Modeling for Crime Busting

Yifu Huang

Summary

The current case requires us to find possible suspected conspirators in the message traffic that exist some certain conspirators and suspicious topic. In fact, our work is to prioritize the value of nodes in a partial classified and value transformed network.

Model, the most importance of our solution:

Though models about probability theory such as Markov chain may be used to analyze our problem, due to the multi-topics feature of conspiracy network, we cannot directly apply Markov chain to it. So we have to compartmentalize conspiracy network by topic first. And then we apply Markov chain to analyze subnet one by one. In this way, we ignore influence of other topics. So we develop an appropriate mechanism to combine all results of subnets. Finally, considering important nodes have higher priority, so we upgrade their ranking.

We assign parameters to each topic to describe the extent of its conspiracy. Meanwhile we just remove known non-conspirators, and give different suspicious parameters to both known conspirators and the rest. We provide a method called CrimeRank to analyze conspiracy transform. In CrimeRank, we provide initial conspiracy matrix and conspiracy transform matrix to describe initial state and state transform of the network. Finally, we assign reasonable priority to important nodes.

By visualization, we test results in all phases, and we find our results are basically the same with the display of visualization. When the weight of topic varies, the ranking of the nodes will change corresponding severely. During CrimeRank, floating point by multiplying may cause overflow, thus it may affect state convergence. So, we select relative state before convergence to eliminate error.

Strengths: Compartmentalize the whole network by the kind of topic can reduce the scope and increase the computing speed. With very large databases of message traffic, we apply MapReduce architecture to our model. Compartmentalizing the whole network is equivalent to Map phase, while combing all results of subnets is equivalent to Reduce phase. So, large scale message traffic can be processed effectively in parallel. The adjustment based on important node can get more reasonable results.

Weaknesses: Sometimes value transform may lead to the weight of node changes. Our model is not suitable in this kind of condition. What's more, the conspiracy weight of topic is uncertain because of lack of original data. So we cannot get more close to reality unless there are experts to assign great value to these parameters.

Content

1 Introduction ……3
1.1 International research background ……..3
1.2 Introduction of relative work ……..3
1.3Motivation and Meaning ……..3

2 Requirement 1 & Requirement 2 …… 3
2.1 Motivations ……. 3
2.2 Assumptions …… 4
2.3 Basic Model …… 4
2.3.1 Formula and Principle …… 4
2.3.2 Model Steps …… 6
2.3.3 Model Metric …….. 8
2.4 Testing & Results …… 8

3 Requirement 3 …… 12

4 Requirement 4 …… 13

5 Strength & Weakness…… 13
5.1 Strengths …… 13
5.2 Weakness ……. 14

6 Future Work …… 14

7 Reference …… 14

1 Introduction

We were all shocked by Bernard Madoff's massive Ponzi scheme that defrauded thousands of investors of billions of dollars [1]. Since then effective tools of investigating crime is required by policy, and data mining and analysis of social network become one of the most popular method under research.

1.1 International research background

Nowadays, intelligent data mining and data analysis has becoming an important method for arresting the criminal group. Lack of needed data is no longer the problem we should face with. We must focus on how to build up models and find out approaches to mine useful information from number of dynamic data in using monitor and capture the criminals. However, the analysis of crime network focus on finding out crucial nodes and extracting the subnets, but ignore ranking the possibility of suspicion.

1.2Introduction of relative work

Computer professional's researches concentrate on developing technique and method in helping analyze criminal network. For instance: The i2 [2] Clarity Platform empowers government agencies and private sector businesses to investigate, predict, prevent and disrupt the world's most sophisticated criminal and terrorist threats. Qiao [3] Mining key members of crime networks based on personality trait simulation E-mail analysis system. WAYNE [4] analyze the social organization of three well-known price-fixing conspiracies in the heavy electrical equipment industry.

1.3Motivation and Meaning

As we speak of before, although researchers merely concentrated on finding out crucial nodes and extracts the subnets while fail to rank the possibility of suspicion. However, on contrary, it may benefit the case. The officials always hope to identify the other members and the leaders before they make arrests. In case that some of them are the senior managers of the company. It would be very helpful to know if any of them are involved in the conspiracy.

