International Journal of Software Engineering
and Knowledge Engineering
Vol. 27, No. 4 (2017) 623–639

DOI: 10.1142/S0218194017500231

World Scientific
www.worldscientific.com

A Proof of Concept Study for Criminal Network Analysis with Interactive Strategies

Peng Zhou*,‡, Yan Liu*,§, Mengjia Zhao*,¶ and Xin Lou†,‖

*School of Software Engineering
Tongji University, Shanghai, P. R. China

†Shanghai iEven Information Technology Co., Ltd.
Shanghai, P. R. China
‡1435855@tongji.edu.cn
§yanliu.sse@tongji.edu.cn
¶1434319@tongji.edu.cn
‖steven.lou@ieven.com.cn

Received 31 August 2016
Revised 5 November 2016
Accepted 6 January 2017

The communication data are becoming increasingly important for criminal network analysis nowadays, and these data provide a digital trace which can be regarded as a hidden clue to support the crack of criminal cases. Additionally, performing a timely and effective analysis on it can predict criminal intents and take efficient actions to restrain and prevent crimes. The primary work of our research is to suggest an analytical process with interactive strategies as a solution to the problem of characterizing criminal groups constructed from the communication data. It is expected to assist law enforcement agencies in the task of discovering the potential suspects and exploring the underlying structures of criminal network hidden behind the communication data. This process allows for network analysis with commonly used metrics to identify the core members. It permits exploration and visualization of the network in the goal of improving the comprehension of interesting microstructures. Most importantly, it also allows to extract community structures in an appropriate level with the label supervision strategy. Our work concludes illustrating the application of our interactive strategies to a real-world criminal investigation with mobile call logs.

Keywords: Criminal network analysis; interactive strategy; community detection.

1. Introduction

Communication data are widely used in criminal network analysis (CNA) to understand direct relationships and identify implicit connections. Many efforts have been devoted to leverage these data in criminal community detection, connection strength evaluation, microstructure discovery, and suspect identification. However,

§Corresponding author.

the skill set required for CNA is complex and diverse, such as the application of empirical study and domain knowledge into the data preprocessing, criminal investigation knowledge, intelligence analysis experience, network visualization layout technique in different network scale, and social network analysis knowledge, which leads to mismatching of expectation and reality on the power of data analytics.

Although many efforts have been used to analyze the CNA, there are no unified processes to support the analyzing task in this domain because of its complexity in interactive process and the integration of techniques in various fields, let alone a standard or a common accepted framework to analyze the Criminal Network (CN). Before this paper, there has been no theoretical study on this issue, which means the framework in this paper can be a preliminarily pioneering study with research value to some extent.

After observing practical workflows which involve detectives, intelligence officers, data engineers, data scientist, and domain experts, combined with the technique of social network analysis and machine learning, we propose an interactive analysis process. According to the important results in each phase, the process can be divided into three phases, namely, (i) *network construction*: the generation of network structure, (ii) *metric design*: the core nodes and relations, (iii) *structure observation*: structures extraction in different levels. In each phase, the process adopted interactive strategies in order to formulate and assess hypotheses in a rapid, iterative manner — thus supporting the exploration of the CNs with the pace of human thought.

Our contributions contain the following: (1) Suggesting a generical analytical process for CNA which can be divided into three phases. (2) Proposing interactive strategies to the analytical process, such as using various visualization layouts to configure the network, reinterpreting different measures from the social network domain to the CN domain, and controlling the community structure level with label supervision strategy. (3) Conducting a proof of concept study using mobile call logs.

The paper is organized as follows: Section 2 describes related works about CNA. Section 3 illustrates how we proposed the interactive analytical process and introduced three visualization layouts for CN. In the following, Secs. 4–6 demonstrate the three phases separately with an example of CN being constructed, analyzed with metrics and explored with detection algorithm. The final section, Sec. 7, points the future work and makes a conclusion according to our work.

2. Related Work

In this section, we provide a background on social network analysis, and we survey existing literature in CNA with a particular focus on work about community structure and community detection. Various research streams focus on finding structural properties of CNs, including in phone call communication networks [1]. Understanding network properties such as the communities present in the network, or the roles that network members play in each community, can help network analysts and police detectives to unveil vulnerabilities and identify potential

opportunities to take destabilizing actions to fight criminal organizations. In the following, we discuss relevant work aimed at detecting network communities, discovering their patterns of interaction, identifying central individuals, and uncovering network organization and structure. Finally, we present a comprehension on the two relevant works about its advantages and limits.

2.1. *Social Network Analysis and CNA*

The development of the Social Network Analysis can be attributed to the well-known small world model proposed by Migram and Travers [2, 3] through analyzing characteristics of the real-life social networks after conducting social experiments in the real world. This model finds out a common property of the social networks that there exists a relatively short path which connects any pair of nodes within the network. In the 70s, Zachary [4] introduced another important concept of the community structure that nodes in such groups are densely interconnected among each other and weakly interconnected with those belonging to other groups. Barabasi *et al.* [5, 6] proved communication networks share the same dynamics of the growth called preferential attachment: new nodes tend to preferentially connect to existing nodes with high degrees rather than lower degrees. Those three crucial ingredients above characterize the structure of social networks and from the foundation of social network analysis [7].

