Finding Conspirators in the Network:
Machine Learning with Resource-allocation Dynamics

Fangjian Guo, Jiang Su, Jian Gao
Team 14531
University of Electronic Science and Technology of China

Key Concepts

Machine learning
Logistic regression
Semantic diffusion
Bipartite graph
Resource-allocation dynamics
Kendall's tau

Key Techniques

Gradient Descent
Revised LeaderRank

Problem Clarification: A conspiracy network is embedded into a network of employees of a company, with its every edge representing a message sent from one node to the other and categorized by topics. Given a few known criminals, non-criminals and suspicious topics, we seek to estimate the probability of criminal involvement for those not identified individuals, and to determine the leader of conspirators. Besides, relevant discussions are also given.

Assumptions (i) Two classes, conspirators and non-conspirators, are linearly separable in the space spanned by local features of a node, which is necessary to machine learning. (ii) A conspirator is reluctant to mention topics related to crime when talking with an outsider. (iii) Conspirators tend not to talk about irrelevant topics frequently with each other. (iv) The leader of conspiracy tries to minimize risk by restricting direct contacts. (v) A non-conspirator has no idea of who are conspirators, thus treating conspirator and non-conspirators equally.

Model Design and Justification The probability of conspiracy for an unidentified node is modeled as a sigmoid function in terms of a linear combination of the node's features (logistic regression), whereas features are formulated from local topological measures and the node's semantic messaging patterns. Parameters of this model are trained by using a subset of identified conspirators and non-conspirators. The performance of the model is enhanced by discovering potential relationship of similarities among topics via topic-word diffusion dynamics on a bipartite graph. Resource-allocation dynamics are also performed to identify the leader of the conspirators, which gains support from empirical evidence in criminal network research.

Strengths and Weaknesses The combination of both the topological properties and semantic affinity among individuals leads to a good performance. As the time complexity is linear, this method is suitable for dealing with large amounts of data. However, our model requires assistance from semantic network analysis to form an expert dictionary. The intrinsic differences with networks may also hinder the portability of the selected features.

Contents

1 Introduction

As illustrated in Figure 1, criminals and conspirators tend to form organizational patterns, interconnected with each other for collaboration, while still maintaining social ties with the outside, thus providing a natural context for description and analysis with networks [1].

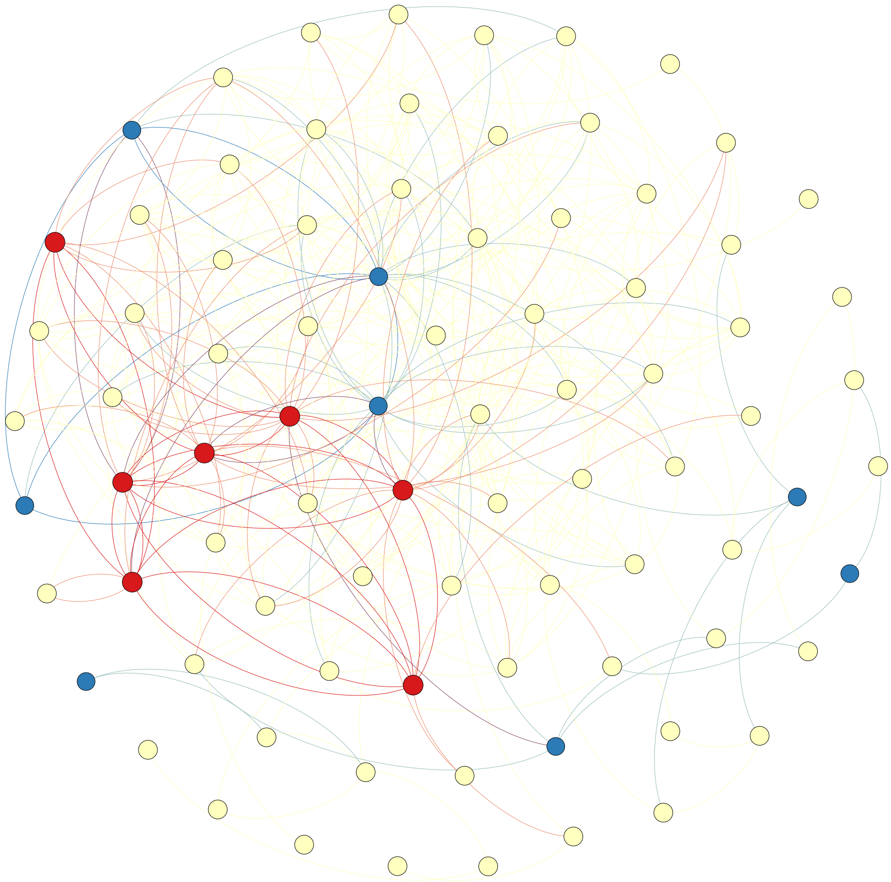

Figure 1: The 83-employee network(red nodes are known conspirators and the blue ones are known non-conpirators)

Criminal networks can be captured from various information, resulting in different types of networks, where each node represents a person, and an edge is present when two nodes collaborate in the same task, share the same family name, etc., or, as in this case, exchange messages [5].

As nodes in this graph can be a mixture of both criminals and non-criminals, it is desirable to determine all the suspicious criminals from topological properties of the network and other prior knowledge, which includes known criminals, known non-criminals and other information related to their interactions. Moreover, it is usually of further interest that a priority list with descending criming likelihood is obtained and the primary leader of the organization is identified, which effectively facilitates law-enforcement by focusing our attention on the most suspicious and the most essential.