2 Requirement 1 and Requirement 2

2.1 Motivations

We are facing the problem that calculating the suspicious level for everyone in the conspiracy network. Though models about theory of probability would be used to analyze our problems, there is no suitable model for analyzing the suspicious level of members in conspiracy network. Due to the feature of conspiracy network, we could model conspiracy network as graph, and try to adjust existed model about probability to construct our own model. One of the most considerable models is the Markov chain. Markov chain has been used to calculate the probability of perfect page in information retrieval—PageRank—which is also a problem of calculating the

probability of graphic data. While the Markov chain is not a consummated model for our problem, Markov chain could not analyze the event that influenced by multi factors. That is to say that single transition probability matrix could not satisfy the current case. So we begin our work by consummate the Markov chain for the current case.

Reducing the scope of the problem is one of the most useful methods of solving complex problem in methodology. According to the fact that transition matrix should be changed by certain factor that influence the event. Our work behind solved this problem.

2.2 Assumptions

We assign w1, w2…w15 to each topic to describe the extent of its conspiracy, and we assume the known suspicious message topics' weights are 1 while others are 0.We set suspicious parameter 1 to known conspirators, 0 to uncertain people, and just remove known non-conspirators. We provide a method resembles CrimeRank to analyze link with a residual probability of d = 0.85.Fianlly, we think important nodes in the crime subnet should get special consideration, so we elevate priority of them in prioritized list.

2.3 Basic Model

2.3.1 Formula and Principle

In the current case, to model such a conspiracy network, we defined directed graph $G(V,E)$ consists of V ; a nonempty set of nodes and V_i represent a person and E , a set of edges. Each edge represents the communication link between a pair of members in the network.

1. According to the graph which represents the conspiracy network, we can start to analyze by the structure. Several centrality measures can be used to identify key members who play important roles in a network. Freeman [5] provided definitions of the three most popular centrality measures: Degree, Betweenness, and Closeness.

(1)Degree measures how active a particular node is. It is defined as the number of direct links a node k has:

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

Where n is the total number of nodes in a network, and a (i, k) is a binary variable indicating whether a link exists between nodes i and k. A network member with a high degree could be the leader or "hub" in a network.

(2)Betweenness measures the extent to which a particular node lays between other nodes in a network. The betweenness of a node k is defined as the number of geodesics (shortest paths between two nodes) passing through it:

$$C_B(k) = \sum_{k}^{n} \sum_{j}^{n} g_{ij}(k),$$

Where gij(k) indicates whether the shortest path between two other nodes i and j passes through the node k. A member with high betweenness may act as a gatekeeper or “broker” in a network for smooth communication or flow of goods (e.g., drugs).

(3)Closeness is the sum of the length of geodesics 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 and is sometimes called farness:

$$C_c(k) = \sum_{i=1}^{n} l(i,k),$$

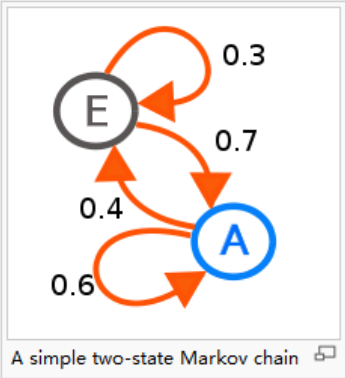
Where l (i, k) is the length of the shortest path connecting nodes i and k.

2. Markov Chain [6]

(1)A Markov chain is a sequence of random variables X1, X2, X3, ... with the Markov property, namely that, given the present state, the future and past states are independent. Formally,

$$P_r(X_{n+1} = x \mid X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = P_r(X_{n+1} = x \mid X_n = x_n).$$

The possible values of Xi form a countable set S called the state space of the chain. Markov chains are often described by a directed graph, where the edges are labeled by the probabilities of going from one state to the other states.



A simple two-state Markov chain [7]

(2)Considering the formulation of the direct graph which showing the message delivery, we decide to use the method of link analyses –PageRank [8].

PageRank is a link analysis algorithm, named after Larry Page and used by the Google Internet search engine, which assigns a numerical weighting to each element of a hyperlinked set of documents, such as the World Wide Web, with the purpose of "measuring" its relative importance within the set.

