0.

Then, we label some nodes as known criminals based on the history crime records. If an entity committed any crimes in the past, it would be labeled as criminal in this graph. The assumption here is that crime activities are likely to occur in repetition. Otherwise, the node is unlabeled, meaning uncertain but still possible to be a potential criminal. The intuition here is that a potential criminal can be learned from known criminals; even though there are no crime records for now, they could be the underground criminal organization.

3.2. Detecting Potential Gangs

The task is formulated as discovering groups or clusters of nodes provided some knowledge about labeled criminals. That is, given some label information, the objective to find clusters within which nodes are close enough and across which nodes are sparse enough; in other words, to discover clusters with the objective to maximize the modularity.

Suppose $\{(x_1,y_c),(x_2,y_c),...,(x_l,y_c)\}$ is labeled data, where y_c is the label for criminal, and l is the number of known labeled data. Let $\{(x_{l+1},y_{l+1}),...,(x_{l+u},y_{l+u})\}$, where $\{y_{l+1},...,y_{l+u}\}$ is the set of unknown labels, and u is the number of unlabeled data.

As common criteria in community discovery, modularity is used to measure the performance of discovered clusters. However, modularity relies greatly on the number of edges in the network, which is relatively sparse in criminal network. Therefore, in this study, we employ modularity density as the optimization objective.

Given network $G=(V,E)$, and its partition $\{G_1(V_1,E_1),G_2(V_2,E_2),...,G_m(V_m,E_m)\}$, and the average modularity of $G_i(V_i,E_i)$ is defined as

$$d(G_i)=d_{in}(G_i)-d_{out}(G_i), \tag{3}$$

where $d_{in}(G_i), d_{out}(G_i)$ are the average in-degree and out-degree for $G_i(V_i,E_i)$.

Suppose V_1, V_2 are two disjoint subsets of nodes, and

$$L(V_1,V_2)=\sum_{i\in V_1, j\in V_2} a_{ij}, \tag{4}$$

where $a_{ij} \in A$.

Then, Equation (3) can be rewritten as:

$$d(G_i)=\frac{L(V_i,V_i)-L(V_i,\overline{V_i})}{|V_i|}. \tag{5}$$

Therefore, the modularity density of a graph partition is defined as the average over modularity of each subgraph, that is:

$$D = \sum_{i=1}^{m} d(G_i) = \sum_{i=1}^{m} \frac{L(V_i, V_i) - L(V_i, \overline{V}_i)}{|V_i|}. \quad (6)$$

Where m is the number of partitions. The larger D is, the better discovered clusters results are. Therefore, maximizing D is the optimization objective in our task.

Our scenario suits the semi-supervised learning process since we have some labeled criminal data. We introduce the idea of Label Propagation Algorithm (LPA) [19] for gang discovery. The basic idea is to learn unknown information using known labels. According to the basic LPA theory, each node propagates its label to its neighbors based on similarity. At each propagation step, each node updates its own label based on the information from neighbors. The more similar it is to the neighbors, the more weights of the labels of neighbors. Finally, similar nodes tend to get the same labels and the labels tend to propagate easily among them. Note that initialized known labels remain unchanged, and are propagate to other unlabeled nodes. When the algorithm stops, the probability distribution of similar nodes tends to be similar as well, and therefore those nodes have the same labels.

Suppose the weights between nodes are:

$$w_{ij} = \exp\left(-\frac{d_{ij}^2}{\sigma^2}\right), \quad (7)$$

Where d_{ij} is the distance between nodes i, j.

In order to measure the probability of label propagation from one node to the other, we define a probability transfer matrix T, where

$$T_{ij} = P(j \to i) = \frac{w_{ij}}{\sum_{k=1}^{l+u} w_{kj}}, \quad (8)$$

Where T_{ij} is the probability of propagation from node j to i.

Define a label matrix $Y_{(l+u)\times c}$, which means there are c kinds of labels. The initialization of Y is based on existing history crime records. The LPA algorithm is described as Figure 2.