In the last 30 years, many efforts have been used in order to analyze the CNs in a more intelligent way. One of the most important researches in the CNA domain is contributed by Malcolm Sparrow [8], who summarized four features of the CN, namely, (i) limited dimension — CNs are often composed of few thousands entities; (ii) incomplete information — CNs are unavoidably incomplete and erroneous; (iii) undefined border — it is difficult to pick out all the relations of each entity; and, (iv) dynamism — new relations always indicate constant evolution of network structure.

Fortunately, this contribution led to a trend that researchers tried to analyze CNs with the techniques in the domain of the social network analysis. For instance, Baker and Faulkner [9] studied illegal networks in the field of electric plants and Klerks [10] concentrated on criminal organizations in the Netherlands. Silke [11] and Brannan *et al.* [12] acknowledged a slow growth in the terrorism network and examined state of the art in CNA. Arquilla and Ronfeldt [13] summarized previous researches and proposed the concept of Netwar with its application to terrorism.

However, in 2006, a popular work by Valdis Krebs [14] applied network analysis in conjunction with network visualization theory to analyze the 2001-09-11 terrorist attacks. This work represents a starting point of a series of academic papers in which the social network analysis methods become applied to a real-world case, differently from previous work where mostly toy models and fictitious networks were used. Krebs paper inspired further research in network analysis for the design of better SNA applications to support intelligence agencies in the fight against terror, and law enforcement agencies in their quest fighting crime.

2.2. *Community structure and community detection*

In the study of complex networks, a network is said to have community structure [20] if the nodes of the network can be easily grouped into (potentially overlapping) sets of nodes such that each set of nodes is densely connected internally. In the particular case of non-overlapping community finding, this implies that the network divides naturally into groups of nodes with dense connections internally and sparser connections between groups. But overlapping communities are also allowed. The more general definition is based on the principle that pairs of nodes are more likely to be connected if they are both members of the same communities, and less likely to be connected if they do not share communities.

Community structure is one of the most common characteristics in the study of networks [15], which refers to the occurrence of groups of nodes in a network that are more densely connected internally than with the rest of the network. There is one widely adopted concept to investigate the quality of these structures in the network called network modularity, which can be expressed as follows: let us consider a network, represented by $G = (V, E)$, which has been partitioned into m communities; its corresponding value Q of network modularity is defined as:

$$Q = \sum_{s=1}^{m} \left[\frac{l_s}{|E|} - \left(\frac{d_s}{2|E|} \right)^2 \right], \tag{1}$$

assuming l_s is the number of links between nodes belonging to the s_{th} community and d_s is the sum of the degrees of the nodes in the s_{th} community. High values of Q indicate high values of l_s for each discovered community, yielding to communities internally densely connected and weakly coupled among each other. The process of detecting community structures in a network is called community detection, and it is still regarded as a computationally difficult task. However, several methods for community detection have been proposed and applied with varying extents of success such as minimum-cut method [16], hierarchical clustering [17], Girvan–Newman algorithm [18], modularity maximization and clique-based methods [19].

Lots of researches have shown that community detection can be demonstrated as a powerful tool to analyze the structure in the CNs. Emilio *et al.* [20] employed Girvan–Newman algorithm and a variant based on modularity optimization called Newmans algorithm to detect and explore the community structures in the CNs reconstructed from phone call logs. Hamed Sarvari *et al.* [21] performed a large-scale analysis with clique-based methods to find patterns and substructures of that network based on a publicly leaked set of customer email addresses. However, these studies and early researches somehow neglected the importance of network visualization and the interactions during the analysis process, laying emphasis on aspects related more to statical network characterization. In our research, we stress these twofold by introducing various visualization layouts and adopting crucial interactive points.

3. Interactive Strategies

At present, the task of analyzing CNs cannot be finished purely relying on the computational analysis or manual analysis. We examined all the analytical processes in these works and found that almost all the processes about CNA are interactive process between humans and computers, but the interactive strategies are weak, limited and even not involved obviously and deeply. In order to gain better understanding of the process with some strong interactive strategies during the CNA, we followed five investigations based on a real criminal cases and also observed a general pattern of work shared by most police officers and intelligence analysts.

After doing induction and summary on the workflow, we proposed an interactive analytical process for CNA. As is shown in Fig. 1, the process can be mainly divided into three phases according to each phase's promising results, namely, (i) *network construction*: the generation of network structure, (ii) *metric design*: the core nodes and relations, (iii) *structure observation*: the extraction of organization structure. This process is supposed to assist investigator to analyze CN in an interactive way.