PageRank is a probability distribution used to represent the likelihood that a person randomly clicking on links will arrive at any particular page. PageRank can be calculated for collections of documents of any size. It is assumed in several research papers that the distribution is evenly divided among all documents in the collection at the beginning of the computational process. The PageRank computations require several passes, called "iterations", through the collection to adjust approximate PageRank values to more closely reflect the theoretical true value.

When calculating PageRank, pages with no outbound links are assumed to link out to all

other pages in the collection. Their PageRank scores are therefore divided evenly among all other pages. In other words, to be fair with pages that are not sinks, these random transitions are added to all nodes in the Web, with a residual probability of usually d = 0.85, estimated from the frequency that an average surfer uses his or her browser's bookmark feature.
So, the equation is as follows:

$$PR(P_i) = \frac{1-d}{N} + d \sum_{P_j \in M(P_i)} \frac{PR(P_j)}{L(p_j)}$$

where p_1,p_2,...,p_N are the pages under consideration, $M(p_i)$ is the set of pages that link to p_i, $L(p_j)$ is the number of outbound links on page p_j, and N is the total number of pages.

The PageRank values are the entries of the dominant eigenvector of the modified adjacency matrix. This makes PageRank a particularly elegant metric: the eigenvector is

$$R = \begin{bmatrix} PR(p_1) \\ PR(p_2) \\ \vdots \\ PR(p_N) \end{bmatrix}$$

Where **R** is the solution of the equation

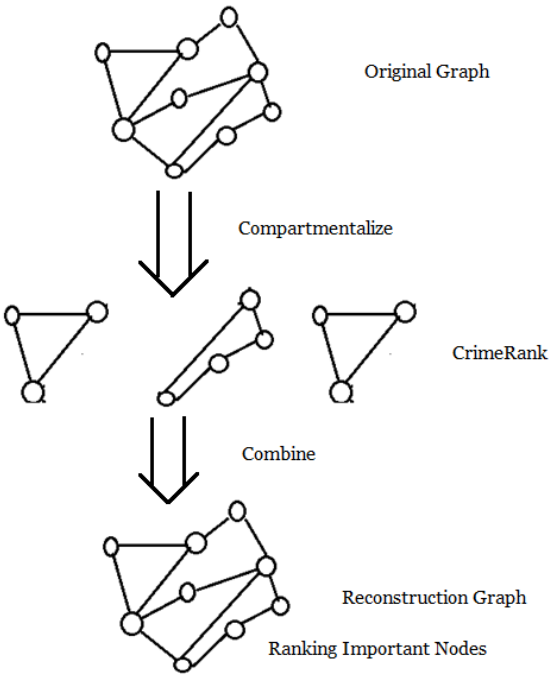
$$R = \begin{bmatrix} (1-d)/N \\ (1-d)/N \\ \vdots \\ (1-d)/N \end{bmatrix} + d \begin{bmatrix} l(p_1,p_1) & l(p_1,p_2 & \cdots & l(p_1,p_N) \\ l(p_2,p_1) & \ddots & & \vdots \\ \vdots & & l(p_i,p_j) & \\ l(p_N,p_1) & \cdots & & l(p_N,p_N \end{bmatrix} R$$

Where the adjacency function $l(p_i, p_j)$ is 0 if page p_j does not link to p_i, and normalized such that, for each j

I.e. the elements of each column sum up to 1, so the matrix is a stochastic matrix (for more details see the computation section below). Thus this is a variant of the eigenvector centralitymeasure used commonly in network analysis. [9]

