Team Control Number

15356

For office use only	For office use only
T1 ________________	F1 ________________
T2 ________________	F2 ________________
T3 ________________	F3 ________________
T4 ________________	F4 ________________

Problem Chosen

C

iRank Model: A New Approach To Criminal Network Detection

Summary

How to detect all the members and the leader of a conspiracy to commit a criminal act has long been the major concern of the Intergalactic Crime Modelers (ICM). The previous method is far from efficient for the current fund fraud conspiracy which involved over 21,000 words of message traffic. Here we will develop a more reliable network analysis model for large volumes of crime conspiracies data and other kinds of network data.

With the assumption of neglecting the time effect of the communication and assigning specific weightings to conspiratorial and irrelevant messages respectively, we develop an iRank rating model to unearth the hidden structure of the criminal network in the current fund fraud conspiracy. It is a modified version of PageRank algorithm which considers both the conspiratorial communication records and the communication density in conspirators network to determine the ranking of conspirators. Also inspired by Social Network Analysis clustering, the model contains a closeness factor to separate the conspirators and non-conspirators and the factor can help us detect the leader of the conspirators. The model outputs the suspicion level of each suspect quantitatively as a priority list.

To further improve the models, we take other elements like time series and contents of the messages into consideration. In the advanced criminal network detection model, we can detect the initiator of all the conspiratorial topics thus to lock the major suspect and avoid suing some innocent people who unconsciously spread conspiratorial in the network. Moreover, we will demonstrate how semantic network analysis and text analysis can improve the accuracy of the judgments by detecting some well-hidden conspirators like Inez and Bob in the first example.

In the final step, we validate the results by setting a critical value to the iRank value through conspirator group size estimation and visualization. Furthermore we will analyze the strengths and weaknesses of the models comparing with PageRank algorithm and Social Network Analysis. In addition, we discuss how the model can be extended to other social network applications like biological systems

Introduction

Detecting criminal network within large amount of data is a well-studied problem in the real world. It is especially important to develop techniques for uncovering conspiracy networks involving white-collar crimes. In most of the cases, such well-organized criminal activities will follow some patterns. Thus we can uncover the structures of this kind of criminal network and nominate the leader of the group by studying reliable data with sophisticated techniques. It would save a lot of endeavors and time for the ICM to conduct their investigation and arrest work in the future.

In the given ICM case, it is known that some conspirators are taking place to embezzle funds from the company and use internet fraud to steal funds from credit cards of people who do business with the company. Here, our goal is to separate the non-conspirators from the ones who are most likely to be conspirators. We will consider:

- the development of criteria and methods to detect the criminal network and the leader of the group
- the application of semantic network analysis and text analysis to improve the method
- further recommendations and other applications of the model

Dataset Observations and Basic Statistical Analysis

To better understand different communication behaviors of conspirators and non-conspirators and to elaborate our assumptions, we conduct a statistical analysis for the given data. In task 1, given that Jean, Alex, Elsie, Paul, Ulf, Yao, and Harvey are conspirators while Darlene, Tran, Jia, Ellin, Gard, Chris, Paige, and Este are innocent, we reach some useful findings for model building. There are two "Elsie's" in the company, No.7 and No.37. As the No.7 obviously has more connections with the other known conspirators, we lock No.7 as the conspirator.

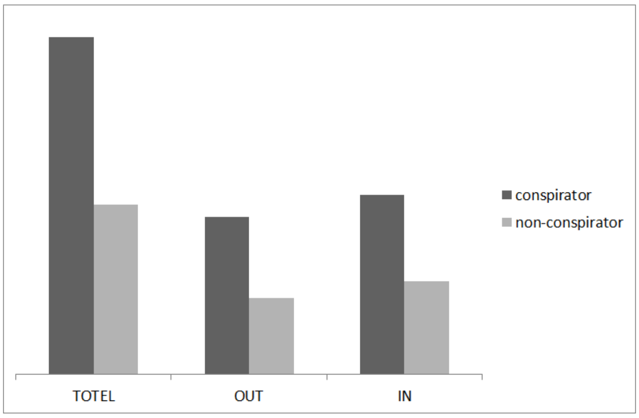


Figure 1: Comparison of number of topics conversed by conspirators or non-conspirators

In Figure 1, from the total number of conversations (left bar), we can see that conspirators are significantly more active than non-conspirators. They tend to communicate more often with other people. The information exchanges among known conspirators groups are also significantly more frequent than that among known non-conspirators.

Carefully examining into the patterns of information exchanges and social connections in the network, we can see that only 24% messages carry conspiratorial information, which seems not systematically significant given that 20% of all the topics are conspiratorial. Therefore, two patterns can be inferred from the statistical results:

- Although conspirators are generally more active than the known innocent people, they exchange irrelevant information with each other. Conspiratorial messages only take a small portion in their message traffic.

- Since the existing 7 conspirators have already involved in spreading about 40% of the total conspiratorial messages, it is very likely that the total number of conspirators is less than 20.

