Team Control Number

For office use only
T1 _____
T2 _____
T3 _____
T4 _____

13215

Problem Chosen

C

For office use only
F1 _____
F2 _____
F3 _____
F4 _____

2012 Mathematical Contest in Modeling (MCM) Summary Sheet
(Attach a copy of this page to each copy of your solution paper.)

Type a summary of your results on this page. Do not include the name of your school, advisor, or team members on this page.

Message Network Modeling for Crime Busting

Abstract

A particularly popular and challenging problem in crime analysis is to identify the conspirators through analysis of message networks. In this paper, using the data of message traffic, we model to prioritize the likelihood of one's being conspirator, and nominate the probable conspiracy leaders.

We note a fact that any conspirator has at least one message communication with other conspirators, and assume that sending or receiving a message has the same effect, and then develop Model 1, 2 and 3 to make a priority list respectively and Model 4 to nominate the conspiracy leader.

In Model 1, we take the amount of one's suspicious messages and one's all messages with known conspirators into account, and define a simple composite index to measure the likelihood of one's being conspirator.

Then, considering probability relevance of all nodes, we develop Model 2 based on *Law of Total Probability*. In this model, probability of one's being conspirator is the weight sum of probabilities of others directly linking to it. And we develop Algorithm 1 to calculate probabilities of all the network nodes as direct calculation is infeasible.

Besides, in order to better quantify one's relationship to the known conspirators, we develop Model 3, which brings in the concept "shortest path" of graph theory to create an indicator evaluating the likelihood of one's being conspirator which can be calculated through Algorithm 2.

As a result, we compare three priority lists and conclude that the overall rankings are similar but quite changes appear in some nodes. Additionally, when altering the given information, we find that the priority list just changes slightly except for a few nodes, so that we validate the models' stability.

Afterwards, by using Freeman's centrality method, we develop Model 4 to nominate three most probable leaders: Paul, Elsie, Dolores (senior manager).

What's more, we make some remarks about the models and discuss what could be done to enhance them in the future work. In addition, we further explain Investigation EZ through text and semantic network analysis, so to illustrate the models' capacity of applying to more complicated cases. Finally, we briefly state the application of our models in other disciplines.

Introduction

ICM is investigating a conspiracy whose members all work for the same noted company which majors in developing and marketing computer software for banks and credit card companies. Conspirators commit crimes by embezzling funds from the company and using internet fraud to steal funds from credit cards. It is a kind of commercial fraud. Fraud is a human endeavor, involving deception, purposeful intent, intensity of desire, risk of apprehension, violation of trust, rationalization, etc. Psychological factors influence the behaviors of fraud perpetrators (Sridhar Ramamoorti, 2008).

ICM provides us the following information that they have mastered

●All 83 office workers' names;

●15 short descriptions of the topics (Topic 7, 11, and 13 have been deemed to be suspicious);

●400 links of the nodes that transmit messages and the topic code numbers;

●7 known conspirators: Jean, Alex, Elsie, Paul, Ulf, Yao, and Harvey;

●8 known non-conspirators: Darlene, Tran, Jia, Ellin, Gard, Chris, Paige and Este;

●Jerome, Delores, and Gretchen are the senior managers of the company.

For crime busting, we develop models to

●Identify all conspirators as accurately as possible, make a priority list that shows the likelihood of one's being conspirator, so that erroneous judgments or miss-judgments won't happen easily;

● Nominate the conspiracy leader.

Declaration of the given data

●"Topics.xls" contains only 15 topics, but "topic 18" appears in line 215 of "Messages.xls". To fix this error, we decide to neglect this invalid data and delete it.

●In page 5, line 2 of "*2012_ICM_Problem*.pdf", it says that "Elsie" is one of the known conspirators. However we find two "Elsie" with node number "7" and "37". Throughout some basic statistics about the message traffic containing suspicious topics, it appears that "7 Elsie" is more likely to be a known conspirator rather than "37 Elsie". Therefore, we assume that "Elsie" in "*2012_ICM_Problem*.pdf" indicates "Elsie" with node number 7 in "names.xls".

●As the problem paper point out, "Delores" is a senior manager. But "Delores" can't be found in "names.xls" while "Dolores" is found. So we consider it as misspelling and replace "Delores" with "Dolores".

●"Gretchen" is also one of the senior managers. But two "Gretchen" are found in "names.xls" with different node number "4" and "32". In consideration of this redundancy, we determine to pick out node 32 for "Gretchen" indicated in the problem paper artificially. In addition, our basic statistics also shows that "32 Gretchen" has more message exchanges than "4 Gretchen", which may imply that "32 Gretchen" is more probably the senior manager than "4 Gretchen" due to managers often contact others more than ordinary office workers.

Problem analysis and assumption

Commercial fraud is committed by those intelligent people who are confident with their professional skills. Meanwhile, this kind of crime couldn't involve only one person, but always need cooperation of a whole group. Thus, communication with other conspirators would be inevitable. However, they obviously know that they are linked together and if one person discloses their secrets, none of them can get off. So they are conscious when they communicate with their colleagues who aren't their companions, especially when they talk about sensitive issues. And the higher intellectual level of perpetrators with rich society experience, the more conscious they are (Zhigang Lin,2010). And ICM can figure out suspicious topic which stands a good chance of being related to the conspiracy by some content analysis method. On the one hand, although Conspirators would try to avoid involving suspicious topics in their messages, they have to convey this kind of information sometimes due to the business or other reason. On the other hand, trust and close relationship play an important role in a conspiracy group, so normal messages exchanges can also reflect the conspiracy relationship.

Based on psychology analysis above, we can state that all conspirators have at least one message communication with other conspirators, whether suspicious or unsuspicious message.

In addition, we make the assumption that sending and receiving messages have same effect when we evaluate the likelihood of one's being conspirator;

Models

Model 1

Establishment of model

According to the analysis of the problem, the likelihood of one's being conspirator is related to various factors, such as what topics are contained in the worker messages, how many messages and suspicious messages are the worker related with, who did the worker contact with, etc. To evaluate the likelihood of one's being conspirators, we use the following equation which combines two quantity indexes:

$$p_i = \frac{1}{2}\left(\frac{n_{1i}}{\max\limits_i\{n_{1i}\}} + \frac{n_{2i}}{\max\limits_i\{n_{2i}\}} \right), i = 0,1,2,...,82 \qquad (1)$$

Where n_{1i} is the suspicious message number that office worker i sent or received and n_{2i} is message number that office worker i sent to or received by known conspirators.