In the first phase, we clean data interactively depending on the empirical rules suggested by the intelligence officers and data engineers, and do some extraction to get the edge tables, provide a format of intermediate result storage and finally perform network visualization by layouts selected by visualization experts. In the second phase, some metrics in the domain of SNA are applied into CNA, and we also need to reinterpret them before employing, then we obtain the ranking results. Additionally, we perform rendering on each entity by its metric value. At the end of this phase, a certain evaluation is defined to suit for our CN to examine the quality of different

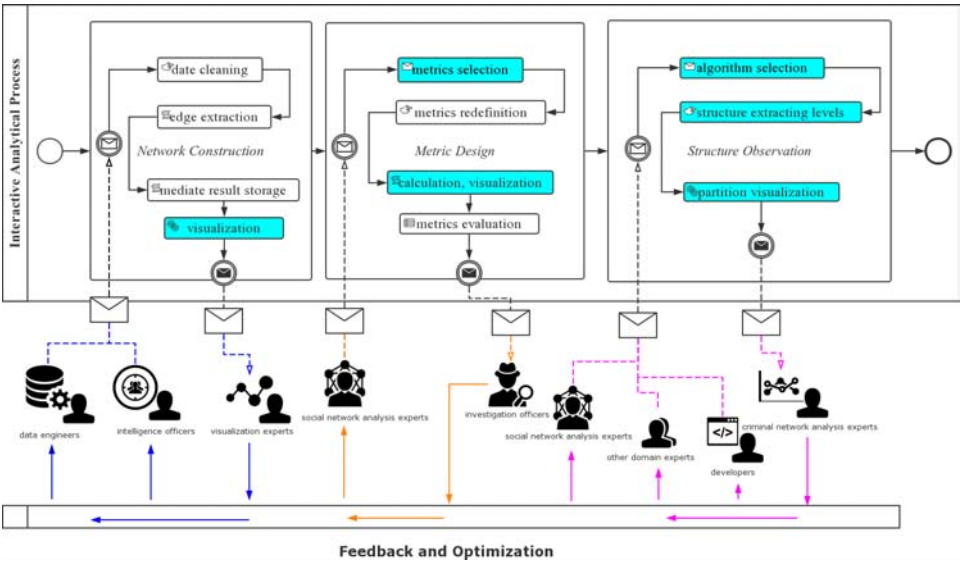


Fig. 1. The interactive analytical process.

metrics. The crucial point related to interaction in this phase is twofold: one is different metric selections which can measure different centralites, another is that we introduce lots of layouts in consideration of the scale of the CN. The final phase is structure observation. Firstly, we choose a detection algorithm taking into consideration, the features of the CNs constructed in phase one, such as network scale, complexity, and so on. So we need the support of SNA experts, other domain experts and programming developers. Then ascertain the extracting structure level to amplify or narrow down graph interests. Finally, a partition visualization result is allowed to examine the quality of the community structure and even find out the hidden knowledge.

Note that the three phases are an integrated whole. The previous stage outputs is regarded as the next phase's input, so each phase results would interfere with the next phase results. Consequently, we suggest another stage — feedback and optimization to ensure the effective communication between various domain experts and optimization of the whole interactive process. As shown in Fig. 1, there are three iterative cycles drawn by different colors, and before starting every phase, some experts would work together to design and carry out their schemes. At the end of each phase, some experts will judge the results and provide the feedback when the result is not good enough. Hence, we call this interactive cycle the interactivity between our collaborators. Since the steps colored by cyan demand rapid interactions and various configurations, we call the interactivity in the analytical process. To sum up, this paper mainly uses these two kinds of interactivties to ensure the accuracy of the final result and diversity of the configurations.

4. Network Construction

In this section, we first give some explanations of the dataset that we used, and then propose a generically interactive workflow to support this phase task. This workflow shows how the phone call logs are preprocessed, transformed, extracted and constructed into a network, as is shown in Fig. 2. Finally, we introduce different visualization layouts to configure the network reconstructed from the phone logs.

4.1. *Dataset*

To illustrate the following works better, we present a brief introduction of the dataset. Our dataset is based on a real investigation case of night burglary. According to the evidences provided by known suspects, (use mobile phones to keep in touch to commit together) the investigator downloaded phone call records from intelligence enforcement, and we did some processing in order to protect the privacy. A brief description of the dataset is shown in Table 1. With a preliminarily discovery on the dataset, we find that the dataset includes 73 detailed phone call tickets, which contains 25 known suspects and 48 persons involved in this investigation case. Phone call tickets are all in the period from January 2014 to June 2015. After summing up the phone call logs from all the call tickets, we obtain totally 1016833 phone logs.

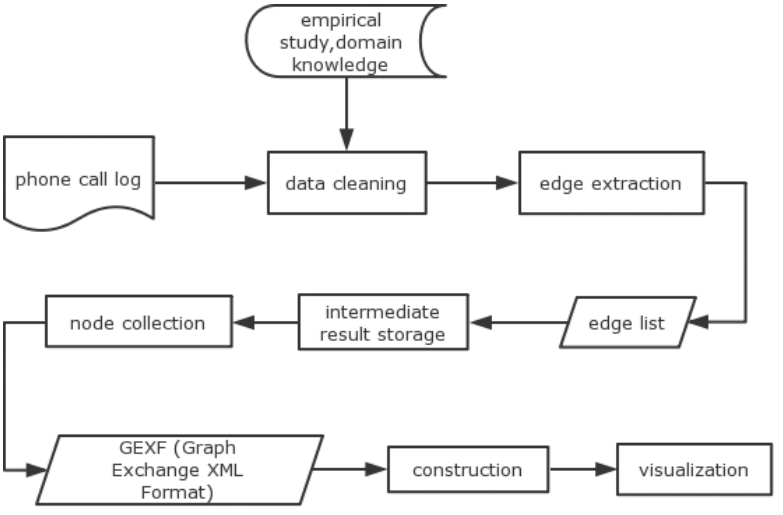


Fig. 2. The data preprocessing workflow.