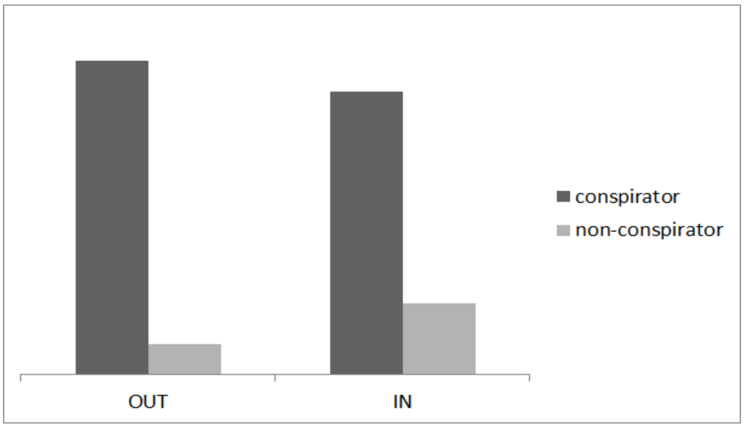


Figure 2: Messages with conspiratorial topics conveyed by conspiracies and non-conspiracies

Another pattern we can derive from the original dataset is the correlation between the involvement of conspiratorial activities and the identity of the worker. We observe a few non-conspirators who have involved in talks with conspiratorial topics. Nevertheless, most of the non-conspirators only receive those messages and seldom give responses to them. Thus, the initiators of such a conversation should have more suspicion. Therefore, we can assume that the motivation of participating in conspiratorial topics is one of the most important indicators of a given worker's identity.

Symbols and Definitions

Symbol	Definition	Formula
G	Society network graph	
N	A set of labeled nodes	
n_{c}	Nodes represent the conspirators	
n_{u}	Nodes represent the unknown	
n_{n}	Nodes represent the non-conspirators	
l_{ij}	The message sent from node *i* to node *j*	
L	A set of labeled links	
L_{c}	Links contain conspiratorial messages	
L_{n}	Links contain irrelevant messages	
D	Degrees of each node	D(n) = Number of links connected to node n
O	Out-degree	O(n)= Out-degree of node n
I	In-degree	I(n) = In-degree of node n
CL	Centrality of each node	CL(n) = Centrality of node n
IR	iRank model value	IR(n) = Ranking weight of node n
α	The link weight between two given nodes	α_{ij} =Link weight between the *i*th and *j*th node
f	The heuristic function	*f*(*l*) = Heuristic function of link *l*
ms_n	Number of messages the node n sends n_c to and receives from n_c	
mr_n	Number of messages the node n receives from n_c	
ts_n	Number of times the node n sends l_c	
tr_n	Number of times the node n receives l_c	
w	Adjusting factor used to standardize the units into a same scale	
hs	Harmonic series of the number of times that node n sends a conspiratorial message to a known conspirator	
hr	Harmonic series of the number of times that node n receives a suspicious message from a known conspirator	
d	Heuristic function of closeness	d(n)= Heuristic function of closeness from node n to n_c and n_n

Assumptions

The criminal network problems can be really complicated if we take every effect into consideration. In Task 1 and Task 2, we simplify the model by assuming that:

- Only the7th, the 11st and the 13rd topics are related to conspiracy in Task 1, and in Task 2 the 1st topic is added to the topics related to conspiracy;

- All the messages are exchanged in a very short period, thus the impact of time can be neglected;

- The contents of the messages can be temporarily ignored, thus all the conspiratorial topics are equally weighted.

- A message that involves k (k>1) topics is equivalent to k messages that each involves 1 corresponding topic. This is valid because we ignore the time effect of communication, and focus on the amount of information exchanged only.

Meanwhile, according to the basic statistical results of the dataset, we can have the following assumptions.

- Non-conspirators do not know about who are conspirators.
- Non-conspirators seldom talk about conspiratorial topics with conspirators.
- Conspirators do talk conspiratorial topics with non-conspirators.
- The identity of an unknown node is determined by its neighboring nodes and the links incident with it.

Task 1

The Mathematical Model — iRank Model

The aim of the task is to obtain a priority list according to the likelihood of being part of the conspiracy and to determine whether any of the senior managers are involved. In this task we develop an ***iRank Model*** which is a combination of PageRank Algorithm and SNA technique. We apply this modified model in this problem as the original PageRank Algorithm cannot deal with links with different weights and the SNA technique does not take the identities of the nodes into consideration(Xu and Chen 2005).

For the likelihood of being a conspirator, intuitively a person's suspicious level relates to the percentage of conspirators he contacts and the percentage of suspicious messages he involves in. Furthermore, a person seems even more suspicious when he sends a suspicious message to a known conspirator, or receives a suspicious message from a known conspirator. Therefore we can consider a function that ranks each suspect by the factors mentioned above as our selection criteria to find out possible conspirators.

In addition, in a normal social group the social activities should be evenly emerged along with the organizational structure to a certain extent. We believe the conspirator group as a sub-group in this company would cause abnormal social activity patterns reflected on their behaviors of communication. Specifically, based on *Small World Theory(Natarajan 2006)* which raised the relationship closeness of any two people among a social group, we pay attention to patterns of all n_u connect to the conspirator group as well as the non-conspirator group. By our model, the abnormal distribution of social activities within the company caused by conspirator group can be tracked and related useful information, which helps us to distinguish people's identities, can also be derived from it.