In order to get each value of n_{1i} and n_{2i}, we make data statistics and draw Figure 1:

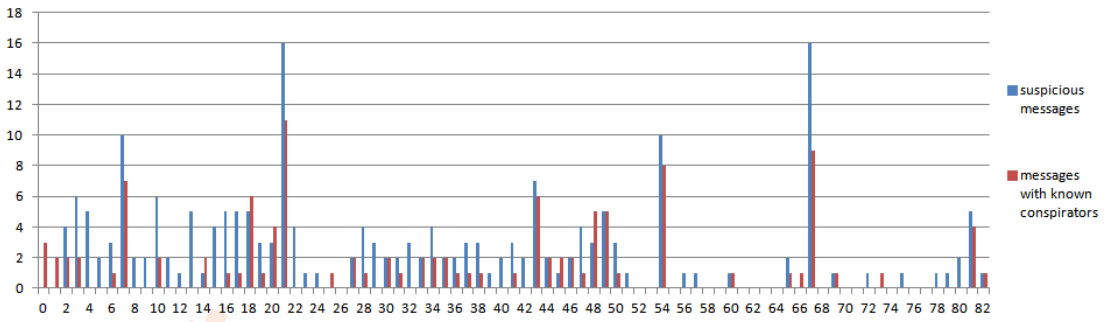


Figure. 1

Result and analysis

Figure 1 shows all the values of n_{1i} and n_{2i}. Using equation (1) we have put forward, we can easily calculate all the values of p_i and make a priority list as Table 1 (note that p_i is not a probability but a metric to evaluate the likelihood, though it value is between 0 and 1)

Table 1

No	node	p	No	node	p	No	node	p	No	node	p
1	**21**	1	21	30	0.1534	43	1	0.0909	57	72	0.0313
2	**67**	0.9091	21	33	0.1534	44	60	0.0767	57	75	0.0313
3	**54**	0.6761	21	35	0.1534	44	69	0.0767	57	*78*	0.0313
4	**7**	0.6307	21	44	0.1534	44	82	0.0767	57	79	0.0313
5	**43**	0.4915	21	46	0.1534	47	5	0.0625	68	26	0
6	**18**	0.429	27	6	0.1392	47	8	0.0625	68	52	0
7	**49**	0.3835	27	19	0.1392	47	9	0.0625	68	53	0
8	81	0.3381	27	37	0.1392	47	11	0.0625	68	55	0
9	*48*	0.321	27	38	0.1392	47	40	0.0625	68	58	0
10	3	0.2784	27	41	0.1392	47	42	0.0625	68	59	0
10	10	0.2784	27	50	0.1392	47	80	0.0625	68	61	0
12	20	0.2756	33	*0*	0.1364	54	25	0.0455	68	62	0
13	*2*	0.2159	34	15	0.125	54	66	0.0455	68	63	0
13	34	0.2159	34	22	0.125	54	73	0.0455	68	*64*	0
15	16	0.2017	36	14	0.1222	57	12	0.0313	68	*68*	0
15	17	0.2017	36	45	0.1222	57	23	0.0313	68	70	0
17	28	0.1705	38	31	0.108	57	24	0.0313	68	71	0
17	47	0.1705	38	36	0.108	57	39	0.0313	68	*74*	0
19	4	0.1563	38	*65*	0.108	57	51	0.0313	68	76	0
19	13	0.1563	41	29	0.0938	57	56	0.0313	68	77	0
21	27	0.1534	41	32	0.0938	57	57	0.0313			

As shown in Table 1, all the known conspirators (heavy tape and red mark) are ranked in the very front of the list, which indicates the model is effective to some extent that it can recognize some workers who is most likely to be conspirators. However, some non-conspirators (green mark and Italic type) are also up at the front,

like node 48 and node 2, which shows that the model has a certain limitation and some wrong recognition.

Model 2

In order to establish an improved model, we make one more assumptions

Except for the known conspirators and non-conspirators, one's probability of being conspirator is relate to those who have direct message contact with him/her. And the probability is both affected by the probability of his/her linking persons and the topic nature of the linking messages.

Introduction of *Law of Total Probability*

In probability theory, the law of total probability or the formula of total probability is a fundamental regulation relating marginal probabilities. It can be described as follows:

if $\{B_n : n = 1, 2, 3, ...\}$ is a finite or countably infinite partition of a sample space and each event B_n in it is measurable, then for any event A of the same probability space:

$$\mathrm{P}(A) = \sum_n \mathrm{P}(A \mid B_n)\,\mathrm{P}(B_n) \tag{2}$$

Establishment of model

According to the material we get hold of , since Topic 7, 11, and 13 have been deemed to be suspicious ,we name S={7,11,13} the suspicious topic set and U={1,2,3,4,5,6,8,9,10,12,14,15} the unsuspicious topic set. In addition, we categorize all 83 office workers into three groups: conspirators, non-conspirators and unknown ones. p_a , p_b and P_j ($j = 0,1,...,83, except\ 15\ known\ persons$) indicate the probability of three kind of office workers commit crime. We have p_a =1, p_b =0 and P_j equaled different unknown numbers which between 0 and 1. The greater probability the unknown one is conspirator, the greater P_j is. A person is much more suspicious if he/she sends or receives suspicious messages more frequently. We can use w_{ji} to represent the suspicious extent and it can be calculated by the following equations:

$$w_{ji} = n_a \times a + n_b \times b,\ i = 1, 2, \cdots \tag{3}$$

Where n_a (n_b)is the number of suspicious(unsuspicious) messages a unknown one sends or receives, a is the weight of elements in the set of S, and b is the weight of elements in the set of U.

Next, we will explain how "probability" works out in the messages network with Figure 2.

