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

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:

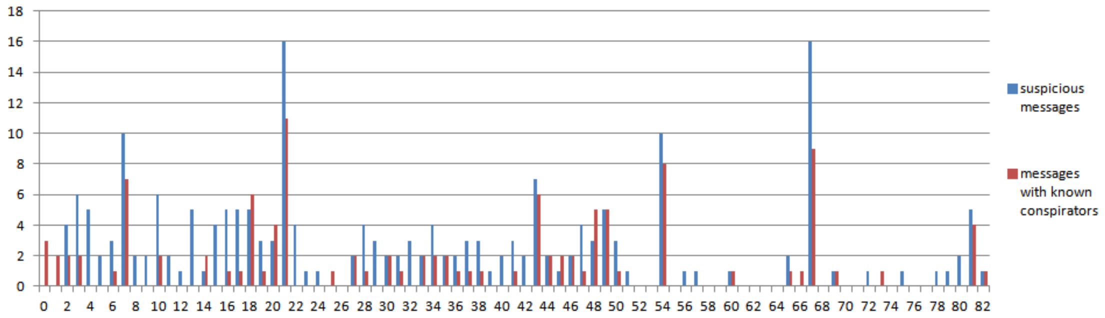

Figure. 1

## 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 $P_j\,(j = 0,1,...,83, \mathit{except}\, 15\, \mathit{known\ persons})$ 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.

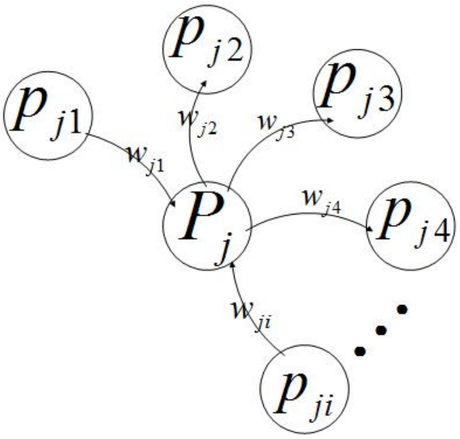

Figure. 2

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