For office use only	Team Control Number	For office use only
T1 ________________	**15141**	F1 ________________
T2 ________________		F2 ________________
T3 ________________	Problem Chosen	F3 ________________
T4 ________________	**C**	F4 ________________

2012 Mathematical Contest in Modeling (MCM) Summary Sheet

(Attach a copy of this page to each copy of your solution paper.)

The Crime Network Discovery

Abstract

The process of solving this problem is divided into three steps each of which is implemented by a particular model, in order to prioritize nodes by likelihood of being conspirator, find the discriminate line and the leader(s).

Firstly, we create Model 1 to determine the weight of each edge in this network. This weight indicates the crime characteristics shared with two nodes of the edge. We abstract three metrics Topics between two direct talkers, Identity of two direct talkers and Frequency of talks to determine the edge weight from information we can get. Since the mutual interdependence between these metrics, the Analytic Network Process (ANP) method is applied to combine all the three metrics. Secondly, after acquiring the formula of the edge weight, we assume the network turns into a social network with edge weights. In Model 2, we create a measure named Edge Weight in Node Level. Another two measures Degree and Closeness from the Social Network Analysis are also chosen. By combining these three measures, we prioritize all the nodes. Finally, in Model 3 we determine the discriminate line and leader(s) by Clustering Analysis with measures derived from Model 2.

We testify our model with the given example. The result matches well with the evidences. In addition, the set of our possible conspirators in Requirement 2 is the subset of that of Requirement 1. Furthermore the leader is the same person in both cases. These facts show our model is good in accuracy and stability.

Our model can be enhanced conveniently. With combining text analysis and semantic network analysis, the model is optimized by using more objective information in ANP model. We also discuss a text analysis algorithm and the use of text analysis software for mass data processing. Although the ANP possesses a certain subjective, our model still performs well considering accuracy, extendibility and succinctness. It provides valuable recommendations to the police.

1. Introduction

Crime has never left along with the human society developing. However forms of crime have also undergone great changes. There is obvious increase in white-collar, high-tech conspiracy crimes. In addition, just as the sociologist Richard Quinney [1] states "crime is a social phenomenon", crime is no longer just the behavior of individuals, but the behavior of a social network. An appropriate framework is required from the perspective of the social network that investigates the identity of members in a particular crime network of conspiracy and provides reasonable recommendations to the police.

So our goal is pretty clear:

- Analyze the information available and determine the metrics for the weight of each edge in the graph which indicates the crime characteristics shared with two direct nodes of an edge.
- Devise a method to combine the metrics for the edge weight.
- Determine suspicion degree for each node by combing the edge weight and disciplines of the social network analysis.
- Use the suspicion degree to sort all the members, find the discriminate line between conspirators and non-conspirators and find the leader of this network.
- Optimize the model by powerful techniques from other areas.

1.1 Deal with the Abnormal Data

After analyzing the given data, some abnormal data is found. Taking into account that human errors or objective random factors may occur during the data acquisition, we decide to amend or delete the abnormal data before it is used in the models. Table 1 in the Appendix shows the basic information of this abnormal data and approaches to deal with it.

1.2 Assumptions

- Assumption 1: Regard the whole network as an undirected graph when determining the Edge Weight.
- Assumption 2: Consider the network as a social network with edge weights after determining the Edge Weight.
- Assumption 3: Assumption to deal with the problem of same names repeating in the original data is made. See Table 1 in the Appendix.

2. Models

2.1 Model 1: Determine Edge Weight Using Analytic Network Process

2.1.1 Metrics for Edge Weight

Two Important Factors

First of all, what should be taken into account is how much available information of this communication network we can get. Two types of information can be obtained from the material provided: identity of part of members and suspicion degree of part of topics which are talked about.

Considering that only part of the topics' suspicion degree is clear, suspicion degree of topics should be a adjustment process of dynamic learning in analysis. This is also the same with suspicion degree of identity of all the members.

Combined with the experience of the Criminal Network Analysis [2], we can conclude that the suspicion degree of topics and the suspicion degree of identity are the two most important factors that affect the crime characteristics of this network in this case. Furthermore, as these two factors should be adjusted in a dynamic learning process, these two factors have interdependence.

Metrics for Edge Weight

Based on Assumption 1 that we regard the whole network as an undirected graph when determining the Edge Weight, we devise a composite measure of the three metrics to determine the Edge Weight.

Topics between two direct talkers: Topic is the obvious and direct evidence for policemen to investigate behavior of a talker. The higher the suspicion degree of the topics which have been talked about is, the greater the likelihood of the talkers being involved in the conspiracy is.

Identity of two direct talkers: The known identity also helps a lot in deciding identity of other members especially when these two people are linked directly in the graph. If any side of a session is a definite conspirator (or non-conspirator), the possibility of the other side being a conspirator gets higher (or lower) relatively.

Frequency of talks: As discussed in Social Network Analysis, the link density plays an important role in determining the key node [3]. Similar to it, frequency of talks also calculates the number of talks between two talkers. The higher frequency of talks an edge possesses, the higher relative possibility of the talkers being conspirators is.

There is mutual influence between the metrics above. Relationship between the metrics is not a simple hierarchical structure, but a more complicated network structure. To determine the Edge Weight formula, a decision-making mechanism for the network structure is needed.

We discuss how to determine Edge Weight using Analytic Network Process in the next section.