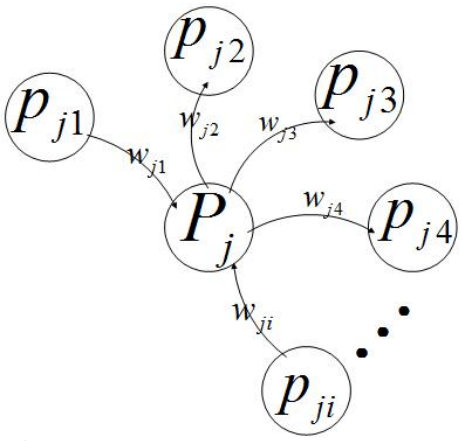


Figure. 2

We consider that the whole network can be separating into a lot of small network like above. Bringing *Law of Total Probability* in our model, we treat the center node (not including nodes in known conspirators or non-conspirators group) as A in description of *Law of Total Probability*, and other nodes directly connected to it as $B_n(n=1,2,3,...)$. So probability of center node is $P_j = \mathrm{P}(A)$, probabilities of other connecting nodes are $p_{ji} = \mathrm{P}(B_i)$, and $\mathrm{P}(A \mid B_n) = \dfrac{w_{ji}}{\sum_i w_{ji}}$.

Based on illustrations we present, we calculate P in the following way:

$$P_j = \sum_i \left(p_{ji} \times \frac{w_{ji}}{\sum_i w_{ji}} \right) = \frac{\sum_i p_{ji} \times w_{ji}}{\sum_i w_{ji}} \tag{4}$$

However, all the probabilities of nodes in the unknown group are uncertain. So it is impossible to use the equation above to calculate all the probabilities directly. As a solution, we develop the following algorithm.

Algorithm 1

All 400 links can constitute a complex relative network, and each office worker can form a simply network centered on himself/herself. Considering the structure of network, we imitated the neural network algorithm but use iterative method to complete the whole relative network:

Step 1: Set iteration times as T, and initialize $P_j^{(0)} = 0(j = 1,2,...,68)$, $t = 1$;

Step 2: Refreshing the network P_j

Loop j from 1 to 68, then utilize equation (4) to calculate each $P_j^{(t)}$;

Step 3: Calculate the quadratic sum of probability errors between last time and present time;

$$e(t) = \sum_{j=1}^{68} [P_j^{(t)} - P_j^{(t-1)}]^2 \tag{5}$$

Step 4: Let $t = t + 1$, if $t > T$, program ends up, else returns to Step 2.

With t increasing, $e(t)$ shows a downward trend. When the value of $e(t)$ tends to become stable, or less equals than a small constant, we can consider all the

present $P_j^{(t)}$ satisfy equation (4) and be the probabilities of being conspirators.

In the end, we can sort P_j form large to small and make a list that shows the likelihood of one's being conspirator, meanwhile, divide the list into two parts based on a critical value p_d which set in advance.

Computing process can be shown in Figure 3, where every circle represents an office worker and the grey level of the circle stands for his/her probability in every iteration. In the end of T times iterations, the grey level of every circle basically no longer changed.

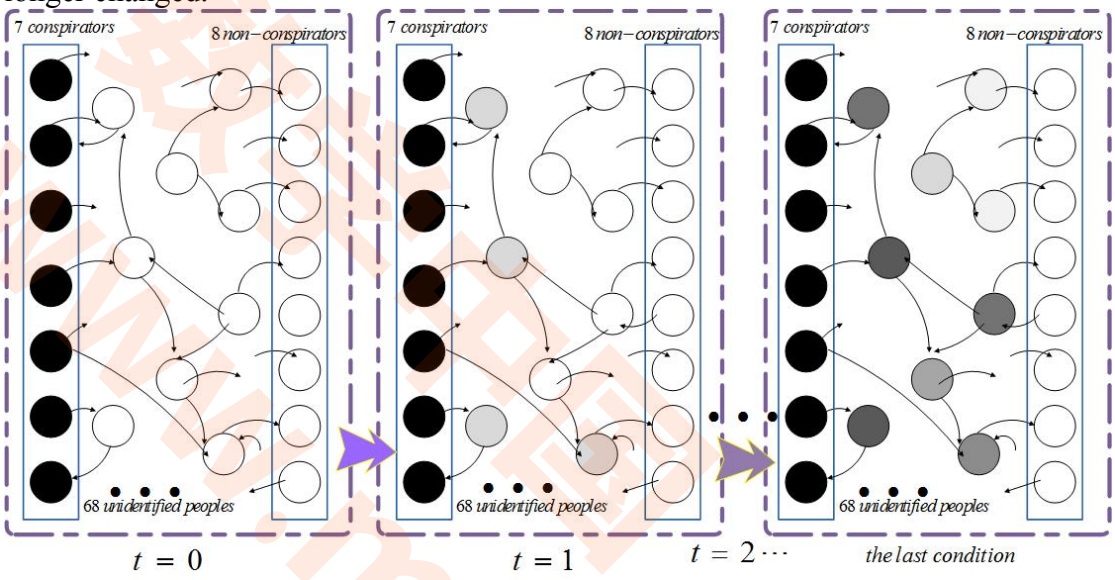


Figure. 3

Result and analysis

According to the given data in "Messages.xls", "Names.xls" and Topics.xls, we can make a priority list that shows the likelihood of one's being conspirator based on Model 2.

We set

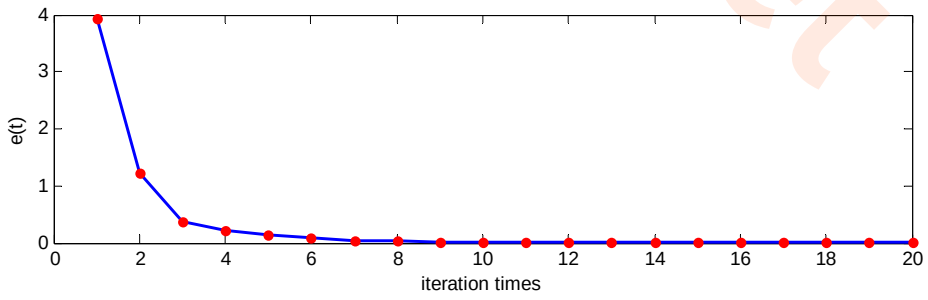
$a = 0.9$,which is the weight of elements in the set of S ;

$b = 0.1$,which is the weight of elements in the set of U ;

$T = 20$,which is the iteration times;

$p_d = 0.5$,which is the critical value separating the conspirator group and non-conspirator group.

