2012 ICM Problem
Modeling for Crime Busting

Your organization, the Intergalactic Crime Modelers (ICM), is investigating a conspiracy to commit a criminal act. The investigators are highly confident they know several members of the conspiracy, but hope to identify the other members and the leaders before they make arrests. The conspirators and the possible suspected conspirators all work for the same company in a large office complex. The company has been growing fast and making a name for itself in developing and marketing computer software for banks and credit card companies. ICM has recently found a small set of messages from a group of 82 workers in the company that they believe will help them find the most likely candidates for the unidentified co-conspirators and unknown leaders. Since the message traffic is for all the office workers in the company, it is very likely that some (maybe many) of the identified communicators in the message traffic are not involved in the conspiracy. In fact, they are certain that they know some people who are not in the conspiracy. The goal of the modeling effort will be to identify people in the office complex who are the most likely conspirators. A priority list would be ideal so ICM could investigate, place under surveillance, and/or interrogate the most likely candidates. A discriminate line separating conspirators from non-conspirators would also be helpful to distinctly categorize the people in each group. It would also be helpful to the DA's office if the model nominated the conspiracy leaders.

Before the data of the current case are given to your crime modeling team, your supervisor gives you the following scenario (called Investigation EZ) that she worked on a few years ago in another city. Even though she is very proud of her work on the EZ case, she says it is just a very small, simple example, but it may help you understand your task. Her data follow:

The ten people she was considering as conspirators were: Anne#, Bob, Carol, Dave*, Ellen, Fred, George*, Harry, Inez, and Jaye#. (* indicates prior known conspirators, # indicate prior known non-conspirators)

Chronology of the 28 messages that she had for her case with a code number for the topic of each message that she assigned based on her analysis of the message:

Anne to Bob: Why were you late today? (1)

Bob to Carol: That darn Anne always watches me. I wasn't late. (1)

Carol to Dave: Anne and Bob are fighting again over Bob's tardiness. (1)

Dave to Ellen: I need to see you this morning. When can you come by? Bring the budget files. (2)

Dave to Fred: I can come by and see you anytime today. Let me know when it is a good time. Should I bring the budget files? (2)

Dave to George: I will see you later --- lots to talk about. I hope the others are ready. It is important to get this right. (3)

Harry to George: You seem stressed. What is going on? Our budget will be fine. (2) (4)

Inez to George: I am real tired today. How are you doing? (5)

Jaye to Inez: Not much going on today. Wanna go to lunch today? (5)

Inez to Jaye: Good thing it is quiet. I am exhausted. Can't do lunch today --- sorry! (5)

George to Dave: Time to talk --- now! (3)

Jaye to Anne: Can you go to lunch today? (5)

Dave to George: I can't. On my way to see Fred. (3)

George to Dave: Get here after that. (3)

Anne to Carol: Who is supposed to watch Bob? He is goofing off all the time. (1)

Carol to Anne: Leave him alone. He is working well with George and Dave. (1)

George to Dave: This is important. Darn Fred. How about Ellen? (3)

Ellen to George: Have you talked with Dave? (3)

George to Ellen: Not yet. Did you? (3)

Bob to Anne: I wasn't late. And just so you know --- I am working through lunch. (1)

Bob to Dave: Tell them I wasn't late. You know me. (1)

Ellen to Carol: Get with Anne and figure out the budget meeting schedule for next week and help me calm George. (2)

Harry to Dave: Did you notice that George is stressed out again today? (4)

Dave to George: Darn Harry thinks you are stressed. Don't get him worried or he will be nosing around. (4)

George to Harry: Just working late and having problems at home. I will be fine. (4)

Ellen to Harry: Would it be OK, if I miss the meeting today? Fred will be there and he knows the budget better than I do. (2)

Harry to Fred: I think next year's budget is stressing out a few people. Maybe we should take time to reassure people today. (2) (4)

Fred to Harry: I think our budget is pretty healthy. I don't see anything to stress over. (2)

END of MESSAGE TRAFFIC

Your supervisor points outs that she assigned and coded only 5 different topics of messages: 1) Bob's tardiness, 2) the budget, 3) important unknown issue but assumed to be part of conspiracy, 4) George's stress, and 5) lunch and other social issues. As seen in the message coding, some messages had two topics assigned because of the content of the messages.

The way your supervisor analyzed her situation was with a network that showed the communication links and the types of messages. The following figure is a model of the message network that resulted with the code for the types of messages annotated on the network graph.

