International Journal of Software Engineering and Knowledge Engineering

RESEARCH NOTES: A PROOF OF CONCEPT STUDY FOR CRIMINAL NETWORK ANALYSIS WITH INTERACTIVE STRATEGIES(116)

Peng ZHOU
School of Software Engineering, Tongji University
Shanghai, China
1435855@tongji.edu.cn

Yan LIU
School of Software Engineering, Tongji University
Shanghai, China
yanliu.sse@tongji.edu.cn

Mengjia ZHAO
School of Software Engineering, Tongji University
Shanghai, China
1434319@tongji.edu.cn

Xin LOU
Shanghai iEven Information Technology Co.,Ltd.
Shanghai, China
steven.lou@ieven.com.cn

Received (Day Month Year)
Revised (Day Month Year)
Accepted (Day Month Year)

The communication data are becoming increasingly important for criminal network analysis nowadays, those 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 a 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; network measure;

1. Introduction

Communication data are widely used in criminal network analysis to understand direct relationships and identify implicit connections. Many efforts have been devoted to leverage those data in criminal community detection, connection strength evaluation, microstructure discovery, and suspect identification. However, the skill set required for criminal network analysis (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 mismatching of expectation and reality on the power of data analytics.

After observing practical workflows which involving detectives, intelligence officers, data engineers, data scientist, and domain experts, combined with the technique of social network analysis and machine learning, we proposed 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 exploration Criminal Networks (CN) with the pace of human thought.

Our contributions contain: 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, rendering various network measures metrics, and controlling the community structure level with label supervision strategy. 3) conducting a proof of concept study using mobile call logs.

2. Related work

In this section, an overview of previous works in the domain of criminal network analysis has been firstly provided, and then comes to the concept of community structure and structure detection method. Finally, we present a comprehension on the two relevant works about its advantages and limits.

2.1. *Criminal Networks Analysis*

Over the last three decedes, many efforts have been made in order to analyse the CNs in a more intelligent way. One of the most important research in the CNA domain is due to Malcolm Sparrow [1], who summarized four features of the Criminal Network, namely, i) limited dimension; ii) incomplete information; iii) undefined border; and, iv) dynamism.

Since then, a new trend arose that researchers tried to analyze Criminal Networks with the techniques in the domain of the social network analysis, thanks to

the contribution of Malcolm Sparrow. For instance, Baker and Faulkner [2] studied illegal networks in the field of electric plants, Klerks [3] concentrated on criminal organizations in the Netherlands. Silke [4] and Brannan et al. [5] acknowledged a slow growth in the terrorism network and examined state of the art in criminal network analysis. Arquilla and Ronfeldt [6] summarized previous researches and proposed the concept of Netwar with its application to terrorism.

However, in 2006, a popular work by Valdis Krebs [7] 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 social network analysis methods become applied to a real-world case, distinguishing from previous work where mostly toy models and fictitious networks were used.

2.2. *Community Structure and Community Detection*

Community structure is one of the most common characteristics in the study of networks [8], 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.

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 [9], hierarchical clustering [10], Girvan-Newman algorithm [11], modularity maximization and clique based methods [12].

Lots of researches have shown that community detection can be demonstrated as a powerful tool to analyze the structure in the criminal networks. Emilio et al. [13] 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. [14] 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.

3. Interactive Strategies

At present, the task of CNA can not purely rely on the computational analysis or manual analysis. After examining all the analytical process in these tasks, we found that almost all the interactive strategies are weak, limited and even not involved obviously and deeply. In order to gain better understanding of the process with strong interactive strategies, we followed 5 investigations based on real world criminal cases and did some inductions and summaries on the workflow.Based on our findings, we proposed an interactive analytical process for CNA. As is shown on the Figure 1. the process can be mainly divided in to 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.

