Network research based on the fuzzy comprehensive evaluation model of natural language

Zhi-hong Ma[1] ， Xun-song He[2]，Hao-xuan Ding[3]

1 Basic Science Department，Tianjin Agricultural University
Tianjin 300384 China

2 Department of Computer Science and Information Engineering，Tianjin Agricultural University
Tianjin, 300384 , China

3 Department of Computer Science and Information Engineering，Tianjin Agricultural University
Tianjin, 300384 , China

Abstract

The size of the possibility of determining the criminal conspiracy helps to survey, monitor or question the most likely suspects. However , we can make some unclear boundary and factors that are not easy to quantitative quantified by using the fuzzy comprehensive evaluation of principle. In this paper, the quantification of the theme of the dialogue of the network crime gang draw a priority list of a criminal conspiracy. Compared with the semantics of message transmission analysis and text analysis, the fuzzy comprehensive evaluation of principle not only makes the theme for the conspiracy more authentic intuitively and improves the efficiency of the infiltration of the core of the criminal gang's conspiracy.

Keywords: *Criminal conspiracy , Fuzzy, comprehensive evaluation , Natural language*

1. Introduction

The current case has 83 nodes, 400 links (some involving more than 1 topic), over 21,000 words of message traffic, 5 topics (3 have been deemed to be suspicious), 7 known conspirators, and 8 known non-conspirators. We want to identify other members of conspirators and their leaders before arrest them. We first figure out other unknown conspirators and then find the leader by using relationships between conspirators.

For identifying other unknown conspirators, we take all topics into consideration. To one topic, it is ambiguous and uncertain for whether it is a conspiracy or not. However, we can calculate the conspiracy probability of each topic through known conspirators' message traffic based on principle of fuzzy mathematics. Then we get conspirator probability of each member by topics of each one discussed. From method of crime and modus object, we find out the keywords connected with conspiracy. Based on text analysis, we calculate weight of each keyword. The node messages contain more topics connected with conspiracy, the node more probable be conspirator. We put the results of two methods together and compare them. Finally we pick out the unknown conspirators.

For the determination of the leaders, we will determine the accomplice out a separate analysis, first construct a network diagram of these co-conspirators from the figure to identify the most wide coverage or degree of the largest point, the point is the leaders.

2. Assumptions

The topics talked between conspirators are mainly to conspiracy.

The key words we find are all related to conspiracy and conspirator.

The crime form of conspirators is fit in conditions.

Conspirators are all discussed conspiracy in statistical information.

3. Symbols And Significance

x_{i1} The times that conspirators send topic of i

x_{i2} The times that conspirators receive topic of i

t_i The appearance times of topic i

x_i He probability of conspiracy for topic i

$y_{i1,j}$ The times of sending topic i from member j

$y_{i2,j}$ The times of receiving topic i from member j

p_i The probability of conspirators for member i

p_j The weights of key words j

4. To EZ Case

The way of identifying people in the office complex who are the most likely conspirators are based on principle of fuzzy mathematics[1].. We adopt the method of combining qualitative and quantitative. First of all, we analyze simple EZ case which had only 10 people (nodes), 27 links (messages), 5 topics, 1 suspicious/conspiracy topic, 2 known conspirators, and 2 known non-conspirators, as Figure1 shown.

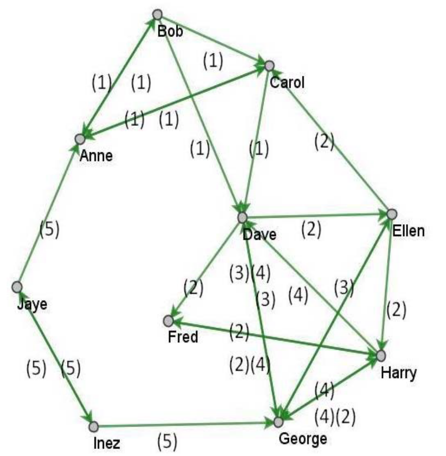

Figure 1 Network of Messages from EZ Case

We number 10 people. The result is shown in Table1

Table 1 Number of 10 Numbers

Name	Anne	Bob	Carol	Dave	Ellen
Number	1	2	3	4	5
Name	Fred	George	Harry	Inez	Jaye
Number	6	7	8	9	10

