A Prioritization Algorithm for Crime Busting based on Centrality Analysis

Yundong Gu[1,a], WentaoLi[2,b], Liwen Zhang[3], Mingke Shen[4], Binglei Xie[5]

[1]School of Mathematics and Physics, North China Electric Power University, Beijing, 102206, China.

[2,3,4]School of Energy Power and Mechanical Engineering, North China Electric Power University, Beijing, 102206, China;

[5]College of Information Science and Technology, Dalian Maritime University, Dalian, 116026, China

[a]guyund@126.com,[b]liwentao90@yahoo.com

Keywords: Crime busting; Social network; Centrality; Sorting

Abstract. Detecting conspirators, which often relates to organized crimes, represents a major problem for many investigation bureaus. A prioritization algorithm based on centrality analysis was introduced. The correlation between suspects was modeled as a social network, and the degree, betweeness and eigenvector centralities were utilized to quantify the suspicion degree of individual conspirators. Due to the analysis, conspirators and non-conspirators were able to be sorted into high-suspected, low-suspected, low-unsuspected and high-unsuspected sections based on their likelihood of involving the conspiracy. A detailed scenario is studied and the efficacy of the given method is verified at the end of this paper.

Introduction

Investigators like federal agents might focus on white-collar criminal activities like money laundering, drug dealings and so on. They usually need to suspect investigate and affirm someone whether participating in the conspiracy or not, which sometimes may consume a lot of time and money, so it is better to start with someone with high possibility. Therefore, a priority sequence sorting method based on the likelihood of involving the conspiracy will greatly contribute to the investigators to start their work from the people highly suspected. By so far, there are many solutions [5-7] available like node removal, specific search algorithm, improved sociological concepts, etc. This paper provides a prioritization algorithm on crime busting with great flexibility. The algorithm can response quickly and ensure the possibility of real-time analyses. The final sequence is sorted into their suspected orders and divided into four suspicion levels.

This paper is organized as follows: the second section describes the crime busting problem and the assumptions. The network model for crime busting problem is constructed in following section. Section four presents a crime busting algorithm. The running-time and flexibility of given algorithm is analyzed in the section five. And a scenario will verify the efficacy of the given method at the end of the paper.

The Problem and Assumptions

The main task of this paper is assigned to identify conspirators by analyzing short message communication network and dig out some useful information concealed in the network. It is already known that some people are surely involved in the crime and some are not. However, most people involving the communication still need to be asserted whether they are participating in the conspiracy or not. There might be all kinds of message between any two people. Some types of message could assert someone involving or not, and some may not even express any influence of the crime, like daily topics etc.

The ultimate goal of our study is to generate a people sequence that everyone is sorted by the suspicion. As the main task of this model is the analysis based on the whole communication network, it is regarded that the semantic analysis results have already been acquired.

Published by Atlantis Press, Paris, France.

Modeling the Crime Busting based on Social Network Analysis

A social network which is an important structure in sociology is made of a set of individuals and ties between them based on communication. Within the scope of graph theory and network analysis, there are four widely used types of measures of the centrality of a vertex within the graph that determine the relative importance of it, i.e. Degree Centrality, Betweenness Centrality, Closeness Centrality[8] and Eigenvector Centrality.

In this paper, we would use the vertices represent the suspicious person and the edge stands for the contact between each pair of people, like short messages, phone calls or e-mails.

In the model the communication among people ought to be classified into separated topics in advanced, for instance, topics involving the conspiracy, topics of daily affairs and the topics that can prove someone not involving in the conspiracy.

The betweeness centrality and degree centrality is adopted in this paper to quantify the priority in each topic. After the combination, the eigenvector centrality can optimize the sequence with the help of nodes already confirmed. Throughout this paper, the value distribution between -1 and +1 is used instead of 0 and +1, with which makes it easier to express the distribution between non-conspirator and conspirator. The more approach of -1 and +1 stands for non-conspiracy and conspiracy respectively. The more the value approaches 0, the less influence the person or topic it represented may possess.

Prioritization algorithm

The prioritization algorithm for crime busting is constructed of three main procedures. Firstly, the affect in each topic is considered separately. Then the second step sums their outputs up with respect to the topic priority. Finally, the combined result is optimized in order to minimum the usual chaos in middle of the sequence.

In the first procedure, the vertices are sorting in sub-graph in each topic. There are two general types of conspirators. One is core conspirators, who communicate frequently with each other about suspicious topics, so they must have close contact with known conspirators. And the other is co-conspirators, who transmit some kind of messages from non-conspirators to conspirators; they are disseminators who play an important role in communication between two nodes. The conspiracy decisions, the suspect topics usually, are decided by the conspirators. In other words, conspirators control the messages of the bad topics. According to the concepts of sociological centrality, Betweenness Centrality is a better means to sort the people in each topic communication control [3]. Meanwhile the same pair of people especially the core conspirator may send different messages of the same topic in several times. So degree centrality is also necessary, it is the number of different messages of the same topic minus one that strength the each topic result. Besides the degree centrality, the percentage of conspirators, non-conspirators and topic priority are also taken into consideration. So the node priority in each topic is

$$p(i)=C_B(i)+\frac{N-2}{2}\times(N_D-1)\times(N_{CON}-N_{NOT})\times p(T)\,. \tag{1}$$

where $C_B(i)$ stands for betweenness centrality, N, N_D, N_{CON}, N_{NOT} are the numbers of the people, different messages of the same topic respectively, conspirators and non-conspirators in each shortest path, $p(T)$ means the topic priority.