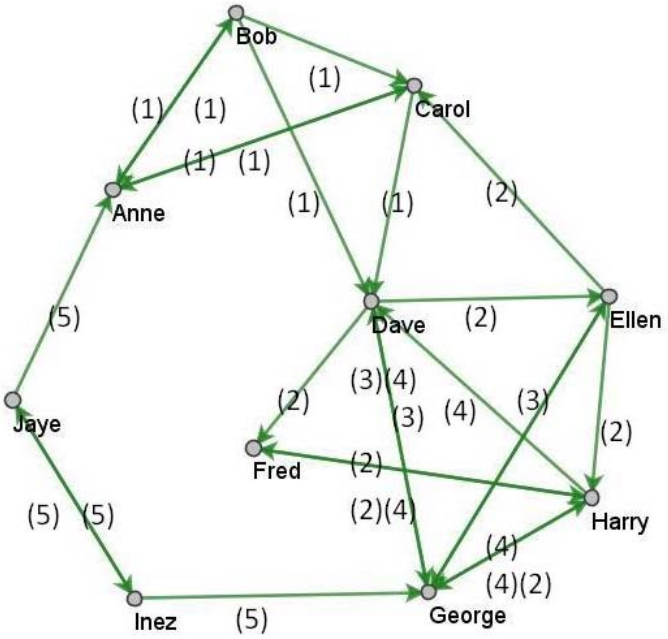

Figure 1: Network of messages from EZ Case

Your supervisor points out that in addition to known conspirators George and Dave, Ellen and Carol were indicted for the conspiracy based on your supervisor's analysis and later Bob self-admitted his involvement in a plea bargain for a reduced sentence, but the charges against Carol were later dropped. Your supervisor is still pretty sure Inez was involved, but the case against her was never established. Your supervisor's advice to your team is identify the guilty parties so that people like Inez don't get off, people like Carol are not falsely accused, and ICM gets the credit so people like Bob do not have the opportunity to get reduced sentences.

The current case:

Your supervisor has put together a network-like database for the current case, which has the same structure, but is a bit larger in scope. The investigators have some indications that a conspiracy is 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. The small example she showed you for case EZ had only 10 people (nodes), 27 links (messages), 5 topics, 1 suspicious/conspiracy topic, 2 known conspirators, and 2

known non-conspirators. So far, the new case has 83 nodes, 400 links (some involving more than 1 topic), over 21,000 words of message traffic, 15 topics (3 have been deemed to be suspicious), 7 known conspirators, and 8 known non-conspirators. These data are given in the attached spreadsheet files: Names.xls, Topics.xls, and Messages.xls. Names.xls contains the key of node number to the office workers' names. Topics.xls contains the code for the 15 topic numbers to a short description of the topics. Because of security and privacy issues, your team will not have direct transcripts of all the message traffic. Messages.xls provides the links of the nodes that transmitted messages and the topic code numbers that the messages contained. Several messages contained up to three topics. To help visualize the message traffic, a network model of the people and message links is provided in Figure 2. In this case, the topics of the messages are not shown in the figure as they were in Figure 1. These topic numbers are given in the file Messages.xls and described in Topics.xls.