LPA algorithm

Input: u unlabeled data and l labeled data
output: labels of unlabeled data
1. Initialization based on the weights in Equation (7);
2. Calculate the propagation probability based on Equation (8);
3. Define the label matrix;
4. Repeat
5. For each node x:
6. Combine the weights from neighbors, and update its propagation probability using
7. $F_{ij} = \sum_{k=1}^{l+u} T_{ij}, 1 \le i \le l+u, 1 \le j \le c.$
8. Reset the propagation probability of labeled data;
9. Until convergence.

Figure 2. Description of LPA Algorithm

When the community structure is obvious, for any node x actually belonging to community L, it is more likely that x is assigned as the label of node within L, say y. Then, it is more likely that node z which is connected to both x and y belongs to L as well, and thus the label of z is updated as that of y. Eventually, nodes within the same community would get the same labels. To avoid large communities, we modify the typical LPA process as follows.

Step 1: label initialization. Label nodes as criminals or unlabeled.

Step 2: update labels of nodes randomly. Suppose we need to update the label for node v_i, first omit the existing label of v_i, and then calculate the increment of modularity density D if it is updated as l_c:

$$\mathrm{D}_{l_c} = \frac{(L(V_{l_c},V_{l_c})+2N_{l_c})-(L(V_{l_c},\overline{V_{l_c}})+k_i-2N_{l_c})}{|V_{l_c}|+1} - \frac{L(V_{l_c},V_{l_c})-L(V_{l_c},\overline{V_{l_c}})}{|V_{l_c}|}, \quad (9)$$

where k_i is the degree of node v_i. Then, the label of v_i is updated as:

$$label(v_i) = \arg\max_{l_c} \mathrm{D}_{l_c}. \quad (10)$$

Step 3: termination condition check. If the termination condition is not met, return to Step 2; otherwise, the algorithm stops.

3.3. Identifying Key Members within Gangs

Now we have all groups of potential criminals, *i.e.*, gangs, and next we aim to identify the key players within each gang. Typically, we employ a SNA based method by ranking the centralities of entity nodes within each group.

However, criminal networks are typically sparse and the connections are relatively limited between criminals. To capture that point, we define the degree of dependence to measure the significance of node, which indicates the consequences if the node is removed from the graph.

Take the example in Figure 3: if any node of a,b,d,g is removed from the graph, the significance value of node c changes slightly. For example, if we measure the node significance by PageRank, the values of node c would change from 27.39 to 25.28 if one of the a,b,d,g is removed. However, the value increases from 27.39 to 58.33 and 61.78 respectively if node e or f is removed. Interestingly, node e or f is not connected to c at all. This implies that the significance of nodes goes beyond the direct connection.

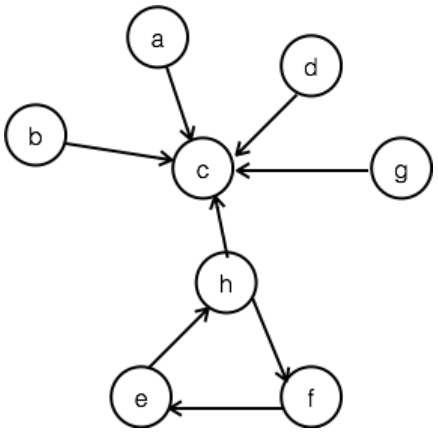

Figure 3. Illustration of Node Significance

Define function $I : V \to R$ to measure node significance, and $I_G(v)$ is the value of node significance. In this work, PageRank algorithm is employed as the implementation of I. Define the degree of dependence of u on v as:

$$r(v \to u) = \frac{|I_G(u) - I_{G-v}(u)|}{I_G(u)}, \tag{11}$$

where $G - v$ denotes the graph with node v removed, $I_G(v)$ is the significance of v in G. If removing v leads to big change of node significance, it means v itself is significant. The significance of node v is calculated as the summation over all nodes:

$$r(v) = \sum_{v_j \in V} r(v \to v_j). \tag{12}$$

Sort the values of $r(v)$ within each discovered groups, and return the top k as the key members.

4. Experiment

4.1. Data Preparation

As mentioned before, our data collection consists of four parts: residential profiles from the public security bureau, the transfer data of banking accounts, the communication data from telecommunication operators, and the social interaction data from social networks. Residential profiles include identification number (which is omitted on purpose for privacy), name, sex, age, address and yearly income. Banking transfer records include account number, recipient account number, timestamp of transfer and the amount of transfer. Mobile communication records include caller number, subscriber name, subscriber number, timestamp of calling or texting, duration of call and content of message. Social networking data is collected via Twitter using Twitter Open API, including user ID, screen name, follower list, friend list, all historical tweets, all contact users and interaction tweets.