2.1.2 Analytic Network Process (ANP)

The ANP generalizes the Analytic Hierarchy Process (AHP), by replacing hierarchies with networks. The AHP is a relatively popular tool for modeling strategic decisions and Saaty [4] suggested the usage of AHP to solve the problems of independence on alternatives or criteria. However, in many cases, there is interdependence among criteria and alternatives. Hence, the ANP can be used as an effective tool in those cases where the interactions among the elements of a system form a network structure [4].

Model of the ANP method is more complex, for it needs to calculate the unweighted super-matrix in the conjugated-comparative matrices, the weighted super-matrix, limit super-matrix of relative weights, and then the final sort of program or score.

Network Structure

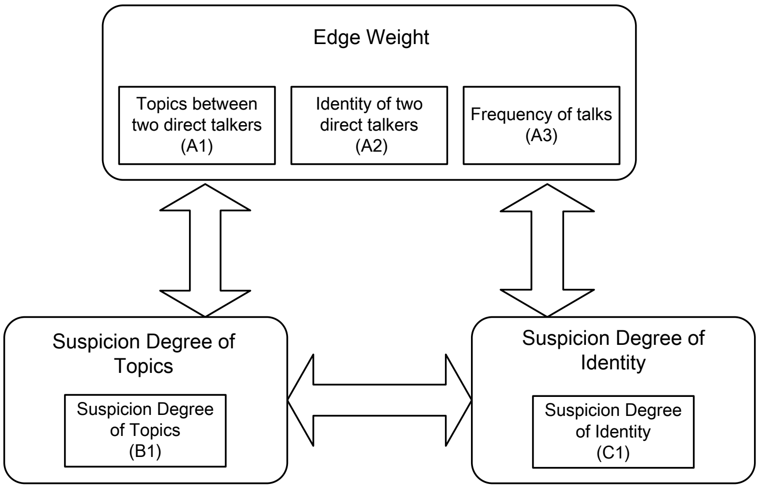
Edge Weight
Topics between two direct talkers (A1)
Identity of two direct talkers (A2)
Frequency of talks (A3)
Suspicion Degree of Topics
Suspicion Degree of Topics (B1)
Suspicion Degree of Identity
Suspicion Degree of Identity (C1)

Figure 1.
Network Structure

As shown in Figure1, Edge Weight, Suspicion Degree of Topics and Suspicion Degree of Identity have interdependence on each other. In this network structure, B1 impacts A1 strongly, influences A2 and A3 slightly; C1 impacts A2 strongly, affects A1 and A3 slightly; In return, Edge Weight affects Suspicion Degree of Topics and Suspicion Degree of Identity.

Conjugated-comparative Matrices

With Satty's Rule, the conjugated-comparative matrices can be constructed. The Consistency Ratio (CR) is applied to measure how consistent the judgments have been relative to large samples of purely random judgments. If the CR is much in excess of 0.1 the judgments are untrustworthy because they are too close for comfort to randomness and the exercise is valueless or must be repeated [4].

Set M1 be the conjugated-comparative matrix under main criteria of Suspicion Degree of Topics and sub-criteria of Edge Weight. Set M2 be the conjugated-comparative matrix under main criteria of Suspicion Degree of Identity and sub-criteria of Edge Weight.

$$M1 = \begin{bmatrix} 1 & 1/3 & 5 \\ 3 & 1 & 7 \\ 1/5 & 1/7 & 1 \end{bmatrix} \qquad M2 = \begin{bmatrix} 1 & 3 & 7 \\ 1/3 & 1 & 5 \\ 1/7 & 1/5 & 1 \end{bmatrix}$$

CR = 0.0624 < 0.01 CR= 0.0624 < 0.01

Let M3 be the conjugated-comparative matrix of all clusters in the network structure under main criteria of Edge Weight. Let M4 be the conjugated-comparative matrix of all clusters in the network structure under main criteria of Suspicion Degree of Topics. And let M5 be the conjugated-comparative matrix of all clusters in the network structure under main criteria of Suspicion Degree of Identity.

$$M3 = \begin{bmatrix} 1 & 2 \\ 1/2 & 1 \end{bmatrix} \qquad M4 = \begin{bmatrix} 1 & 2 \\ 1/2 & 1 \end{bmatrix} \qquad M5 = \begin{bmatrix} 1 & 2 \\ 1/2 & 1 \end{bmatrix}$$

CR=0.0000 < 0.01 CR=0.0000 < 0.01 CR=0.0000 < 0.01

Since all the value of CR is less than 0.01, the judgments can be considered as trustworthy.

Unweighed Super-matrix

Assume a network structure is composed of hierarchy $C_h(h = 1,2,\dots,m)$. For each hierarchy, assume there exist elements $e_{h1}, e_{h2}, \dots e_{hm_k}$, so the influence of $C_h(h = 1,2,\dots,m)$ can be denoted as below:

$$W = \begin{bmatrix} W_{11} & W_{12} & \cdots & W_{1m} \\ W_{21} & W_{22} & \cdots & W_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ W_{m1} & W_{m2} & \cdots & W_{mm} \end{bmatrix}$$

Thus, the super-matrix can be formed as the following matrix:

$$W = \begin{array}{c|ccc} & C_1 & C_2 & C_3 \\ \hline C_1 & 0 & W_{12} & W_{13} \\ C_2 & W_{21} & 0 & W_{23} \\ C_3 & W_{31} & W_{32} & 0 \end{array}$$

where C_1 represents the cluster of Edge Weight, C_2 represents the cluster of Suspicion Degree of Topics and C_3 represents the cluster of Suspicion Degree of Identity. W_{ij} indicates the influence of each element of the ith hierarchy on jth hierarchy, which is named a block of a super-matrix.

