Ramapo College of New Jersey

# Identifying Threat:

# Classifying Malicious Intent via Networks and Machine Learning

Ritwik Katiyar

December 20th, 2023

# Introduction

In the complex world of white-collar crime, the Intergalactic Crime Modelers (ICM) face the daunting task of unraveling conspiracies within large organizations. This project presents a challenging scenario where the ICM investigates a conspiracy involving the embezzlement of funds and internet fraud within a rapidly growing software company. Drawing inspiration from a past case known as Investigation EZ, the team is armed with a set of 82 messages exchanged among the office workers, aiming to identify potential conspirators and leaders. Building upon the lessons learned from the EZ case, the ICM supervisor introduces a network-based approach to modeling the communication links and message topics, emphasizing the need for a discriminating model to distinguish between actual conspirators and innocent employees. With the current case involving 83 nodes, 400 links, and over 21,000 words of message traffic, the project outlines the requirements to develop a prioritized list of likely conspirators, adapt to new information, leverage semantic and text analyses, and present a comprehensive report for legal proceedings. The investigation unfolds with known conspirators and non-conspirators, introducing new challenges such as suspicious message topics and the involvement of senior managers. As the ICM navigates this intricate web of communication, the project not only aims to crack the current case but also envisions a methodology that could be applied globally to solve high-tech conspiracy crimes with extensive databases. In this endeavor, the report explores the potential of network modeling techniques to identify, prioritize, and categorize nodes within various types of networks, extending beyond crime conspiracies to applications such as identifying infected cells in biological networks. The ICM's mission is not only to crack the present case but also to contribute to the global fight against complex white-collar crimes, leveraging advanced computational methods and methodologies to make a lasting impact.

# Problem Statement

Background:

- The Intergalactic Crime Modelers (ICM) are investigating a conspiracy within a rapidly growing software development company.
- The goal is to identify potential conspirators, prioritize them, and discern leaders before making arrests.

Previous Case (Investigation EZ):

- The supervisor's analysis involved a network model based on message traffic.
- Identified conspirators (George, Dave, Ellen, Carol) and non-conspirators (Bob, Anne).
- Emphasized the need for accuracy to avoid false accusations and ensure justice.

Current Case Overview:

- Larger scale with 83 nodes, 400 links, and 15 topics, involving embezzlement and internet fraud.
- 7 known conspirators (Jean, Alex, Elsie, Paul, Ulf, Yao, Harvey) and 8 known non-conspirators (Darlene, Tran, Jia, Ellin, Gard, Chris, Paige, Este).
- 3 suspicious message topics (7, 11, 13) with details in Topics.xls.
- At most, three topics were discussed within the same message between two individuals.

Build a Model and Algorithm:

- Prioritize 83 nodes based on the likelihood of involvement in the conspiracy.
- Known conspirators and non-conspirators provided.
- Three suspicious topics were identified (7, 11, 13).
- Senior managers (Jerome, Delores, Gretchen) involvement needs clarification.

Account for Priority List with New Information:

- Assess the impact on the priority list if Topic 1 is linked to the conspiracy, and Chris is a conspirator.

Semantic and Text Analysis:

- Explain how semantic network analysis and text analysis could enhance modeling.
- Explore the potential use of these techniques on the provided topic descriptions (Topics.xls).

Report to DA and Methodology:

- Include the possibility of large-scale application to high-tech conspiracy crimes globally.
- Discuss how network modeling techniques can be adapted beyond crime scenarios.

By addressing these requirements, the ICM team aims to develop an effective model to uncover and prioritize potential conspirators, contributing to the fight against white-collar crimes.