Table 1. Dateset summary.

Detailed ticket	Known suspect profile	Time period	Total call logs
73	25	2014.1 2015.6	1016833

4.2. *Data preprocessing and network construction*

Performing a good preprocessing on the dataset can get rid of the noise as much as possible, and it permits the network structure to be constructed easily without cost of quality. It allows the network metrics to depict a member more exactly and make visualization closer to the nature of the criminal communication network. Additionally, it likewise can reduce the following analytical errors and the workload of the structure analysis phase. Owing to the great importance of data preprocessing, we think highly of this stage.

The data preprocessing stage is shown in Fig. 2, and the properties of the raw data were displayed in Fig. 3(a). According to the suggestion of the investigator, we need to select some of them for the following works. The data cleaning comes firstly to our work, the empirical study and domain knowledge that we used are as follows:

(a) Eliminate the call logs containing bank notifications, communication providers and other public service providers. For instance, the number 10086 is a communications service providers in China, and 95595 is the call number of Agricultural Bank of China. (b) Remove redundant logs produced by peer-to-peer calling. When one entity calls another entity, both of their phone detailed tickets will generate almost the same log with a certain time difference and different call types that one is calling and another is called. (c) Wipe off the call logs whose call duration is zero.

Fig. 3. The data samples and network modeling.

This type of logs indicate that it is not a successful connection and should not be considered in the following works.

After data cleaning, the workflow extracts the data with few noise to build the edge list of the network. The edge list sample is shown in Fig. 3(b). In this step, we need to change call type according to the corresponding call type in order to identify the source nodes and the target nodes. When finishing edge extracting, our work chooses neo4j — a famous graph database as our intermediate result storage which is shown in Fig. 4. Then we write a python script to collect the nodes to build the node set and importing the node sets and edge list by the python-igraph package and constructing our CNs, the network modeling is shown in Fig. 3(c). Finally, we export GEXF file in order to visualize by a software called Gephi. At length, we construct a burglary CN which has totally 7840 nodes and 1016833 links.

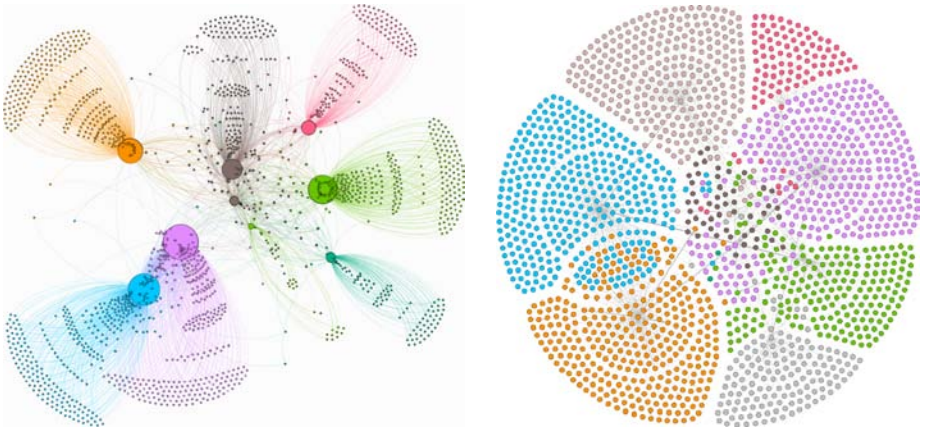
Fig. 4. The mediate results in the Neo4j.

4.3. *Visualization layouts*

Considering the large scale and complexity of the CN, we choose serveral layouts below to configure our network. ForceAtlas2 [22] is a force vector algorithm proposed in the Gephi software, appreciated for its simplicity and for the readability of the networks it helps to visualize — its distinctive features, its energy-model and the way it optimizes the speed versus precision approximation to allow quick convergence. We also claim that ForceAtlas2 is handy because the force vector principle is unaffected by optimizations, offering a smooth and accurate experience to user, as is shown in Fig. 5(a).

The Fisheye layout is a local linear enlargement technique that, without modifying the size of the visualization canvas, allows us to enhance the region surrounding the focus, while compressing the remote neighboring regions. The overall structure of the network is nevertheless maintained, as is shown in Fig. 5(a).

Normally, graphs are depicted with their vertices as points in a plane and their edges as line or curve segments connecting those points. There are different styles of representation, suited to different types of graphs or different purposes of presentation. The Fruchterman Reingold layouts [23] concentrate on the most general class of graphs: undirected graphs, drawn with straight edges. In our paper, we introduce the FR layout that attempts to produce aesthetically-pleasing, two-dimensional pictures of graphs by doing simplified simulations of physical systems. This layout is concerned with drawing undirected graphs according to the fact that some generally do well at distributing vertices evenly, making edge lengths uniform, and reflecting symmetry, as is shown in Fig. 5(b). Our goals for the implementation are speed and simplicity.



