Research on the Application of Cluster Analysis in Criminal Community Detection

Tian Junjing¹ *
¹The Chinese People’s Armed Police Forces Academy, Lang fang, Hebei, 065000, China
* a88241@163.com

Abstract. With the development of China's economy and technology, the mobility of people enhances rapidly, the trend of crime towards to community and organization. Criminal behaviour is a kind of social behaviour. Many malignant cases, especially those that are highly harmful to the lives and property of the citizens and have bad social influence, are not executed by a single person, but need teamwork to complete. At present, the city's population base is large, the police force is insufficient, and it is difficult to expedite the investigation of suspicious personnel by conventional methods in a short time. This paper establishes the social network model according to the local telephone call records, and then uses the cluster analysis algorithm looking for people with close links to the criminal suspect to identify suspected criminal accomplices. The cluster analysis algorithm proposed in this paper can help the public security organs filter suspicious people rapidly, narrow the scope of possible criminal gangs, and improve the efficiency of detected cases.

1. Introduction
Criminal social network is not separated but consists of a series of associations. Community detection technology in social network is an important method for mining criminal gangs, and the quality of community detection technology depends mainly on clustering algorithms. At present, there have been a large number of researches on clustering analysis in academic circles. However, these algorithms are mostly too theoretical and have not been specifically optimized in combination with actual application scenarios. In particular, there are only a handful of optimizations for police crimes [1]. Moreover, the public security department has limited manpower in criminal investigation, and it takes a lot of time to investigate criminal gangs related to criminal suspects. Therefore, it is very necessary to use the cluster analysis algorithm to detect the suspect's society, determine close suspects, reduce the inspection targets, and provide intelligent decision support for the staff in the criminal investigation department.

2. Clustering algorithms in social network
It is of great significance to grasp the network structure of criminal gangs for the investigation of crimes and the prevention of crimes in criminal investigation work. Therefore, a social network analysis (SNA) method can be used to build a communication network, and then a cluster analysis algorithm can be used to determine the degree of closeness between the community members, to identify associated gangs according to specific suspects, and provide the public security department with a list of suspicious criminal gangs to help them effectively prevent and fight crimes [2].
2.1. Community Division in Social Network

The social network can be represented as $SN = (V, E)$, $V$ is a set of nodes that are non-empty, $E$ is the edge set connecting the two nodes in the network. Social networks are divided into undirected networks and directed networks. If any two node pairs $(i, j)$ and $(j, i)$ correspond to the same edge, the network is called an undirected network, otherwise it is directed network [3]. In an undirected network, if each edge has a corresponding weight, the network can be called a weighted network; otherwise it is called an un-weighted network. In this paper, communication network that the nodes communicate with each other belongs to undirected network. The two connected nodes in the network have their call duration and number of calls. Therefore, the communication network is a weighted undirected network.

Community division is an important technology in the current social network analysis. It can extract some closely-associated communities from the social network. The result of this extraction is that the nodes in the community are closely connected and the nodes between different communities are loosely connected. Based on this characteristic, those who are related to known criminals in the community can be classified as key monitoring objects in the criminal investigation. Thereby, Community division can greatly reduce the burden of surveillance, and quickly discover potential criminal partners. The methods of community division in social network mainly include the division methods based on graph theory and the division methods based on hierarchical clustering algorithm[4][5]. The representative methods based on graph theory mainly include the spectral bisection method, Kernighan-Lin algorithm and so on. The algorithms based on hierarchical clustering include decomposed hierarchical clustering algorithms and agglomerative hierarchical clustering algorithms.

2.2. The Process of Cluster Analysis

The process of cluster analysis is shown in Figure 1. Generally, it includes the following steps:

(1) Data Preprocessing

In the data preprocessing stage, the validity of the data is mainly checked to identify whether the data is null or error data and so on. In some cases, it is also necessary to perform operations such as summation and averaging in the data preprocessing stage. In order to find the frequently contacted community in this paper, the duration of calls and the number of calls between two nodes that are two people communicating with each other in a social network should be summed up [6].

(2) Extraction of Clustering Attributes

The data in the data set may contain many attributes, such as name, gender, call duration, and so on. Some of these attributes are continuous attribute, and some are binary discrete attribute. Choosing the attributes of the data is a key issue in the process of cluster analysis. The selected attributes should be sufficiently representative of the data sample in the data set. Different attributes need to adopt different node similarity measurement methods. This paper selects two attributes that can reflect the density of the two nodes, which are the duration of the calls and the number of calls.
(3) Selection of Clustering Algorithm
Clustering algorithms include hierarchical clustering algorithms, partitioned clustering algorithms, grid-based clustering algorithms and many other algorithms [7]. According to the application and data set, this paper chooses the nearest neighbor hierarchical clustering algorithm.