Data preprocess is performed before further analysis. First, since the data sources and collection methods are different, there might be missing, redundant and even wrong data records. For example, the format of identification number of residents is fixed, which could be used to filter wrong records at the first step. Besides, the format of dates and timestamps should be unified for all data sources. Then, mapping between identification number, mobile number, bank account number and Twitter user ID to make sure each qualified data entry is associated with four sources with a unique ID of a person as the primary key. Originally, we have over 50 millions phone data, 10 millions banking transfer data, and 800 millions tweets data. After the preprocessing and mapping, we get 1,856 entries ready for analysis.

The algorithm is implemented using Java, and the environment setting is shown in Table 1.

Table 1. Experimental Environment

CPU	Intel Core i7 2.2GHz
Memory	8 GB 1600 MHz DDR3
Operating system	CentOS Core 2.6.18
Running environment	Java Runtime Environment 1.8

4.2. Evaluation Metrics

Since we have the ground truth labels for criminals before modeling, the first metric we use is Normalized Mutual Information (NMI), to measure the difference between ground truth and the modeling results.

Given two communities i, j, the numbers of nodes in them are n_i, n_j respectively, and the number of nodes that belong to both i and j is n_{ij}. NMI is defined as:

$$NMI = \frac{-2\sum_{i=1}^{p}\sum_{j=1}^{g} \frac{n_{ij}}{n} \log \frac{n_{ij}}{n_i n_j}}{\sum_{i=1}^{p} n_i \log \frac{n_i}{n} + \sum_{j=1}^{g} n_j \log \frac{n_j}{n}}, \quad (13)$$

Where n is the total number of nodes. The range of NMI is [0,1]. If NMI=0, it means the results is completely inconsistent with the real data. If NMI=1, it means the results fit the ground truth labels completely. Obviously, larger NMI is better.

Besides, modularity is another common metric for community discovery when the network structure is unknown.

$$Q = \frac{1}{2m}\sum_{i,j} (A - \frac{k_i k_j}{2m})\delta(i, j), \quad (14)$$

Where m is the number of edges in G, A is the adjacent matrix, k_i is the degree of node i, c_i is the community of i, and $\delta(c_i, c_j) = 1$ only if $c_i = c_j$. The larger Q is, the better performance.

4.3. Experimental Results

In order to evaluate our community discovery method for criminals, we compare the proposed method with k-means and LPA in Table 2. All results are obtained by the average of ten runs. We can see that our proposed method outperforms other algorithms in both NMI and modularity measurement, even though the execution time is slightly longer.

Table 2. Performance Results of Criminal Groups Detection

	NMI	Q	Execution time (s)
K-means	0.69	12.01	25.87
LPA	0.38	18.27	31.65
Proposed method	0.72	22.64	33.95

Besides, Figure 4 gives the node significance distribution. We can observe that the number of nodes with high significance value is relatively very small compared to the total population.

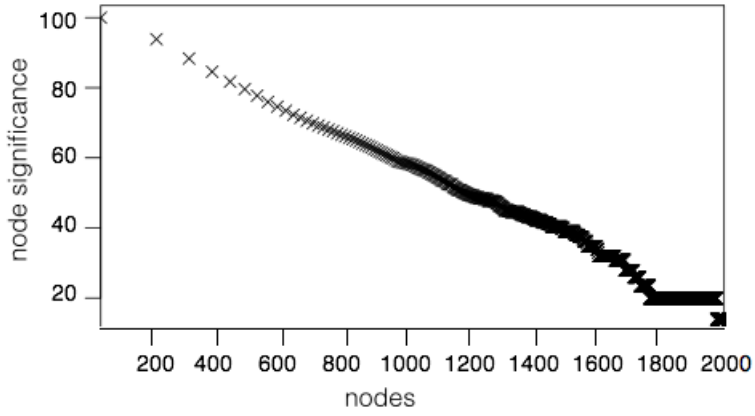

Figure 4. Illustration of Node Significance

5. Conclusion

We provide the initial attempt to conduct criminal data analysis using data fusion from multiple sources, and specifically in this work we focus on the community detection problem, that is, gangs of criminals. However, there are some simplifications here, which would be further investigated in future works. First, we simplify the social networking data as Twitter data, and only use timeline tweets for analysis. Indeed, event or keyword based search results from social networking sites might also be interesting for crime investigation. Second, we simplify the ground truth of criminals based on the history crime records. That is, as long as one has a dirty crime records, he/she is labeled as criminal in our learning model. In future, we would differentiate different kinds of crimes to make things more understandable.