(a) FA2 and Fisheye layout (b) FR layout

Fig. 5. The left figure shows a combination of ForceAtlas2 with Fisheye layout. The right part of the figure is FR layout.

5. Metric Design

In order to exactly capture the influence of core members and its roles playing in the criminal organizations, we introduce a series of measures from the domain of Social Network Analysis and interpret them in the context of our certain CN. After calculating these metrics, we render the network view with different color intensity and size based upon the metric value of each entity and spotlight certain nodes.

5.1. *Network measures reinterpretation in CNs*

Degree centrality measures the activity and influence by calculating the number of direct relations a node has. It is convinced that a call number with a high degree could be regarded as a hub which acts as an import information channel in the communication network and a node with high activity. It is defined as:

$$Dgr_i = \frac{d_i}{S-1} = \frac{\sum_{i \in M} m_{ij}}{S-1}, \tag{2}$$

where d_i directly links to other call numbers of a call number i, and m_{ij} is the ijth element of the adjacency matrix M and S is the sum of the whole call numbers. $S-1$ is the normalization factor.

Betweenness centrality is another indicator to measure nodes centrality, which equals the number of the shortest paths from all nodes to all other nodes which pass through that node. In the context of CN, a call number with high betweenness can be regarded as broker of messages, indicating the great importance of this call number in the information transfer. Most importantly, this kind of call numbers usually connects two or more densely call number clusters, and their removal might lead to the destabilization, abruption and even destruction of the CNs. The betweenness centrality of a call number (i) is redefined as:

$$Btw_i = \frac{\sum_{j<k \in N} \frac{p_{jk}(i)}{p_{jk}}}{(S-1)(S-2)}, \tag{3}$$

where p_{jk} is the total number of the shortest paths from call number j to call number k and the $p_{jk}(i)$ is the number of those paths that pass through call number i, the $(S-1)(S-2)$ is the normalization factor.

Closeness centrality is the reciprocal of the sum of the shortest paths between that node and all other nodes in the network. In the analysis of our CN, a high closeness indicates whether this call number can reach majority of the other nodes easily and faster. The closeness centrality of a call number (i) is redefined by the expression:

$$Clsn_i = (H_i)^{-1} = \frac{S-1}{\sum_{j \in N} d(i,j)}, \tag{4}$$

where $d(i,j)$ is the distance between node i with node j, H_i is the normalized distance.

Eigenvector centrality defines another centrality with the consideration of the importance of their neighbors. In the framework of our CN, a call number with high eigenvector centrality means that this call number can reach a group of other call numbers easily and quickly. The eigenvector centrality of a node (i) is redefined as:

$$Eign_i = \frac{1}{\alpha} \sum_{j \in L(i)} m_j = \frac{1}{\alpha} \sum_{j \in N} a_{ij} x_j, \tag{5}$$

where $L(i)$ is the direct link set of call number i, α is a constant, and a_{ij} is the ijth element of the adjacency matrix M.

Clustering coefficient is a measure of the degree to the aggregation of nodes in a graph. In the context of our CN, a call number with high clustering coefficient means a high likelihood of that the direct links of the given call number can reach each other. It is redefined as:

$$Clst_i = \frac{|e_{jk}|}{l_i(l_i - 1)}, \tag{6}$$

where e_{jk} is the link existing in the neighbors of call number i and l_i is the number of direct links of the call number i.

PageRank is a variant of the eigenvector centrality measure in some degree. Our work employs this measure to the importance of a certain call number globally. It is redefined as:

$$\text{Page}(i) = (1 - f) + f * \sum_{j \in H(i)} \frac{P_i(j)}{L_j}, \tag{7}$$

where $H(i)$ is the set of call numbers directly linking to the call number i, L_j is the number of outgoing links in j and f is the damping factor.

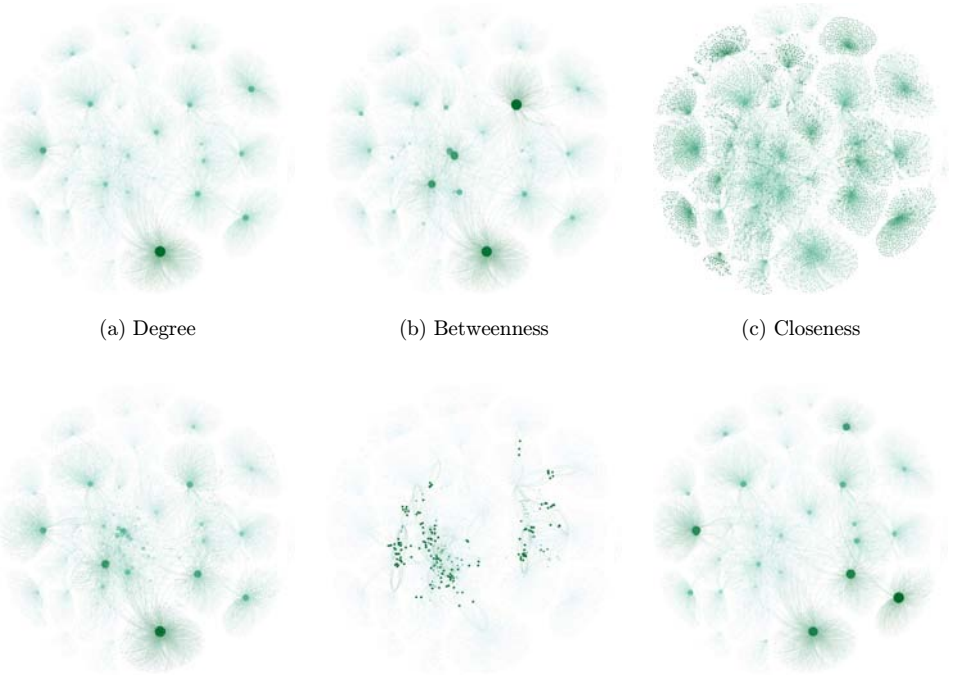
5.2. *Metrics visualization and ranking interpretation*