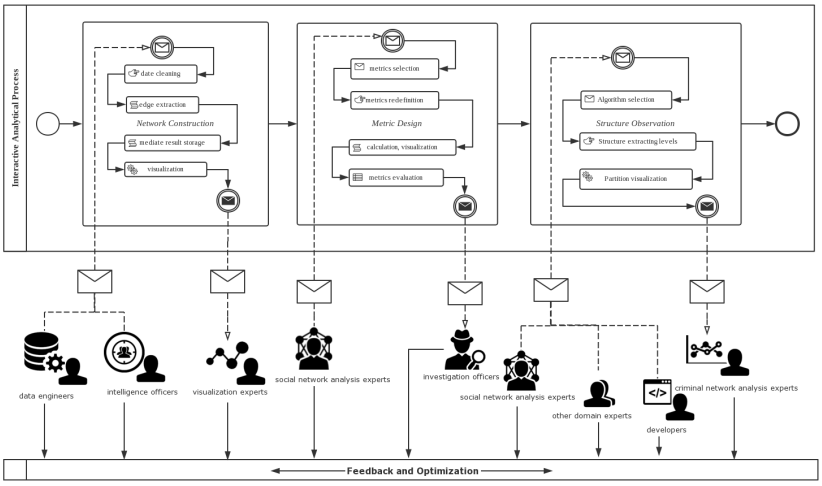


Fig. 1. The interactive analytical process

In the first phase, we clean data interactively depending on the empirical rules, which is suggested by 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 them, 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 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 the features of the CNs which constructed in the phase one into consideration, 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 results is allowed to examine the quality of the community structure and even find out the hidden knowledge. Note that the three phase is 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.

4. Network Construction

In order to assure the quality of the network construction, we proposed a generical interactive workflow to support this phase task. This workflow shows how the phone call logs are preprocessed, transformed, extracted and constructed into a network.

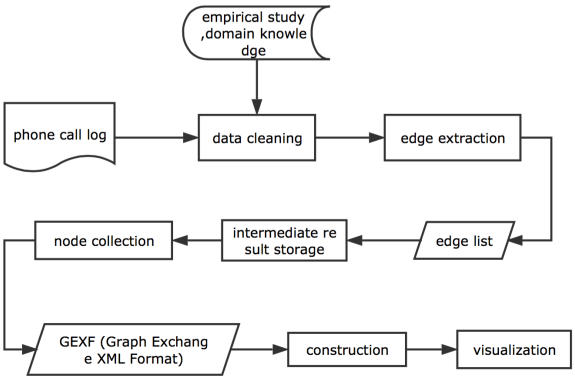


Fig. 2. The data preprocessing workflow

The data preprocessing stage was shown on the Figure 2. Our work begins with the data cleaning. In this step, the empirical study and domain knowledge that we used are as followed: *a)* eliminate the call logs containing bank notifications, communication providers and other public service providers. For instance, the number 10086 is a communications service provider 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. This type of logs indicate that is not a successful connection and should not be considered in the following works. At length we construct a burglary criminal network, which has totally 7840 nodes and 1016833 links.

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 SNA and interpret them in the context of our certain CN. After calculating these metrics, we render the network view with different layouts.

Our visualization results with the combination view of Fruchterman Reingold layout and Fisheye layout was shown on the Figure 3, and the color intensity and the size of the nodes both are rendered by their corresponding metric value. To check

Table 1. Network Measures Reinterpretation in CNs

Metrics	**Function**
$Dgr_i = \frac{d_i}{S-1} = \frac{\sum_{i \in M} m_{ij}}{S-1}$	the activity and influence of call numbers
$Btw_i = \frac{\sum_{j<k \in N} \frac{p_{jk}(i)}{p_{jk}}}{(S-1)(S-2)}$	the speed of information transfermation
$Clsn_i = (H_i)^{-1} = \frac{S-1}{\sum_{j \in N} d_(i,j)}$	the difficulty connceting other numbers
$Eign_i = \frac{1}{\alpha} \sum_{j \in L(i)} m_j = \frac{1}{\alpha} \sum_{j \in N} a_{ij} x_j$	the importance of their neighbours
$Clst_i = \frac{\lvert e_{jk} \rvert}{l_i(l_i-1)}$	the likelihood of neighbours reaching each other
$Page(i) = (1-f) + f * \sum_{j \in H(i)} \frac{P_i(j)}{L_j}$	the global importance of a certain call number

Note: parameter description

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

[2]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.

[3]where $d_(i,j)$ is the distance between node i with node j, H_i is the normalized distance.

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

[5]where e_{jk} is the link existing in the neighbours of call number i and l_i is the number of direct links of the call number i.