After calculating, the unweighed super-matrix is demonstrated in Table 2 in the Appendix.

Weighted Super-matrix

After forming the super-matrix, the weighted super-matrix is derived by transforming all columns sum to unity exactly. This step is much similar to the concept of Markov chain for ensuring the sum of these probabilities of all states equals to 1 [5].

By method Markov chain Monte Carlo [6], we get the weighed super-matrix which is shown in Table 3 in the Appendix.

Limited Super-matrix

What we wish to obtain is the priorities along each possible path in a super-matrix, namely the final influence an element on the highest goal. This kind of result can be acquired by solving,

$$\overline{W}^{\infty} = \lim_{k \to \infty} \overline{W}^{k} \quad (1)$$

By raising the weighted super-matrix to limiting powers such as Eq.(1), we can get the global priority vector or called weights.

Table 4 in the Appendix shows the limited super-matrix of this Model.

2.1.3 Determine Weights

After normalizing data in the limited super-matrix above, weights of the three factors of Edge Weight can be obtained. See Table 1 below.

Table 1.
The overall synthesized priorities.

Name	Ideals	Normals	Raw
Topics between two direct talkers	1.000000	0.494878	0.197951
Identity of two direct talkers	0.875357	0.433195	0.173278
Frequency of talks	0.145343	0.071927	0.028771

The 'Raw' column in Table 1 is the raw data derived directly from limited super-matrix, and the 'Normals' column is the weights that we need. It can be seen from the Table 1 that queue sorted by the degree of influence on the Edge Weight is Topics between two direct talkers > Identity of two direct talkers > Frequency of talks.

2.1.4 Formula

With all the process of calculation above, we can form this formula that evaluates Edge Weight:

$$EW\left(i,k\right) = 0.494878 \times TT + 0.433195 \times IT + 0.071927 \times FT$$

where i and k indicate two talkers and the other notations are defined in Table 2 below:

Table 2.
Symbols used.

Abbreviations	Meaning
EW	Edge Weight

TT	Topics between two direct talkers
IT	Identity of two direct talkers
FT	Frequency of talks
SDT	Suspicion Degree of Topics
SDI	Suspicion Degree of Identity

Now determine the value of factor TT, IT, and FT.

(1) TT= SDI of one node + SDI of the other node

Table 3 below shows how to determine SDI value.

Table 3.
Value Determination of SDI.

Definite non-conspirator	Unknown	Definite conspirator
-1	0	1

(2) IT = SDT

Table 4 below shows how to determine SDT Value.

Table 4.
Value Determination of SDT.

Unknown	Definite Topic Related to Conspiracy
0	1

(3) FT is the number of talks between two talkers which are linked directly.

2.2 Model 2: Prioritize Members Using Methods in Social Network Analysis

Based on Assumption 2, after determining Edge Weight we consider the network as a social network with edge weights. However, Edge Weight characterizes the criminal characteristics of an edge shared by two nodes of this edge. To sort all the nodes, it is necessary to adopt new features and standards in node level.

As mentioned, this network has been transformed to a social network with edge weights, which means several quantitative methods and measures from the area of social network analysis (SNA) can be employed to evaluate criminal characteristics of a node.

2.2.1 Methods in SNA

Basic but essential measure in social network analysis is centrality. Centrality is often used to indicate the importance of a member within a group or a network. Various centrality measures have been proposed and they have different interpretations and implications. Freeman defines the three most popular centrality measures: degree, betweenness, and closeness [7].

Degree measures how active a particular node is. It is defined as the number of direct links a node k has. An individual with a high degree could be the leader in his group or the network. However, degree is not a reliable indicator of leadership role in criminal networks. We therefore use this measure to only represent an individual’s activeness.

$$C_D\left(k\right)=\sum_{i=1}^{n}a\left(i,k\right)$$

where n is the total number of nodes in a group or a network and a(i, k) is a binary variable indicating whether a link exists between nodes i and k. a(i, k) = 0 indicates there is no direct link between nodes i and k [2].

Closeness is the sum of the length of geodesics (shortest paths between two nodes) between a particular node k and all the other nodes in a network. It actually measures how far away one node is from other nodes. It indicates how easily an individual connects to other members [2].

$$C_C\left(k\right)=\sum_{i=1}^{n}l\left(i,k\right)$$

where n is the total number of nodes in a group or a network and l(i, k) is the length of the shortest path connecting nodes i and k.

Betweenness measures the extent to which a particular node lies between other nodes in a network or a group. The betweenness of a node k is defined as the number of geodesics passing through it [2]. Yet, considering that the edge weight should be involved in the extent as greater as possible and also the computational complexity, a novel measure Edge Weight in Node Level which is also related to the edge weight is designed to replace betweenness in this case.

Edge Weight in Node Level (EWN) is the sum of weights of all the edges which are linked to a particular node directly. EWN measures the extent to the possibility of crime in this network for each particular node.

$$C_{EWN}\left(k\right)=\sum_{i=1}^{m}EW\left(i,k\right)$$

where m is the total number of nodes that are linked to node k directly and EW(i,k) is the Edge Weight of the edge between node i and node k.

2.2.2 Priority Index for Prioritizing Members

We already have three measures Degree, Closeness and EWN. Firstly normalize all this three measures, and then we define the Priority Index (PI) which determines the priority of all the members in the network as follows:

$$PI\left(k\right)=\sqrt{C_D{}^2\left(k\right)+C_C{}^2\left(k\right)+C_{EWN}{}^2\left(k\right)}$$

A node with higher PI is regarded as a man possessing high possibility of being a conspirator. Now we can get the priority list according to the value of PI.