Our visualization results with the combination view of Fruchterman Reingold layout and Fisheye layout are shown in Fig. 6, and the color intensity and the size of the nodes are both rendered by their corresponding metric value. To check the usability of different metrics, we defined the hit rate of the known burglary suspect in the top 25 and top 100 ranks as:

$$h = \frac{n}{K} * 100\%, \tag{8}$$

where n is the number of known suspects in the top K ranking of the corresponding metric. The hit rate results were shown in Table 2.

To begin with, we found that the degree, betweenness, eigenvector and PageRank have a similar and excellent layout view. After checking the hit rate of these metrics, it indicated that 96% of suspect belongs to the top 25 of the degree, 88% to betweenness, 84% to eigenvector, 96% to PageRank. Therefore, in the top 100 ranking,



(a) Degree (b) Betweenness (c) Closeness

(d) Eigenvector (e) Clustering (f) PageRank

Fig. 6. Network metrics.

Table 2. Different network metrics of the known suspect.

Dgr	Btw	Clsn	Eigvt	Clst	PgRk
9699*	1002*	7071	5636*	5117	7777*
0691*	9699*	1083	8662*	3229	9699*
1509*	5636*	8890	0691*	4903	8662*
7777*	2406*	5703	1002*	5064	0691*
8662*	8662*	4848	7777*	7037	7789*
8199*	4751*	7515	5676*	9868	1002*
1002*	8199*	8986	8199*	1115	8199*
2223*	6526*	9549	9699*	7777*	9250*
1359*	7777*	5362	2406*	8318	3605*
5636*	0691*	4097	2223*	5782	5636*
9250*	1509*	9891	0399*	9699*	2223*
7789*	2223*	8582	4076*	6409	1509*
3605*	7970*	3401	9250*	0691*	8391*
8391*	8391*	0206	7789*	5435	1359*
2406*	8318	1223	4494	8662*	2406*
5676*	1359*	2601	8391*	5347	6526*
7970*	5676*	6179	1509*	7071	7970*
7574*	4494	0393	1359*	8199*	5676*
6526*	7789*	3012	3605*	1083	7574*

Table 2. (*Continued*)

Dgr	Btw	Clsn	Eigvt	Clst	PgRk
3157*	3605*	0722	6859	8890	7603*
4076*	7588	5894	6526*	7789*	4076*
7692	9250*	3040	8033	1002*	0399*
2647*	3157*	5049	4751*	9250*	3157*
0399*	4076*	2602	2397	3605*	2647*
4751*	7574*	4892	1050	5636*	7692

Table 3. The hit rate of top-K ranking.

Rank	Dgr (%)	Btw (%)	Clsn (%)	Eigvt (%)	Clst (%)	PgRk (%)
Top 25	96	88	0	84	60	96
Top 100	100	100	0	100	100	100

they consequently got 100%. These four metrics were proved to be valid measures for burglary group CN. However, the remaining two metrics were not good enough and even no one suspect was ranked in the top 25 of closeness metric, the details was shown in Table 3. Although the hit rate of clustering coefficient only got 60% in the top 25, we performed a manual check on the top 100, and it worked well. But for closeness centrality, it still got 0% in the top 100. The most likely factor resulting in this phenomenon is the way of data collection.

In summary, the metrics selection should take the way of the data collecting into consideration, ensuring that the effective metrics for measuring the CNs would be adopted. In the context of burglary group phone CN, these four indicators which are degree, betweenness, eigenvector, and PageRank, are good enough to analyze the network in an effective way. The clustering coefficient might work well if we broaden the ranking scope.

6. Structure Observation

This section focuses on the structure analysis of the CN and is expected to select a suitable algorithm to finish the task of community extraction rapidly and interactively. After the metric design phase, we have got the core entities and important relations, but the interaction of these members is still unknown. So, this section is focusing on the structure analysis of the CN and is expected to select a suitable algorithm to finish the task of community extraction rapidly and interactively.