To determine the conspiracy leaders, we will iteratively review how a person makes influence on the conspirator groups, or *the degree of centrality* we defined as follow, to find out a person's impact among known conspirators.

The iRank Model includes two steps: initialization and iteration.

Step 1: Initialization

The initialization offers a initial suspicious level to all nodes with unknown identity n_u. Consider the iRank value IR(n) as the suspicious level of node n. Intuitively we have:

IR(n) = Suspicion raised by the frequency of contacting n_c
+Suspicion raised by the frequency of exchanging l_c
+ Suspicion raised by the communication distance to n_c (1.1)

Let ms_n, mr_n denote the number of messages the n-th node sends to and receives from n_c, respectively, and let ts_n , tr_n denote the number of times the n-th node sends and receives l_c, respectively. Also let C_n be the centrality of node n and $CL(n_c)$, $CL(n_n)$ denote the closeness from node n to n_c and n_n.

Assume each part in (1.1) plays an equal importance in detecting conspiracy. Let w_{ms}, w_{mr}, w_{ts}, and w_{tr} denote the adjusting factor used to standardize the units into a same scale so that each part is assigned a same weight in IR(n). Therefore we can set up the following ***iRank function*** to assign *an initial weight* to each node.

$$IR(n) = \begin{cases} 0, & n \text{ is not a conspirator} \\ 1, & n \text{ is a conspirator} \\ \dfrac{max\left(\frac{ms_n}{O(n)} \times w_{ms}, hs(n)\right) + max\left(\frac{mr_n}{I(n)} \times w_{mr}, hr(n)\right) + \frac{ts_n}{O(n)} \times w_{ts} + \frac{tr_n}{I(n)} \times w_{tr} + d(n)}{5}, & \text{otherwise} \end{cases}$$

where *hs(n)*, *hr(n)* are the harmonic series of the number of times that node n sends a suspicious message to a known conspirator, or receives a suspicious message from a known conspirator, and

$$d(n) = C_n \times \frac{CLn_c}{CLn_n}$$

$$\text{where } C(n_i) = \left[\sum_{j=1}^{g} d\,ist\,(n_i, n_j)\right]^{-1}$$

is a heuristic function of closeness from node n to n_c and n_n. Statistical analysis shows that the strong positive correlation of closeness to conspirator group and the closeness to non-conspirator group, expect that a few nodes demonstrating significantly more closeness towards conspirator group against non-conspirator group as following graph shows.

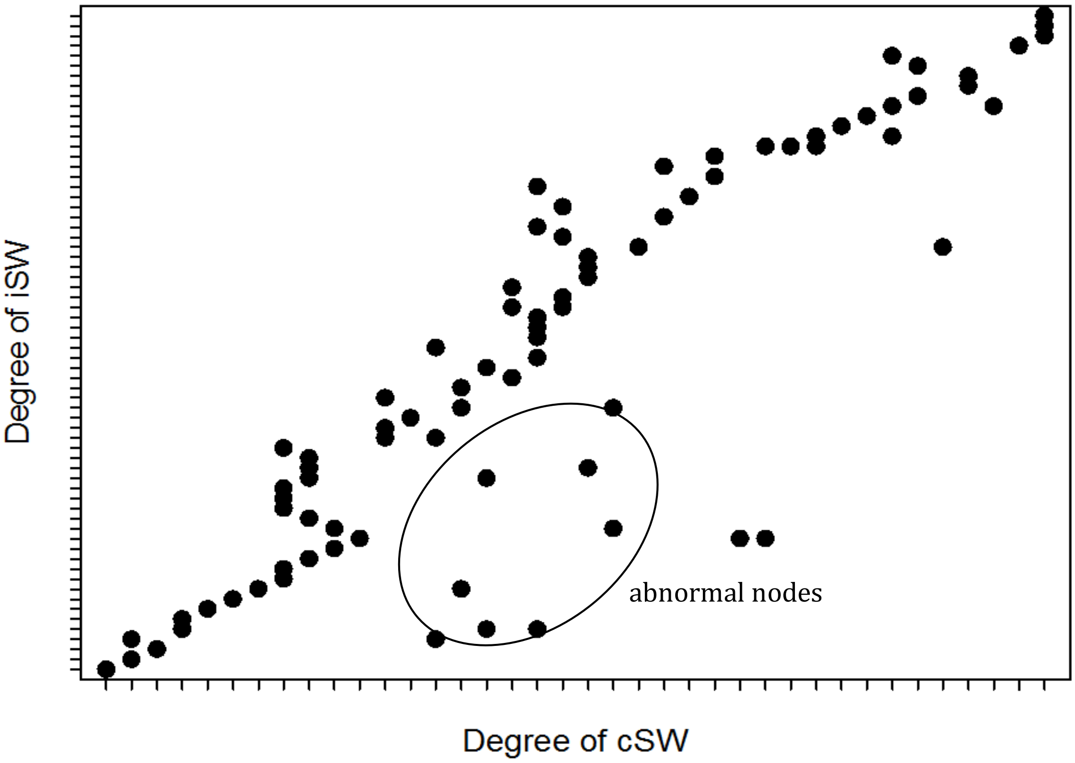


Figure 3 Abnormal nodes observed by correlation of closeness to conspirator group and non-conspirator group