# Message Categories

This section aims to display the various message categories that were provided to us. The messages have been simplified to a specific topic number based on what the discussion was generally about. There are fifteen different topics outlined as follows:

1. Discussion of company finances → Stock Price, yearly earnings, prospects, and sales.
2. Discussion of new product line for the company.
3. Complaints about the cleanliness and maintenance of the office building.
4. Reflections on last week's office party and its after-effects. (Alex and Elise's argument with Paige)
5. Complaints about building a computer database security system.
6. Discussion/ Worry about the company politics and who would be promoted to a vacant vice president position.
7. Invitations and discussion around a private meeting that was going to be held at Paul's house in the evening. [Considered as a malicious or incriminating message if a message possesses this label].
8. Planning a ski trip in the next two weeks.
9. Debates upon the merits and performance of the local high school football team.
10. Steps to open a new satellite business office in Mexico. Relocation and co-ordination of current employees.
11. Messages regarding the accounting, credit card, and auditing packages used by the company. [Considered to be possibly conspiratorial by its content and context]
12. Best restaurants to go to lunch.
13. The discussion was about the rollout of the new computers and how to establish a new network for the office. The crux of the discussion was centered on when and for how long would Paige, Ellin, and Chris be offline; in addition to discussions on how to make the offline time longer. [Considered key to the conspiracy plan]
14. Discussants of this topic were concerned about the high price of a new product introduced by the company.
15. Concerns regarding the amount of computer security training the company was requiring the employees to undergo.

# Assumptions

In this section outlined some of the general assumptions that needed to be made to solve this problem:

- The messages were labeled correctly based on the topics specified. Some messages that were in Spanish were also labeled correctly despite being in a different language.
- The suspicious messages are all equally incriminating. No message is more incriminating than another.
- There were no hidden messages and all messages that are non-conspiratorial are equally benign in nature.
- The people indicated to be part of the plot as well as those considered innocent are in fact a part of the conspiracy or were not involved.
- There were no more than three topics of discussion that occurred at once between two individuals.

# Model Development and Results

In a nutshell, our approach was to utilize machine learning to generate probabilities for an individual's involvement based on the messages sent. Then, with the use of network centrality measures, sus out potential co-conspirators or individuals of interest that may require further investigation. As outlined we begin by assigning weights to the edges of our networks based on the types of messages that were shared.

## **Data Processing**:

In preparation for the determination of the edge weights, a data analysis and clearing process was executed to ensure the absence of errors or flaws within the provided dataset. Furthermore, specific simplifications were applied to the data to enhance its integration into machine learning algorithms.

1. Some employees had the same first names. Since no last names were provided the second worker in the list with the same name was marked with 'B' after their first name.
2. A new 'incrimination' column was added where the presence of an incrementing message as one of the topics within the messages sent by the user would flag the user as suspicious represented as 1's or non-suspicious; as 0's.
3. We observed that due to the application of our incrimination column, there were far more 0's than 1's within our dataset (97 1's and 303 0's). Hence, we used random oversampling to balance our class distribution to obtain the best results from our machine-learning

algorithms. With the random oversampling, we now had 303 0’s and 303 1’s within our dataset.

## Machine Learning:

To generate a probability of an individual being suspicious based on the topics of the messages that they sent, we first considered the application of logistic regression; regarded as one of the most basic statistical algorithms for classification problems. Logistic regression is used to determine the relationship between variables such as (James):

$$p(X) = \Pr(Y = 1|X)$$

Equation 1: Shows the basic idea behind logistic regression (James).

Where the probability of the independent variable equals the probability of the dependent variable being ‘1’ given X. Mathematically this can be presented as

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

Equation 2: Shows the logistic function used to generate probabilities of a function.

Given the equation, the probability of X or the independent variable can be determined via the natural exponent of the intercept summed with the slope of the first independent variable. Over one plus the natural exponent of the intercept is summed with the slope of the first independent variable. For our specific problem, this can be represented as (James):

$$p(suspecious) = \frac{e^{\beta_0+'Text\ Topic\ 1\ X\ +\ Text\ Topix\ 2\ X\ +\ Text\ Topic\ 3\ X}}{1\ +\ e^{\beta_0+'Text\ Topic\ 1\ X\ +\ Text\ Topix\ 2\ X\ +\ Text\ Topic\ 3\ X}}$$

Equation 3: Shows the logistic function written in terms of our problem.

With the logistic function, we were able to obtain an accuracy of just 0.675 suggesting that the model couldn’t accurately determine the probability of an individual being suspicious. Additionally, we can observe from our confusion matrix that our model failed to classify individuals from our data set.