After the iterative computation, $e(t)$ tends to be 0 as Figure 4 shows



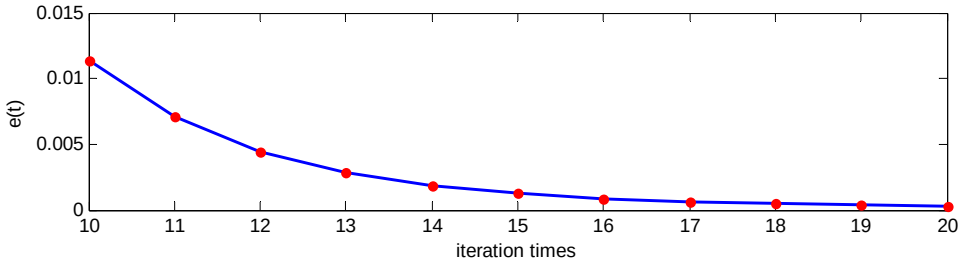


Figure. 4

As shown in Figure 4, after 20 times' iteration, $e(t)$ is less than 0.001. So we can consider $P_j(j=1,2,...,68)$, which represents probability of being conspirator of person in the unknown group, is so stable that all P_j can perfectly satisfy the equation $P_j=\frac{\sum_i p_i \times w_i}{\sum_i w_i}$ which is presented in model 1. Size $P_j(j=1,2,...,68)$ down, we get the priority list as Table 2 shows

Table 2

No	Node	Pro	No	Node	Pro	No	Node	Pro	No	Node	Pro
1	**7**	1	22	1	0.6248	43	22	0.5022	64	55	0.2658
1	**18**	1	23	31	0.6172	44	5	0.4969	65	82	0.254
1	**21**	1	24	14	0.6083	45	8	0.4906	66	52	0.2209
1	**43**	1	25	27	0.584	46	24	0.49	67	56	0.201
1	**49**	1	26	6	0.5809	47	26	0.4792	67	57	0.201
1	**54**	1	27	42	0.5798	48	72	0.476	69	63	0.1989
1	**67**	1	28	11	0.5765	49	20	0.4741	70	79	0.1978
1	73	1	29	46	0.5762	50	4	0.4591	71	77	0.1907
9	81	0.9773	30	69	0.5736	51	40	0.4407	72	23	0.1891
10	60	0.9367	31	45	0.5728	52	32	0.4192	73	80	0.1608
11	59	0.9366	32	16	0.5728	52	58	0.4192	74	76	0.1001
12	33	0.8078	33	28	0.5632	54	53	0.3987	75	75	0.01
13	30	0.7769	34	41	0.5626	55	62	0.3977	76	*0*	0
14	36	0.7561	35	13	0.5444	56	66	0.3964	76	*2*	0
15	37	0.6846	36	44	0.5412	57	61	0.3963	76	*48*	0
16	50	0.6683	37	39	0.539	58	35	0.3915	76	*64*	0
17	38	0.6591	38	47	0.5386	59	3	0.3672	76	*65*	0
18	10	0.6344	39	34	0.5366	60	29	0.3543	76	*68*	0
19	51	0.6344	40	15	0.5364	61	19	0.3313	76	*74*	0
20	9	0.6288	41	25	0.5327	62	71	0.2883	76	*78*	0
21	17	0.6254	42	12	0.5065	63	70	0.2766			

As for discriminate line, we previously have $p_d=0.5$, we can distinguish that the 43 previous workers in the list is categorized as conspirators while the others are non-conspirators.

Model 3

Introduction of graph theory

In mathematics and computer science, graph theory is the study of graphs, mathematical structures used to model pairwise relations between objects from a certain collection. A “graph” in this context is a collection of “vertices” or “nodes” and a collection of edges that connect pairs of vertices. The “shortest path” represents a path between two vertices (or nodes) in a graph such that the sum of the weights of its constituent edges is minimized. And Dijkstra's algorithm, is a graph search algorithm that solves the single-source shortest path problem for a graph with nonnegative edge path costs, producing a shortest path tree.

Establishment and Algorithm of the model

We develop a model based on graph theory which is good at dealing with network problem. As far as the second figure ICM shows us in the problem paper, we use a graph $G=\{V(G),E(G)\}$ to visualize the message traffic. A set of vertices $V(G)$ represents office workers, and a set of edges $E(G)$ represents messages. A set of known conspirators is named $V_0(G)$ while a set of known non-conspirators is named $V_n(G)$.

A member communicates with another through several paths in the conspirators' message network. However，in order to reduce the intercepted risk in the process of information transfer, they usually choose the shortest one. So the shortest path is much more important than the longer one in the assessment of the suspicion. Of course, suspicious messages are more important than the unsuspicious one as well. So the probability of one's being conspirator depends on the type and quantity of his message traffic, as well as the shortest distance between him and known conspirator group. It means the shortest distance between this vertex v_i and any element of $V_0(G)$ in the network. We use $d(v_i,V_0(G))$ to represent it, and its value is the quantity of the edges.

$$d(v_i,V_0(G))=\min_k\{d(v_i,v_k\},\ v_k\in V_0(G) \tag{6}$$

In conclusion, a suspicious index named $Score$ is used to decide whether a member is worth to be suspicious. For each vertex, in $V_0(G)$, $Score_i$=10 ,while in $V_N(G)$, $Score_i$=0. And the other vertices can be computed by

$$Score_i=\sum_j\frac{w_j}{d(v_i,V_0(G))} \tag{7}$$

Where w_j represents the weight of edge e_i , which is linked with the vertex v_i directly; the summation symbol is for all the edges linked directly with the vertex v_i. The larger $Score_i$ is, the more suspicious the office worker i is.

The calculating process of $Score_i$ can be described as Figure 5.