Step 2: Iteration

After obtaining the initial value, we can iteratively adjust the ranking weight of each node to get a more precise iRank value because for a node n its suspicious level *IR(n)* changes as the iRank values of its neighboring nodes have changed. Consider a rating system that contacting with a more suspicious node will results in a higher *IR(n)*, we can set up the following rating function:

$$IR(n) = \begin{cases} 0, \quad n\ is\ not\ a\ conspirator \\ 1, \quad n\ is\ a\ conspirator \\ \sum_{x \in adj(n)} IR(x) \times \alpha_{n-x} + \sum_{n \in adj(x)} IR(x) \times \alpha_{x-n}\ , otherwise \end{cases}$$

where adj(n) denotes a node that receives a message from the n-th node, and α_{i-j} denotes the weight of the message from i-th node to j-th node that satisfies $\sum_{i \in adj(j)} \alpha_{i-j} = 1$.

By Markov property, for every $n \in N$, *IR(n)* will eventually reach the limit after a large number of iterations, and the final *IR(n)* will be a credible estimate of the suspicious level of the node.

Estimation of Parameters

By analyzing the sample data, we have the following statistical results:

Task 1 Statistics							
Sender	Topic	Receiver	Count	Sender	Topic	Receiver	Count
n_c	--	n_c	26	n_c	l_c	--	31
n_c	--	n_n	6	n_n	l_c	--	3
n_n	--	n_c	5	--	l_c	n_c	28
n_n	--	n_n	10	--	l_c	n_n	7

Table 1: The statistics of suspicious action counts in Task 1

Assume that the sample distribution is coherent with the total distribution, based on the observation on the sample data, we can find out

$$\begin{cases} w_{ms} = 1.2 \\ w_{mr} = 1.2 \\ w_{ts} = 1.1 \\ w_{tr} = 1.25 \end{cases}$$

As the messages including suspicious topics are more useful for our detection of conspirators than irrelevant messages, we can define

$$\alpha_{n-x} = \frac{1}{\mathrm{O(n)} + \mathrm{ms_n}} + \frac{\mathrm{I_{n-x}}}{\mathrm{O(n)} + \mathrm{ms_n}}$$

where $\mathrm{I_{n-x}}$ is the indicator function of whether the message from node n to node x contains suspicious topic, that is

$$\mathrm{I_{n-x}} = \begin{cases} 1, & \text{the message is suspicious} \\ 0, & \text{the message is not suspicious} \end{cases}.$$

Output and Evaluation

Node #	81	51	16	33	57	60	28	79	10
iR	0.8571	0.6244	0.6142	0.5765	0.5667	0.551	0.5377	0.5141	0.4790