2.3 Model3: Discriminate Conspirators and Non-conspirators Using Clustering

Clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters. Finding the discriminate line between conspirators and non-conspirators is equivalent to assigning all the members into two groups. We can get the discriminate line by finding the appropriate threshold in the process of clustering.

Clustering can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. We choose the method to include groups with low distances among the cluster members.

Mahalanobis distance [8] is employed in this clustering, for it takes into account the correlations of the data set and is scale-invariant. By taking into consideration the three measures Degree, Closeness and EWN as mentioned above, the Mahalanobis distance in three dimensions can classify members. Subsequently an appropriate threshold is picked to arrange members into two clusters: the cluster of conspirators and the cluster of non-conspirators. This threshold is the discriminate line that we need.

3. Results and Analysis

3.1 Testify the Models with the Example Provided

There are 10 nodes in the example. As mentioned, topic 3 is assumed to be related with conspiracy. We use MatLab to draw hierarchical clustering tree. The hierarchical clustering is as follows:

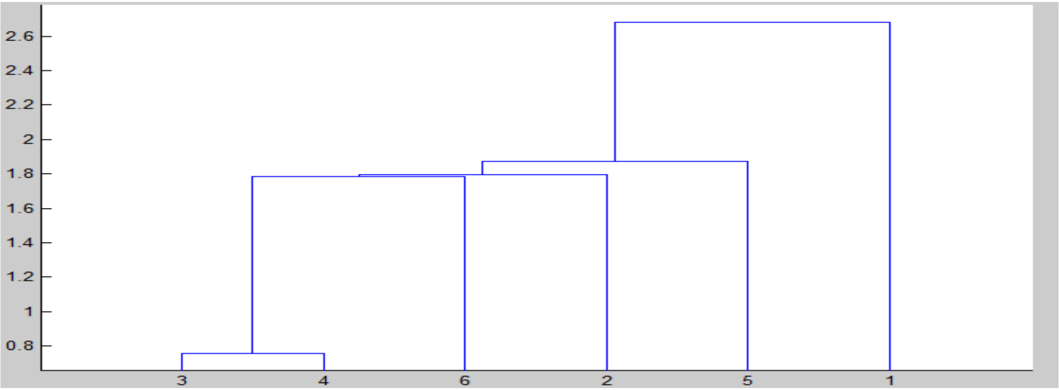

Figure 2.

Hierarchical Clustering Tree of Example.

We take the midpoint of the scale in the figure as the threshold. So the threshold is 1.9.

Now we get:
Group1: 3, 4, 6, 2
Group2: 5, 1
(5, 1 in this matrix in MatLab correspond to No.8, No.4 in the actual scenario)

No. 8 is Inez and No. 4 is Ellen. As mentioned, they are all conspirators. Maybe due to the lack of information, we are not able to predict Bob is a conspirator. But our model accurately predicts other conspirators. So this test can still show that the performance of our model is pretty good.

3.2 Results of Requirement 1

After introducing data into the model, we get the results of Requirement 1 as follows.

Priority List

By calculating PI, we can get the priority list. The higher PI value a node owns, the higher possibility of being a conspirator the node has.

Table 5.

Priority List of Requirement 1

Rank	No	PI
1	22	2.55493;
2	16	2.483;
3	47	2.483;
4	3	2.41969;
5	81	2.33915;
6	24	2.339;
7	57	2.21251;
8	32	2.11251;
9	40	1.70903;
10	8	1.70903;
11	15	1.70903;
12	10	1.6371;
13	20	1.56517;
14	29	1.56517;
15	27	1.42132;
16	13	1.42132;
17	19	1.34939;
18	37	1.34939;
19	9	1.27746;
20	11	1.14222;
21	35	0.998367;
22	31	0.998367;
23	38	0.998367;
24	34	0.998367;
25	44	0.92644;
26	6	0.92644;
27	46	0.92644;
28	50	0.92644;
29	45	0.854513;
30	26	0.791197;
31	42	0.782586;
32	39	0.782586;
33	28	0.782586;
34	23	0.710659;
35	60	0.710659;
36	33	0.710659;
37	69	0.710659;
38	12	0.647343;
39	72	0.638732;
40	79	0.638732;
41	56	0.566805;
42	51	0.566805;
43	30	0.503489;
44	82	0.503489;
45	14	0.503489;
46	1	0.503489;
47	41	0.431562;
48	17	0.431562;
49	25	0.359635;
50	36	0.359635;
51	4	0.215781;
52	52	0.215781;
53	70	0.215781;
54	71	0.215781;
55	5	0.143854;
56	55	0.143854;
57	75	0.143854;
58	77	0.143854;
59	62	0.071927;
60	58	0.071927;
61	61	0.071927;
62	63	0.071927;
63	66	0.071927;
64	73	0.071927;
65	76	0.071927;
66	80	0.071927;
67	53	0;
68	59	0;