Despite the discrepancy with network types, several general methods have been proposed by researchers.

Most notably, many authors have adopted centrality measures of the graph for analyzing the characteristics of criminals. It has been found that criminals with high betweenness centrality are usually brokers, while those with high degree centrality enjoy better profit by taking higher risk [5]. And Morselli et al. proposed that leaders of a criminal organization tend to balance profit and risk by making a careful trade-off between degree centrality and betweenness centrality [7].

However, centrality approaches, which utilize local properties, tend to overlook the complex topology with the whole networks. Therefore, social network analysis (SNA) methods including subgroup detection and block-modeling have been introduced, which try to discover the hidden topological patterns by partitioning the big network into small closely connected cliques [11]. Despite the light they shed upon the internal structures of criminal networks, these methods still suffer from intimidating complexity with large databases [10].

In this paper, we carefully combine the local-feature-based methods with approaches related to global topology of conspiracy networks. We propose a machine learning scheme to leverage local features, so as to estimate each node's likelihood of conspiracy involvement. And dynamics-based methods, which are less computationally expensive than most of other topology-based approaches, are adopted to help find out the leader of conspirators and to discover semantic connections between topics.

We start with the formulation of useful local features of a node in the network, which then lead to the machine learning scheme. By feeding a subset of known conspirators and non-conspirators as training samples into the learning algorithm, the classification hypothesis is formed. We then use it to estimate the probability of being a conspirator for every unidentified individual in the network.

As highly suspicious topics are essential to the performance of machine learning, we then try to discover similarities between topics, by performing simple source-allocation dynamics on the bipartite semantic network made up of topics and sensitive words. Those findings expanded our knowledge on suspicious topics, thus enhancing the accuracy of our machine learning model.

Motivated by the goal of finding criminal leaders, we applied a dynamics-based ranking algorithms on a subgraph extracted from the network. Our findings are in agreement with empirical knowledge on the centrality balance of criminal leaders.

Finally, sensitivity analysis is performed to test the robustness of our approach, followed by further discussions.

2 A Machine Learning Solution to Criminal Priority

Machine learning is carefully selected by us to play the key role in the entire the strategy mainly for consideration on its capability of adaptiveness and reorganization, which simulate human beings on actions of study to obtain fresh knowledge. Such character is quite important since now we encounter a problem which is usually done by people through just the same method: deduction, reasoning and reorganization our structure of knowledge to get over it. Especially met with such deduction task based on hundreds of thousands of data or big amount of information, people are helpless and their ability so terribly limited that we have to turn to machines.

In this section, we will describe the whole construction process of our machine learning framework in detail including feature formulation, core learning methods and experimental results. Through statistical analysis on the results, we propose our enhancement based on semantic diffusion.

We commence with several necessary assumptions:

• We assume that all the data and information about the EZ case network and 83-node network are relatively stable in a long period, rather than from coincident observation, to guarantee the representability of the results from the aspect of data origin.

• Based on necessary observation on the network, which will be exhaustively described in main body, we put forward our assumption that the contents of the communication among conspirators tends to be relevant about suspicious topics or some formal issues, rather than gossip.

• We assume that both networks in EZ case and in more complicated case obey the same information transmission rule that ensure the analogy about some core mechanism could stand.

2.1 Feature formulation

- **Centrality**

  We exploit three types of centrality including degree, betweenness and closeness centrality to determine the center of the suspicious network from different aspects:

▶ *Degree centrality* [3] indicates activeness of a member, i.e. the member who tends to have more links to its surroundings. As explained in [12], degree centrality is not quite reliable to indicate the team leader in a criminal network. For a graph $G(V, E)$, the normalized degree centrality of node i is as follows:

$$C_D(i) = \frac{\sum_{j=1}^{|V|} \nu(i,j)}{|V|-1}, i \neq j \tag{1}$$

Where ν is a binary indicator showing whether exists a link between two nodes. Considering the graph is directed in our case, we separately calculate the in-degree and out-degree of every node.

▶ *Betweenness centrality* [3] describes how much a node tends to be on the shortest path of other nodes. A node with large betweenness centrality does not necessarily induce its large degree, but illustrates its role of "gatekeeper", who is more possibly to be a intermediary when any other two members transmit information between themselves. The normalized betweenness centrality is defined as:

$$C_B(i) = \frac{\sum_{j=1}^{|V|} \sum_{k<j}^{|V|} \omega_{j,k}(i)}{|V|-1}, k \neq i \tag{2}$$

where $\omega_{j,k}(i)$ indicates whether the shortest path between node j and node k passes through node i.

▶ *Closeness centrality* [8] is usually utilized to measure how far away one node is from the others. Closeness of a node is defined as the inverse of the sum of its distances to all other nodes and can be treat as a measure of efficiency when spreading information from itself to all other nodes sequentially. It indicates how easily an individual connects with other members. The normalized closeness centrality is defined as:

$$C_c(i) = \frac{\sum_{j=1}^{|V|} \rho(i,j) - C_c min}{C_c max - C_c min}, i \neq j \tag{3}$$

where $\rho(i, j)$ is the length of the shortest path connecting nodes i and j. $C_c min$ and $C_c max$ are the minimum and maximum lengths of the shortest paths respectively.