Table 2: Significant suspects ranked by IR in Task 1

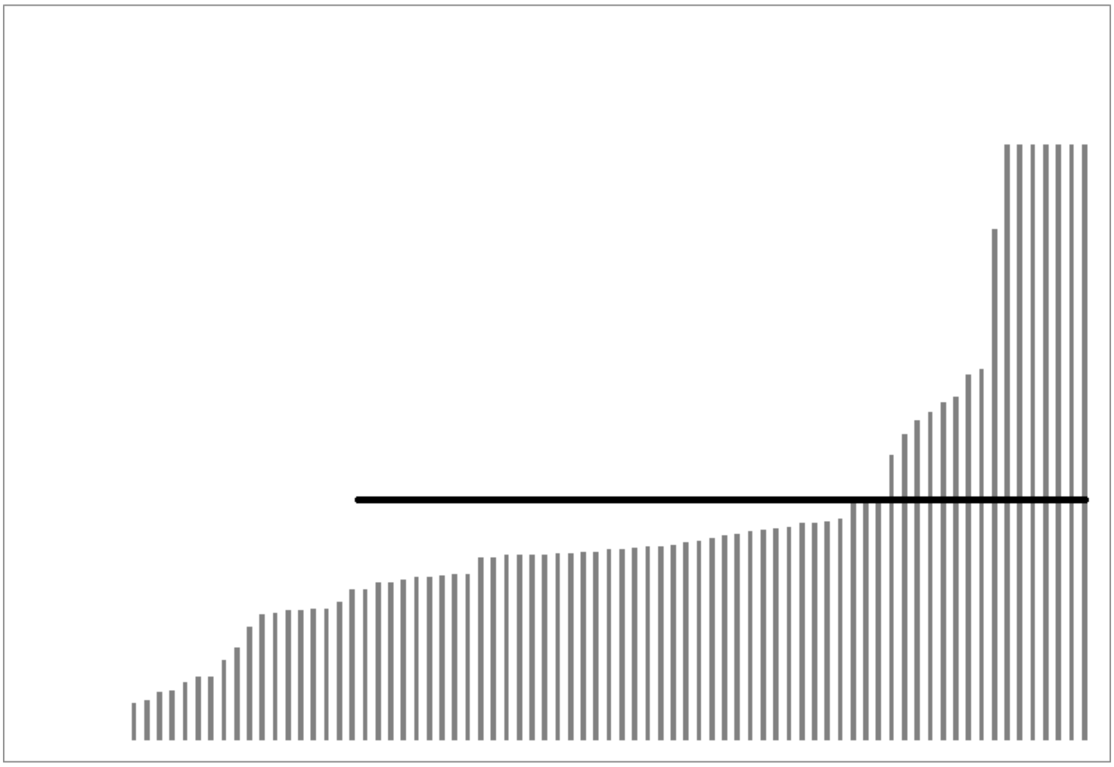


Figure 4: Suspicious level shown by IR

To distinguish leaders from the conspirator group we found by our iRank model, we further develope the analytical model to demonstrate the leadership within the group. Firstly, we make following assumptions about the behavior of the leader in a group:

- The leader usually acts as an intermediate node to connect different functional sub-groups

- The leader prefers to communicate with heads in different functional sub-groups rather than common members.

- Sub-group heads, as an intermediate node among the leader and the other members, can access all of their group members.

From the above assumptions, the following facts can be inferred:

- Normally, the leader can achieve one of the highest neighborhood connectivity among all members, since the leader can connect to all members through those sub-group heads.

- The leader may not have the smallest average shortest path length since the number of members in different sub-groups may differ greatly.

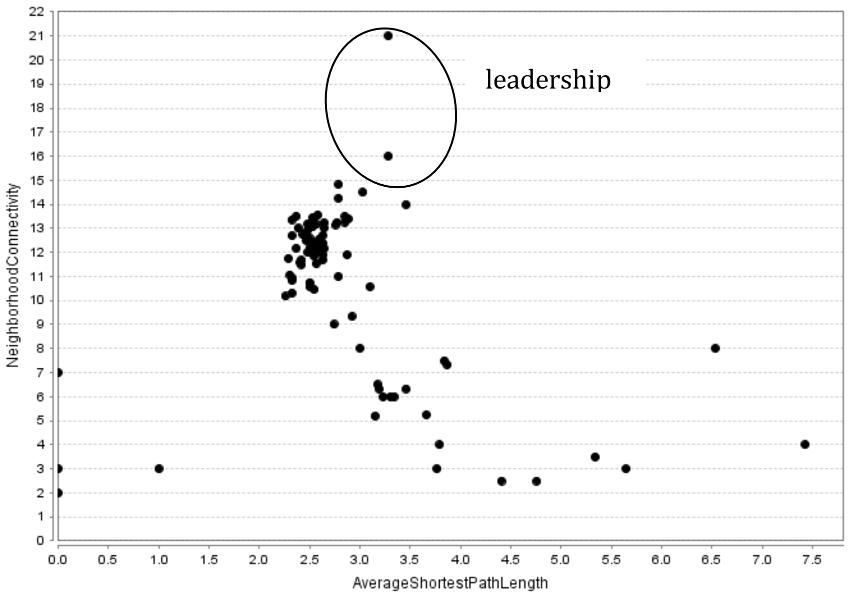


Figure 5: the correlation between neighborhood connectivity and average shortest path length

Obviously, from the chart we infer that those two nodes showing abnormal patterns are very likely the leaders of the whole group. There are No.16 Jerome and No.10 Dolores.

Task2

Adjusting to the iRank Model

We can apply the same model illustrated in Task 1, but we should calculate the new parameters according to the new condition added.

Task 2 Statistics							
Sender	Topic	Receiver	Count	Sender	Topic	Receiver	Count
n_c	--	n_c	29	n_c	l_c	--	38
n_c	--	n_n	9	n_n	l_c	--	5
n_n	--	n_c	6	--	l_c	n_c	33
n_n	--	n_n	3	--	l_c	n_n	9

Table 3: The statistics of suspicious action counts in Task 1

From the sample statistics we can find

$$\begin{cases} \mathrm{w_{ms}} = 1.3 \\ \mathrm{w_{mr}} = 1.2 \\ w_{ts} = 1.13 \\ w_{tr} = 1.27 \end{cases}$$

Output

By iterating the iRank function 1000 times, we obtain the following results.