Hierarchical Clustering Tree

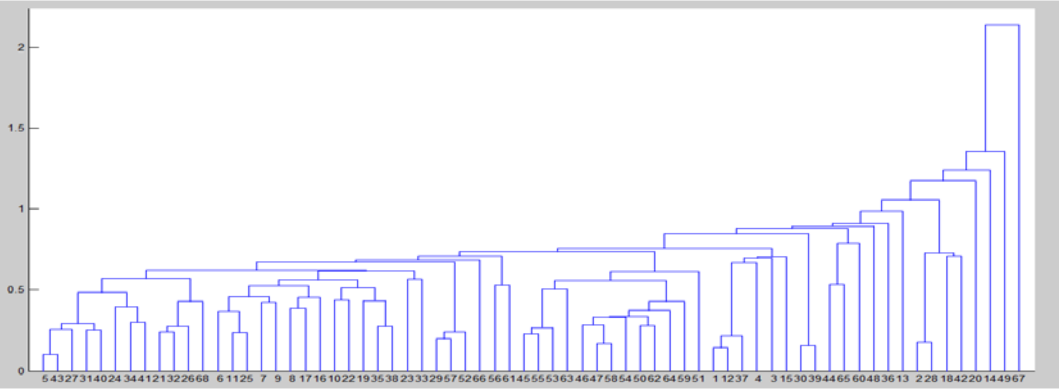

Figure 3.

Hierarchical Clustering Tree of Requirment 1.

We take the midpoint of the scale in the figure as the threshold. So the threshold is 1.

Now we get:

Group1: 2, 28, 18, 42, 20, 14, 49, 67

(2, 28, 18, 42, 20, 14, 49, 67 in this matrix in MatLab correspond to No.3, No.32, No.22, No.47, No.24, No.16, No.57, No.81 in the actual scenario)

Conclusion of this section

The discriminate line between conspirator and non-conspirator is 1. And No. 3, No. 32, No.22, No.47, No.24, No.16, No.57 and No.81 are the members with the higher possibility of being a conspirator.

Leader(s)

Considering a crime gang leader may be one or several, we don't take the one with the largest PI value as the leader. We use conspirators' PI value to take a hierarchical clustering. To find out the leader of the conspirators, we only consider No.3, No.32, No.22, No.47, No.24, No.16, No.57, No.81 and then re-analyze this case.

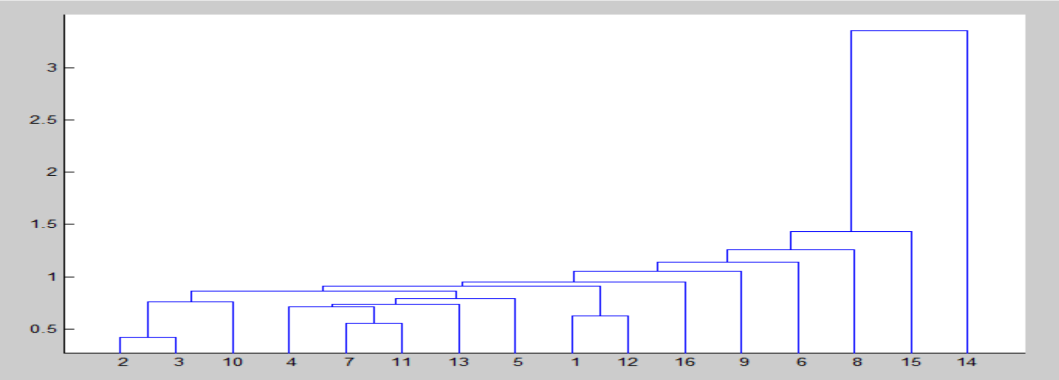

Figure 4.

Hierarchical Clustering Tree 2 of Requirment 1.

Now we get

Group 1: 14

Group 2: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 15, 16

(14 in this matrix in MatLab correspond to No.57 in the actual scenario)

So No.57 has the largest possibility to be the crime gang leader. In addition, No. 32 is Gretchen_manager (as there are two men named Gretchen, we adjust by the disciplines given by us. See Table in the Appendix), who is one of the senior manager of the company, is highly considered to be conspirator.

Conclusion of this section:

No.57 has the largest possibility to be the crime gang leader.

3.3 Results of Requirement 2

New information comes to light that Topic 1 is also connected to the conspiracy and that Chris is one of the conspirators. After introducing data into the models, we get the results of Requirement 2 as follows.

Priority List

By calculating PI, we can get the priority list. The higher PI value a node owns, the higher possibility of guilt the node has.

Table 6.
Priority List of Requirement 2

Rank	No.	PI	Rank	No.	PI	Rank	No.	PI
1	3	3.40944;	26	35	0.998367;	51	1	0.503489;
2	32	3.34227;	27	38	0.998367;	52	30	0.503489;
3	57	2.3016;	28	82	0.998367;	53	36	0.359635;
4	81	2.27915;	29	6	0.92644;	54	4	0.215781;
5	16	2.223;	30	41	0.92644;	55	52	0.215781;
6	40	2.2139;	31	44	0.92644;	56	70	0.215781;
7	24	2.202;	32	46	0.92644;	57	5	0.143854;
8	15	2.20203;	33	50	0.92644;	58	75	0.143854;
9	10	2.13198;	34	25	0.854513;	59	77	0.143854;
10	13	1.9162;	35	45	0.854513;	60	58	0.071927;
11	20	1.56517;	36	28	0.782586;	61	61	0.071927;
12	29	1.56517;	37	39	0.782586;	62	63	0.071927;
13	14	1.49324;	38	42	0.782586;	63	66	0.071927;
14	31	1.49324;	39	23	0.710659;	64	73	0.071927;
15	17	1.42132;	40	33	0.710659;	65	76	0.071927;
16	27	1.42132;	41	60	0.710659;	66	80	0.071927;
17	19	1.34939;	42	69	0.710659;	67	53	0;
18	37	1.34939;	43	71	0.710659;	68	59	0;
19	26	1.28608;	44	12	0.647343;			
20	9	1.27746;	45	55	0.638732;			
21	47	1.263;	46	72	0.638732;			
22	22	1.25493;	47	79	0.638732;			
23	8	1.2039;	48	51	0.566805;			