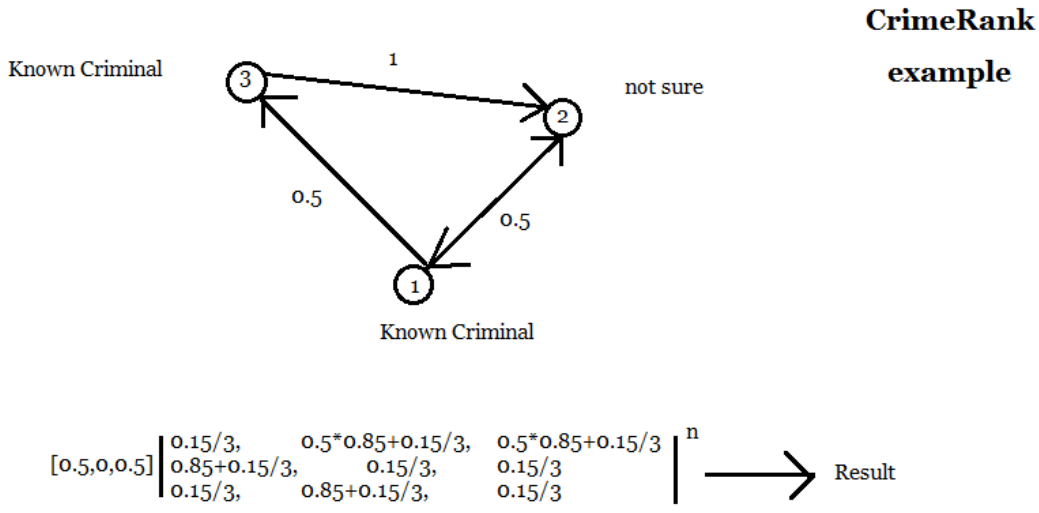
3. It is the method of representing the topics in social network with edge weight which reduce the time of clustering the topics of network. We assume people talking about the same topics have the familiar trend, or even to be a group. And then we analyzed the possibility of each member in the group.

2.3.2 Model Steps



A. We establish such a model, facing this problem: The whole network was classified to 15 sub graph according to category of the topics, and $N_i(i=0,1..15)$ represent one of reduced networks – sub graph in our model. We assign the parameters $W_i(i=0,1..15)$ for each sub graph for further work. Here we let W_7 , W_{11} , W_{13} be 1 .Then we proceed with Link Analysis towards N_7, N_{11} , N_{13} and let the value of conspirator be 1. We eliminate the good man from our graph and assign 0 to the value of others .Then we got the initial state of transition probability matrix.

B. In the next stage, we calculate transition probability matrix based on feature of the network and the value of each node by means of $M_n.n$ multiple $M_1.n$ to stimulate the process of topic transition .Until $M_n.n$ reach the stable point.



C. Though Preliminary result would provide a rank of conspiracy probability, the result didn't the fact that the central of crime organize most probable crime and the leader of the organization. So, we introduce Degree, Betweenness, and Closeness theory to get a new rank of the probability in the preliminary result.

2.3.3 Model Metric

To access the function of the model:

A. People who are familiar with internal information of the company or specialize in economic crime can adjust the weight w1, w2… w15 on his own to achieve better returns.

B. Compartmentalize the network into corresponding subnets can increase efficiency by focus on top of the list as soon as possible.

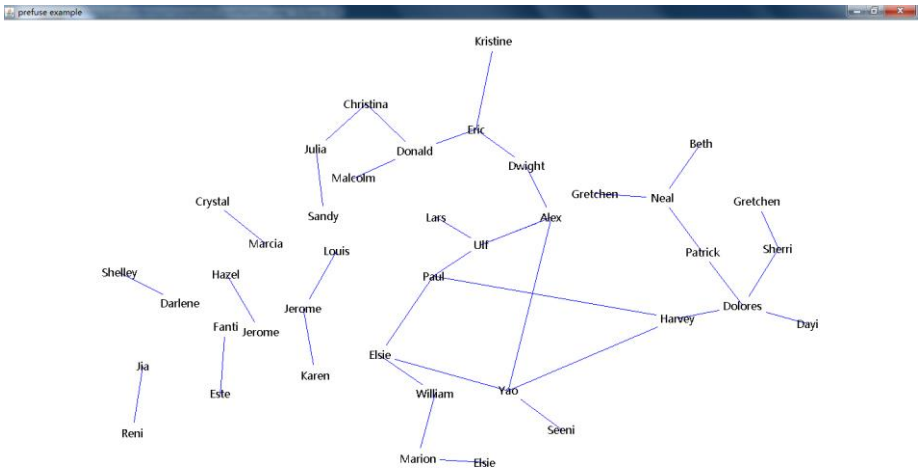
C.With very large databases of message traffic (thousands of people with tens of thousands of messages and possibly millions of words), we use MapReduce framework, which step A equals to Map process and step B equals to Reduce process. Thus, we can process mass data simultaneously with distributed structure and improve performance greatly.