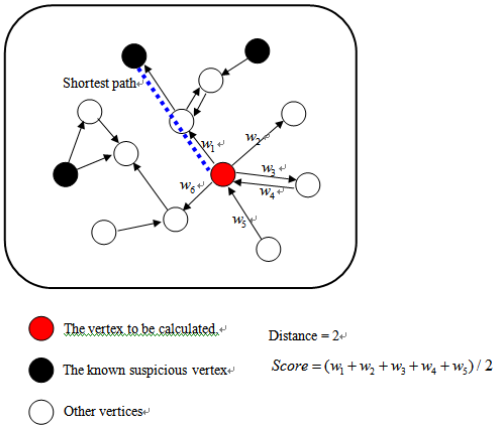


Figure 5

Algorithm 2

Step 1: Create the set of known conspirators $V_0(G)$ and the set of non-conspirators $V_n(G)$;

Step 2: Compute the shortest distance from all the vertices to the $V_0(G)$;

Based on message network, create an adjacency matrix with the connected value equaled to 1 and the unconnected value equaled to 0. Initialize $d(v_k, V_0(G)) = 0$ in the set of $V_0(G)$, and $d(v_k, V_0(G)) = +\infty$ in $V_n(G)$,

1) Start the vertices of $V_0(G)$,then search for any other vertex in matrix which connected with it directly and consist of a new set $V_1(G)$ at the same time , have its value equal to 1,
2) Continue to search down. But if one vertex has been visited, its value will not be assigned again. The loop will not stop until all the vertices are accessed;

Step 3: Visit all the edges, assign a value to their weights w_j, and according to the equation(6) to calculate $d(v_i, V_0(G))$. And let its two vertices cumulative $\frac{w_j}{d(v_i, V_0(G))}$;

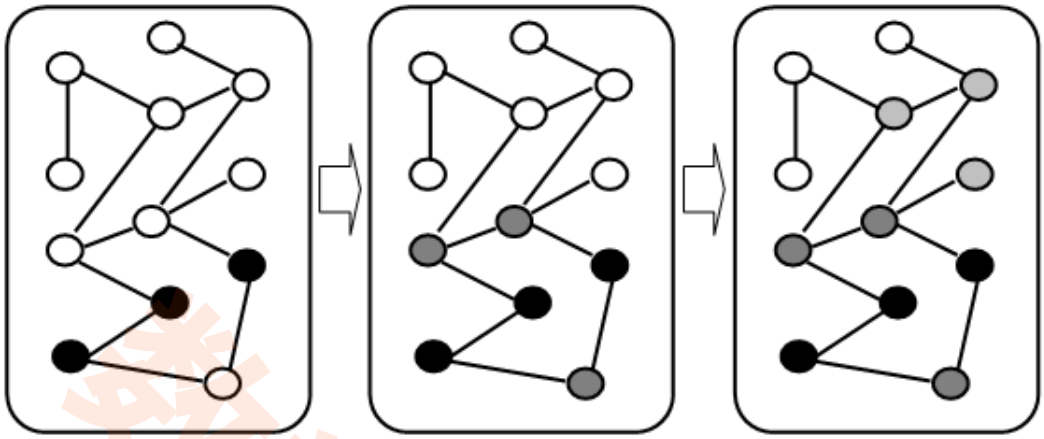


Figure 6

d is the distance where searching the vertices. A known conspirator is represented by a black round

Step 4: As for the vertices in $V_0(G)$, $Score$ is assigned to 10.0 ; for the vertices in $V_n(G)$, $Score$ is assigned to 0.0 ;for the other vertices, $Score$ can also be calculated.

Result and analysis

In the model, we set

$$w_j = \begin{cases} 1.0 & if\ the\ jth edg\ is\ a\ suspicious\ message \\ 0.1 & if\ the\ jth edg\ is\ a\ unsuspicious\ message \end{cases}$$

It means the effect of 10 unsuspicious messages is equal to 1 suspicious message in the suspicious evaluation.

Based on the ***Algorithm 2***, compute each $Score_i$, size $Score_i (i = 1,2,3,\cdots,83)$ down, and list them on table 3:

Table 3

No	node	Score	No	node	Score	No	node	Score	No	node	Score
1	**7**	10.000	22	44	3.700	42	40	1.600	63	71	0.200
1	**18**	10.000	22	50	3.800	44	1	1.500	65	62	0.133
1	**21**	10.000	24	30	3.300	44	5	1.500	65	70	0.133
1	**43**	10.000	24	31	3.300	44	69	1.500	67	55	0.100
1	**49**	10.000	26	35	3.100	47	9	1.400	67	73	0.100
1	**54**	10.000	27	4	3.050	47	42	1.400	67	76	0.100
1	**67**	10.000	28	27	3.000	49	12	1.200	70	52	0.075
8	3	8.200	29	15	2.900	49	60	1.200	71	53	0.067
9	10	7.100	30	13	2.850	51	80	1.100	72	59	0.050
10	17	6.500	30	32	2.850	52	25	1.000	72	63	0.050
11	34	6.000	32	36	2.800	53	39	0.850	74	58	0.033
12	16	5.500	33	46	2.700	54	23	0.750	75	61	0.025
13	81	5.100	34	22	2.650	55	26	0.700	76	0	0.000
14	47	4.900	35	14	2.300	56	79	0.550	76	2	0.000

15	28	4.600	36	33	2.200	57	51	0.500	76	48	0.000
16	19	4.200	36	45	2.200	58	72	0.400	76	64	0.000
16	20	4.200	38	29	2.000	59	66	0.300	76	65	0.000
16	37	4.200	39	82	1.800	59	75	0.300	76	68	0.000
19	6	4.000	40	8	1.750	61	56	0.275	76	74	0.000
19	38	4.000	41	24	1.650	62	77	0.250	76	78	0.000
19	41	4.000	42	11	1.600	63	57	0.200			

Furthermore, we set critical value as $Score = 2$. It is equivalent to contact with the known conspirator group by 2 messages. It is reasonable to consider an office worker is a conspirator. Draw a line on table, which show that the large probability of being conspirator comes before the number of critical value. Through analyzing, we conclude that the 38 previous workers in the list are categorized as conspirators while the others are non-conspirators.

Results comparison of Model 1, 2 and 3

Putting three lists we get respectively in three models together, we can draw Figure 7 to compare the consistency of all three model results.