4.1 Basic Model

As we all known, the more frequent the topic be said by conspirator know， the more probable it is conspiracy. Because Dave and George are known conspirators (NO.4 and NO.7), we can calculate the probability of a topic for whether it is a conspiracy according to the emerged probability of NO.4 and NO.7 in 5 topics. We define the times that conspirators send topic of i as x_{i1} and receive topic as x_{i2} .The number of occurrences for topic i is t_i . Then the probability of conspiracy for topic i we defined is x_i. The relation between x_{i1} , x_{i2} , t_i and x_i is shown as follow:

$$x_i = \frac{x_{i1} + x_{i2}}{t_i} \qquad i = 1,2,3,4,5$$

So, the topic for conspiracy of probability matrix is:

$$X = \left[x_1,\ x_2,\ x_3,\ x_4,\ x_5 \right]$$

We use y_{i1j} expressing the times of sending topic i from member j and y_{i2j} expressing the times of receiving topic i from member j. So the times for every member discussion of each topic y_{ij} is the sum of y_{i1j} and y_{i2j}.

So the matrix for times of discussion is:

$$Y = \begin{bmatrix} y_{11} & \cdots & y_{1j} \\ \vdots & \ddots & \vdots \\ y_{i1} & \cdots & y_{ij} \end{bmatrix} \qquad i = 1,2...,5, j = 1,2,...,10$$

According to the matrix Y above, we can gain the total times of each member sending and receiving topic:

$$Z = \left[\sum_{i=1}^{5} y_{i1}, \sum_{i=2}^{5} y_{i2}, \cdots\cdots, \sum_{i=5}^{5} y_{ij} \right]$$

If we let p_j represent as the probability of conspirators for member j then we can get: $p_j = \dfrac{\sum_{i=1}^{5} y_{ij} \times x_i}{\sum_{i=1}^{5} y_{ij}}$

$$\mathrm{P} = \left[p_1, p_2, p_3, \ldots\ldots, p_j \right]$$

Normalization processing:

$$P' = \frac{p_j - \min(P)}{\max(P) - \min(P)}$$

The result is shown in Table2

Table 2 Probability of 10 Members on Basic Model

Number	1	2	3	4	5
Probability	0.205	0.220	0.233	0.750	0.703
Number	6	7	8	9	10
Probability	0.416	1	0.455	0	0

From Table2, we can get a conclusion that member 4, 5, 7 are conspirators, member 1, 2, 3, 9, 10 are non-conspirators and member 6, 8 are unsure. However, it is not fit the fact that member 2, 4, 5, 7, 9 are conspirators, member 1, 3, 10 are non- conspirators and member 6, 8 are unsure.

4.2 Improved Method

The weakness which we calculated the probability of conspiracy in topic before is that we considered sending messages together with receiving messages. In fact, the effect of two aspects is different. So we should consider them separately. We assume α as the degree of effect on conspiracy in topic when conspirators send messages and β as the degree of effect on conspiracy in topic when conspirators receive messages. Obviously, we can gain $\alpha+\beta=1$ .The conspiracy probability of topic i changes into:

$$x_i = \frac{\alpha x_{i1} + \beta x_{i2}}{t_i} \qquad i = 1,2,3,4,5$$

To estimate parameter α, β we regulate the value of them by calculating P' and comparing P' with truth. Try many times. We conclude that it is appropriate when $\alpha = 0.2$ and $\beta = 0.8$. So we conclude:

$$x_i = \frac{0.2x_{i1} + 0.8x_{i2}}{t_i}$$

The result is shown in Table3

Table 3 Probability of 10 Members on Improved Model

Number	1	2	3	4	5
Probability	0.101	0.112	0.090	0.771	0.538
Number	6	7	8	9	10
Probability	0	1	0.424	0.056	0.056

The actual results of Bob, Dave, Ellen, George, Inez is an accomplice, Anne, Carol, Jaye is not an accomplice, comparison is not particularly close to the results., This is because the data is not enough.