D. Amend the list we get primarily with freeman dimension value we can uncover the leader of crime family and active malefactors efficiently.

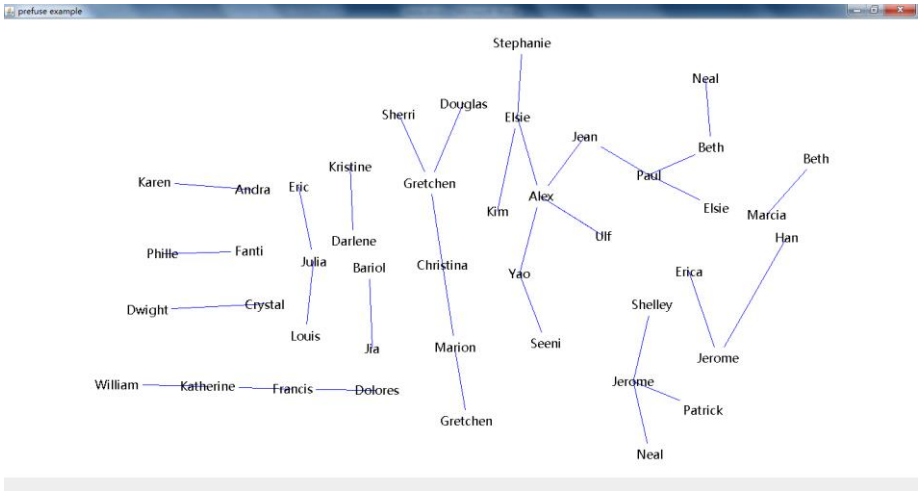
E.Based on consideration of requirement 2, our model can maximize the reuse of previous outcome with the condition of variation, which avoid repeated operation and reduce computational cost.