- **Number of known neighboring conspirators**

We consider the number of known neighboring conspirators of a node as a significant feature. The interaction among known conspirators in message network suggests a much stronger connectivity than the one among the known non-conspirators. This phenomenon reasonably reveal that a conspirator is more likely to communicate with his or her accomplice rather than a outlier and, on the contrary, non-conspirators lack such consciousness. As shown in Figure 2, we calculate the ratio of known conspirators among one's adjacent neighbors, which illustrates his or her proximity with known accomplices: the value is 1 if it connects with all the known conspirators and 0 means no conspirators is adjacent to it. The known suspicious clique obviously represents a compacter connectivity. Therefore, the more known conspirators being a node's neighbors, the more possibly the node itself is a accomplice.

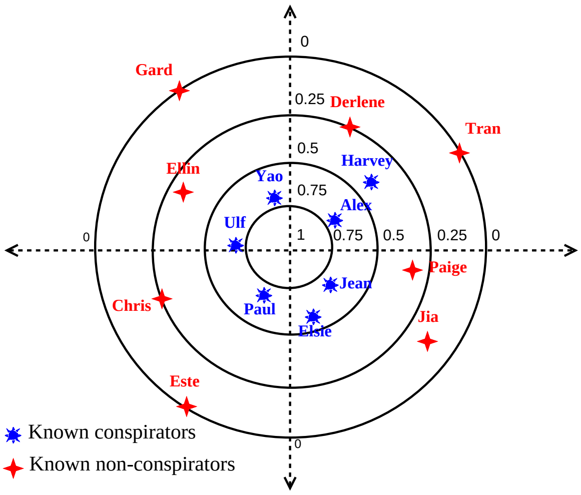

Figure 2: The ratio of known conspirators among one's neighbors. Values in both axes are the number of known conspirators that is connected to a person over the total number of his or her neighbors. The closer a person is to the center, then more possibly he or she is a conspirator.

- **Number of current non-suspicious messages from the known conspirators**

Table 1 is the topics mentioned between known conspirators.[1] It is obvious that a known conspirator rarely talks about irrelevant topics, i.e. topics irrelevant to their conspiracy, with his or her accomplices even though some

[1] Topic No. 16 in the raw data is regarded as wrong and thus discarded.

unknown topics appear among them, which accounts for a very small proportion. If the information received from a known conspirator is most irrelevant, the receiver is much probably to be an outlier. So it is quite reasonable to take such argument as a feature.

	Jean	Alex	Elsie	Poul	Ulf	Yao	Harvey
Jean		**11***			8		14
Alex			1	**13***	**11***	3,**7***	
Elsie		**11***			**13***		
Poul	**11***		**7***		**7***		4
Ulf		**7***, **11***, **13***				**13***	
Yao	**13***	**7***, **11***, **13***	**7***,9		**13***		2, **7***
Harvey						**13***	

Table 1: Topics among known conspirators (known conspiratorial topics are those with star and highlighted in blue)

2.2 Methods

We use the L-2 regularized logistic regression to model the probability of a node being involved in the conspiracy, and the parameters of the model are obtained by solving an optimization problem related to training set by gradient ascent algorithm.

2.2.1 Logistic regression

We consider a training set of size m: $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \cdots, (x^{(m)}, y^{(m)})\}$, where $x^{(i)}$ is an n-dimensional feature vector, and $y^{(i)}$ indicates the classification of the agent, i.e. $y^{(i)} = 1$ for conspirators and $y^{(i)} = 0$ for non-conspirators. All the nodes in the training set are drawn from the 15 known conspirators and non-conspirators.

As a descendant of generalized linear model for Bernoulli distribution, logistic regression tries to estimate the probability of being a conspirators as

$$P(y = 1|x; \theta) = h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}, \tag{4}$$

where $\theta \in \mathbb{R}^n$ is the parameter vector.

Then, under the framework of generalized linear model, the *maximum a posteriori* (MAP) estimate of the parameter θ is given by

$$\min_\theta J(\theta), \tag{5}$$

where the cost function is given by

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}[-y^{(i)}\log(h_\theta(x)^{(i)}) - (1-y^{(i)})\log(1-h_\theta(x^{(i)}))] + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2, \quad (6)$$

with λ being the regularization parameter.

2.2.2 Gradient descent

The cost function $J(\theta)$ is minimized by using the algorithm of gradient descent, which always drives θ down the locally steepest slope, in hope to reach the global minimum of the cost function.

At every iteration before convergence, new θ is replaced by the old θ as

$$\theta := \theta - \alpha\nabla_\theta J(\theta), \quad (7)$$

where α is a small positive constant.

2.2.3 Leave-one-out cross validation

As we are only informed of the correct classification of N_0 nodes ($N_0 = 15$ in this case), at a given round, we only use $(N_0 - 1)$ of them as the training set, while leaving one out for cross validation (C-V). At every round, the next correctly classified node is left out and the others serve as the training set, then the trained hypothesis is tested on the left-out node. In this way, by averaging N_0 rounds without overlapping, the error for both the training set and the cross validation set can be evaluated.

Supposing, for example, in the j-th round, sample $(x^{(j)}, y^{(j)})$ is left out, and the training set is given by

$$S_j = \{(x^{(l)}, y^{(l)})|l = 1, 2, \cdots, j-1, j+1, \cdots, N_0\}. \quad (8)$$

Using this training set, parameter vector $\theta^{(j)}$ is obtained, and the corresponding hypothesis is tested on both S_j and the left-out $(x^{(j)}, y^{(j)})$, arriving at this round's training error ε_{S_j} and C-V error ε_j respectively.