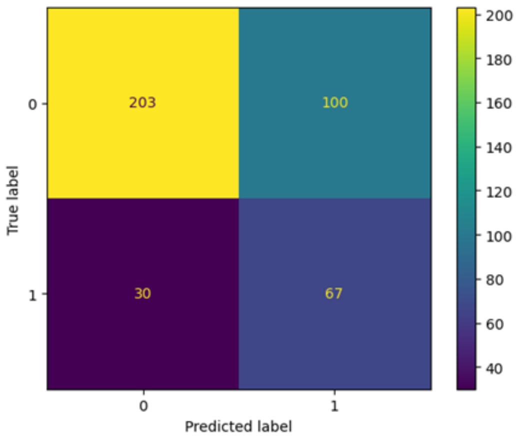

Figure 1: Illustrates the confusion matrix obtained from the logistic regression model.

Based on our assumption that sending malicious messages equates to suspicious behavior. We need all individuals marked accurately. Hence, we decided to employ the random forest algorithm which is a far more sophisticated algorithm than logistic regression, however, there is a significant drawback in the sense that it can be hard to determine how exactly the algorithm arrived at conclusions that it did (black box model).

Decision trees are non-parametric supervised learning methods that are commonly used for classification and regression. The objective of a decision tree is to build a model that can predict the value of the target by learning simple decision rules from the features provided. Most decision trees use the Gini index to determine the nodes it needs to select. The Gini score is a measure of the impurity of the node. A node is considered pure if all training instances it applies to belong to the same class. The score can be defined as (James):

$$G_i = 1 - \sum_{k=1}^{n} P_{i,k}^2$$

Equation 4: Gini Index/ Gini Score used by decision trees to generate appropriate splits in nodes within a tree(James)

Gini score impurity score: where Pi,k is the ratio of class k instances in the ith node. Although with the implementation of this technique, if the same impurity score is shared between several features during a split, the split can be selected at random. Resulting in a level of randomness within trees, which is why not every decision tree is the same. Hence, Random Forest are largely considered over decision trees, as a random forest is simply a collection or an ensemble of decision trees. Each decision tree within the ensemble is built upon a random subset of input features, and eventually, the majority vote among the trees is taken to predict results (James).

With the random forest classifier, we were able to obtain an accuracy of 1.0 Suggesting that our model was able to accurately classify everyone. However, we noticed that the probability generated was about 100% for each value. Suggesting that our model may have overfit and since we are attempting to generate probabilities based on the messages having probabilities of 1's and 0's isn't ideal. Hence, to prevent overfitting of our random forest model we added the criteria to limit the max-depth of the tree to be seven splits' deep. With the addition of our max depth, we were able to generate probabilities lower than 1.0 and maintain the accuracy of 1.0. Furthermore, we can observe the confusion matrix below for the predicted vs. true results.

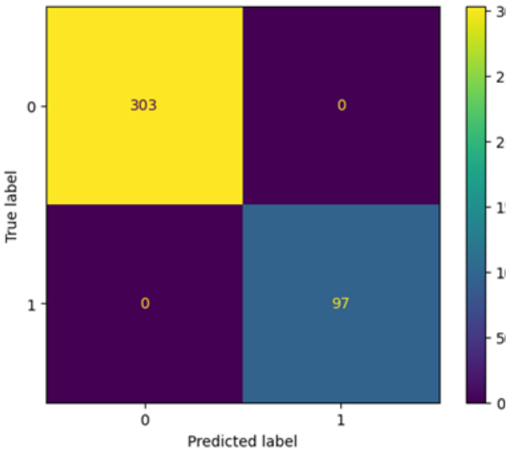

Figure 2: Shows the confusion matrix obtained from the random forest classifier.

We can observe the weights as follows; for example: A message that contained all three suspicious topics (7,11,13) was assigned a 99% probability of being suspicious vs. 90% probability when a message contained one suspicious message as its second topic (2, 7, null).

With these obtained probabilities we can now construct a weighted multidirectional network graph; with the probabilities as edge weights and utilize centrality measures to highlight and identify conspirators within our network.