2.4 Testing & Results

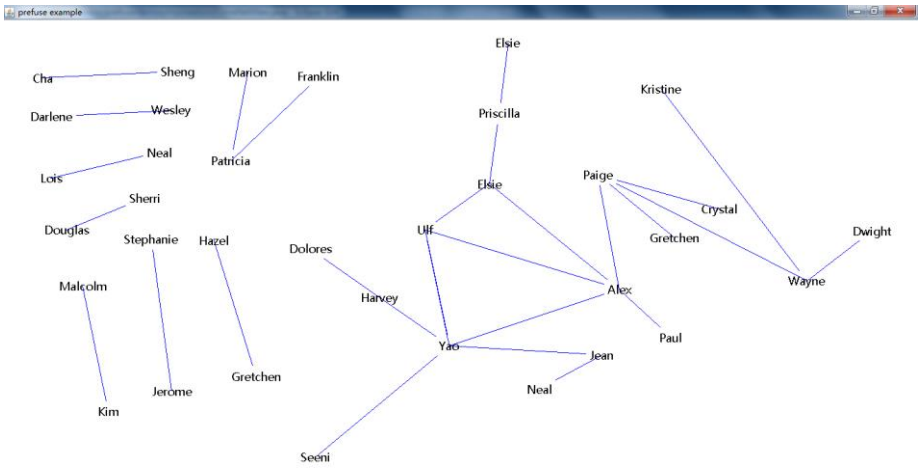
Visualization of Topic 7 subnet



Visualization of Topic 11 subnet



Visualization of Topic 13 subnet



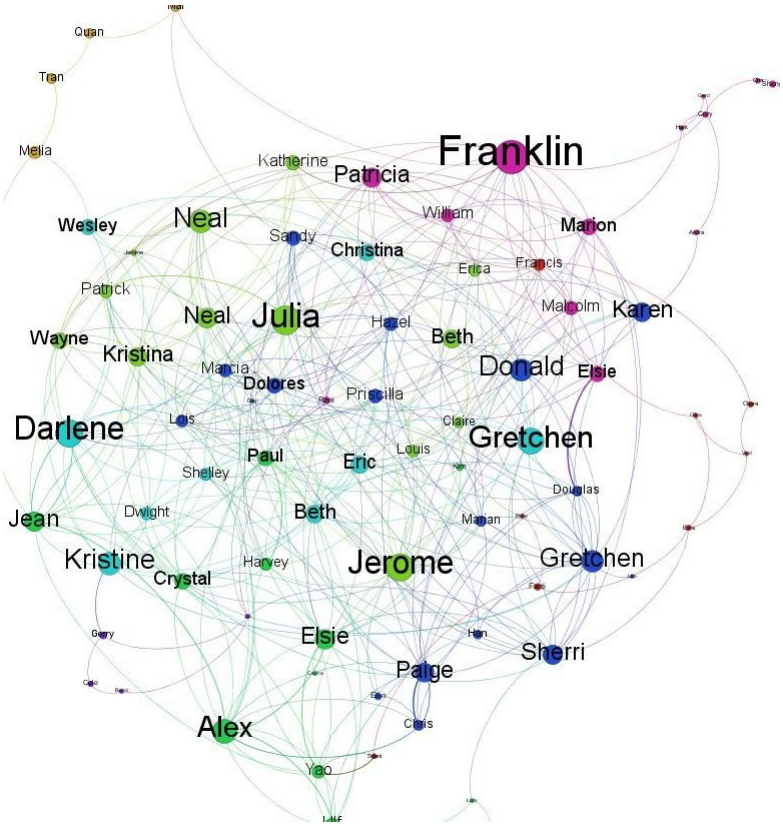
topic 7		
Id	Name	CrimeRank
21	Alex	0.093763553
17	Neal	0.076158744
54	Ulf	0.06330312
67	Yao	0.060192339
6	Patrick	0.054694923
13	Marion	0.05212868
10	Dolores	0.051454398
50	William	0.037455885
49	Harvey	0.031130681
7	Elsie	0.031130681
5	Karen	0.028243615
41	Donald	0.028088177
12	Sandy	0.02427742
80	Fanti	0.020311225
82	Reni	0.020311225
20	Crystal	0.020311225
34	Jerome	0.020311225
16	Jerome	0.020311225
48	Darlene	0.020311225
15	Julia	0.015645054

topic 11		
Id	Name	CrimeRank
21	Alex	0.159994151
54	Ulf	0.146578814
50	William	0.034084698
4	Gretchen	0.032238879
34	Jerome	0.02887857
46	Louis	0.027515255
42	Katherine	0.027515255
32	Gretchen	0.023651139
38	Beth	0.022817539
15	Julia	0.019787024
7	Elsie	0.019787024
11	Francis	0.019787024
48	Darlene	0.019787024
14	Beth	0.019787024
75	Bariol	0.019787024
5	Karen	0.019787024
28	Dwight	0.019787024
80	Fanti	0.019787024
30	Stephanie	0.019105366
3	Sherri	0.01860156

topic 13		
Id	Name	CrimeRank
2	Paige	0.0805482
67	Yao	0.0536512
21	Alex	0.0522648
17	Neal	0.0438713
24	Franklin	0.0430486
43	Paul	0.0389467
54	Ulf	0.0366807
36	Priscilla	0.0356968
18	Jean	0.03194
44	Patricia	0.0309561
4	Gretchen	0.0309561
57	Sheng	0.0309561
31	Neal	0.0309561
29	Wayne	0.0309561
48	Darlene	0.0309561
30	Stephanie	0.0309561
33	Kim	0.0309561
3	Sherri	0.0309561
28	Dwight	0.0298913
10	Dolores	0.0238451

Id	Names	CrimeRank+
21	Alex	2.02312
54	Ulf	1.76426
17	Neal	1.36068
67	Yao	1.33211
10	Dolores	1.21898
13	Marion	1.21368
2	Paige	1.20378
4	Gretchen	1.19906
50	William	1.17907
48	Darlene	1.17775
6	Patrick	1.17063
7	Elsie	1.16856
28	Dwight	1.15816
43	Paul	1.1498
3	Sherri	1.14957
32	Gretchen	1.12555
30	Stephanie	1.12218
34	Jerome	1.11993
5	Karen	1.11694
49	Harvey	1.11652
20	Crystal	1.1162
16	Jerome	1.1162
18	Jean	1.11088
24	Franklin	1.1042
31	Neal	1.10066
33	Kim	1.10066

Table topic 7, topic 11, topic 13 are CrimeRank of subnets, and the last table is the composite result. Then, we get important nodes visualization.