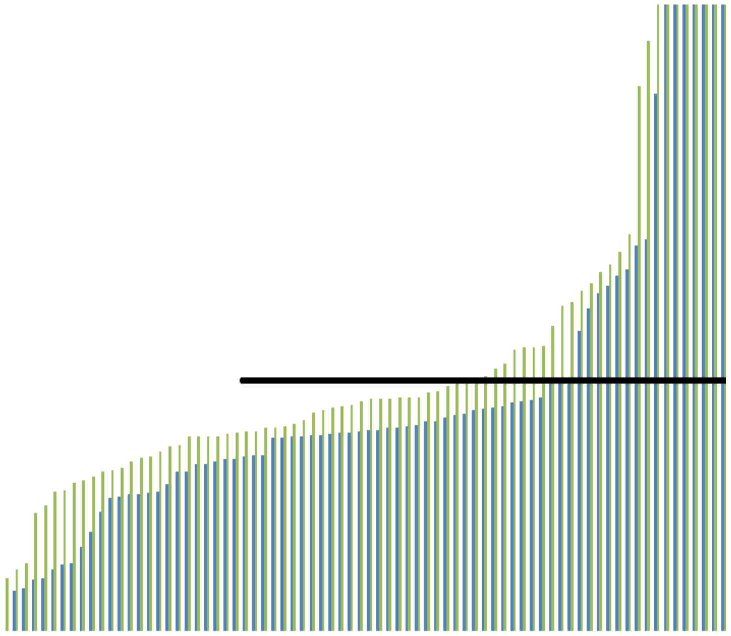
Node #	iR	
16	0.94099	
81	0.86962	
51	0.63188	
56	0.60522	
33	0.58404	
57	0.57165	
60	0.55354	
28	0.54328	
10	0.52434	
79	0.51788	
69	0.48713	
13	0.45334	
17	0.45170	
20	0.45149	
22	0.44896	
3	0.42529	
15	0.41772	

Table 4: The significant suspects ranked by iR in Task 2

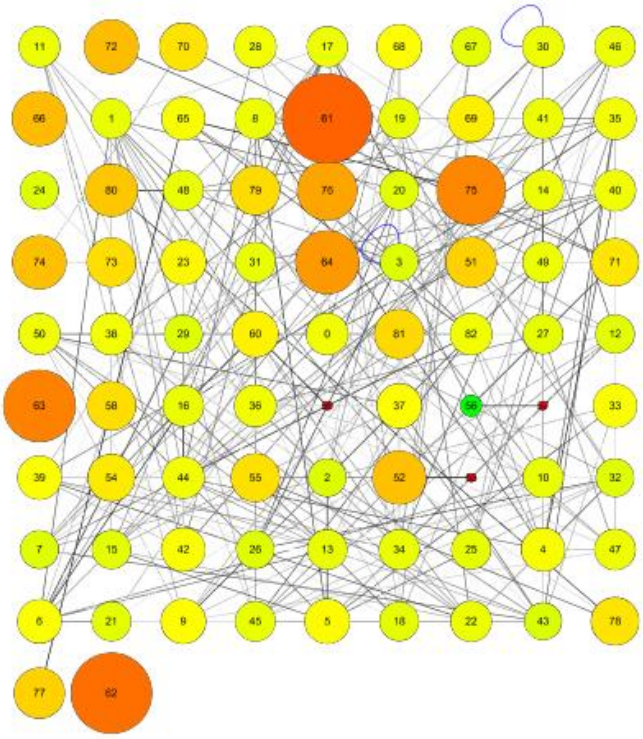


Figure 6: Visualization of the Criminal Network Based on Task 2 Results

Evaluation and Discussion

Strength

- The iRank model perfectly distinguishes every different node and ranks the suspicion level of all nodes quantitatively because *IR(n)* considers both the suspicious communication made by node n and the communication density of node n in a social network. For example, in the data set both Node 16 and Node 34 are called Jerome, but the model indicates that Node 16 is the senior manager and further shows that Node 16 is involved in the conspiracy.

- The iRank model generates an appropriate initial value for each node using all the information known from the data set, which is better than the original Page Rank algorithm that generates the initial value randomly(Graham and Tsiatas 2010).

- The iRank model keeps track of the information flow by following the numerical node weightings and link weightings, which is not considered in general Social Network Analysis clustering(Coffman, Greenblatt et al. 2004).

- The iRank model is highly efficient in time complexity and space complexity because it can dynamically adjust the ranking of each node by iteration without performing high dimensional matrix operations by iterations.

Weakness

- The data structure of communication ignores the timing and the sequence of messages, causing the information loss at the beginning stage. Our iRank model purely regards that, all conversations within the social group are organized as a static directed path of which a node denotes a group member and an edge denotes a message. Obviously, the information of timing and message sequence is fairly helpful in busting up crime syndicates, e.g. it is believed that one initiating a message carrying suspicious topics is more conspiratorial than one replying it.

- Another major weakness of iRank model is that, our model is not able to indicate a critical value of conspirators and non-conspirators before reviewing the result of the priority list. Actually, to decide whether a person is a conspirator or not appropriately, we have to go over the data of results in detail and set the critical value manually based on our assumptions.

Task 3

Improvements on the Criminal Detection Model

In the above mathematical model, we assume every irrelevant topic is equally important, and we may ignore some underlying correlations between any two topics. Next we will improve our model using semantic network analysis and text analysis.

Semantic Network Analysis and Text Analysis

Semantic Network Analysis is a technique in which the content of a message is extracted from text and represented as a network of semantic relations between actors and issues, which can be queried to look for specific patterns and answer various research questions."(Morselli 2010) In our crime busting model, we can apply this technique to help us extract critical words or messages from the heavy message traffic.