As the absolute betweenness centrality and the absolute degree centrality are not in the same order of magnitude. We multiply half of the number of people minus two referring to their definition. After normalizing it, the node priority in each topic is ready for the next procedure.

Secondly, the separated sequences are merger into one. Through the semantic analysis based on detailed messages, it is possible to quantify the likelihood of topics involving in the conspiracy by a fuzzy variable. Meanwhile, it is also able to calculate the edge in-degree and out-degree of each topic against the conspirators and non-conspirators in each topic graph, which is another expression of degree centrality in sociology. The more contacts with the conspirator already identified, the higher possibility the topic involving in the conspiracy may have. So the value (2) expresses the priority of each topic, where the EI_{con}& EI_{non} represent edge in-degree related with conspirator and non-conspirator, EO_{con}& EO_{non} means the out-degree types.

$$(EI_{con} - EI_{non}) - (EO_{con} - EO_{non}) \qquad (2)$$

Then distribute this value between -1 and +1 normally according to the values gained so far. Then, with the help of the polarized function which is a type of function makes the value more approaching to -1, 0 and 1. And 1 stands for the high possibility of topics involved in the conspiracy and -1 means the contrary. The approach of 0 leads to minimum the affect of regular topics. The section above shows that the value also participates in the results in each topic which strengths the results.

And as well there would be another essential step to make the sorting results of separated topics in the same order of magnitude, distributing them between -1 and +1 respectively again is a simple and effective method, and in that way the topics priority sequence has been deduced. It would also be possible to combine the semantic analyses results in our model. The reserved semantic interface can adjust combination of different topics. In fact it is optimization of topic priority sequence that affects the final combination.

Equipped with the topic priority, it is easy to merge the separated results in each topic by weighted summing. After that a first-hand priority sequence has been gained but it seems that there are some other kinds of work ought to be done.

Finally in the last procedure, the result is optimized by the eigenvector centrality. In the first-hand priority sequence, we regard the array from the first one in the list to the last bad conspirator whom we've already known as the black-area. At the same time, we regard the array from the first non-conspirator in the list to last person as the white-area. As the priority of good and bad has been strengthened, the black-area and white-area ought to have no intersection; in other words, it would be possible to eliminate the intersection through strengthening the effect of topics priority and numbers of conspirator and non-conspirator between the black and white areas. Under such condition, there would be a grey-area between the black-area and white-area. The priority sequence in the grey-area may not be quite clear or satisfying because of strongly concentrating on the strength of the two-side polarization between good and bad which resulting in the chaos of the mixture of transition sequence. And then it's the ShowTime of Eigenvector Centrality. The eigenvector centrality of a node is deduced by others. It is also feasible to assume the eigenvector with the same constant at the very beginning of convergence iteration algorithm or it had better use the already-known value in the black and white area. With the help of adjacent matrix, we can optimize the sequence in the grey-area affected by black and white area, when the sequence approaches to a fixed one or reach a acceptable constrain, the final result is gained and the iteration can terminate. For better understanding, the pseudocode of this procedure is presented.

```
CONVERGENCE-PROCEDURE
greystart = Find-the-First-in-Grey(grey_array)
greyend = Find-the-Last-in-Grey(grey_array)
One_of_People_Priority_Increment= Positive Infinity
while FindMax(People_Priority_Increment_in_each) > Constrain_Expected do
    for i = greystart to greyend do
        Initalize and reorder the People_Priority sequence
        for j = each person not in grey
            People_Priority_in_grey_(i) +=Adjacency_Matrix(j,i)×People_Priority_not_in_grey_(i)
    Normalizating-Array(Array_in_grey) between last in white and first in black's priorities
    Calculate the People_Priority_Increment_in_each
return People_Priority
```

Besides, the adjacent matrix we used here is obtained by calculating the connectedness in each topic's communication graph in 0-1 logic. And use the logical OR to multiply up the same cell in different topic's 0-1 matrices. Equipped all above, we can apply the convergence iteration algorithm to the grey-area array. After that a satisfied sequence will be gained.

Algorithm Analysis

In a programmer perspective, the separating and merging is quite efficient. The whole time consuming is $\Theta(V^3)$ and $O(V^2 \lg V + VE)$ for a sparse communication graph implemented by a Fibonacci heap as one of its components. The running-time of the convergence iteration procedure mainly depends on the adjacency matrix, however in usual cases, it can guarantee that it's between $\Omega(V^2)$ and $O(V^3)$. Since the likelihood is an empirical value, it's hard to estimate the regular convergence iteration's running-time. Even so it's also possible to quit the iteration when the sorting result does not change instead of the likelihood value which is a practical way on real mass problem solving [1].