## **Graph Construction**:

To construct the graphical representation of our dataset, we employed the Gephi program for visualization and interactive analysis. Specifically, we utilized the ForceAtlas2 algorithm with Lin-Log mode activated to facilitate a rapid and straightforward visualization process. It is noteworthy that ForceAtlas2 operates as a force-directed layout algorithm, simulating a physical system to spatialize a network. In this simulation, nodes behave akin to charged particles, repelling each other, while edges exert an attractive force on their connected nodes, resembling springs. The algorithm strategically employs these forces to achieve a balanced spatial arrangement of the network, ultimately enhancing data interpretation and visualization. (Jacomy)

It is pertinent to mention that ForceAtlas2 accommodates weighted edges, allowing for the consideration of edge weights in the visualization process. This feature proves valuable in grouping nodes based on the provided edge weights, thereby contributing to a more nuanced representation of the network structure. The incorporation of Lin-Log mode in the algorithm represents an adjustment aimed at improving the legibility of node placement within the network. Without this mode, the visualization may become overly intricate and congested, which can hinder the ability to make meaningful inferences from the data. (Jacomy)

In light of these distinctive characteristics, the decision to adopt ForceAtlas2 over alternative visualization techniques available on Gephi was made. Moreover, after the base visualization was implemented, the edges were colored, and the size of the edges was based on the weight to offer more visual clarity to the user. Figure 3 shows the resulting network that was generated from Gephi.

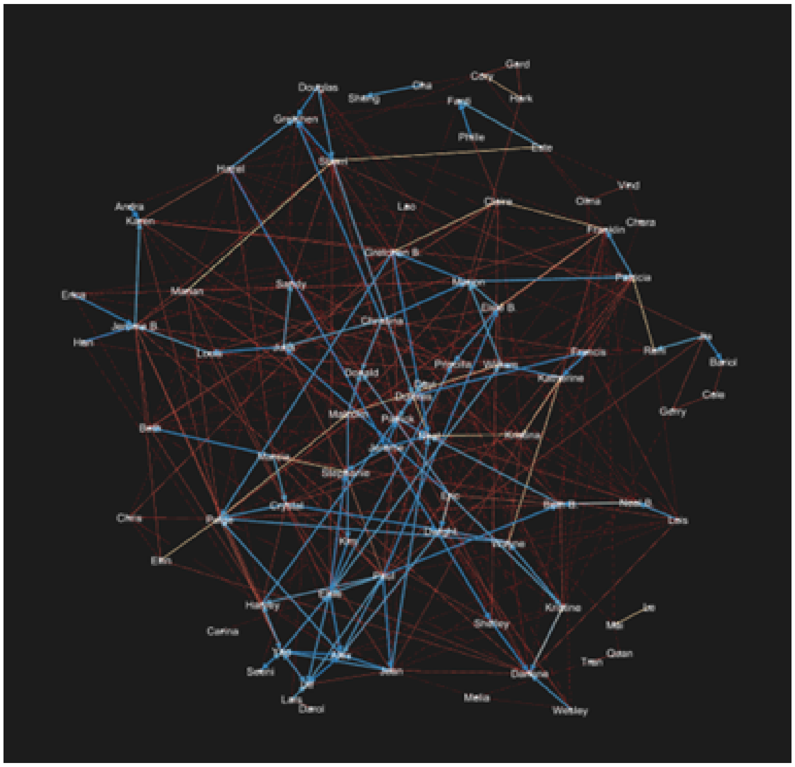

Figure 3: Illustrates the network generated from Gephi. Blue color represents a higher weight as opposed to yellow and red being lower weights. Furthermore, we can observe that the conspirators: Elise, Harvey, Jean, Yao, Alex, Ulf, and Paul are all grouped at the bottom left of the network; with the majority of their messages being marked as suspicious.

With the basic framework of our model established, various centrality measures can now be employed to sus out co-conspirators. However, from our basic plot, we can potentially mark Seeni and Lars as suspicious as they had several incriminating messages shared between Yao and Ulf respectively.