5. To The Current Case

The current case has 83 nodes, 400 links (some involving more than 1 topic), over 21,000 words of message traffic,15 topics (3 have been deemed to be suspicious), 7 known conspirators, and 8 known non-conspirators

5.1 Using Improved Model

According to our improved model, we calculate the probability of each member being conspirator. The outcome is sorted in ascending order and shown in Table4

Table 4 Probability of 83 Members on Improved Model

N	P	N	P	N	P	N	P	N	P
53	0.000	41	0.33	50	0.409	23	0.481	43	0.577
57	0.000	25	0.331	62	0.412	12	0.485	37	0.579
59	0.000	82	0.331	64	0.412	39	0.487	13	0.621
61	0.093	24	0.341	42	0.412	79	0.49	16	0.646
77	0.228	55	0.346	5	0.419	20	0.491	22	0.647
80	0.228	36	0.355	15	0.422	3	0.503	49	0.655
68	0.258	10	0.358	0	0.425	29	0.517	9	0.679
58	0.259	1	0.369	35	0.425	8	0.522	7	0.685
63	0.259	6	0.371	32	0.448	31	0.522	47	0.697
74	0.289	28	0.395	2	0.453	65	0.524	21	0.706
17	0.293	66	0.398	71	0.455	19	0.525	67	0.779
70	0.301	73	0.398	34	0.461	40	0.533	54	0.823
14	0.308	11	0.399	52	0.461	78	0.537	81	0.844
30	0.317	48	0.399	60	0.468	18	0.557	51	0.980
76	0.318	4	0.401	45	0.474	33	0.561	56	1.000
26	0.325	38	0.401	75	0.477	44	0.568		
69	0.326	72	0.407	46	0.478	27	0.576		

(N: on behalf of serial number. P: on behalf of conspiracy probability. The shaded means known conspirators)

From Table4 we know that the conspiracy probability of known conspirators is bigger than others. That is consistent with fact.

5.2 The Semantic Network Analysis Model

In the previous method, we stared from known conspirators and figured out the probability of each topic being conspiracy. And then we in turn calculated the probability of who may be conspirator. The next we reconsider from aspects of words on message traffic. If someone's messages are mostly connected to the conspiracy, he is more likely to be a conspirator. This idea comes from the method of text analysis.

5.2.1 Background

With the development of computer network technology, the exchange between people becomes more and more convenient. The semantic analysis and text analysis become increasingly important and difficult. Recently text analysis has focused on text representation model selection and the selection of feature selection algorithm. Our model is based on the semantic network analysis.

5.2.2 Analysis Model

By known conditions, a conspiracy is taking place to embezzle funds from the company and use internet fraud to steal funds from credit cards of people who do business with the company. From the details about the topics in file Topics.xls, we find key words from suspicious message topics. We first find three key words: Spanish, Paige and Compute. If someone has more keywords in his message, he has more conspiracy probability.

We use formula of TFIDF (term frequency–inverse document frequency) to figure out weights of keywords that is based on principle of semantic analysis for solving weight [2]-[3]. The formula:

$$W(i,j) = LW(i,j) * GWT(i)$$

$LW(i,j) = tf_{ij}$ as local weight, $GWT(i) = df_i / N$ as global weight, tf_{ij} as frequency of key word i appearing in message j, df_i as amounts of message appearing key word i and N as total number of message.

Because the more frequent key words appearing the more probable being conspirator, we gain the conspirator probability p_k of member k:

$$p_k = \frac{\sum W(i,j) * t_{kj}}{\sum t_{kj}}$$

t_{kj} as the sending or receiving times for topic j. Here we have some innovation.