In addition, such model is of great flexibility which can not only be applied on crime busting but also be adjusted to analyze on other realms. It is also practical to apply the model to an economic corporate networks, journal citation networks and technological innovation network with solely few slight adjustments. Moreover, by means of efficient semantic analysis, we have reasonable ground to believe that the models we develop are endowed with broad prospects.

Implementation

For the convenience of comprehension, a simple scenario result in 2012 ICM [9] and its intermediate data are presented to illustrate the model. For detailed description you can refer to the websites in the references.

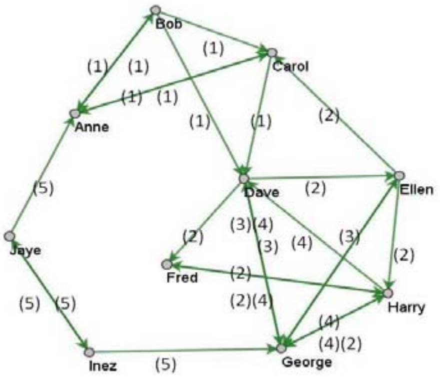

Figure 1 Graph of the communication network

The vertices priorities in each topic are the normalized results from Eq.1. After the merging step, the sub-sequence in grey is slightly in chaos. The iteration procedure optimizes it and generates a satisfying result finally. And the table is sorted into final result sequence.

Table 1 Final result and it intermediate data

Name	Normalized vertices priority in each topic					First-hand result	Final result	Sections
	Topic 1	Topic 2	Topic	Topic 4	Topic			
George	0	0	1	1	0	1.000657	1.000657	High-Suspect
Ellen	0	0.8	1	0	0	0.995266	0.995266	
Dave	0	1	1	1	0	0.99474	0.99474	
Inez	0	0	0	0	0.8	-0.8	0.89474	Low-Suspect
Bob	0.833333	0	0	0	0	-0.81168	0.404318	
Fred	0	0.8	0	0.333333	0	-0.00451	-0.41044	Low-Unsuspected
Harry	0	1	0	1	0	-0.00526	-0.41044	
Carol	0.833333	0.2	0	0	0	-0.81287	-0.9	
Jaye	0	0	0	0	1	-1	-1	High-Unsuspected
Anne	1	0	0	0	0.4	-1.37402	-1.37402	
Priority	-0.97402	-0.00592	1	0.000657	-1			

Published by Atlantis Press, Paris, France.

Table 2 Adjacency matrix used

1	1	1	1	0	0	0	0	0	0
1	1	1	1	0	0	0	0	0	0
1	1	1	1	0	0	0	0	0	0
0	0	1	1	1	1	1	1	0	0
0	0	1	1	1	1	1	1	0	0
0	0	0	0	0	1	1	1	0	0
0	0	0	1	1	1	1	1	0	0
0	0	0	1	0	1	1	1	0	0
1	0	0	0	0	0	1	0	1	1
1	0	0	0	0	0	1	0	1	1

Conclusion

In this paper, a prioritization algorithm on crime busting based on centrality analyses is constructed. The algorithm is quite fast, and can be implemented to analyze mass communication and give a real-time result.

Acknowledgment

This paper is supported by the National Natural Science Foundation of China (Grant No. 71171080) and the Fundamental Research Funds for the Central Universities (12MS84).

References

[1] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, Clifford Stein: Introduction to Algorithms, third edition, The MIT Press 2009.

[2] John Scott: Social Network Analysis a handbook, second edition, SAGE Publications Ltd, United Kingdom, 26 January 2000.

[3] Freeman, R White Douglas, A Kimball: Social Networks and the Structure Experiment in Romney Research Methods in Social Network Analysis. New Brunswick, New Jersey, Transaction Publishers. 1992.

[4] Rasmus Rosenqvist Petersen, Christopher J. Rhodes, Uffe Kock Wiil: Node Removal in Criminal Networks. 2011 European Intelligence and Security Informatics Conference.

[5] Jennifer J. Xu, Hsinchun Chen: Fighting organized crimes: using shortest-path algorithms to identify associations in criminal networks. Decision Support Systems. Vol.38(2004), p.473– 487.

[6] SONG Wenjun, LIU Hongxing, WANG Chongjun, et al: Core nodes detection based on frequent itemsets of graph. Journal of Frontiers of Computer Science and Technology. Vol.4(2010),p.82-88.

[7] Walter Didimo, Giuseppe Liotta, Fabrizio Montecchiani, Pietro Palladino: An Advanced Network Visualization System for Financial Crime Detection. IEEE Pacific Visualisation Symposium 2011 1 - 4 March, Hong Kong, China.

[8] Linton C: Freeman Centrality in Social Networks and Conceptual Clarification. Lehigh University 1978/79 Elsevier Sequoia S.A. Lausanne printed in Netherlands Swiss.

[9] 2012 ICM Problem: Modeling for Crime Busting. http://www.comap.com/undergraduate/contests/mcm/contests/2012/problems/.

[10] Freeman, L. C: (1977) A set of measures of centrality based on betweenness. Sociometry 40, 35-41.