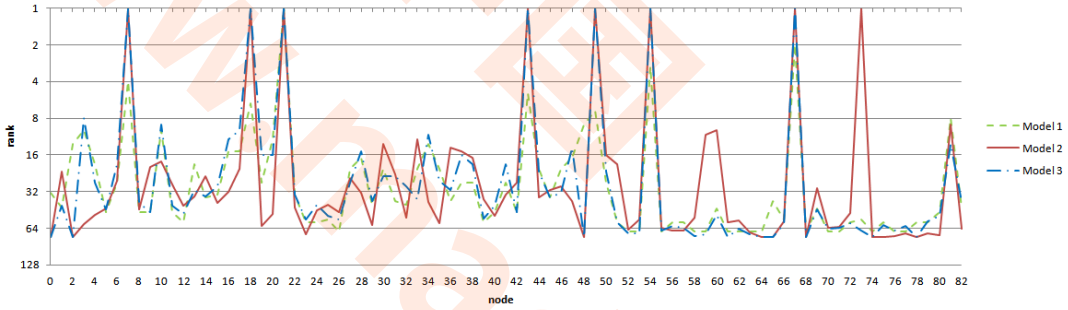


Figure. 7

Note that we use logarithmic scale for the vertical axis as rank changes are much more important in high-ranking part than low-ranking part.

And from Figure 7 we can intuitively see certain changes in some particular nodes are apparent but the whole ranking distributions of three models in all 83 nodes are similar.

Model 4

Relevant work

Centrality is a common Social network analysis which can be used to solve criminal network models (Peng Chen ,2011). Linton C. Freeman put forward a set of calculation method to find out the importance of any member in network (Freeman 1979). He explained some terms such as point centrality

Point degree of Point centrality: In a social network, if a conspirator has direct contact with other conspirators, this conspirator is in the central part of the network and has much more control power. Thus the importance of a point could be weighed by the number of points linked to it.

$$C_D(n_i) = d(n_i) \tag{8}$$

Where $d(n_i)$ is the number of points that member n_i has contact with. The higher the point centrality degree, the more likely the conspirators is a leader.

Betweenness of Point centrality: A point that falls on the shortest communication paths between other points exhibits a potential for control of their communication. It is this potential for control that defines the centrality of these points.

$$C_B(n_i) = \sum_{j<k} g_{jk}(n_i) / g_{jk} \tag{9}$$

Where g_{jk} is the number of the shortest paths linking arbitrary conspirator n_j and arbitrary n_k, but not including n_i. $g_{jk}(n_i)$ is the number of paths linking n_j and n_k that contain n_i

Closeness of Point centrality: Concerning with either independence or efficiency during the delivering of messages, closeness of point centrality is necessarily to be measured since it grows as points are far apart.

$$C_C(n_i) = \frac{1}{\sum_{j=1}^{g} d(n_i, n_j)} \tag{10}$$

Where $d(n_i, n_j)$ is the shortest distances between conspirator n_i and n_i. $C_C(n_i)$ is an inverse of sum of the shortest distance from one conspirator to others. The larger its value is, the closer the relationship he/she have with others.

Establishment of model

In order to comprehensively evaluate the probability of one conspirator's being leader, we define an aggregative index number C

$$C = \beta_1 C_D^* + \beta_2 C_B^* + \beta_3 C_C^* \tag{11}$$

Where β_1, β_2 and β_3 represent weight coefficient of point degree, betweenness and closeness respectively. C_D^*, C_B^* and C_C^* are normalization of C_D, C_B and C_C.

In addition, we define another aggregative index reflecting the network's structure called network density, to assess the average of shortest distances of any two conspirators. The larger the value of the index is, the safer but less efficient the organization is. It can be calculated by:

$$\rho = \frac{\sum_{i,j \in G, i \neq j} d_{ij}}{\frac{1}{2} N(N-1)} \tag{12}$$

Result and analysis

Start a new social network of 43 conspirators located in model 2. Then we have $\rho = 2.1971$ after computing, which means that a conspirator can transmit information to another conspirator through only 2.1971 messages in average. That's to say, the structure are too close to ensure safety. If a conspirator were caught, the others would be found out in no time, which make the conspiracy at great risk. However, conspiracy has great operation efficiency. The result shows that a commercial fraud often need close cooperation.

Compute C_D, C_B and C_C, respectively. Then set $w_1 = w_2 = w_3 = 1/3$, compute

C , then we have Table 4 .

Table 4

rank	1	2	3	4	5	6	7	8	9	10
C_D	Paul	Elsie	Alex	Yao	Neal	Julia	Jerome	Stephanie	Dolores	Beth
C_B	Elsie	Paul	Ulf	Jean	Dolores	William	Yao	Neal	Alex	Lars
C_C	Paul	Elsie	Jean	Neal	Beth	Dolores	Alex	Kristina	Jerome	William
C	Paul	Elsie	Dolores	Jean	Neal	Alex	Ulf	Yao	Jerome	Beth

Table 4 shows us Paul, Elsie and Dolores are three most probable conspiracy leaders. Since Delores is one of the senior managers in the company, it may be an important intelligence.

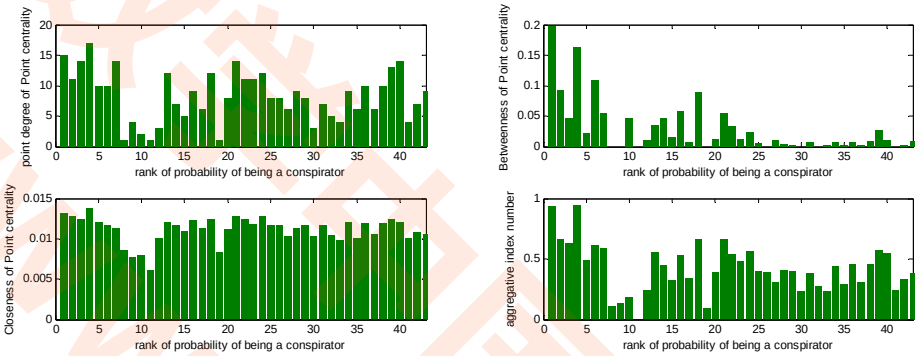


Figure. 8

In Figure 8, the sequence number in horizontal axis indicates the rank of probability of being a conspirator, and the smaller the number is, the greater probability of being a conspirator. We consider these three index have remarkable correction after we compute their Pearson correlation coefficient:

$$r(C_D, C_B) = 0.600,\ r(C_D, C_C) = 0.861 \quad and \quad r(C_B, C_C) = 0.498\ ,$$

which all pass significant test at 0.1% level. Moreover, betweenness of Point centrality varies in each conspirator is more significant than point degree of point centrality while closeness of point centrality is the least important one.