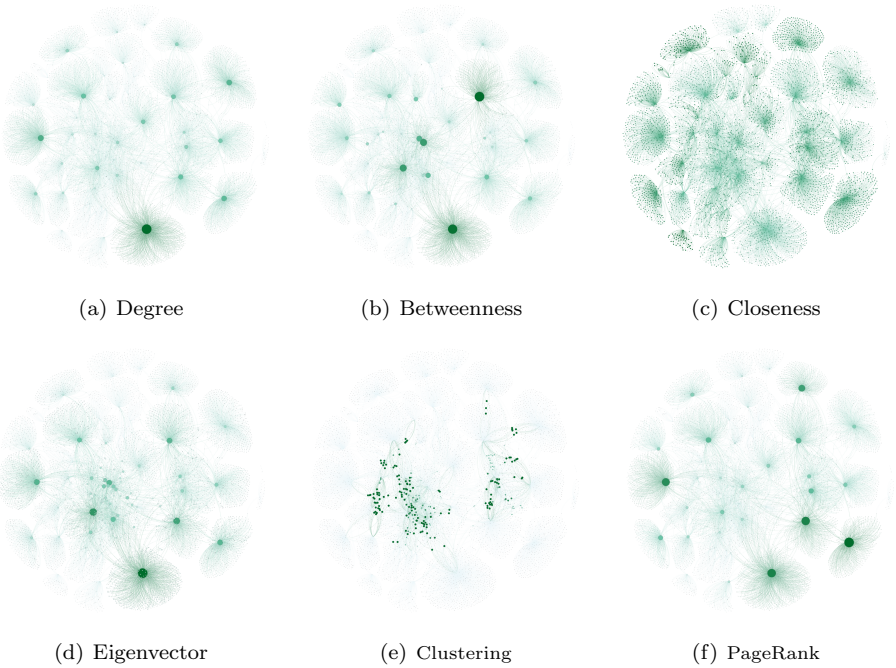
[6]where $H(i)$ are the set of call numbers directly linking to the call number i , L_j is the number of outgoing links in j and f is the damping factor.

the usability of different metrics, we defined the hit rate of the known burglary suspect in the top 25 and top 100 rank as:

$$h = \frac{n}{K} * 100\% \tag{1}$$

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



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

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

Fig. 3. Network metrics

ranking, 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. Although the hit rate of clustering coefficient only got 60% in the top 25, we performed a manual check on the top 100, it works 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.

Table 2. 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%

In summary, these four indicators which are degree, betweenness, eigenvector, and PageRank, are good enough to analysis 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.

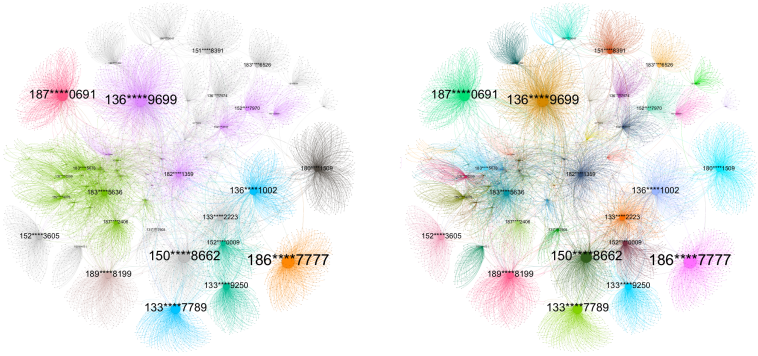
6.1. *Fast Unfolding with Label Supervision Strategy*

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 the section III, Fast Unfolding [15] has been adopted to detect the communities structure in burglary CN. Besides, we provide a label supervision strategy to put priori 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.

However, it is unknown when we should stop reduction or increasements of 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 algorithm. According to detection algorithm initial stage, it can be divided into forward and reverse strategy. If the detection starts from one community to appropriate mounts of community, it means that the process of the detection is one to more, and the label supervision should inspect the partition of the nodes labeled in the same level, we called this type 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. 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 neighbours j of i, put node i to its neighbour 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. 3) take the communities found during the step 1 as a nodes and build a new network, then back to step 1.