## Centrality Measures:

There is a plethora of centrality measures that can be applied depending on the problem. Some of the measures considered in this study are as follows:

Weighted Out-Degree Centrality:

Generally, the degree centrality is a measure that simply counts the number of neighbors a node constitutes of. But, for directed graphs where we are trying to determine involvement within a conspiracy, it would be beneficial to look at the out-degree or the measure of outgoing links from a particular node. The weighted measure accounts for the edge weights associated with the node. With the weights, we could flag individuals who were sending out malicious messages the most. The weight aspect simply adds the weight of the edge during the out-degree centrality calculation (Candeloro). Moreover, mathematically the Out-Degree centrality (Ci) of a node *i* is given by (Degree):

$$C_I(i) = \sum_{i=1}^{n} A_{ij}$$

Equation 3: Here the out-centrality of a node is calculated by adding across the columns *i* instead of adding down the columns j for in-degree centrality. (Degree)

Page Rank Centrality:

Page Rank assesses a node's significance on the quantity and quality of connections. Nodes that receive connections from other important nodes contribute to a higher score. This score can help identify co-conspirators that have a higher score due to the provided weights based on their suspicious messaging (Ghazaryan)(Disney). Mathematical page rank can be represented as

$$PR(A) = (1-d) + d[\frac{PR(T_1)}{L(T_1)} + \ldots + \frac{PR(T_n)}{L(T_n)}]$$

Equation 4: Where page rank (PR) or page A can be obtained by taking the page rank of all pages that link to A (T), divided by the number of outgoing links (L). The parameter d is referred to as the 'dampening factor', which was set to 0.85. (Generally meant to simulate randomness) (Moor).

Hub Centrality:

Finally, hub centrality or Kleinberg centrality measure is a measure that ranks nodes that point to many important pages. In other words, a strong hub is a node that points to many other stronger nodes. Once again this is a measure that in conjunction with other centrality measures flags potential co-conspirators as individuals and would be considered a good hub if they sent malicious messages to other individuals who are also conspirators. In terms of a mathematical formula hub centrality can be shown as (Kleinberg):

$$y_i = \beta \sum_{i \epsilon N} w_{ji} x_i$$

Equation 5: Here the hub centrality ($y_i$ ) of a node j, $w_{ij}$ is the sum of weights of links from node I to node j times a constant (β) (Kleinberg).

## <u>Implementation</u>:

Once the centrality measures were acquired for determining co-conspirators an average score of all the centrality scores was obtained and implemented onto the graph. Figure 4 showcases the results that were obtained on the network that was constructed in Figure 3.

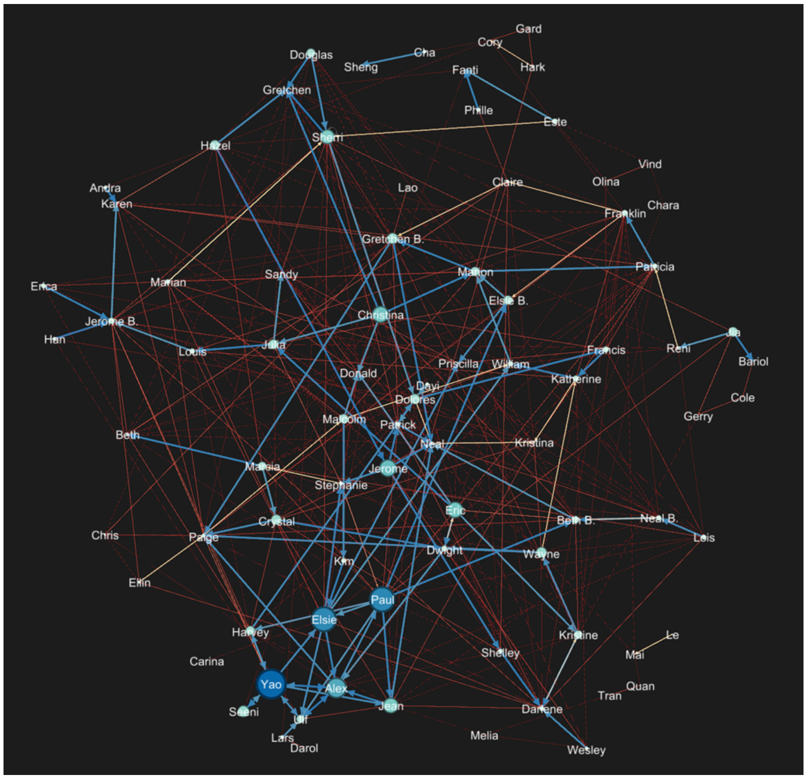

Figure 4: Illustrates the Network after the application of our centrality measures. Larger size and deeper color represent the more important nodes vs. smaller size and smaller nodes represent less important nodes determined by our measures.