The result is not ideal through this way, for example Jean and Yao who are known conspirators but out of conspirators list in our result. Analysis again to our model, we find that the reason is our key words is too less. So we add up another key word "finance" which is from Topic 1 (one of condition changes). Try again as before, the result is shown in Table5

Table 5 Probability of 83 Members on Semantic Network Analysis Model

N	P	N	P	N	P	N	P	N	P
53	0.000	26	0.338	5	0.428	45	0.499	27	0.596
57	0.000	41	0.339	50	0.431	3	0.510	44	0.597
59	0.000	62	0.340	15	0.441	52	0.516	13	0.647
61	0.093	64	0.340	42	0.443	20	0.52	16	0.651
77	0.228	82	0.340	71	0.446	39	0.522	49	0.66
80	0.228	69	0.341	0	0.451	78	0.537	22	0.671
68	0.247	24	0.356	35	0.458	8	0.543	21	0.685
58	0.259	10	0.360	32	0.465	19	0.545	7	0.696
63	0.259	36	0.378	60	0.483	65	0.547	9	0.702
74	0.289	1	0.392	34	0.483	12	0.548	47	0.715
17	0.293	28	0.395	46	0.487	29	0.549	67	0.805
14	0.303	6	0.397	79	0.49	31	0.552	54	0.808
55	0.310	72	0.407	23	0.496	40	0.566	81	0.844
70	0.314	38	0.414	66	0.497	18	0.576	51	0.980
76	0.318	11	0.418	73	0.497	43	0.577	56	1.000
30	0.321	48	0.425	2	0.498	33	0.591		
25	0.324	4	0.428	75	0.499	37	0.593		

(N: on behalf of serial number. P: on behalf of conspiracy probability. The shaded means known conspirators)

If $P_K \geq \min\{P_7, P_{18}, P_{21}, P_{37}, P_{43}, P_{49}, P_{54}, P_{67}\}$,then k is an accomplice.

From Table5, we conclude that conspirators are Elsie, Malcolm, Marion, Jerome, Jean, Alex, Eric, Marcia, Elsie, Paul, Christina, Harvey, Dayi, Ulf, Cha, Yao and Seeni. For Jean, Alex, Elsie, Paul, Ulf, Yao and Harvey are known conspirators.

From the Names.xls, we find some members have same name, for example NO.16 and NO.34 have the same name

"Jerome", NO.4 and NO.32 are all named "Gretchen". To solving the problem, we express all names in other way such as Jerome16, Jerome34, Delores10, Gretchen4 and Gretchen32. Their probability respectively: 0.646, 0.461, 0.358, 0.401, 0.448. Since Jerome16 is highest so we can say Jerome16 is most likely to be conspirator.

5.3 Explore The Model on Conditions Change

The conditions change: Topic 1 is also connected to the conspiracy and that Chris is one of the conspirators. The same method we used as former, the result is shown in Table6

Table 6 Probability of 83 Members on Conditions Change

N	P	N	P	N	P	N	P	N	P
53	0.000	82	0.286	64	0.369	34	0.436	16	0.549
57	0.000	25	0.305	15	0.374	0	0.437	43	0.557
59	0.000	14	0.308	10	0.374	50	0.439	13	0.561
61	0.053	69	0.316	24	0.382	12	0.45	49	0.561
55	0.210	76	0.316	71	0.386	48	0.454	27	0.579
68	0.210	26	0.325	75	0.395	52	0.456	7	0.605
58	0.211	42	0.329	38	0.399	40	0.489	9	0.605
63	0.211	70	0.333	19	0.400	60	0.491	22	0.618
17	0.228	45	0.337	2	0.405	3	0.508	21	0.632
36	0.253	28	0.342	20	0.408	33	0.509	47	0.662
30	0.257	72	0.342	6	0.412	31	0.511	54	0.716
41	0.263	35	0.353	46	0.412	44	0.518	67	0.750
66	0.263	4	0.369	32	0.415	65	0.518	81	0.768
73	0.263	5	0.369	39	0.421	78	0.526	56	0.790
74	0.263	11	0.369	79	0.421	8	0.542	51	1.000
77	0.263	23	0.369	1	0.429	18	0.542		
80	0.263	62	0.369	29	0.434	37	0.547		

(N: on behalf of serial number. P: on behalf of conspiracy probability. The shaded means known conspirators)

Compared to the probability in different conditions, we can see they have little difference. The result is shown in Figure2.