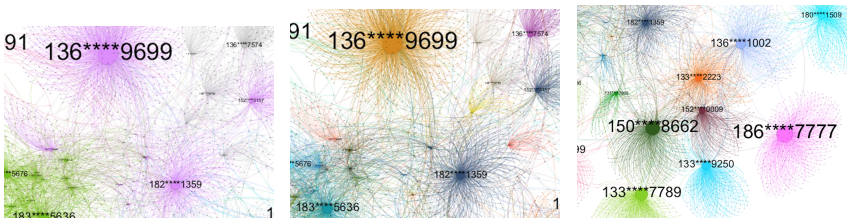
6.2. *Structure Partition Results*

The Figure 5 (a) has shown the result after performing FU on the burglary criminal network 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 communities, including nodes and edges, were rendered by different colors and the size of the node varies with the community scale. In order to gain a appropriately structural levels, we marked the known suspects call number 136****9699 and 182*****1359 shown on the Figure 6 (a) and put this knowledge into each iteration of the Fast Unfolding algorithm with a step of 0.1 reduction



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

Fig. 4. 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. 5. The sample of label supervision method

of the resolution. After 83 iterations, we examined these call numbers which were assigned into different communities shown on the Figure 6 (b), where the resolution is 0.18, modularity is 0.778 also a high value and the number of communities is 41. This partition result is shown on the Figure 5 (b).

An another goal of this phase is to find out potential suspects in burglary gang. In the Figure 6 (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, 133****9250. Consequently, this entity maybe an another member of the burglary group, so we recommend such entities to officer in support of the following investigation.

7. Conclusion

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 helps a lot during the whole process. Most importantly, our framework can extract community structures in a appropriate level with the application of label supervision into the Fast Unfolding algorithm.

References

[1] M. K. Sparrow, "The application of network analysis to criminal intelligence: An assessment of the prospects," *Social networks*, vol. 13, no. 3, pp. 251–274, 1991.

[2] W. E. Baker and R. R. Faulkner, "The social organization of conspiracy: Illegal networks in the heavy electrical equipment industry," *American sociological review*, pp. 837–860, 1993.

[3] P. Klerks, "The network paradigm applied to criminal organizations: Theoretical nitpicking or a relevant doctrine for investigators? recent developments in the netherlands," *Connections*, vol. 24, no. 3, pp. 53–65, 2001.

[4] A. Silke, "The devil you know: Continuing problems with research on terrorism," *Terrorism and Political Violence*, vol. 13, no. 4, pp. 1–14, 2001.

[5] 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," *Studies in Conflict and Terrorism*, vol. 24, no. 1, pp. 3–24, 2001.

[6] J. Arquilla and D. Ronfeldt, *Networks and netwars: The future of terror, crime, and militancy*. Rand Corporation, 2001.

[7] V. E. Krebs, "Mapping networks of terrorist cells," *Connections*, vol. 24, no. 3, pp. 43–52, 2002.

[8] M. A. Porter, J.-P. Onnela, and P. J. Mucha, "Communities in networks," *Notices of the AMS*, vol. 56, no. 9, pp. 1082–1097, 2009.

[9] M. E. Newman, "Detecting community structure in networks," *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 38, no. 2, pp. 321–330, 2004.

[10] A. J. Alvarez, C. E. Sanz-Rodríguez, and J. L. Cabrera, "Weighting dissimilarities to detect communities in networks," *Phil. Trans. R. Soc. A*, vol. 373, no. 2056, p. 20150108, 2015.

[11] M. E. Newman, "Fast algorithm for detecting community structure in networks," *Physical review E*, vol. 69, no. 6, p. 066133, 2004.

[12] T. S. Evans, "Clique graphs and overlapping communities," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2010, no. 12, p. P12037, 2010.

[13] E. Ferrara, P. De Meo, S. Catanese, and G. Fiumara, "Detecting criminal organizations in mobile phone networks," *Expert Systems with Applications*, vol. 41, no. 13, pp. 5733–5750, 2014.

[14] H. Sarvari, E. Abozinadah, A. Mbaziira, and D. McCoy, "Constructing and analyzing criminal networks," in *Security and Privacy Workshops (SPW), 2014 IEEE*. IEEE, 2014, pp. 84–91.

[15] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, p. P10008, 2008.