Validation

● Based on the statement in the Introduction part, we can suppose that every worker in conspiracy group must have message contact with other conspirator more or less. So that a message network made up by all the conspirators should be a “connected graph”, which is a term in graph theory means there is a path from any point to any other point in the graph. That is to say, every conspirator has at least one message exchange with another conspirator, or no single conspirator who has no message exchange will exist. Take the “connected graph” method above as a validation method for Model 2 and 3. After the computation, we find that both Model 2 and Model 3 satisfy the “connected graph” condition.

● Both in Model 2 & 3, we have some model parameters that we set manually. In order to validate the model stability when these parameters change to a small extent, we vary the value of these parameters, run the model program again, and see how much the result changes. In turns out that the most workers’ rank in the priority list remain invariant or change a little, which illustrate that our models are capable to

tolerate certain changes in model parameters and our manually given value of these parameter are relatively reasonable.

Impact of information changes

Information collected by investigation organizations is not always precise and sometimes during the process of investigation, new information would add in and old information would be revised. Thus a robust model is which, when input data is modified or some intrinsic parameters are changed a little, its results should not change a lot.

In this case, we assume the next situation: new information comes to light that Topic 1 is also connected to the conspiracy and that Chris is one of the conspirators. Using Model 2 with some corresponding modification as an example, we will see how would the priority lists we got before changed (using Model 3 is similar).

New priority list generated by Model 2:

Table 5

No	node	pro	No	node	pro	No	node	pro	No	node	pro
1	↑0	1	22	27	0.6177	43	22	0.508	64	↓61	0.2804
1	7	1	23	17	0.6176	44	8	0.5027	65	63	0.2553
1	18	1	24	9	0.6132	45	26	0.4984	66	71	0.2345
1	21	1	25	31	0.6046	46	↓44	0.4968	67	70	0.2257
1	43	1	26	1	0.599	47	↓12	0.4927	68	56	0.2225
1	49	1	27	↑28	0.5882	48	5	0.4887	69	57	0.2225
1	54	1	28	6	0.58	49	4	0.4747	70	79	0.2031
1	67	1	29	↓14	0.5795	50	↑55	0.4704	71	77	0.1906
1	73	1	30	↓50	0.5777	51	72	0.4633	72	23	0.1655
10	81	0.9774	31	↑20	0.5752	52	32	0.4492	73	80	0.164
11	60	0.9404	32	16	0.5721	52	58	0.4492	74	↓66	0.106
12	59	0.9403	33	46	0.5705	54	↓13	0.4478	75	76	0.1
13	33	0.8004	34	↑25	0.5575	55	53	0.4188	76	75	0.01
14	30	0.7766	35	41	0.5499	56	3	0.404	77	*2*	0
15	36	0.7497	36	34	0.5428	57	35	0.3925	77	*48*	0
16	↑69	0.7486	37	↓11	0.5421	58	↑82	0.3667	77	*64*	0
17	↑45	0.7002	38	15	0.5349	59	29	0.3602	77	*65*	0
18	37	0.6573	39	39	0.531	60	↓24	0.3597	77	*68*	0
19	38	0.656	40	47	0.5296	61	19	0.3312	77	*74*	0
20	10	0.6364	41	↓42	0.5238	62	52	0.2955	77	*78*	0
20	51	0.6364	42	↑40	0.5145	63	↓62	0.2837			

Comparing to the list we got before, there are totally 20 office workers whose change of ranking exceeds more than 5 positions, which is shown with arrow symbols in the list above. And we consider our Model 2 is stable throughout the comparison.

Strengths and weaknesses

In this paper, we develop three models (Model 1, 2 & 3) to make a priority list which shows the likelihood of one's being conspirator, and one model (Model 4) to nominate the conspiracy leader.

● Metric equation given in Model 1 is so concise and clear that it gives us the basic ideas and strategy to weigh the likelihood of one's being conspirator and its method and result are worth being referenced. But as it's far too simple to include more factors, we develop two more complex models to solve this problem at its basis.

● Model 2, which is based on *probability theory*, regards the whole nodes network linking through various kinds of messages as a "probability network". It skillfully links every node together using the *Law of Total Probability*, which makes every office worker's probability of being a conspirator related to each other by message type and number.

● Model 3, which is based on *graph theory*, take the "distance" between an unknown office worker and a known conspirator into account starting from the structure of the network graph. The "distance" we define well indicates the closeness of relationship with known conspirators. The shorter the "distance", the more likely the unknown one is a conspirator.

● Model 4 determines one's possibility of being the conspiracy leader from the prospective of centrality, which conforms to the characteristic of leaders in real criminal organization.

In order to enhance our model and increase the accuracy of our model results, some shortcoming and additional aspects about our models need to be pointed out so that further improvement and refinement can be made:

● Some model parameters such as weight coefficients and discriminate line in our models are mostly determined manually, which could cause uncertainty of our model solutions and the priority will change to some extent if different constant is set. Although we have varied the value of these constants and validate the models' stability, we still can't explain why the way we set these values can guarantee the accordance of our model results to the reality. Improvement methods are that we figure out a most suitable value for these model parameters by more means such as comparison with other model's result, more actual cases, etc.

● In all of our models, we consider that sending and receiving a message have the same effect for our analysis, which means the network graph is deem to be a undirected graph. But in fact, this assumption may have certain irrationality in some cases. So a certain modifications about the issue can be made to improve our models through further analysis in our future work.

● When analyzing messages exchanges in our models, we actually only focus on the 15 topics of the messages into account. More precisely, only two kinds of topics, suspicious and unsuspicious, are we taking into account while modeling. And it is better if we have more detailed classification about the topics. Moreover, if we can get the specific content of these messages, we can collect more information by identify if anyone is involved in the message content so that we can enhance our models.

Further discussion

Text and semantic network analysis

Text analysis

Text analysis is the analysis of text taking advantage of algorithmic techniques and usually be divided into three kinds according to algorithmic–exploratory spectrum: Concording, Content analysis, Statistical analysis.