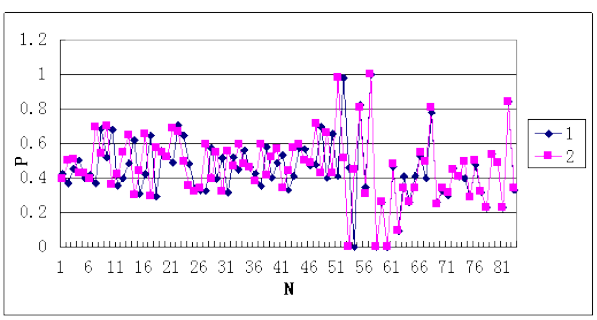

Figure 2 The Probability in Different Conditions

6. The Network Sociology Analysis

6.1 Background

1930, W. L. Warner pointed out that the social structure of a modern community type constitute by many sub-groups, such as family, church and classes.[4]
1972, Bruce Kapferer successfully predicted a strike. That greatly improved the level of theory and practice of network sociology. [5]
Network sociology may be subordinated to the future independent discipline network science (Weizhi Deng 2001) [6]
We propose a network sociology model to nominate the conspiracy leaders that is based on network analysis. The model is run in UCINET software which is one of most popular simple software of social network analysis at the present time.
Known by the common sense, it is very useful to combat the conspiracy leaders for fighting against criminal gangs. The leading figure is the hub of the network for

information exchange based on the social network model. If someone's degree centrality and betweenness centrality is rank in front of sequence we think he is the conspiracy leader. So we calculate the parameter of degree centrality and betweenness centrality to find the conspiracy leaders.

6.2 The Relationships Matrix and Network Diagram

We use 1 and 0 to describe the two whether linked or not and build the relationships matrix for conspirators. Then we draw the relationship network diagram by using UCINET software. The diagram is shown in Figure3

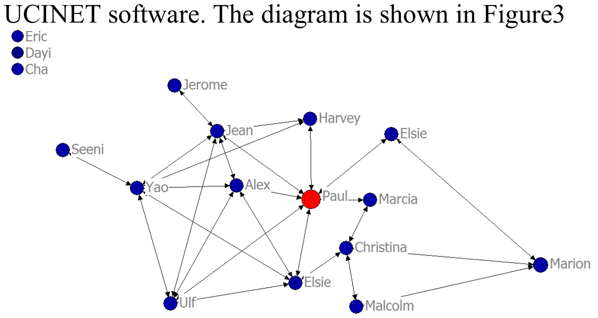

Figure 3 Network Diagram of Conspirators

6.3 The Degree Centrality and The Betweenness Centrality

For a network node, degree is the most basic connectivity metric parameters. Degree is expressed by the node and other nodes connection number d and divided into in-degree and out-degree. 1974, Nieminen put forward the calculation formula of degree centrality:[7]

$$C_D = \sum_{i=1}^{n} \alpha(p_i, p_k)$$

Table 7 The Degree Centrality of Conspirators

Number	Name	Degree	Number	Name	Degree
10	Paul	7.000	9	Elsie	2.000
5	Jean	6.000	8	Marcia	2.000
16	Yao	6.000	2	Malcolm	2.000
1	Elsie	5.000	17	Seeni	1.000
6	Alex	5.000	4	Jerome	1.000
14	Ulf	5.000	7	Eric	0.000
11	Christina	4.000	15	Cha	0.000
3	Marion	3.000	13	Dayi	0.000
12	Harvey	3.000			

We can see that Paul, Jean and Yao are the top three. Paul is the most probable conspirator leader whose degree centrality is 7.00.

Table 8 The Betweenness Centrality of Conspirators

Number	Name	Betweenness	Number	Name	Betweenness
10	Paul	25.050	8	Marcia	2.150
1	Elsie	17.400	12	Harvey	1.133
16	Yao	14.983	7	Eric	0.000
11	Christina	14.917	13	Dayi	0.000
5	Jean	13.800	2	Malcolm	0.000
9	Elsie	5.917	15	Cha	0.000
3	Marion	2.917	4	Jerome	0.000
6	Alex	2.367	17	Seeni	0.000
14	Ulf	2.367			

From the table above, Paul is also in the top and his betweenness centrality is 20.05.
Because Paul's degree centrality and the betweenness centrality is the highest from others.Paul is the most probable to be conspirator leader. In summary, Paul is the conspirator leader.