24	11	1.14222;	49	56	0.566805;
25	34	0.998367;	50	62	0.566805;

Hierarchical Clustering Tree

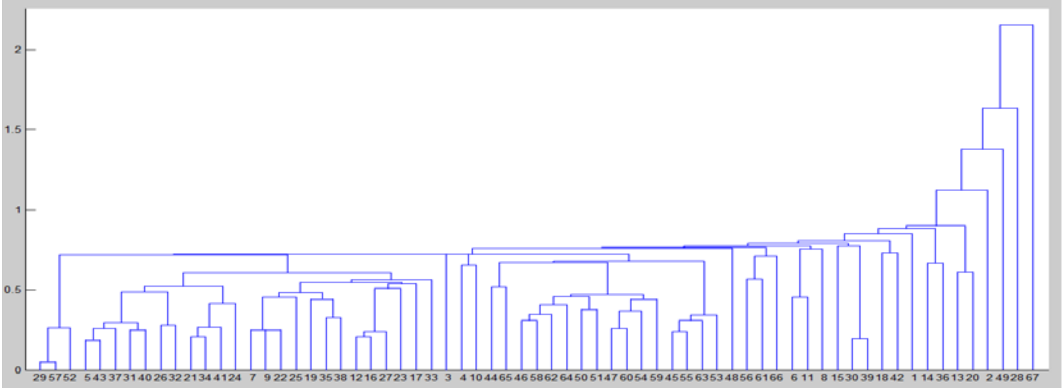

Figure 5.

Hierarchical Clustering Tree of Requirment 1.

We take the midpoint of the scale in the figure as the threshold. So the threshold is 1.

Now we get:

Group1: 2, 49, 28, 67

(2, 49, 28, 67 in this figure corresponds to No.3, No.57, No.32, No.81 in the actual scenario)

Conclusion of this section:

The discriminate line between conspirator and non-conspirator is 1. And No.3, No.57, No.32, No.81 are the members with the highest possibility of being a conspirator.

Leader(s)

So as the former case, to catch the leader we get hierarchical clustering tree.

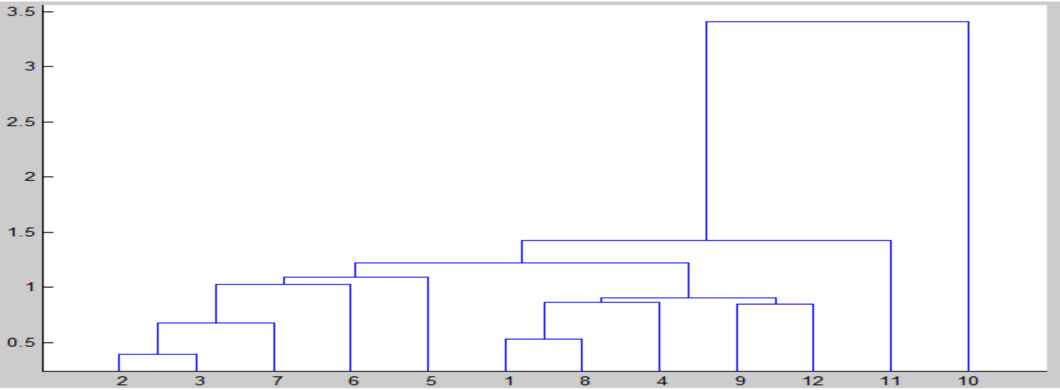

Figure 6.

Hierarchical Clustering Tree 2 of Requirment 1.

Now we get

Group 1: 10

Group 2: 1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 12

(No.10 in this figure corresponds to the member of No. 57.)

So No.57 still has the largest possibility to be the crime gang leader. In addition, No. 32 is Gretchen_manager who is one of the senior managers of the company, is still highly considered to be conspirator.

Conclusion of this section:

No.57 has the largest possibility to be the crime gang leader.

3.4 Comparison between Requirement 1 and Requirement 2

Chris (No.0) is considered to be a non-conspirator in the first case, and in the second case changed to be conspirator. Besides, Topic 1 is connected to the conspiracy in the second case.

In the second case, the group of our possible conspirators is the subset of the first case. And both of them get the results that No. 57 has the largest possibility to be the crime gang leader and one of the senior managers Gretchen considered to be involved in the conspiracy.

So we can reach the conclusion that our model can get reliable information and has a good accuracy.

4. Evaluation and Optimization

4.1 Strengths and Limitations of Our Model

Strengths

Our models exhibit the following positive characteristics:

Comprehensiveness. Not only do we abstract the definition Edge Weight from the available data to mark the network's unique crime characteristics, but also analyze the network combined with some universal characteristics from the perspective of the Social Network Analysis. Both the generality and particularity are taken into account.

Dynamics. We make full use of the data available to form several reasonable metrics which indicates characteristics of this network. Furthermore relationship between the metrics is not a simple hierarchical structure, but a more complicated network structure. By using the classic ANP method, we successfully implement the dynamic nature of these metrics.

Extendibility. Another strength of our model is that it would be relatively easy to add metrics to our model, which can improve accuracy. By analyzing the new information of the network, we can get new metrics. As long added to Model 1, the new metrics are fully considered in the whole analysis process. Thus, we can enhance our models conveniently.