And we rank important nodes higher, can get final result. In our result, the senior managers of the company told in Requirements 1 have high conspiracy.

We got the preliminary result by Markov chain form each topic subset. Referring to the figure, we could easily find out that the node has higher degree, the person represented by this node will rank more front. This result is reasonable, for the more a person talked about the suspicious the more he would like to be a conspirator.

However, to get a more accurate result, we should also consider the structure of the crime group. There are several network topologies such as chain structure, wheel structure, and complete structure [Evan 1972; Ronfeldt and Arquilla2001]. And the most probable conspirator would associate with more nodes in the network then other people. So the rank of the suspected person should be changed by centrality of each person.

Requirement 2

For requirement 2, ordinary thought and method is changing the input condition and proceed with repeated computation integrally. For performing the entire computation with mass data is costly, we reuse our previous results on the condition of input varying and simply recalculate the altering parts. The steps are as follows:

A. for additional suspect, examine the existing networks and determine whether it should add the node. If necessary, we conduct PageRank link analysis with it. Otherwise, the previous results are reused.

B. Conduct PageRank link analysis after Add n1 subnet to criminal network.

C. Add result state of the four subnets and divided by four, and then take off those results equal to 0 to get the final state.

D. After step C, we already get a rank about criminal degree. But the active nodes and leader nodes are disorder. Thus we merge the subnets into a bigger net, estimate the importance of these nodes primarily on degree, betweenness and closeness with Freeman dimension value. Adjust the ranking of these important nodes in the former list.

The result of Requirement 2 is similar to Requirement 1, so we just omit it.

3 Requirement3

Semantic network analysis is a powerful technique to obtain, understand and deduce text information. It is an application with analysis of semantic model which built upon machine learning. General information is usually redundant, which inessential adjectives are round key words. To obtain key words, analyze structure of sentence and characteristic of words are required. For classical problem like synonymy and polysemy, it is hard to extract the unobvious key words. Thus, the correct way is distinguish and classify the words based on accurate understanding. Furthermore, to extract and establish semantic entity need not only better comprehension but reasoning capability.

Text analysis is the chief means of NLP which covered an extensive range such as encoding transform, tokenization, eliminating stop word, token normalization, stemming, lemmatization, semi-supervised text classification, non-supervised text clustering, and topic extraction.

We apply above mechanism to original messages:

Employ experts who are familiar with the affairs of company or have certain study on economic crime to operate text classification manually. With this semi-supervised method, topics are well described and their conspiracy degrees are well estimated. In this way, our model can perform better. Otherwise, we apply non-supervised text clustering mechanism to cluster the original information hierarchically and produce cluster labeling. During the process, corresponding semantic models are established. Use methods with semantic analysis, we obtain, understand and deduce well described topics and corresponding conspiracy degrees which also drive our model carry out well.

For the content described in Topics.xsl file, we made further progress:

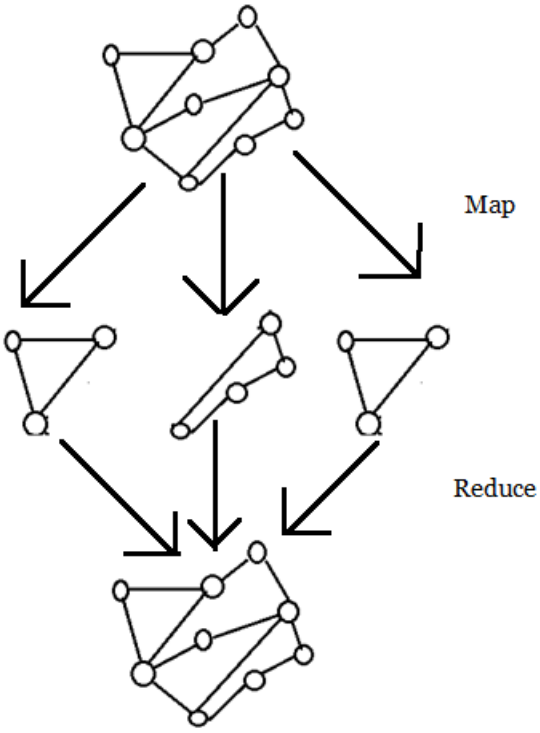
Considering the topics are small in quantity, we choose manual method to extract key words from all the topics. Then replace original text with appropriate part of speech and more accurate words to simplify the topic. After that, assign weights on topics according to the similarity between uncertain conspiratorial topics and certain conspiratorial topics. With the model established before, we get optimized result. However, in reality, the number of topics is quite large. It is unreasonable to process it manually.