(4) Clustering Process
In the clustering process, similarity is calculated for the nodes according to the distance between the two nodes in the social network. According to the input parameters and the selected clustering method, the nodes are clustered.

(5) Evaluation of Clustering Result
The clustering result is evaluated to determine the rationality in the evaluation process. If the clustering result is reasonable, the entire clustering process is ended. If the clustering result is not reasonable, the clustering algorithm and parameters should be adjusted, and the clustering process should be repeated until a reasonable clustering result is obtained.

3. Clustering Analysis Based on Call Records

3.1. Construction of Social Network
When the public security department conducts a case investigation, the relationship strength and relationship model between the parties can be revealed by the telephone records analysis. In the data analysis and extraction stage in this paper, call records of the users are analysed, key information is extracted, and the information such as the calling telephone number, the called telephone number, and the call duration are obtained, and a sum operation is performed on the call duration and the number of calls. Call duration and the number of calls reflect the strength of the relationship between the two contacts. Substituting the duration of the calls and the number of calls into the formula can obtain the weight of the edge between the two nodes. Each person is abstracted to a node. The connection between two people is abstracted to the edge in the network, thus a weighted undirected graph is formed.

3.2. Detection and Division of Criminal Associations
This paper constructs a weighted and undirected graph which shows the calling network of the users. Each edge in the network needs to be assigned a weight value. The weight value is inseparable from the call situation of the two users. When determining the weight value, not only the duration of two users' calls but also the number of calls made by two users must be taken into consideration. Therefore, the weight value of the edge is defined as \( \text{WE}(i,j) \). The value of \( \text{WE}(i,j) \) can be calculated by Equation (1).

\[
\text{WE}(i,j)=M \times t(i,j)+N \times c(i,j)
\]

(1)

\( t(i,j) \) is the weight value of the call duration between user \( i \) and user \( j \), \( c(i,j) \) is the weight value of the number of calls between user \( i \) and user \( j \). \( M \) is the influence factor of the call duration and is used to determine the influence of the call duration on the edge weight. \( N \) is the influence factor of the call number and used to determine the influence of the number of calls on the edge weight. \( M+N=1 \), \( M \in [0,1] \), \( N \in [0,1] \).

The value of \( t(i,j) \) can be calculated by Equation (2).

\[
t(i,j) = \frac{t_{ij}}{t_i + t_j}
\]

(2)

\( t_{ij} \) represents the call duration between user \( i \) and user \( j \), \( t_i \) represents the total call duration between user \( i \) and everyone, \( t_j \) represents the total call duration between user \( j \) and everyone. If the call duration between user \( i \) and user \( j \) is a large proportion, it means that the relationship between them is relatively close.

The value of \( c(i,j) \) can be calculated by Equation (3).

\[
c(i,j) = \frac{c_{ij}}{c_i + c_j}
\]

(3)
c_j represents the number of calls between user i and user j, c_i represents the total number of calls between user i and everyone, c_j represents the total number of calls between user j and everyone. If the number of calls between user i and user j is a large proportion, it means that the relationship between them is relatively close.

In the process of clustering analysis, the similarity between node i and node j is related to the distance between the two nodes. The closer the distance between them, the higher the similarity is. The distance D(i,j) between node i and node j is related to the edge weight. The calculation of the distance between them should consider the following three situations:

1) If the two nodes are directly connected, the distance D(i,j) between them is calculated according to Equation (4).

\[
D(i,j) = \frac{1}{1+WE(i,j)}
\]  

(4)

2) If the two nodes are not directly connected but connected through other nodes, the distance between the two nodes is the shortest path between them. In this case, the distance between the two nodes is calculated by Equation (5).

\[
D(i,j)=D(i,k)+D(k,p)+\cdots+D(o,j)
\]  

(5)

D(i,k), D(k,p), D(o,j) can be calculated according to Equation (4), node k, node p, and node o are the intermediate nodes on the shortest path between node i and node j.

3) If the two nodes are not connected by any path, the distance between them is \(D(i,j)=+\infty\).

3.3. Cluster Analysis Based on Node Distance

During the process of community detection, the suspect nodes are assigned by the user, which are used as the initial cluster centers, and then the distances between the other nodes in the social network and these cluster centers are calculated. These nodes are assigned to the community where the nearest cluster center is located. A threshold can be set to control the size of the community. If the threshold is not set, the size of the community is determined by the algorithm.

The process of cluster analysis is as follows:

Input parameters: an weighted and undirected social network G=<V, E>, where V is a set of nodes in the network and E is a set of edges in the network;

Output result: M communities;

Follow these steps for cluster analysis:

Step1: Initialize the network and use the M (\(M \geq 1\)) nodes set by the user as the initial clustering center(s).

Step2: Calculate the distances between two nodes in the social network according to Equation (4) and Equation (5), and store the results in an n-dimensional distance matrix.