Hence, by averaging them over j, the training error is given by

$$\varepsilon_S = \frac{1}{N_0}\sum_{j=1}^{N_0}\varepsilon_{S_j}, \quad (9)$$

and the cross validation error is given by

$$\varepsilon = \frac{1}{N_0}\sum_{j=1}^{N_0}\varepsilon_j. \quad (10)$$

2.2.4 Selecting regularization parameter

The regularization parameter λ ($\lambda > 0$) is selected optimally as to minimize the cross validation error, i.e.

$$\lambda = \arg\min_{\lambda>0} \varepsilon. \tag{11}$$

2.3 Results

By training the logistic regression with our leave-one-out cross validation strategy, λ is optimally set to 1.9 and the overall C-V error $\varepsilon = 0.27$ (training error $\varepsilon_S = 0$). Then, while fixing the chosen λ, the hypothesis is finally retrained on the maximum training set, making full use of every known conspirators and non-conspirators.

Name	Dolores ⋆	Crystal	Jerome ⋆	Sherri	Neal
Node No.	10	20	34	3	17
Probability of conspiracy	0.555	0.508	0.388	0.316	0.299
Name	Christina	Jerome	William	Dwight	Beth
Node No.	47	16	50	28	38
Probability of conspiracy	0.267	0.252	0.245	0.242	0.233

Table 2: Top 10 in the priority list (known conspirators excluded)

The trained hypothesis gives us the estimated probability for node i being a conspirator, resulting in a priority list of suspicious individuals, ranked in descent order of criminal likelihood. The top 10 suspicious are shown in Table 2, where managers are marked by a star ($\star$).

Figure 3 illustrates the probability of criminal involvement estimated by $h_\theta(x)$ versus the corresponding rank in the priority list, where three managers (Jerome, Dolores and Gretchen)[2] are marked by circles.

Dolores (manager) is indeed the person deserving highest suspicion, and Jerome (manager) is also likely to be involved in conspiracy.

2.4 Semantic model enhancement

When talking about enhancement of the accuracy, we should reduce the difference the performance with those who could deal with it more flexibly and exactly:

[2] As more than one nodes are named either Gretchen or Jerome, we select those with bigger out-degrees to be managers, i.e. Manager Gretchen is Node No. 32 and Manager Jerome is Node No. 34

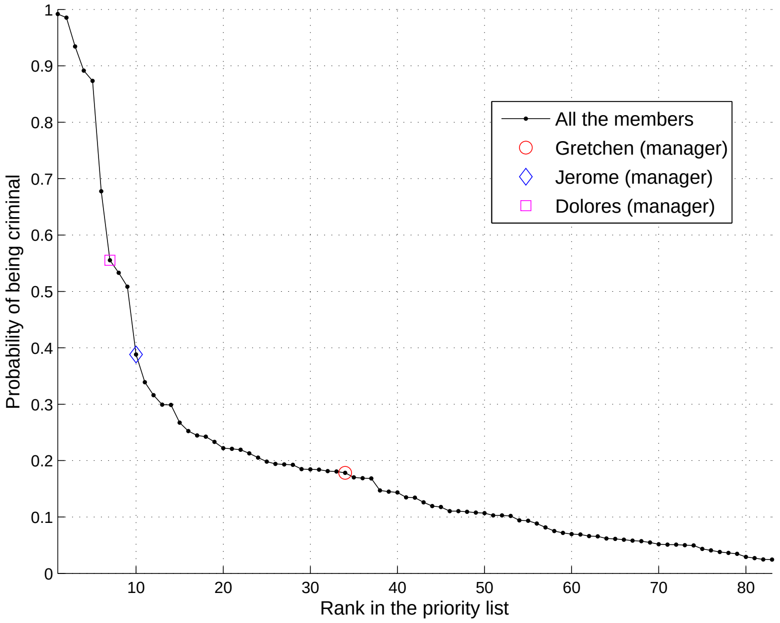

Figure 3: Probability of conspiracy versus the corresponding rank in priority list

human being. A very important way we humans solve cases is analysis the communication contents including messages and records. However, just as aforementioned, our ability so quite limited when handling huge amount of information that we have to utilize machine to help us. Therefore, semantic information is more important for humans rather than extremely complicated topology structure. For example,through analysis into the information in message traffic, we could discover several interesting and helpful phenomena.

As is in EZ case, Some similar text information in the dialog motivate us to discover that Inez represents some attributes that are quite similar to George, who is definitely a conspirator. For instance, the word "tired" when describing Inez and the word "stressed" when describing George. Similar case can be also found in the 83-people network case such as the word "Spanish" from known conspiratorial topic 7 is highly suspectable and appear in other unknown topics (e.g.topic 2 and 12) repeatedly. The contents about "computer security" which is treated as part of the key in the whole conspiracy also keep active in many other unknown topics like 5 and 15. Above relativity in information may easily cause humans' vigilance. Hence it is natural to train a computer to find a method that could measure similarities among topics and reveal some potential information.