We can observe that most of the conspirators (Bottom Left) seem to have much larger nodes with deeper colors, to indicate that they were in fact rightfully identified by our measures and the network. Moreover, we can now apply the lowest score given to a known conspirator as the minimum score an individual needs to have to not be considered suspicious. With the application of this range, we have determined the following suspects:

| Name | Score |
|---|---|
| Yao | 2.418860997649717419 |
| Paul | 2.0437656626 |
| Elsie | 2.0241483034 |
| Alex | 1.623990076517416638 |
| Jerome | 1.3720356068603785752 |
| Christina | 1.3241734709 |
| Eric | 1.280216549381706488 |
| Jean | 1.215627264920906172 |
| Sherri | 1.1561184195 |
| Seeni | 0.9113366509441017078 |
| Gretchen B. | 0.8833004976 |
| Wayne | 0.8616189322 |
| Crystal | 0.8268368545 |
| Dolores | 0.823337514714543202 |
| Hazel | 0.8070719408 |
| Julia | 0.7797634051 |
| Jia | 0.767910173091081453 |
| Malcolm | 0.75777401767547149 |
| Harvey | 0.734422511640470404 |
| Marion | 0.7304809664 |
| Douglas | 0.7278376413 |
| Elsie B. | 0.7230066229 |
| Marcia | 0.701229281273719067 |
| Kristine | 0.68361559461840859 |
| Ulf | 0.663060788394897812 |

Table 1: List of suspects that fall within the range of 2.418 and 0.663 scores of all the known conspirators (shaded in blue). Jia is known to be innocent hence she was shaded in red to indicate an error with our technique. Furthermore, Jermon, Gretchen, and Dolores are senior managers of the company and are hence highlighted in green.

# Discussion & Conclusion

Based on the outcomes of our analysis, our model successfully identified a roster of 18 potential suspects through the application of centrality scores, emphasizing the prevalence of incriminating weighted messages associated with these individuals. Notably, our algorithm flagged all three senior managers, a result congruent with the paradigm shift advocated in scholarly reports such as '*The Organization of Corporate Crime*' and '*Corporate Crime and Punishment: An Empirical Study*.' These references underscore the imperative for a comprehensive perspective on corporate crime, implicating managerial responsibility in preventing or halting illicit activities (Lund) (Van). While acknowledging that our algorithmic flags do not conclusively establish the managers' guilt, we contend that their identification warrants a thorough investigation given their potential awareness of the crime.

It is essential, however, to exercise caution in our interpretation of flagged activities. The algorithm detects suspicious behavior tied to individuals involved in or discussing critical topics. The absence of contextual information poses a challenge in definitively discerning whether the mention of an event in a message is genuinely harmful or benign. For instance, an employee inquiring about an incident should not automatically trigger suspicion, as opposed to a conversation between two employees engaged in malicious discussions. Therefore, a nuanced evaluation of contextual cues is indispensable to avoid mischaracterizing interactions in our investigative process. With the addition of sentiment analysis our algorithm for determining suspicious messages doesn't need to be changed drastically. The main change would be to our 'incrimination' column that we added in the beginning during the data processing stage.

The versatility of our model extends beyond its current application, showcasing potential efficacy in diverse fields, such as tracking the propagation of viruses within biological or computational networks. Notably, our model draws inspiration from the paper '*Centrality Measures for Graph-Based Bot Detection Using Machine Learning*.' (Shinan) Despite this foundation, we have implemented critical modifications to tailor our model, differentiating it to help with the problem at hand. In the realm of cybersecurity, our model holds promise for detecting the spread of computer viruses across networks. By employing network analysis, it becomes a valuable tool to pinpoint individuals responsible for the dissemination of a virus within a network. The adaptability of our approach is evident in the context of biological systems. Within biological networks, our algorithm can identify pivotal nodes, such as cells or processes, that exhibit heightened secretion of specific chemicals and hormones.