Accuracy. After introducing data into the models, the results we get are compendious and in line with expectations. What is more, we use our model to re-analyze the previous example given by the supervisor, and the results match well with the actual situation. This also proves that our model has good accuracy.

Succinctness. The process to solve this complex problem is divided into three steps.

Each step employs a model to complete only one goal. The whole process retain succinct and understandable.

Limitations

Our model also shows the following drawbacks:

Due to the limited nature of the data, we find three metrics affecting directly the edge weights. The smaller number of metrics may to some extent affect the accuracy of the model. If more information is provided, we can abstract more metrics to reflect the network characteristics.

The ANP method in Model 1 is a good combination of qualitative and quantitative analysis, and it gives the weights conveniently. But it possesses a certain subjective.

In order to simplify the problem, the three normalized metrics are considered to be in the status on determining the priority. Yet, these three metrics have ambiguous relationship in practical applications based on the specific circumstances. This may affect the accuracy of the ranking of priority.

In the studies of Crime Network Analysis, the feather is often discussed that the network changes over time. It has been found the stereotypical impression of hierarchical organizations within organized crime is being replaced by an image of more fluid and flattened networks. However, because useful information in this aspect is not provided, our models do not look into this aspect [2].

4.2 Optimize Models Using Text and Semantic Network Analysis

We successfully model the Edge Weight between two talkers by the ANP, using three metrics that are Identity of two direct talkers, Topics between two direct talkers and Frequency of talks. We treat one person as one node , change the Edge Weight into PI. The correctness of the whole models largely depends on the accuracy of Model 1. However, due to the limit nature of the information available, correctness of Model 1 is inevitably undermined. If we can take advantage of the knowledge of the semantic network analysis and text analysis which is bound to help us get more information for modeling of Model 1, the accuracy of the model will be improved. The content below will discuss this aspect.

The term text analytics describes a set of linguistic, statistical and machine learning techniques that model and structure the information content of textual sources for business intelligence, exploratory data analysis, research, or investigation. General, it was used for the combination of text content profiles, found hidden in the text object structure and similarity.

A semantic network is a network which represents semantic relations between concepts. This is often used as a form of knowledge representation. It is a directed or undirected graph consisting of vertices, which represent concepts and edges [9]. The Semantic Network is an important step in text analysis, by providing the text a concise and accurate summary, keyword and extracting text useful information.

To optimize out model with these two techniques, we can use the following steps:

- Select suitable textual materials: Because the text of the greater amount of information, the more meaningful information we can get. We don’t use one message as a text to analyze. We take all the message Communicated between two people as a session. And we use the session as the text to analyze .In this way, we can get more meaningful information about two people’s Communication.

- Extracting textual materials useful information: Use the semantic network to get the text’s useful information such as accurate summary and text’s keyword.
- Get similarity and found hidden between textual materials: we can use the General method of text analysis to get similarity and found hidden between sessions.
- Mining sensitive words：We can get the “sensitive words library” provided by the police. Since we get the keyword at step (2), we can refer to the “sensitive words library”, as important feature of the textual material.

With the step above, we can use the information we get to optimize Model 1 by updating the network structure. Figure 7 shows the optimized network structure of Module 1. (We use the information as the session’s overall merit)

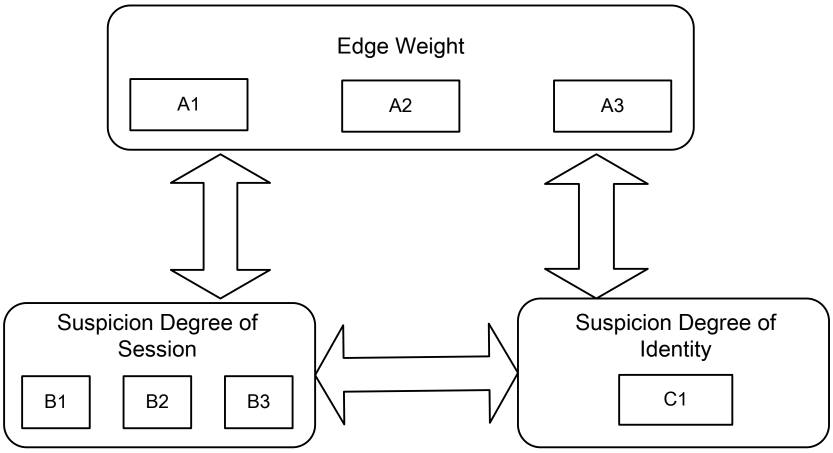

Figure 7.
The Optimized Network Structure.

where notations are defined in Table 7 below

Table 7.
Symbols used.

Abbreviations	Meaning
A1	Session’s overall merit
A2	Identity of two direct talkers
A3	Frequency of talks
B1	Sensitive words
B2	Similarity and found hidden between texts
B3	Sensitive words
C1	Suspicion Degree of Identity

4.3 Optimize Model for Future Use

In fact, most of the algorithm of text analysis is based on the clustering. We discuss an outstanding clustering algorithm which is called DBSCAN. DBSCAN has a good performance in clustering spatial data, it is designed to discover clusters of arbitrary shape, and requires only one input parameter and supports the user in determining an appropriate value of it. The average run time complexity of DBSCAN is only $O(nlgn)$(n is the number of objects in the database). Some software such as TRS can support 10000 textual materials analysis per minute. In multi-machine distributed cluster environment, it can support to process one billion of the amount of data.