As the original messages are not given, here we will just demonstrate our method following the process below. Meanwhile, we will show in detail how this improvement to the criminal network detection model can help lock Inez and Bob in the first example.

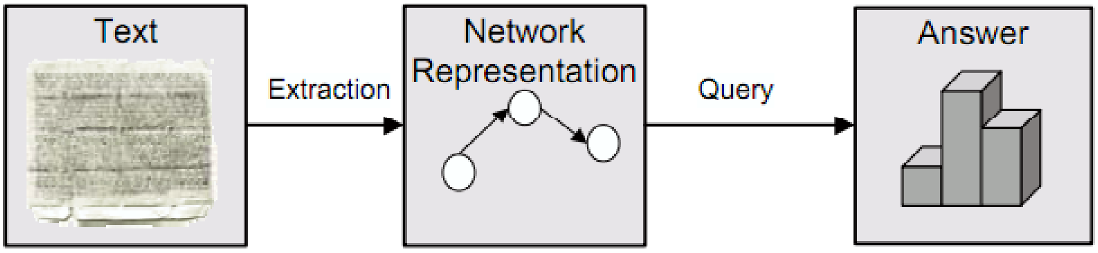


Figure 7: Semantic Network Analysis Work Flow

In the first step, we will extract some conspiratorial or informative messages out from the message traffic.

- According to some basic criminal psychology knowledge, we can assume that conspirators are usually under more pressures. We can ask the model to extract any phrases or words that can reveal the abnormal emotions of certain people. For example, in the first case, Inez mentioned two times that she was "tired" or "exhausted", while Jaye did not have "much going on". Harry also detected that George was stressed.

- Secondly, we should also extract messages in other language or which have some ambiguous statements. It is likely that those are used as codes within the conspirators.

- Contents or conversations which show high exclusiveness should be pay attention to, including the invitings to some private talks or meetings.

- Some messages which contain strong feelings should be extracted ant analyzed.

- Also, if the conversation or message has mentioned other people, we will extract the names and the related activities or descriptive words.

The Mathematical Model

The model applied is similar to the iRank model in Task 1, but link weight of link *l* $\propto_l$ is determined by a text analysis function f(m) instead of a constant that is related to topic involved only.

The text analysis function f(m) is judged by comparing the similarity with the message sent by n_c or n_i. Inspired by the principal of supervised learning(Wiil, Memon et al. 2010), we can set the initial link weight of l_c sent or received by n_c to 1, and set the initial link weight of l_i sent or received by n_n is -1. In this way f(m) is a value from continuous interval [-1,1], and a larger f(m) value implies a greater likelihood of being a suspicious message.

Hence we can rewrite the model in the iterative step as:

$$\mathrm{PR(n)} = \begin{cases} 0, \quad \text{n is not aconspirator} \\ 1, \quad \text{n is a conspirator} \\ \dfrac{\sum_{l_{nx}\in L} PR(x) \times \alpha_{nx} + \sum_{l_{xn}\in L} PR(x) \times \alpha_{xn}}{2}, \text{otherwise} \end{cases}$$

where:

$$\begin{cases} \alpha_{nx} = \max\left(\dfrac{\mathrm{f(m)}}{\sum_{\mathrm{l_{nx}\in L}} \mathrm{f(l)}}, 0\right) \\ \displaystyle\sum_{\mathrm{l_{nx}\in L}} \alpha_{nx} = 1 \end{cases}$$

Evaluation and Discussion

We believe the semantic network analysis and the text analysis can efficiently enhance our model by assigning an appropriate weight for each message according to its own message content rather than assigning a same weight on messages of different importance. For one hand, as a more irrelevant message owns a smaller weight and a more suspicious message owns a greater weight, the interactions of suspicious message flow will be clearer. On the other hand, an effective text analysis take the correlation among messages into account, which can provide more accurate link weights and help us find out the underlying relationship among the communication network so that we can get a more credible result.

The influence of a suspicious link on the corresponding node is strengthened in the improved model, which can be shown in the following simple numeric example. In the picture below we can see that in previous model the significant message only contributes 0.5 to IR(n), equal to the contribution of irrelevant messages, whereas in the improved model the significant message contributes 0.9 to IR(n) under same circumstance.