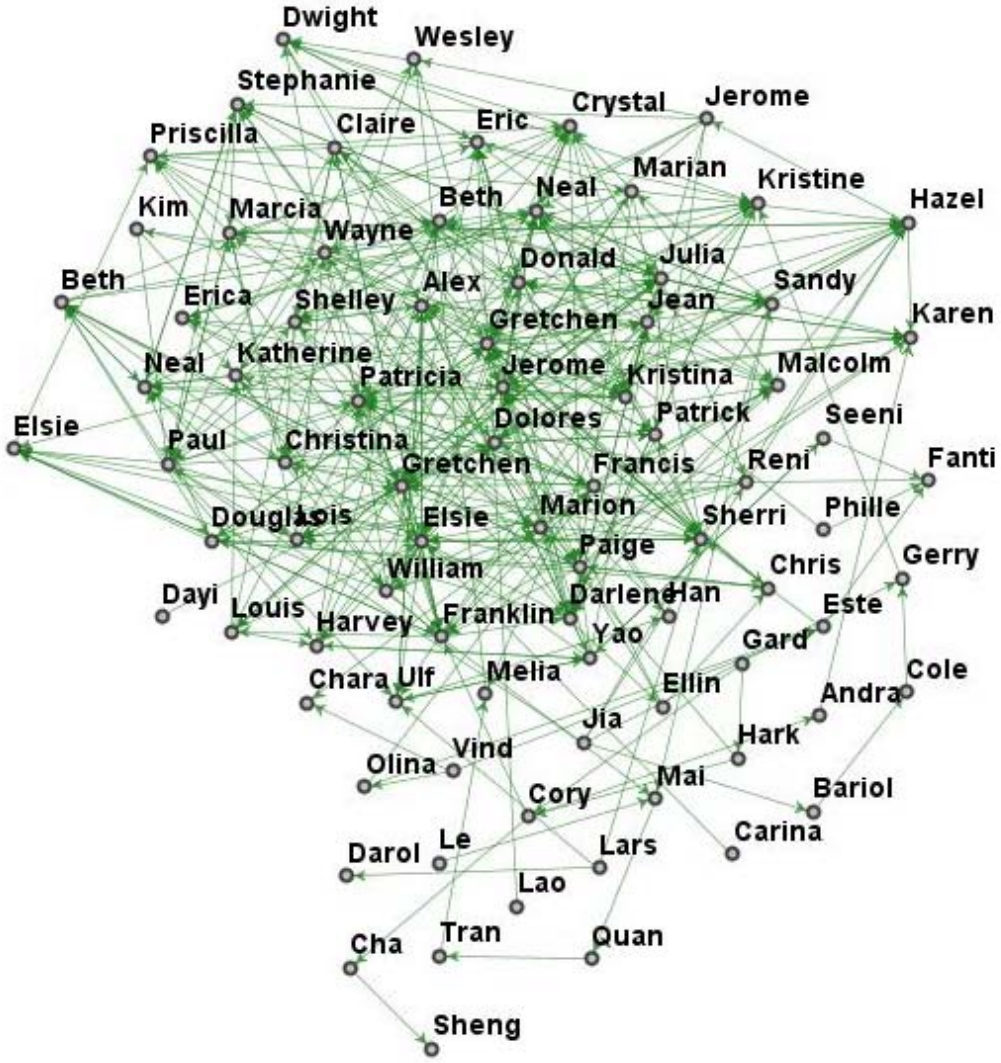

Figure 2: Visual of the network model of the 83 people (nodes) and 400 messages between these people (links).

Requirements:

Requirement 1: So far, it is known that Jean, Alex, Elsie, Paul, Ulf, Yao, and Harvey are conspirators. Also, it is known that Darlene, Tran, Jia, Ellin, Gard, Chris, Paige, and Este are not conspirators. The three known suspicious message topics are 7, 11, and 13. There is more detail about the topics in file Topics.xls. Build a model and algorithm to prioritize the 83 nodes by likelihood of being part of the conspiracy and explain your model and metrics. Jerome, Delores, and Gretchen are the senior managers of the company. It would be very helpful to know if any of them are involved in the conspiracy.

Requirement 2: How would the priority list change if new information comes to light that Topic 1 is also connected to the conspiracy and that Chris is one of the conspirators?

Requirement 3: A powerful technique to obtain and understand text information similar to this message traffic is called semantic network analysis; as a methodology in artificial intelligence and computational linguistics, it provides a structure and process for reasoning about knowledge or language. Another computational linguistics capability in natural language processing is text analysis. For our crime busting scenario, explain how semantic and text analyses of the content and context of the message traffic (if you could obtain the original messages) could empower your team to develop even better models and categorizations of the office personnel. Did you use any of these capabilities on the topic descriptions in file Topics.xls to enhance your model?

Requirement 4: Your complete report will eventually go to the DA so it must be detailed and clearly state your assumptions and methodology, but cannot exceed 20 pages of write up. You may include your programs as appendices in separate files that do not count in your page restriction, but including these programs is not necessary. Your supervisor wants ICM to be the world's best in solving white-collar, high-tech conspiracy crimes and hopes your methodology will contribute to solving important cases around the world, especially those with very large databases of message traffic (thousands of people with tens of thousands of messages and possibly millions of words). She specifically asked you to include a discussion on how more thorough network, semantic, and text analyses of the message contents could help with your model and recommendations. As part of your report to her, explain the network modeling techniques you have used and why and how they can be used to identify, prioritize, and categorize similar nodes in a network database of any type, not just crime conspiracies and message data. For instance, could your method find the infected or diseased cells in a biological network where you had various kinds of image or chemical data for the nodes indicating infection probabilities and already identified some infected nodes?

***Your ICM submission should consist of a 1 page Summary Sheet and your solution cannot exceed 20 pages for a maximum of 21 pages.**