Semantic network analysis

Semantic network analysis is both a research method and a theoretical framework, which focuses on the structure of a system based on shared meaning as well as cognitive processes. It requires a content analysis of textual data to determine the most frequently used symbols previously.

Apply Text and Semantic network analysis to crime case, here's some procedure we should taken:

(1) Keywords mining

We use some mathematical methods or computer technique to achieve keywords mining in the text database, like Frequency Statistics, Association Rule Analysis and Data Mining. The processing flow is as follows:

Step 1: Mining the high-frequency keywords

We perform the frequency Statistics of keywords in the message or text and consider the high frequency keywords as High-frequency Keywords. Take supervior's simply case as an example, the frequency of "budget" is 7/ 28 , "late" is 6/ 28 and "stress is 5/28 .

Step 2: Identify Suspicious Keywords preliminarily

Study the implementation of the conspiracy. Indentify its content or topics. Finish the classification of the High-frequency keywords and find out Suspicious Keywords.

However, it is defective to consider a message with only one Suspicious Keywords as a Suspicious Message, because some Suspicious Keywords may usually exit in the normal business, such as "budget" in the cases of economic crime.

Step 3: Mining the new Suspicious Keywords:

We can use Association Rule Analysis to mining the new Suspicious Keywords using the known Suspicious Keywords. If some word combinations' support and confidence are bigger than the minimum support and minimum confidence we give and they contain a known Suspicious Keywords, we can successfully to mining out the new Suspicious Keywords.

(2) The classification of topics:

The keywords and their combinations can determine a topic. Besides, the suspicious messages depend on the classification of topics. It is more reliable to consider a message containing two or more Suspicious Keywords as a Suspicious Message. Thus, some non-conspirators can avoid falsely being accused. In the current case, a non-conspirator Paige sent 4 suspicious messages. It shows that there are some defects in the topic classification of the messages.

On the other hand, criminal activities sometimes may be found and suspected. In the communication messages, someone who questions them is likely to be a non-conspirator. For example in the case of Investigation EZ, Anne questioned Bob's tardiness, Harry questioned George' s stress and Harry question the status of budget. In fact, they are all non-conspirators.

(3) Categorize social group

Look out names occurring on the message whose owners don't contact directly. However these people have some relationship with both contacts. We can analyze emotional relationship from semantic and other aspects from similarity of characters in life as well work, then make sure their closeness so that we can categorize social

group.

As for the case, we semantic analyze social relationship, and indicate

Table 6

content	Close relationship	Distant relationship
Quarrel for Bob' tardiness		Anne、Bob
Operation and interactive argue	Bob、Dave、George	
Similar character of stress and tired	Inez、George	
Misunderstanding and suspicion		Harry、George
Cooperation and trust	Anne、Carol	

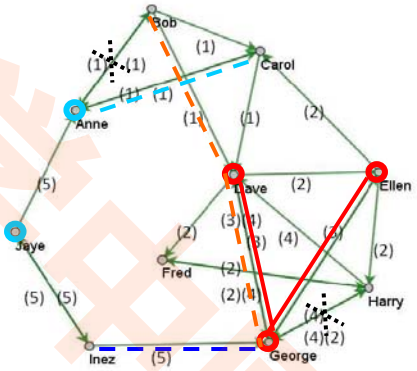


Figure. 9 The expanded relationship network

In the Figure 9 ,red rounds represent known conspirators and someone with the Suspicious Message communication; blue rounds represent the known non-conspirators; red solid lines represent suspicious messages; dotted lines with different colors represent different close relationship; crosses represent distant relationship.

As the Figure 9 shows, the expanded relationship network can be well explained by the actual case result.

Based on step-by-step procedure we introduce above, a program could be written to analyze message traffic and the main idea of each traffic could be figured out. Semantic and text analysis can help us with large data sets easily. Thus, even several million messages need to be analyzed, we also can take advantage of these two techniques and make categorization for message traffic.

Model application in other disciplines

In this paper, the models we establish for this specific crime case are actually not restricted to solve this kind of problem. A lot of knowledge and techniques we use in these models, such as probability theory, iterative algorithm, graph theory and so on, are widely applied in many different disciplines and domains. And the models we build are also particular suitable for analyze network database of many types to identify and prioritize some nodes with some specific characteristic.

For example, identification of infected or diseased cells in cell network is one of the hot topics in biological terms. The probability of one cell's being infected is much like the probability of one office worker's being conspirator in our Model 2. That is to say, whether a cell's being infected or not is connected to the situation of those cells directly around it. Through modern medical imaging technology and other medical

laboratory technologies, we can get the similar data as that we got in this crime case, such as how many and where are the cells infected already. Thus we can apply our model to calculate the infected probability of other cells of unknown state by modify some model setting.

In all, concept like "node" we use in our models can be materialized to lots of specific entities in reality, and then our models can be properly altered, improved and adopted to solve a lot of practical problem but not just in criminology and biology.

Reference

Sridhar Ramamoorti. *2008.*The Psychology and Sociology of Fraud: Integrating the Behavioral Sciences Component Into Fraud and Forensic Accounting Curricula. *Issues in Accounting Education (November): pp. 521–533*

Zhigang Lin. 2010.psychology analysis of corruption and bribery crime. *Huaihaiwenhui(Chinese Journal)*(03)

Zwillinger, D, Kokoska, S. 2000. CRC Standard Probability and Statistics Tables and Formulae, CRC Press

Shuhe Wang. 1990.Graph theory and algorithm.Hefei,China:Press of University of Science and Technology of China

Peng Chen,Hongyong Yuan. 2011.Social network analysis of crime organization structures.*J Tsinghua Univ(Sci &Tech)*51(8):1097-1101

Freeman L C. 1979.Centrality in social networks: Conceptual clarification. *Social Networks*(1):215-239

Method in text-analysis: An introduction. http://www.cch.kcl.ac.uk/legacy/teaching/av1000/textanalysis/method.html

Marya L.Doerfel.1999.A Semantic Network Analysis of the International Communication Association.*Human Communication Reseach*(June)

Chang' an Yuan.2009.Data Mining Theory and Application of SPSS Clemetine.Publishing House of Electronics Industry