Lexical ambiguity broadly exists among words and they always contain different meanings depending on particular scenarios. Therefore, it is not wise to abandon human intelligence and only depend on particular algorithms to crack a criminal

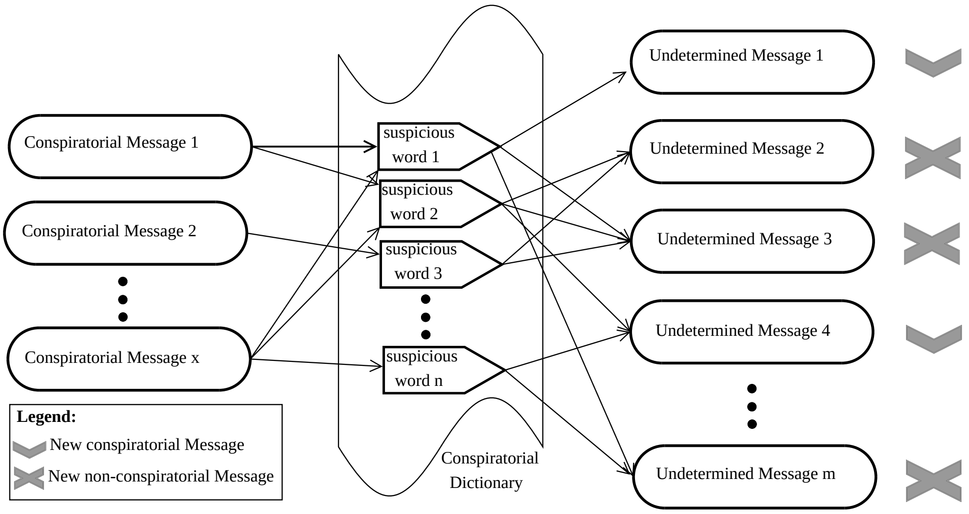

Figure 4: Framework of topic semantic diffusion

case during the detection period. Detectors' reasoning plays a indispensable role through out the entire process. Therefore, we draw the problem of topic semantic diffusion into a topic similarity measurement task based on expert dictionary. As seen in Figure 4, a conspiratorial dictionary is firstly constructed from the conspiratorial messages about known suspicious topics. Resource allocation mechanism on bipartite network outperforms in extracting the hidden information of networks [13], which is exploited by us to unfold the similarity among different topics after a bipartite network is constructed (see Figure 4) between the conspiratorial dictionary and all the information in message traffic.

The bipartite network is modeled by $G = (D, T, E)$. E is an edge set, indicating the relationship between key word set D of expert dictionary and topic set T , where $D = d_1, d_2...d_n$ and $T = t_1, t_2...t_m$. Then, we arrange all the topics with 0 resource except each known conspiratorial topic with one unit of resource and commence with the first allocation from set T to set D:

$$f(t_l) = \sum_{i=1}^{n} \frac{a_{il} f(d_i)}{D(d_i)}. \tag{12}$$

Equation 12 expresses the calculation of the resource held by $t(l)$ after the first step : $f(t_l)$. $D(d_i)$ indicates the degree of the node d_i and a_{il} is defined as follows:

$$a_{il} = \begin{cases} 1, & d_i t_l \in E \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

Intuitive explanation of step 1 is to arrange the resource averagely by degree of t_i from T to D if t_i owns resource. The second step is to reflect the resource flow back to T from D obeying the same rule. So the resource finally locates on ti satisfies :

$$f'(t_i) = \sum_{l=1}^{m} \frac{a_{il} f(d_l)}{D(d_l)} = \sum_{l=1}^{m} \frac{a_{il}}{D(d_l)} \sum_{j=1}^{n} \frac{a_{ji} f(t_j)}{D(t_j)}. \tag{14}$$

After this two-fold method, the amount of resource held by every element in T could be seen as a score of similarity. The rank of all topics according to such score represents the degree of their similarity to the information from dictionary,i.e. the higher this score is, the topic is more likely to be a newly found conspiratorial topic.

Since we set $D = \{'spanish', 'system', 'network', 'computer', 'meeting'\}$ as the conspiratorial dictionary, Table 3 illustrates the final result of all the 15 topics in 83-people network case. The known suspicious topic numbers 7,11,13, which is our fundamental basis for further development, are naturally to be top three and topic 5 is also very suspicious than other unknown topics. 2,12,15 are among the group with the second highest possibility in unknowns and the left ones tends to be irrelevant topics to the conspiracy.

Rank	Topic Number	Similarity to known suspicious topics
1	**11***	0.750
2	**7***	0.667
3	**13***	0.667
4	5	0.417
5	2	0.167
6	12	0.167
7	15	0.167
8	1,3,4,6,8,9,10,14	0

Table 3: Rank of all topics based on similarity to known suspicious topics.(known conspiratorial topics are those with star and highlighted in blue)

We then append topic 4 into the set of known conspiratorial topic set and train the model again, the overall C-V error decrease from former 0.27 to current value of 0.13. As the confidence degree of topic 2,12,15 is low as shown in Table 3, there is not obvious influence on the detection correctness. The limited resource and the impressive performance here indicate that if we absorb enough key words into conspiratorial dictionary and more topics with abundant contents, such method is much likely to perform better. However, when dealing with huge amounts of information, it will become a problem to get valuable words into dictionary as human wisdom become helpless.

On the other hand, if we utilize the speaker instead of the key words to construct a bipartite graph with the topics, we will also get similarity among topics based on speaker who transmit them. However, the determination of the relationship between different results under these two standards, even more standards, is definitely beyond this paper.

After comprehending the actual meaning of the topics, we find the rank result is quite reasonable and valuable. Meanwhile, it is not only its reliable result impresses us a lot, but also its high efficiency and low complexity of implementation will give it another good performance in huge amounts of data, for this method is only of linear time complexity $O(n)$.

3 Identifying the leader of the conspiracy

Our machine learning scheme tries to estimate the likelihood of a node committing conspiracy, however, the likelihood does not proportionally indicate the leadership inside the network, for the identification of leaders is further complicated by its topology.