Step3: Calculate the distances between each node and M cluster centers according to the n-dimensional distance matrix, then assign the nodes into the community represented by the nearest cluster center, and delete these nodes in the node set. If the distance between a node and two cluster centers is equal, a modularity function is used to determine the node belongs to which community.

Step4: If the node set is empty, the algorithm ends. If the node set is not empty go back to step3 to continue.

4. Presentation of the Output Result

In the criminal network, different roles have their own contact characteristics. Contacting members or gangs have higher frequency of contact before committing the crime. The links between closely-connected members are generally characterized by high frequency, long average talk time, and contact at dinner time or at night.

4.1. Call data set

First of all, the call data set should be pre-processed. The total call duration of two users and the total number of calls should be calculated. The edge WE(S, T) and distance D(S, T) also should be
calculated according to Equation (1)-Equation (5). When constructing the social network and detecting the communities, clustering analysis is executed based on the distance between two nodes. The call data set after pre-processing is shown in Table 1.

**Table 1. Call data set**

|   | S_138****2321 | S_132****8890 | ...... | S_158****3421 | S_139****4580 |
|---|---------------|---------------|-------|---------------|---------------|
| T | 183****7659   | 137****0909   | ...... | 139****4578   | 158****4664   |
| S_Duration | 1056432 | 423098 | ...... | 1276580 | 265780 |
| T_Duration | 876534 | 456522 | ...... | 956412 | 42383 |
| Tol_Duration | 324784 | 200514 | ...... | 78902 | 33454 |
| S_Count | 1235 | 432 | ...... | 2876 | 7890 |
| T_Count | 879 | 321 | ...... | 1983 | 341 |
| Tol_Count | 544 | 243 | ...... | 106 | 246 |
| M | 0.5 | 0.5 | ...... | 0.5 | 0.5 |
| N | 0.5 | 0.5 | ...... | 0.5 | 0.5 |
| WE(S, T) | 0.213 | 0.276 | ...... | 0.065 | 0.139 |
| D(S, T) | 0.824 | 0.784 | ...... | 0.939 | 0.878 |

4.2. Analysis of Output Result

After the Clustering analysis, the clustering result is shown in Figure 2.

The communities in the social network are detected and the divisions are shown in Figure 3. According to the division result, the social network should be divided into two clusters that are represented by two colors. Each cluster represents a community. The centrality of the nodes is different in each community. The shape of the nodes with a large centrality is relatively small, and the shape of the nodes with a small centrality is relatively large. A node with a large centrality indicates that it is of greater importance in social network and it belongs to core nodes in the community.

According to the result of community division in Figure 3, the closeness of these nodes is exhibited. If it is determined that a certain mobile phone number belongs to a criminal suspect, then according to the clustering result, the contact person related to him can be found. And, according to the weight values of the edges in the social network, the closeness of the contact and the suspect can be judged so as to judge possible criminal gangs.
5. Conclusion
Sometimes people are not concerned about the structure of the entire network in real life, but only need to know the structure of part of the network. That is, people only need to know the situation of the group you care about in the network. Based on the characteristics of communication data, this paper uses the communication records of the criminal suspects to establish social networks. Based on the shortest distance core mining method, the nearest neighbour hierarchical clustering method is used to detect communities and identify possible criminal gangs automatically. Clustering analysis can facilitate criminal investigation personnel of the Public Security Bureau to conduct targeted investigations of suspicious individuals and speed up the detection of cases. In addition, the core members of possible criminal gangs can also be automatically identified. Although the core member does not necessarily have to be a leader within a possible gang, his influence in the gang is very large, and he(or she) is likely to be a key figure in the gang.

References
[1] Wu Shaozhong 2016. Potential Terrorist Gang Discovery Techniques in Counterterrorism Intelligence Based on Cluster Analysis. Police Technology. 6 18-21
[2] PanFang 2009. Criminal Network Analysis Based on Fuzzy Hierarchical Clustering. Journal of Southwest China Normal University (Natural Science Edition). 34 210-214
[3] Li Liang 2008 The Recognition System of Crime Syndicate Based on Social Network Analysis (Shanghai: School of Software Shanghai Jiaotong University) pp 6-9
[4] Wang Wei 2015 Research on Social Network Community Division Method Based on Clustering Analysis (Nanjing: Nanjing University of Posts and Telecommunications) pp 17-25
[5] Shi Taibin 2017 Research on Community Detection Algorithm and Community Profile Based on User Call Record (Zhejiang: Zhejiang University) pp 4-5
[6] Ao Wenjing 2013 Study of mining social network based on call records (South China University of Technology) pp 7-11
[7] Zhang Yuchao 2015 The Research on Cluster Analysis for Target Group of the Police (Jinan: Shandong University) pp 5-14