No.10 Xitucheng Road
Haidian District, Beijing 100876
(010) 58828308

June 6, 2015

Mr. Kailin Tang
Beijing University of Posts and Telecommunications
Haidian District, Beijing 100876
2013213057@bupt.edu.cn

Research Report for Finding Conspirators in the Network via Machine Learning

INTRODUCTORY SUMMARY

As shown in **Figure 1**, criminals and conspirators tend to form organizational patterns, interconnected with one another for collaboration, while still maintaining social ties with the outside, thus providing a natural context for description and analysis via networks [Baker and Faulkner 1993].

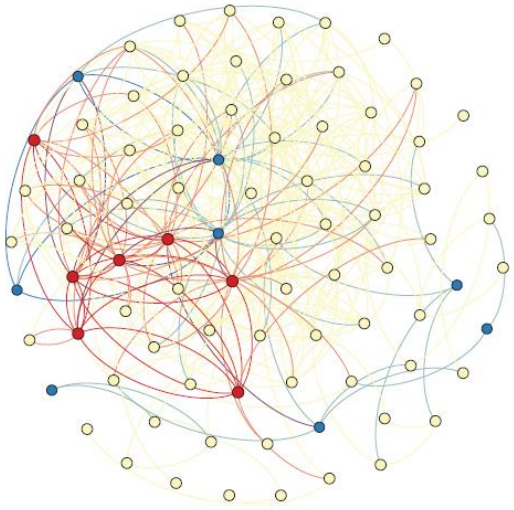

Figure 1. The 83-employee network. Red (darker gray) nodes are known conspirators.

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, or (as in this case) exchange messages [Krebs 2002].

Since the nodes in the graph can be a mixture of both criminals and noncriminal, it is desirable to determine suspected criminals from topological properties of the network and other prior knowledge, which includes known criminals, known non-criminals, and information related to their interactions. Moreover, we desire a priority list of descending criminal likelihood so as to identify the primary leader of the organization.

To find the leader of the conspirators, we apply a dynamics-based ranking algorithm on a subgraph extracted from the network. Our findings are in agreement with empirical knowledge about the centrality balance of criminal leaders.
Finally, we perform sensitivity analysis to test the robustness of our approach.

PROBLEM CLARIFICATION

A conspiracy network is embedded in a network of employees of a company, with each edge representing a message sent from one employee (node) to another and categorized by topics. Given a few known criminals, a few known non-criminals, and suspicious topics, we seek to estimate the probability of criminal involvement for other individuals and to determine the leader of the conspirators.

MODEL DESIGN AND JUSTIFICATION

For an unidentified node (an employee not identified as a conspirator or non-conspirator), we model the probability of conspiracy as a sigmoid function of a linear combination of the node's features (logistic regression). Those features are formulated from local topological measures and the node's semantic messaging patterns. Parameters of the model are trained on a subset of identified conspirators and non-conspirators. The performance of the model is enhanced by discovering potential similarities among topics via topic-word diffusion dynamics on a bipartite graph. We also perform resource allocation dynamics to identify the leader of the conspirators; the identification is supported by empirical evidence in criminal network research.

STRENGTHS AND WEAKNESSES

The combination of topological properties and semantic affinity among individuals leads to good performance. The time complexity of the algorithm is linear, so the method is suitable for large amounts of data. However, our model requires assistance from semantic network analysis to form an expert dictionary. Also, intrinsic differences among networks may hinder portability of the model's features.

A MACHINE LEARNING SOLUTION

We use machine learning mainly because of its reorganization, which simulate humans' actions to obtain fresh knowledge.

We describe the construction of our machine learning framework in detail, including feature formulation, core learning methods, and experimental results. Through statistical analysis on the results, we propose an enhancement based on semantic diffusion.

We commence with several necessary assumptions:

• All the data and information about the EZ case network and the 83-node network are relatively stable over a long period.

• The contents of the communication among conspirators tends to be relevant about suspicious topics or some formal issues, rather than gossip.