1. stock price, earning, sale
2. product line
3. cleaness , maintainess
4. office party
5. security system , permissions
6. president, Paige
7. private meeting
8. ski trip
9. football team
10. satellite business
11. accounting, credit card, and auditing packages
12. best restaurants, lunch
13. off-line, computers, network
14. high price
15. computer security

4 Requirement4

In the area of exploring the conspirator by social network, three main methods are used to analyze the crime network. Two of them are based on the frequency of communication to discover the close group using minimum spanning tree and cluster by the familiarity. The third one is also a model based on the suspicious degree. While analyzing the current case, we believe that a fast and extensive method should be proposed. So partition the whole network into several subnets by topics is required, which enable us get high performance by parallel computation. Then we define our model by PageRank, which calculate the stage result from the preliminary result by …method. At last, by crime theory, we attempt to use the theory of Freeman to find out the center of the suspected conspirators.

Considering the huge scope of data, we decide to implement the MapReduce framework to accelerate the process of data analysis. In our model, Step1 refer to the map step of MapReduce . Step 2 refer to the step of reduce in MapReduce. Therefore we could finish the analysis by distributed system, which enable us handle the data parallel computation and provide excellent performance. To work in coordination with further investigate of the crime, we could introduce AI to our model, train the computer and get a better result. And regression may be considered in the future. We have talked about semantic and text analysis above. Partition of network by topic enable us to analyze the nodes belongs to same category and avoid t re-computation When introduce one suspected topics. Our link analyze have ability of stimulation of association transition and Freeman method is widely used theory of finding the centrality.



5 Strength & Weakness

5.1 Strengths

1. Compartmentalize the whole network by the kind of topic can reduce the scope and increase the computing speed. Real network is always very huge, so it cannot be directly computed but must be compartmentalized to solve.

2. With very large databases of message traffic, we apply MapReduce architecture to our model. Compartmentalizing the whole network is equivalent to Map phase, while combing all results of subnets is equivalent to reduce phase. So, large scale message traffic can be processed effectively in parallel. Hadoop[9] is a framework that allows for the distributed processing of large data sets across clusters of computers using a simple programming model. We can use it to implement our idea.

3. The adjustment based on important node can get more reasonable results. It looks like recommendation system, and provides the preferences of the audience for sorting nodes.

5.2 Weakness

1. Sometimes value transform may lead to the weight of node changes. Our model is not suitable in this kind of condition. For example, the transmission of the virus between cells in tissues needs to consider the influence of time.

2. What's more, the conspiracy weight of topic is uncertain because of lack of original data. So we cannot get more close to reality unless there are experts to assign great value to these parameters. A heuristic method to get better parameters need to use machine learning methods on large scale data.

6 Future Works

1. Develop more consummated theory to adapt weight variant conditions.
2. Provide some heuristic methods to get better parameters of topic weight.
3. Enhance the effect of visualization on the basis of conspiracy list.
4. Apply more models such as machine learning, regression and co-clustering etc. to test data.

7 References

[1] http://en.wikipedia.org/wiki/Bernard_Madoff
[2] www.i2group.com
[3] Mining Key Members of Crime Networks Based on Personality Trait Simulation Email Analysis System
[4] The Social Organization of Conspiracy: Illegal Networks in the Heavy Electrical Equipment Industry
[5] FREEMAN, L. 1979. Centrality in social networks: Conceptual clarification. Soc. Netw. 1, 215–239.
[6] http://en.wikipedia.org/wiki/Markov_chain
[7] ^ "Google Press Center: Fun Facts". www.google.com. Archived from the original on 2009-04-24.

[8]http://en.wikipedia.org/wiki/PageRank
[9]http://hadoop.apache.org/