In additional, our model only forces attention on the Edge Weight. If someone wants to use this model on other case, he only need to find a relation between each node such as atoms bound and the node’s Identity such as healthy cells or unhealthy cells and our model would works very well.

5. Conclusions

This article aims to explain a method using limited information to model the Criminal Network identify complicit in the crimes and criminal leaders. We believe that the combined knowledge of graph theory and mathematical modeling is a good method to solve the problem. At the view of Graph theory, we first model the weights in the graph. In order to characterize the extent of contact between the talkers and the crime information, the identity of both sides, as well as the topic of conversation influence the weights, and these two factors also influence each other, our model uses Analytic Network Process to objectively reflect the actual situation. After getting the right value, we use the theory of centrality which is widely well known for the field of Social Network. Degree, Closeness and Betweenness are combined as a threshold, and Cluster Analysis is used to obtained criminal conspiracy; reusing the weights of edge which is direct to every node, and take it as a standard, Sort of non-suspects, thus identify the criminal leaders, the model takes full advantage of the limited conditions, which can dynamically change with the conditions change, given the discussion under the increase of criminals and suspected topic. Next, we discuss the role of semantic networks, and text analysis in modeling. Finally, use them to improve the model. In this paper, we discuss the complexity of the algorithm and the promotion of the network model finally.

Appendix

Table 1.
Abnormal Data and Approaches to Deal with It.

Location	Description	Problem	Approach
File: Messages.xls Row: 215	start node:14 end node:25 topic 18	There is no topic 18 in Topics.xls	Delete this item.
File: Messages.xls Row: 322	start node:3 end node:3 topic 11	Talking to oneself does not make sense	Delete this item.
File: Messages.xls Row: 67	start node:30 end node:30 topic 8	Talking to oneself does not make sense	Delete this item.
File: Names.xls Row: 6&24	Cretchen	Same Names repeat	As it is known that Cretchen is a senior manager, assume that the one with a larger node degree in the graph is the senior manager. With calculation, assume No.32 is Gretchen_leader, andNo.6 is Gretchen_civilian.
File: Names.xls Row: 18&36	Jerome	Same Names repeat	As it is known that Jerome is a senior manager, assume that the one with a larger node degree in the graph is the senior manager. With calculation, assume No.34 is Jerome_leader, and No.16 is Jerome_civilian.
File: Names.xls Row: 9&39	Elsie	Same Names repeat	As it is known that Elsie is a conspirator, assume that the one is considered as the conspirator who talks to well-known conspirators and talks about suspicious topics frequently. With calculation, assume No.7 is Elsie_conspirator, and No.37 is Elsie_civilian.

Table 2.

Unweighted Super-matrix of Model 1.

Unweighted Super-matrix	TT	IT	FT	SDT	SDI
TT	0.00000	0.00000	0.00000	0.64912	0.27894
IT	0.00000	0.00000	0.00000	0.27895	0.64913
FT	0.00000	0.00000	0.00000	0.07193	0.07193
SDT	1.00000	1.00000	1.00000	1.00000	1.00000
SDI	1.00000	1.00000	1.00000	1.00000	1.00000

Table 3.

Weighted Super-matrix of Model 1.

Weighted Super-matrix	TT	IT	FT	SDT	SDI
TT	0.00000	0.00000	0.00000	0.43275	0.18596
IT	0.00000	0.00000	0.00000	0.18579	0.43276
FT	0.00000	0.00000	0.00000	0.04795	0.04795
SDT	0.66667	0.66667	0.66667	0.00000	0.33333
SDI	0.33333	0.33333	0.33333	0.33333	0.00000

Table 4.

Limited Super-matrix of Model 1.

Limited Super-matrix	TT	IT	FT	SDT	SDI
TT	0.19795	0.197951	0.197951	0.197951	0.197951
IT	0.17327	0.173278	0.173278	0.173278	0.173278
FT	0.02877	0. 02877	0. 02877	0. 02877	0. 02877
SDT	0.35000	0.350000	0.350000	0.350000	0.350000
SDI	0.25000	0.250000	0.250000	0.250000	0.250000

References

[1] Richard_Quinney , The Social Reality of Crime, Little, Brown, 1974.

[2] Jennifer Xu, Byron Marshall, Siddharth Kaza, and Hsinchun Chen, Analyzing and Visualizing Criminal Network Dynamics:A Case Study, INTELLIGENCE AND SECURITY INFORMATICS Lecture Notes in Computer Science, 2004, Volume 3073/2004, 359-377, DOI: 10.1007/978-3-540-25952-7_27

[3]Jennifer Xu , HsinchunChen, Criminal network analysis and visualization, Communications of the ACM ,
Vol. 48 No. 6, Pages 100-107 10.1145/1064830.1064834

[4] Saaty T.L., Decision Making With Dependence Feedback: The Analytic Network Process, RWS Publications, Pittsburgh, 2001.

[5] Rachung Yu , Gwo-Hshiung Tzeng ,A soft computing method for multi-criteria decision making with dependence and feedback, Applied Mathematics and Computation 180 (2006) 63–75

[6] Gilks W R , Richardson S , Spiegelhalter D J ,Markov Chain Monte Carlo in Practice. London :Chapman and Hall ,1996.1

[7] Freeman, L.C. (1979). Centrality in social networks: Conceptual clarification. Social Networks, 1 , 215-240.

[8] http://en.wikipedia.org/wiki/Mahalanobis_distance

[9] John F. Sowa (1987). "Semantic Networks". In Stuart C Shapiro. Encyclopedia of Artificial Intelligence. Retrieved 2008-04-29.).