6.1. *Fast unfolding with label supervision strategy*

In the consideration of the scale of our burglary CN, which contains 7840 nodes and 103043 links, and the demands for interactive point mentioned in Sec. 3, Fast Unfolding [24] has been adopted to detect the communities structure in burglary CN. Besides, we provide a label supervision strategy to put prior knowledge and evidence

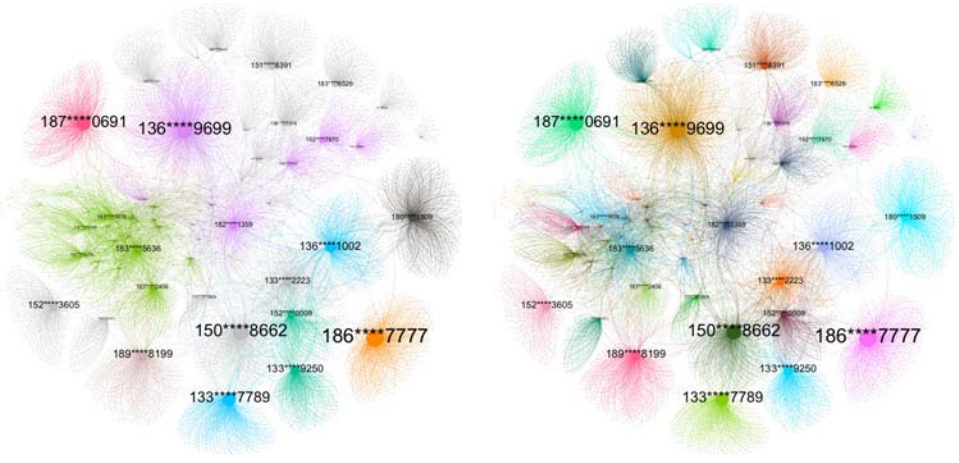
into the structure extraction. Thus, the level of the community structure can be well controlled. Finally, we visualize our community results with combination of FR layout and Fisheye layout. Besides the extremely high speed, another reason why we choose Fast Unfolding algorithm is that this heuristic method provides a parameter of resolution to control the scale of the communities detected. In most situations, the default value of the resolution is 1, if the value is lower, then the size of the communities detected would become smaller, which also means that it will produce more communities. If we set a higher value than the default, it means bigger structure and smaller number of communities. Therefore, in this way, our process can gradually increase or decrease the value of resolution in order to extract subgroups with appropriate scales in the burglary CN.

According to the analysis of the workflow shared by most police officers, we simulate the process of putting the evidence obtained into the following investigation and combine it with the philosophy of supervision in machine learning technique. However, it is unknown when we should stop reduction or increase the resolution to gain the appropriate structure level. So, we proposed a general method based on label supervision to solve this question which can be suitable for all the detection algorithms. According to the initial stage of the detection algorithm, it can be divided into forward and reverse strategies. If the detection starts from one community to appropriate amounts of community, it means that the process of detection is one to more, and the label supervision should inspect the partition of the nodes labeled in the same level, so we called this type as forward strategy; but if the algorithm is from more to less, the label supervision should inspect the mergence of the nodes in the same level. Hence, this type called reverse strategy. The following steps describe the process of the FU with the reverse label supervision strategies: (1) assign each node to a different community, for each node i, consider the neighbors j of i, put node i to its neighbor j when reaching the maximum positive gain of modularity, (2) check the nodes labeled in the same level whether be assign into the same structure, and (3) take the communities found during step 1 as nodes and build a new network, then back to step 1.

In the context of the CNA, when we know several entities in a same structure level, then we label this node and examine them after each phase of the Fast Unfolding working on the network till that they are arranged into different communities, and this detection can be seen as an effective analysis.

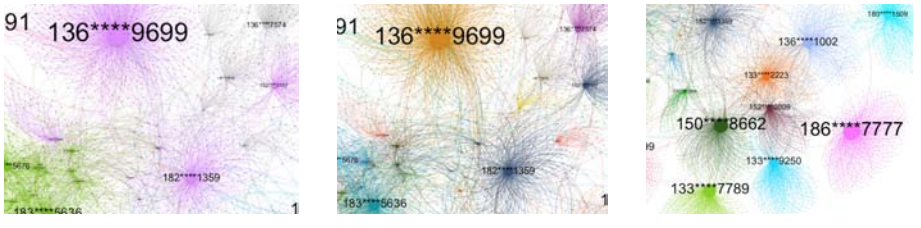
6.2. *Structure partition results*

Figure 7(a) has shown the result after performing FU on the burglary CN with the default resolution value, where the modularity value Q is 0.813 and the number of the communities is 22. The network was partitioned into different substructures and each community, including nodes and edges which were rendered by different colors and the size of the node varies with the community scale. In order to gain a appropriately structural level, we marked the known suspects call number 136****9699



(a) r = 1 m = 0.813 nc = 22 (b) r = 0.18 m = 0.778 nc = 41

Fig. 7. The left figure shows a community detection with resolution = 1 and the number of communities is 22. The right part of the figure is the situation with resolution of 0.18.



(a) Label supervision process-1 (b) Label supervision process-2 (c) The protential suspect

Fig. 8. The sample of label supervision method.

and 182*****1359 shown in Fig. 8(a) and put this knowledge into each iteration of the Fast Unfolding algorithm with a step of 0.1 reduction of the resolution. After 83 iterations, we examined these call numbers which were assigned into different communities shown in Fig. 8(b), where the resolution is 0.18, modularity is 0.778 which is also a high value and the number of communities is 41. This partition result is shown in Fig. 7(b).