Thus we adopt LeaderRank, a node ranking algorithm closely related to network topology, to find the leader of the criminal group. Meanwhile, a subgraph connected by known suspicious topics is extracted from the network, in order to decouple the structure with company employees. Besides, because of its robustness against random noise, LeaderRank is also appropriate for addressing criminal network problems, which usually suffer from incompleteness and incorrectness.

3.1 LeaderRank

LeaderRank algorithm is a state-of-art achievement on node ranking, which is more tolerant of noisy data and robust against manipulations than more traditional algorithms including HITS and PageRank [6]. This method is mathematically equivalent to random walk mechanism on the directed network with adaptive probability, leading to a parameter-free algorithm readily applicable to any type of graph. A ground node, who connects with every node through newly added bidirectional links, is arranged into the topology in order to make the entire network a strongly connected one and hence the random walk will definitely converge into a homoeostasis process.

For a graph $G = (V, E)$, every node in the graph obtains 1 unit of resource except the ground node. After the commence of voting process, node i at step t

will get an adaptive voting score $\nu(t)$ according to the voting from its neighbors:

$$\nu_i(t+1) = \sum_{j=1}^{|V|+1} \frac{\mu_{ij}}{D_{out}(j)} \nu_i(t) \tag{15}$$

Where μ_{ij} is a binary indicator with value 1 if node i points to j and 0 otherwise. $D_{out}(j)$ denotes the out-degree of node j. The fraction of above two arguments could be considered as the probability that a random walker at i goes to j in the next step. Finally, the leadership score of node i is proved to be $\nu_i(T_c)+\nu_{gn}(T_c)/|V|$, where $\nu_{gn}(T_c)$ is the score of the ground node at steady state.

3.2 Suspicious topic sub-network extraction

As the criminal network in embedded in a network of company employees, we extract the sub-network G_{T_S} connected by suspicious topics only, so as to minimize the coupling of the company's hierarchical structure to the conspiracy relations.

Supposing T_{ij} denotes the set of topics mentioned by messages from node i to node j, and T_S denotes the set of known suspicious topics ($T_S = \{7, 11, 13\}$). Then G_{T_S} is the maximum subgraph of the original graph G, whereas

$$T_{ij} \subseteq T_S, \text{for all } (i,j) \subseteq E_{T_S} \tag{16}$$

3.3 Edge reverse

Because the original LeaderRank deals with finding leaders in Internet social networks (SNS), where the direction of an edge has a dissimilar meaning from our case, i.e. if A points to (follows) B in twitter, then B is considered to be a leader of A. However, in our communication network, an edge pointing from A to B suggests A has sent B a message. Therefore, if assuming that a leader in a criminal network tends to be the sender of a message rather than receiver by issuing commands, then each edge in G_{T_S} has to be reversed to be compatible with LeaderRank's original design. The reversed sub-network induced by suspicious topics is denoted by G'_{T_S}.

3.4 Results

By running LeaderRank on G'_{T_S} , a ranking score is assigned to every node in this subgraph, which generates a list of possible leaders ranked in descent order, as shown in Table 4.

Yao (node number 67) is ranked as the chief leader of the conspiracy organization.

Name	LeaderRank score
Yao	2.67
Alex	2.21
Paul	1.92
Elsie	1.62

Table 4: Partial results of LeaderRank on G'_{T_S}

3.5 Empirical support

Empirical analysis of criminal networks has found that a leader of a criminal organization tends to carefully balancing his or her degree centrality and betweenness centrality. It has been proposed that the leader usually maintains a high betweenness centrality but a relatively low degree centrality, for enhancing efficiency and meanwhile ensuring safety [7].

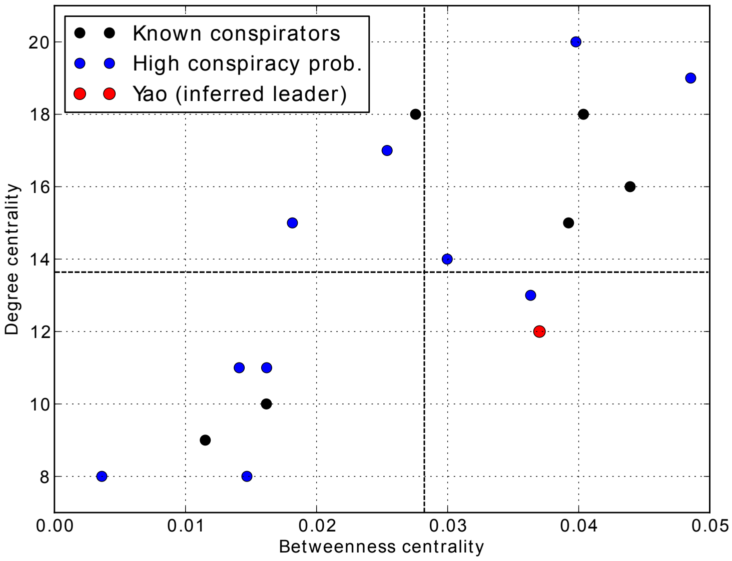

Figure 5: The joint distribution of betweenness centrality and degree centrality

And our inference that Yao is the leader is thus empirically supported. Figure 5 illustrates the joint distribution of betweenness centrality C_B and degree centrality $(D_{in}+D_{out})$ for 7 known conspirators and 10 nodes with high conspiracy likelihood, where two dashed lines mark average values of the displayed nodes. Yao's high betweenness centrality with relatively low degree centrality accord with the identity of a leader.