The beauty of network analysis lies in its broad applicability, transcending specific problem domains. Our technique's potential spans various fields, offering a unique perspective and contributing to solutions beyond its initial development context. As we explore these applications, we aspire to provide innovative insights and contribute to addressing challenges in diverse domains

# Sensitivity Analysis

Parameter Changes:

In response to our organization's request to assess the adaptability of our model to changes in priority and the incorporation of new information, we conducted a test by introducing Topic 1 to the catalog of malicious messages. Additionally, we identified Chris as a co-conspirator, augmenting the original list of seven suspects (Alex, Jean, Elsie, Paul, Ulf, Yao, and Harvey). The evaluation of these adjustments revealed noteworthy observations. To accommodate the inclusion of Topic 1 and the addition of Chris to the list of co-conspirators, we found it necessary to modify our random forest model. Specifically, the maximum depth parameter was adjusted from 7 to 8, ensuring a sustained accuracy rate of 1.0.

The application of centrality metrics and the subsequent construction of the network produced discernible changes in the results. As we analyze the updated network, it becomes imperative to scrutinize the evolving patterns of interactions and their implications for suspect identification. These modifications underscore the model's flexibility in adapting to shifts in priorities and the infusion of new data, demonstrating its robustness in maintaining accurate assessments even amidst dynamic scenarios.

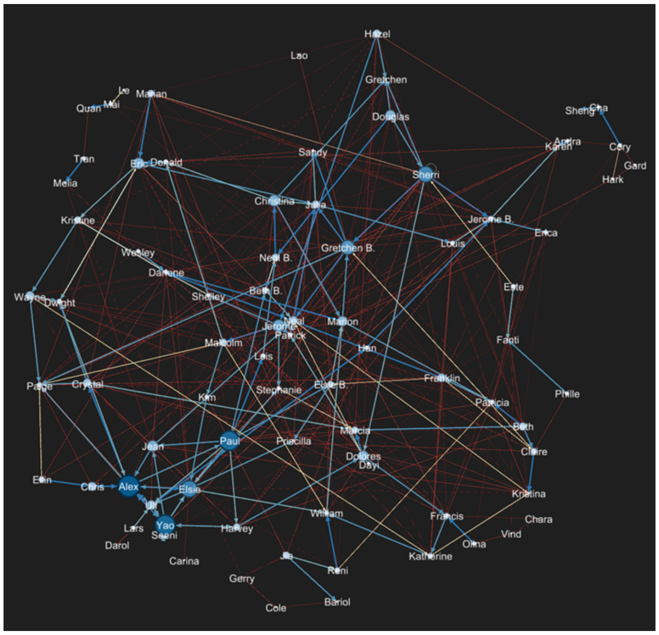

Figure 5: Displays the network with topic 1 included as an incriminating message and Chris added as a co-conspirator. We can observe that Chris is now grouped closer to the other conspirators within the network. Furthermore, Chris seems to have more importance than his previously observed importance when topic 1 was not considered to be incriminating.

| Name | Score |
|---|---|
| Alex | 2.5255537604 |
| Yao | 2.300812271910039864 |
| Paul | 2.237116984152789156 |
| Elsie | 1.888344191894441095 |
| Sherri | 1.778792979366574365 |
| Gretchen B. | 1.5780718024 |
| Jerome | 1.3238176411 |
| Christina | 1.23612472559862601 |
| Eric | 1.184676695104435375 |
| Jean | 1.1764653010 |
| Dolores | 1.1287071782 |
| Douglas | 1.0699700119 |
| Marion | 1.0440949821 |
| Ulf | 0.9773267256 |
| Chris | 0.8937460540 |
| Wayne | 0.8849157167 |
| Neal | 0.884312332060338051 |
| Seeni | 0.8624311292 |
| Franklin | 0.8610253261 |
| Crystal | 0.8217981719 |
| Hazel | 0.773534717885552535 |
| Julia | 0.7646864064 |
| Beth | 0.7571316055 |
| Neal B. | 0.7465184694 |
| Malcolm | 0.735391493499992306 |
| Harvey | 0.7077683712 |

Table 2: Shows the new list of employees that are implicated as conspirators by our model. We can observe that once again every senior employee has been implicated due to the importance each of them holds.

In summary, our model exhibits exceptional flexibility, showcasing its adaptability across diverse scenarios for identifying suspicious individuals. Notably, the incorporation of Topic 1 into the assessment and the inclusion of Chris as a co-conspirator resulted in our model assigning him a higher score. This outcome underscores the model's versatility in effectively capturing nuanced patterns of involvement and content relevance.