7. Model Evaluation

7.1 Model Promotion

The former models all took one aspect of factor into consideration. Basic model and improved model only consider "Conspirators". The semantic network analysis model and the network sociology model only take "Key words" into account. To improving our model, we consider four factors "Conspirators", "Key words", "Non-conspirators", "Suspicious topics". We make a comprehensive evaluation with four factors. The improvement ideas graph is shown in Figure4

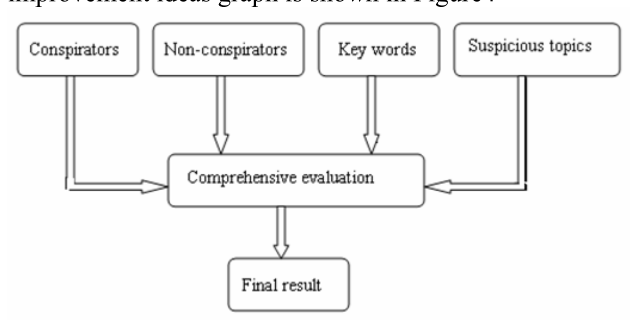

Figure 4 The Improvement Ideas Graph

The model improved could apply to the assessment of product quality, evaluation of the quality of hotel services, and also could apply to cluster analysis.

7.2 Strength

Basic model and improved model: we adopt principle of fuzzy mathematics. We start from known conspirators. We deal with the data and calculate the weight of topic contain conspirators node's boundary x_i,we regard this probability as conspiracy topic. Then we calculate in turn the conspirator probability of nodes. So, we quantify each node and it is easy to sorting, comparison and screening conspirators.

The semantic network analysis mode: we take conspiracy topic into account and pick out key word connected with conspiracy. We calculate weight with semantic analysis. Taking consideration from key words is close to our purpose and avoiding leaving out some conspirator when people discuss too much insignificant topic.

7.3 Weakness

Basic model and improved model: It is a bit one-sided to estimate the probability of conspiracy by frequency of conspiracy occurrence. If we comprehensive evaluate three factors "Conspirators", "Non-conspirators", and "Suspicious topics", the result will be better.

The semantic network analysis mode: The keywords selection and number determined directly determine the accuracy of the results. If the first step of selection keywords are wrong it could result in incalculable error on the overall.

Reference

[1] Guangwu Meng. The Basic Theory and Its Application of Fuzzy Mathematics(1)——Overview of The Emergence and Development of Fuzzy Math. Liaocheng Teachers University (Natural Science Edition), 1998, (6)

[2] Yunfeng Liu, Huan Qi, Xiang'en Hu and Zhiqiang Cai. Weight Aalculated By Latent Semantic Analysis Improvements. Journal of Chinese Information Processing, Vol. 19 No.6: 64-69

[3] Lei Chen, Bidan Dong and Yanping Zhao. Hidden Within The Social Network Analysis in The Semantic - based Enterprise Relationship Detection. Computer and Digital Engineering, 2009, (9): 58-63

[4] John Scott. Social Network Analysis [M]. Chong Qing: Chongqing University Press, 2007

[5] Martin kilduff, Wenpin Tsai. Social Networks and Organizations [M]. Sage Publications of London, 2003

[6] Weizhi Deng, Jianwei Fan and Weisheng Shi. On The Establishment of The Network of Community Schools [J]. Jianghai Tribune Phase IV, 2001

[7] Freeman, L.C. Centrality in Social Networks I: Conceptual Clarification [J]. Social Networks, 1979, (1):215-239.

[8] Zeng Xianzhao. Network Science :Volume II. Bei Jing: Military Science Press. 112-162(2008).

First Author: Zhi-hong Ma (1975-), male, born in Ningxia, Associate Professor of the Tianjin agricultural University, master's degree, mainly engaged in the teaching of mathematics and applied mathematics research.

Second Author Xun-song He is a student of Tianjin Agricultural University .

Third Author Hao-xuan Ding is a student of Tianjin Agricultural University .