3.6 Discussion

The leader of the criminal network is identified by performing LeaderRank on the extracted, edge-reversed, suspicious-topic-connected subgraph. And our findings are strengthened by empirical research results.

LeaderRank, as an algorithm that ranks nodes by performing source reallocation dynamics on the network, is generally more computationally inexpensive compared to traditional methods like block-modeling. Our scheme accommodates large databases with higher efficiency.

4 Evaluating the Model

4.1 Sensitivity analysis

Considering the usual incompleteness, imprecision and even inconsistency with criminal social networks [11], the method for inferring criminality or conspiracy should be robust enough to cope with minor alternations of the network. Otherwise, small flaws or incompleteness of the network would possibly lead to mistaken accusations or connivance of criminals. Therefore, a sensitivity analysis is performed for our approach.

Requirement 2 provides an appropriate scenario for such a test: while other conditions remain unchanged, new information indicates that Topic 1 is also connected to criminal activity, and Chris, who was considered innocent before, has now proved guilty.

4.1.1 Priority list

By applying our methods to these altered conditions, we find out that among the top-10 of the previous priority list (7 known conspirator excluded), 7 of them are still top-10 holders of the current list, while the remaining three find their new places at 12th, 14th and 16th respectively, as illustrated in Table 5.

A more sophisticated measurement of the sensitivity of priority list is *Kendall's tau* coefficient τ [9]. Given two priority lists $\{p_k\} = \{p_1, p_2, \cdots, p_n\}$ and $\{q_k\} = \{q_1, q_2, \cdots, q_n\}$ (for example, $p_2 = 5$ means node 2 is ranked 5th by $\{p_k\}$ list), then $(i, j), i \neq j$ is said to be a concordant pair if their rankings agree in two lists, i.e. $p_i > p_j, q_i > q_j$ or $p_i < p_j, q_i < q_j$; (i, j) is said to be a discordant pair if their rankings disagree in two lists.

Name	Dolores	Crystal	Jerome	Sherri	Neal	Christina	Jerome	William	Dwight	Beth
Rank (previous)	1	2	3	4	5	6	7	8	9	10
Prob. of conspiracy (previous)	0.555	0.508	0.388	0.316	0.299	0.555	0.508	0.388	0.316	0.299
Rank (new)	2	1	6	9	3	14	5	4	12	16
Prob. of conspiracy (new)	0.621	0.629	0.405	0.393	0.504	0.335	0.407	0.409	0.352	0.334

Table 5: Change with top 10 in the priority list (known conspirators excluded)

Then *Kendall's tau* is defined as

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{\frac{1}{2}n(n-1)}. \tag{17}$$

τ lies in the range of $[-1, 1]$, whereas 1 for perfect ranking agreement, -1 for utter disagreement.

The *Kendall's tau* between two priority lists obtained in Requirement 1 and Requirement 2 is $\tau = 0.86$, justifying the robustness of the machine learning approach.

If we assume those known conspirators and non-conspirators are independently wrongly classified with certain probability, then the expected value of τ between our computed priority list and the real priority list would vary with that probability. Figure 6 depicts the expected Kendall's tau versus the misclassification probability of conspirator set and non-conspiracy set separately.

As can be seen from Figure 6, even if the misclassification error occurs with probability as big as 0.5, the Kendall's tau does not drop below 0.80, substantially proving the inherent stability of our methods.

4.1.2 Probability inflation

Figure 7 illustrates the change with estimated conspiracy probability due to modified conditions in Requirement 2, with the previous value as x-axis, and the new as y-axis. Generally, most nodes exhibit a small "inflation" in criminal probability, as indicated by the distribution of dots skewed from the diagonal line. The

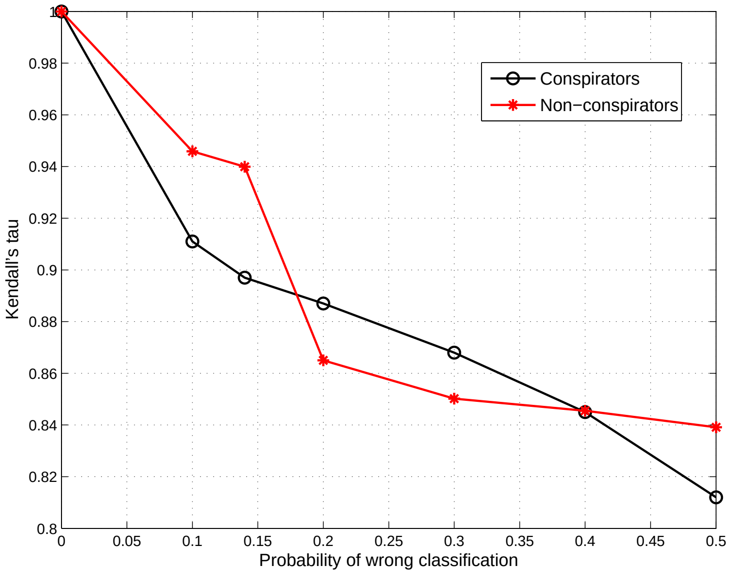

Figure 6: The expected Kendall's tau declines as misclassification probability increases

augmented probability is compatible with the new information that expands both the set of suspicious topics and known conspirators.

The analysis suggests that our machine learning method is insensitive to minor alternations with known conditions, while still able to produce new, reasonable results implied by newly introduced information.