• The two networks feature similar core mechanisms for communication transmission.

Number of known neighboring conspirators:

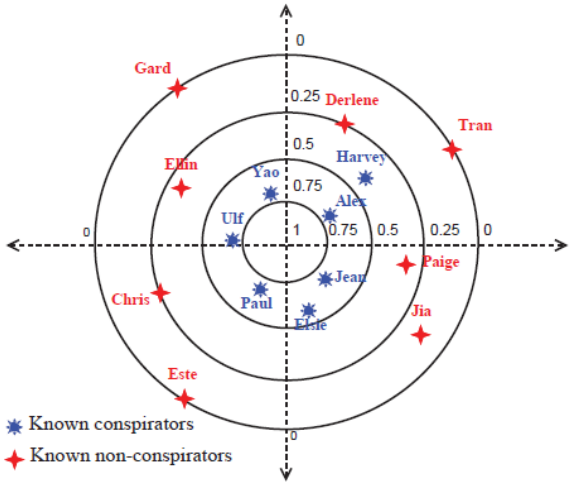

Figure 2.Ratio of known conspirators among adjacent neighbors.

We consider as a significant feature the number of known neighboring conspirators of a node. The interaction among conspirators in a message network suggests a much stronger connectivity than that among non-conspirators: A conspirator is more likely to communicate with an accomplice. As shown in **Figure 2**, we calculate the ratio of known conspirators among one's adjacent neighbors, which measures proximity with known accomplices: The value is 1 if the individual connects with all the known conspirators, and 0 means that no conspirators connect to the individual. The known suspicious clique obviously represents a more compact connectivity. Therefore, the more known conspirators among an individual's neighbors, the greater the possibility that the individual is an accomplice.

Number of current non-suspicious messages from known conspirators:

Figure 3 shows the topics mentioned between known conspirators. A known conspirator rarely talks with accomplices about topics irrelevant to their conspiracy, though a very small proportion of unknown topics appear. If most of the information received from a known conspirator is irrelevant, the receiver is probably not a conspirator.

	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*	

Figure 3. Topics among known conspirators. Known conspiratorial topics have an asterisk.

CONCLUSION

The trained hypothesis gives the estimated probability for node being a conspirator, resulting in a priority list of suspects, ranked in descent order of criminal likelihood. **Figure 4** illustrates the probability of criminal involvement estimated by $h(x)$ versus the corresponding rank in the priority list, where three managers (Jerome, Dolores, and Gretchen) are marked by circles. Dolores (manager) is indeed the person deserving highest suspicion, and Jerome (manager) is also likely to be involved in conspiracy.

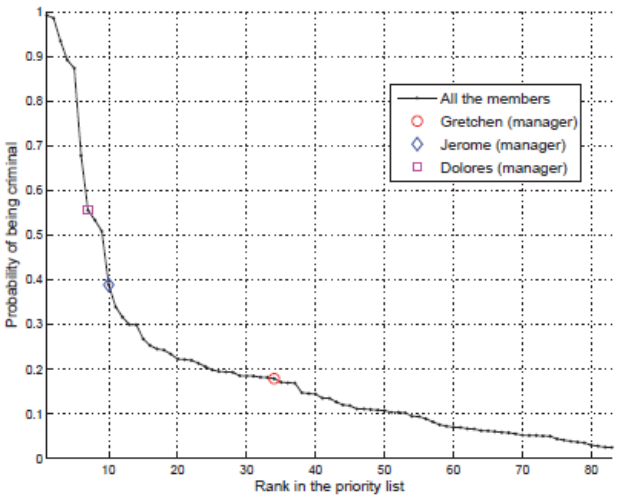

Figure 4.Probability of conspiracy vs. corresponding rank in the priority list.

I am very glad that you can read my research report. If there is any shortage in my report, I hope you can make some valuable suggestions for improvement!

Sincerely,

Kailin Tang

Kailin Tang