Another goal of this phase is to find out potential suspects in burglary gang. In Fig. 8(c), we can find the entity with phone number of 152****0099 which was rendered by brown kept large amounts of relation with top ranking criminal entities, which are 150****8662, 186****7777, 133****2223 and 133****9250. Consequently, this entity maybe another member of the burglary group, so we recommend such entities to officer in support of the following investigation.

7. Conclusion

Future Work: Our framework purely concentrates on the mobile phone record, but there are lots of heterogeneous heterologous data which can depict a suspect profile more completely, such as location-based service dataset from LBS provider and traces information from traffic service provider, which can reconstruct a more holonomic network dynamically with matching on the GIS platform, such as ArcGIS, Google Map. Consequently, we hope that our framework can analyze CN from a perspective of spatial and temporal dimension in order to capture more interests inside the CNs dynamically.

In this paper, we introduce an interactive analytical process to explore the CNA with a case study using mobile phone logs. This process generally works well with the CN in our case study. In detail, the core members of the burglary group usually get a high ranking in network metrics, but not all measures are effective for our analysis, such as closeness centrality and clustering coefficient, and three visualization layouts help a lot during the whole process. Most importantly, our framework can extract community structures as an appropriate level with the application of label supervision into the Fast Unfolding algorithm.

References

1. D. V. Canter and L. Alison, *The Social Psychology of Crime: Groups, Teams and Networks* (Ashgate, 2000).
2. S. Milgram, The small world problem, *Psychology Today* **2**(1) (1967) 60–67.
3. J. Travers and S. Milgram, An experimental study of the small world problem, *Sociometry* **32**(4) (1969) 425–443.
4. W. W. Zachary, An information flow model for conflict and fission in small groups, *J. Anthropol. Res.* **33**(4) (1977) 452–473.
5. R. Albert, H. Jeong and A.-L. Barabási, Internet: Diameter of the world-wide web, *Nature* **401**(6749) (1999) 130–131.
6. R. Albert and A.-L. Barabási, Statistical mechanics of complex networks, *Rev. Mod. Phys.* **74**(1) (2002) 47.
7. E. Ferrara and G. Fiumara, Topological features of online social networks, *Commun. Appl. Int. Math.* **9**(2) (2012) e381.
8. M. K. Sparrow, The application of network analysis to criminal intelligence: An assessment of the prospects, *Social Networks* **13**(3) (1991) 251–274.
9. W. E. Baker and R. R. Faulkner, The social organization of conspiracy: Illegal networks in the heavy electrical equipment industry, *Am. Sociol. Rev.* **58**(6) (1993) 837–860.
10. P. Klerks, The network paradigm applied to criminal organizations: Theoretical nitpicking or a relevant doctrine for investigators? Recent developments in the Netherlands, *Connections* **24**(3) (2001) 53–65.
11. A. Silke, The devil you know: Continuing problems with research on terrorism, *Terror. Polit. Violence.* **13**(4) (2001) 1–14.
12. D. W. Brannan, P. F. Esler and N. Anders Strindberg, Talking to "terrorists": Towards an independent analytical framework for the study of violent substate activism, *Stud. Confl. Terror.* **24**(1) (2001) 3–24.

13. J. Arquilla and D. Ronfeldt, *Networks and Netwars: The Future of Terror, Crime, and Militancy* (Rand Corporation, 2001), Vol. 10 No. 2.
14. V. E. Krebs, Mapping networks of terrorist cells, *Connections* **24**(3) (2002) 43–52.
15. M. A. Porter, J.-P. Onnela and P. J. Mucha, Communities in networks, *Notices of the AMS* **56**(9) (2009) 1082–1097.
16. M. E. Newman, Detecting community structure in networks, *Eur. Phys. J. B* **38**(2) (2004) 321–330.
17. A. J. Alvarez, C. E. Sanz-Rodríguez and J. L. Cabrera, Weighting dissimilarities to detect communities in networks, *Philos. Trans. R. Soc. Lond. A* **373**(2056) (2015) 20150108.
18. M. E. Newman, Fast algorithm for detecting community structure in networks, *Phys. Rev. E* **69**(6) (2004) 066133.
19. T. S. Evans, Clique graphs and overlapping communities, *J. Stat. Mech. T* **2010**(12) (2010) P12037.
20. E. Ferrara, P. De Meo, S. Catanese and G. Fiumara, Detecting criminal organizations in mobile phone networks, *Exp. Syst. Appl.* **41**(13) (2014) 5733–5750.
21. H. Sarvari, E. Abozinadah, A. Mbaziira and D. McCoy, Constructing and analyzing criminal networks, in *Security and Privacy Workshops*, 2014, pp. 84–91.
22. M. Jacomy, T. Venturini, S. Heymann and M. Bastian, Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software, *PloS One* **9**(6) (2014) e98679.
23. T. M. Fruchterman and E. M. Reingold, Graph drawing by force-directed placement, *Software: Practice and Experience* **21**(11) (1991) 1129–1164.
24. V. D. Blondel, J.-L. Guillaume, R. Lambiotte and E. Lefebvre, Fast unfolding of communities in large networks, *J. Stat. Mech.* **2008**(10) (2008) P10008.