5 Discussion and Conclusion

5.1 Summary

We draw the problem of predicting conspiracy in a company on a multi-feature machine learning problem. 5 features are selected for their representation of centrality in topology, conspiratorial level of connectivity and communication contents. Experimental results hits about 73% correct prediction rate in the basic model.

Then, we offer an algorithm, which refers to resource allocation mechanism in bipartite graph, to reveal potential similarity among topics and discover new conspiratorial topics to feedback in the optimized learning process. Obvious enhancement lead us to some deeper discussion about background of large data and extended standards for measuring similarity among topics.

In the following section, Revised LeaderRank scheme is proposed in searching for the leader, required by DA. We additionally validate our result by known

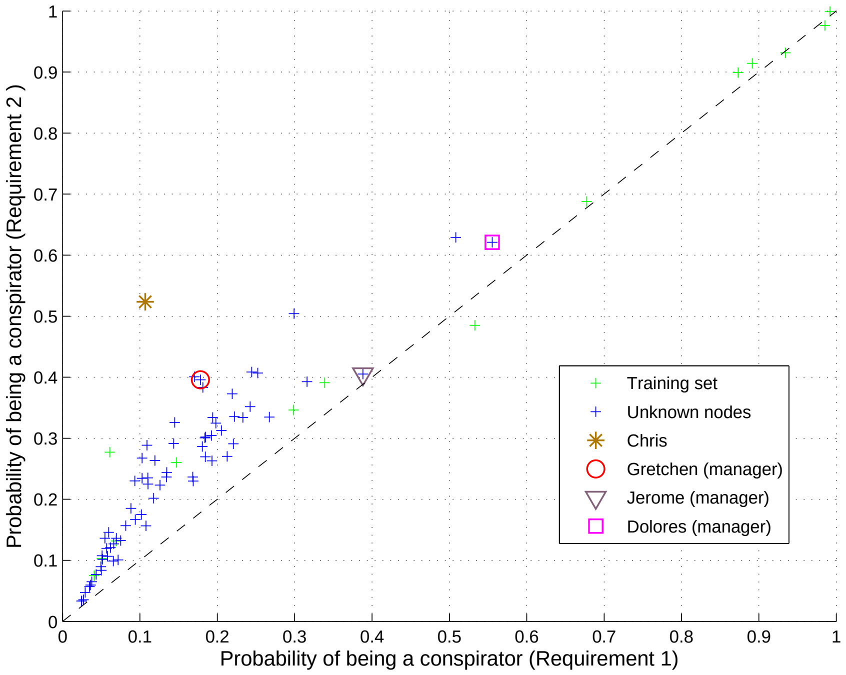

Figure 7: Criminal probabilities before and after the change of conditions

empirical patterns of the topological properties of a criminal leader.

Finally, the high value of *Kendall's tau* illustrates the non-sensitive property of the model. The weakness of our model is mainly related to features, as different networks may not share features with similar effects. Therefore, the features used in a particular network may be less useful when applying them to another network. Furthermore, huge loads of information might limit the capability of the resource allocation strategy because the construction of a bipartite graph itself becomes a challenge especially when dealing with large amounts of data with noise.

5.2 Further discussion

When measuring the portability of a model in different situations, we shall focus on two aspects: the similarity of two networks and their inevitable difference. Some common patterns that appear in a wide range of networks, including biological, social etc., are the small-world property, power-law degree distributions, network transitivity and community structures [4]. Those topological features shared by disparate networks can help us build a model with good portability.

However, on the other hand, the fundamental difference in the mechanism of information diffusion and network growth prevents one model from being universal. While people find out clues through evidence and related materials to break down the invisibility of a criminal network based on reasoning and deduction, this co-occurrence-based approach can hardly work for biological networks [2], although they do share some common patterns. However, some other methods may yield satisfactory results by utilizing a revised version of disease spreading dynamics for predicting dysfunctional or diseased biological components.

References

[1] Wayne E. Baker and Robert R. Faulkner. The Social Organization of Conspiracy: Illegal Networks in the Heavy Electrical Equipment Industry, 1993.

[2] Hao Chen and Burt M Sharp. Content-rich biological network constructed by mining PubMed abstracts. *BMC bioinformatics*, 5:147, October 2004.

[3] LC Freeman. Centrality in social networks conceptual clarification. *Social networks*, 1979.

[4] M Girvan. Community structure in social and biological networks. *PNAS*, 2001.

[5] Valdis E Krebs. Mapping Networks of Terrorist Cells. *Connections*, 24(3):43–52, 2002.

[6] Linyuan Lü, Yi-Cheng Zhang, Chi Ho Yeung, and Tao Zhou. Leaders in social networks, the Delicious case. *PloS one*, 6(6):e21202, January 2011.

[7] C. Morselli. Assessing Vulnerable and Strategic Positions in a Criminal Network. *Journal of Contemporary Criminal Justice*, 26(4):382–392, September 2010.

[8] Gert Sabidussi. The centrality index of a graph. *Psychometrika*, 1966.

[9] Kumar Pranad Sen. Estimates of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association*, 1968.

[10] Christopher Wheat. Algorithmic Complexity and Structural Models of Social Networks. *Management*, pages 1–38, 2007.

[11] J Xu. Criminal network analysis and visualization. *Communications of the ACM*, 48(6), 2005.

[12] Jennifer Xu and Hsinchun Chen. Untangling Criminal Networks : A Case Study. *Intelligence and Security Informatics*, pages 232–248, 2003.

[13] Tao Zhou. Bipartite network projection and personal recommendation. *Physical Review E*, 76(4):1–7, 2007.