## Comprehensive Overview:

In conducting a sensitivity analysis, the model's robustness and adaptability to variations in parameters and input data were thoroughly examined. The initial model development utilized logistic regression and random forest algorithms to assign probabilities of suspicious behavior based on message topics. The model's accuracy was initially limited, prompting the adoption of the more sophisticated random forest algorithm, which yielded a perfect accuracy of 1.0 after addressing overfitting concerns.

Subsequently, centrality measures, including weighted out-degree centrality, Page Rank centrality, and hub centrality, were employed to identify potential conspirators within the network. The resulting network graph, visualized using Gephi, effectively highlighted individuals of interest based on their centrality scores. The model demonstrated its flexibility by successfully incorporating new information, such as the addition of Topic 1 as an incriminating message and identifying Chris as a co-conspirator. Despite these changes, the model maintained its adaptability, producing nuanced and accurate results in the updated scenario. The centrality measures, combined with the network analysis, successfully identified 18 potential suspects, including all three senior managers, illustrating the model's ability to respond to variations in input data and priorities dynamically.

Furthermore, the sensitivity analysis revealed that adjustments to the random forest model's parameters, such as modifying the maximum depth from 7 to 8, were necessary to accommodate the changes in the dataset. The resulting network visualization showed clear distinctions in the patterns of interactions, reflecting the evolving dynamics introduced by the inclusion of Topic 1 and the identification of Chris as a co-conspirator. Notably, the model's adaptability was evident in consistently assigning higher scores to Chris and accurately capturing the shifts in the network structure. This comprehensive sensitivity analysis underscores the model's reliability in responding to dynamic scenarios, offering a robust tool for identifying potential conspirators and contributing to the broader fight against white-collar crimes.

# References

1. (Candeloro) Candeloro, Luca, et al. “A New Weighted Degree Centrality Measure: The Application in an Animal Disease Epidemic.” *PLOS ONE*, vol. 11, no. 11, 2016, doi:10.1371/journal.pone.0165781.
2. (Degree)*DegreeCentrality*, www.sci.unich.it/~francesc/teaching/network/degree.html#:~:text=Degree%20is%20a%20simple%20centrality,the%20number%20of%20successor%20nodes.
3. (Disney) Disney, Andrew. “Social Network Analysis 101: Centrality Measures Explained.” *Cambridge Intelligence*, 14 Aug. 2023, cambridge-intelligence.com/keylines-faqs-social-network-analysis/.
4. (Ghazaryan) Ghazaryan, Ani. “Betweenness Centrality and Other Essential Centrality Measures in Network Analysis.” *Betweenness Centrality*, Memgraph.Com, 1 Sept. 2023, memgraph.com/blog/betweenness-centrality-and-other-centrality-measures-network-analysis.
5. (Jacomy) Jacomy, Mathieu, et al. “FORCEATLAS2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software.” *PLoS ONE*, vol. 9, no. 6, 2014, doi:10.1371/journal.pone.0098679.
6. (James) James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). An Introduction to Statistical Learning: with Applications in R. Springer.
7. (Kleinberg)Kleinberg centrality. (n.d.). https://www.sci.unich.it/~francesc/teaching/network/kleinberg.html
8. (Lund) Lund, Dorothy S. and Sarin, Natasha, "Corporate Crime and Punishment: An Empirical Study" (2021). Faculty Scholarship at Penn Carey Law. 2147. https://scholarship.law.upenn.edu/faculty_scholarship/2147
9. (Moor) Moor, Brian. “Mathematics Behind Google’s Pagerank Algorithm.” *Texas Woman’s University*, Texas Woman’s University, 2018, pp. 10–12.
10. (Shinan) Shinan, Khlood, et al. “Botsward: Centrality Measures for Graph-Based Bot Detection Using Machine Learning.” Computers, Materials &amp;amp; Continua, vol. 74, no. 1, 2023, pp. 693–714, doi:10.32604/cmc.2023.031641.
11. (Van) Van Erp, Judith. “The Organization of Corporate Crime: Introduction to Special Issue of Administrative Sciences.” Administrative Sciences, vol. 8, no. 3, 2018, p. 36, doi:10.3390/admsci8030036.