References

[1] J. C. Peter, J. Scott and S. Wasserman, "eds. Models and methods in social network analysis", Cambridge university press, vol. 28, **(2005)**.
[2] G. B. John, "Digging for the Digital Dirt: Discovery and Use of Evidence from Social Media Sites", SMU Sci. & Tech. L. Rev., vol. 14, **(2010)**, pp. 465.
[3] L. C. Freeman, "Centrality in social networks conceptual clarification", Social Networks, vol. 1, no. 3, **(1978)**, pp. 215-239.
[4] J. J. Xu and H. Chen, "Crimenet Explorer: A Framework for Criminal Network Knowledge Discovery", ACM Transactions on Information Systems Tois Homepage, **(2005)**, pp. 201-226.
[5] P. Klerks and E. Smeets, "The Network Paradigm Applied to Criminal Organizations: Theoretical nitpicking or a relevant doctrine for investigators? Recent developments in the Netherlands", //Connections, **(2001)**, pp. 53-65.
[6] V. E. Krebs, "Mapping Networks of Terrorist Cells", Connections, vol. 24, no. 3, **(2002)**, pp. 43-52.
[7] G. Jianqiang and T. J. C. Yongfa, "A gang of social network communication traces based on analysis model", Computer applications and software, vol. 29, no. 3, **(2012)**, pp. 206-208.
[8] Q. C. Tang and J. Peng, "Crime is the core of the network computer", Journal of mining based on personality trait simulation email analysis system, vol. 31, no. 10, **(2008)**, pp. 1795-1803.
[9] M. Fang, "Crime network analysis, social network analysis in studies of organized crime in the application", Southwest University of political science and law, vol. 14, no. 2, **(2012)**, pp. 34-43.
[10] L. Guocheng and Q. X. Xiao, "Based on social network analysis of the financial criminal gangs detection", Financial and economic: the academic version of the, no. 12, **(2013)**, pp. 84-86.
[11] L. Page, S. Brin and R. Motwani, "The PageRank Citation Ranking: Bringing Order to the Web", Stanford Infolab, **(1999)**.
[12] W. Xindong, "Data mining with big data", Knowledge and Data Engineering, IEEE Transactions, vol. 26, no. 1, **(2014)**, pp. 97-107.
[13] W. T. Robert, E. J. Fritsch and J. Liederbach, "Digital crime and digital terrorism", Prentice Hall Press, **(2014)**.
[14] T. Mark, "Forensic investigation of social networking applications", Network Security, vol. 2014, no. 11, **(2014)**, pp. 9-16.
[15] H. Eric, "Social Networking and Identity Theft in the Digital Society", ICDS 2014, the Eighth

International Conference on Digital Society, **(2014)**.
[16] B. Roderic, "An Analysis of the Nature of Groups Engaged in Cyber Crime", An Analysis of the Nature of Groups engaged in Cyber Crime, International Journal of Cyber Criminology, vol. 8, no. 1, **(2014)**, pp. 1-20.
[17] K. M. Richard, D. C. Pyrooz and S. H. Decker, "From 'What the F#@% is a Facebook?'to 'Who Doesn't Use Facebook?': The role of criminal lifestyles in the adoption and use of the Internet", Social science research, vol. 42, no. 6, **(2013)**, pp. 1411-1421.
[18] C. P. David, S. H. Decker and R. K. Moule Jr., "Criminal and routine activities in online settings: Gangs, offenders, and the Internet", Justice Quarterly ahead-of-print, **(2013)**, pp. 1-29.
[19] J. Xie and B. K. Szymanski, "Community Detection Using A Neighborhood Strength Driven Label Propagation Algorithm", IEEE Nsw, **(2011)**, pp. 188 - 195.