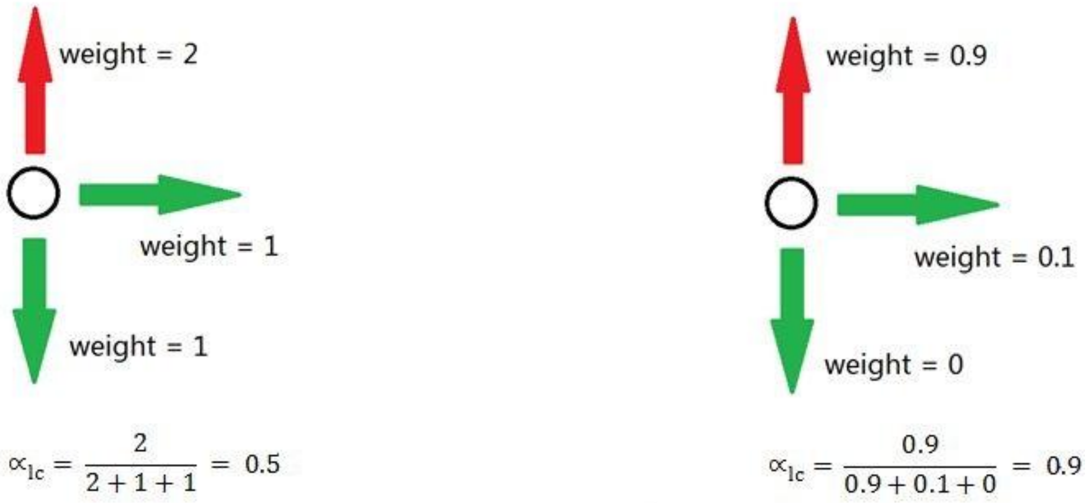

weight = 2
weight = 1
weight = 1
$\propto_{lc} = \frac{2}{2+1+1} = 0.5$
weight = 0.9
weight = 0.1
weight = 0
$\propto_{lc} = \frac{0.9}{0.9+0.1+0} = 0.9$

Figure 8: The suspicious link contributes more on IR(n) in the improved model

The correlation between messages can be found via semantic network analysis and text analysis. For example, in the topic description given, we can find that the suspicious topic 7 involves Spanish words as codes, and we may further induce that the Spanish words in topic 2 and topic 12 can also be suspicious. Also in topic 4, 5, and 6 we can see some negative feelings like anxiety and complaints, which might infer that the sender is suffering from guilty conscience.

Task 4

Recommendations and Future Development

The IR model performs well in separating the non-conspirators and the conspirators as well as detecting the leader in the criminal network. However, our models can be further improved by considering the following:

- Build a thorough network with more messages in the traffic with more links between any two nodes. It will increase the accuracy of the results of the model by considering more explicit interactions between the nodes.

- Introduce time series into the model. A clear timeline may help us detect the initiators of certain highly conspiratorial topics. It will also show the pattern changes in the network before and after a conspiracy occurred.

- Apply text analysis to deal with large volumes of data. Text analysis can help us in detecting conspiratorial messages or some abnormal expressions efficiently when the dataset is large.

- Introduce the semantic network analysis. With accumulation of database, we can uncover some usual tactics in high-tech conspiracy crimes. For example, some sudden changes in attitudes and conversational styles between two workers may indicate a conspiracy. Also, the increasing frequency of some anxious or stressed words may suggest a conspiratorial event is taking place in the company.

Other Applications

Besides the study of criminal network detection, we can use this model to deal with various network problems by adjusting the weighting parameters or adding new constraint equations. Here is an example about how this model can be implemented to find infected or diseased cell in biological network.

- The probability of getting infected is inversely proportional to the distance between one infected cell and other healthy cells(Chen, Ding et al. 2009), so the weight of being infected between two cells α_i can be seen as 1/distance.

- Given some known infected cells, we can assume the infection ability of different cells have some different probabilities IR(n).

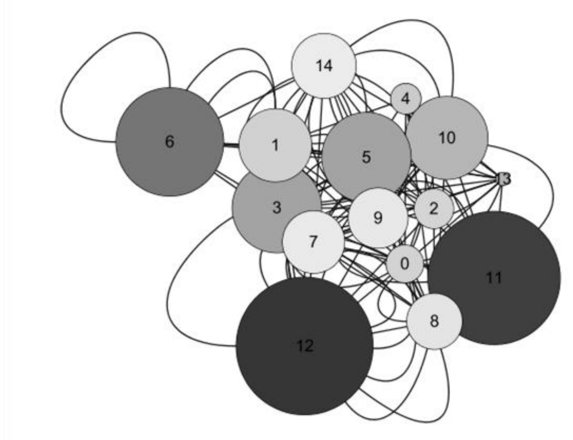


Figure 7: The Model of the Infection Cells Detection

References

Chen, H., L. Ding, et al. (2009). "Semantic web for integrated network analysis in biomedicine." Briefings in Bioinformatics **10**(2): 177-192.

Coffman, T., S. Greenblatt, et al. (2004). "Graph-based technologies for intelligence analysis." Communications of the ACM **47**(3): 45.

Graham, F. and A. Tsiatas (2010). Finding and Visualizing Graph Clusters Using PageRank Optimization
Algorithms and Models for the Web-Graph. R. Kumar and D. Sivakumar, Springer Berlin / Heidelberg. **6516:** 86-97.

Morselli, C. (2010). "Assessing Vulnerable and Strategic Positions in a Criminal Network." Journal of Contemporary Criminal Justice **26**(4): 382-392.

Natarajan, M. (2006). "Understanding the Structure of a Large Heroin Distribution Network: A Quantitative Analysis of Qualitative Data." Journal of Quantitative Criminology **22**(2): 171-192.

Wiil, U. K., N. Memon, et al. (2010). "Detecting New Trends in Terrorist Networks." 435-440.

Xu, J. and H. Chen (2005). "Criminal network analysis and visualization." Communications of the ACM **48**(6): 100-107.