For office use only T1 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ T2 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ T3 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ T4 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Team Control Number 14531 Problem Chosen C For office use only F1 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ F2 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ F3 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ F4 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 2012 Mathematical Contest in Modeling (MCM) Summary Sheet (Attach a copy of this page to each copy of your solution paper.) Type a summary of your results on this page. Do not include the name of your school, advisor, or team members on this page. Finding Conspirators in the Network: Machine Learning with Resource-allocation Dynamics KEY CONCEPT 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 are fundamentally asked to estimate the probability of criminals involvement for the not identified individuals , and to clarify the leader of conspirators. Besides, relevant discussions are suggested. 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 bipartite dynamics. Resource-allocation dynamics are performed to identify the leader of the conspirators, which win theoretical evidence in criminal network research. Results and Sensitivity Analysis：(i)The accuracy of the machine learning scheme is measured by its performance on leave-one-out cross validation. Basic solution gets 73% prediction accuracy and semantic enhanced solution win 87% correct rate. (ii)The insensitivity of priority conspirator list is manifested by analyzing Kendall's tau. This argument is 0.86 illustrates high stability of the model performance.(iii)The leader we predicted tends to be Yao and the top three in priority list are Dolores, Crystal and Jerome (known conspirators excluded). Strengths and Weaknesses Discussion： The combination of both the topology properties and semantic affinity among individuals leads to a good performance. The time complexity is linear in the whole process in mining of semantic potential information, which is suitable with large amounts of data. However, when facing with large amounts of data, our model prefer obtaining assistance from semantic network analysis to form the expert dictionary. Such features might also meaningless when change a new network background. Machine learning Logistic regression Semantic diffusion Bipartite graph Resource-allocation dynamics Kendall's tau coefficient KEY TECHNIQUES Gradient descent Revised Leader Rank Bipartite graph transmission 更多数学建模资料请关注微店店铺"数学建模学习交流" https://k.weidian.com/RHO6PSpA

Team # 14531 Page 1 of 19 Contents 1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2 2 A Machine Learning Solution to Criminal Priority . . . . . . . . 4 2.1 Feature formulation . . . . . . . . . . . . . . . . . . . . . . . . . . . 4 2.2 Methods . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6 2.2.1 Logistic regression . . . . . . . . . . . . . . . . . . . . . . . 7 2.2.2 Gradient descent . . . . . . . . . . . . . . . . . . . . . . . . 7 2.2.3 Leave-one-out cross validation . . . . . . . . . . . . . . . . . 8 2.2.4 Selecting regularization parameter . . . . . . . . . . . . . . . 8 2.3 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8 2.4 Semantic model enhancement . . . . . . . . . . . . . . . . . . . . . 10 3 Identifying the leader of the conspiracy . . . . . . . . . . . . . . . 12 3.1 LeaderRank . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 13 3.2 Suspicious topic sub-network extraction . . . . . . . . . . . . . . . . 13 3.3 Edge reverse . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 13 3.4 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 14 3.5 Empirical support . . . . . . . . . . . . . . . . . . . . . . . . . . . . 14 3.6 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 15 4 Evaluating the Model . . . . . . . . . . . . . . . . . . . . . . . . . . 15 4.1 Sensitivity analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . 15 4.1.1 Priority list . . . . . . . . . . . . . . . . . . . . . . . . . . . 15 4.1.2 Probability inﬂation . . . . . . . . . . . . . . . . . . . . . . 16 5 Discussion and Conclusion . . . . . . . . . . . . . . . . . . . . . . 18 5.1 Summary . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 18 5.2 Further discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . 18

Team # 14531 Page 2 of 19 1 Introduction As illustrated in Figure 1, criminals and conspirators tend to form organization- al 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 [Baker & Faulkner, 1993]. 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 d- iﬀerent 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 [Krebs, 2002]. 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

Team # 14531 Page 3 of 19 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 [Krebs, 2002]. And Morselli et al. proposed that leaders of a criminal organization tend to balance proﬁt and risk by making a careful trade-oﬀbetween degree centrality and betweenness centrality [Morselli, 2010]. 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 intro- duced, which try to discover the hidden topological patterns by partitioning the big network into small closely connected cliques [Xu, 2005]. Despite the light they shed upon the internal structures of criminal networks, these methods still suﬀer from intimidating complexity with large databases [Wheat, 2007]. In this paper, we carefully combine the local-feature-based methods with ap- proaches 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 compu- tationally 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 sen- sitive 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.

Team # 14531 Page 4 of 19 2 A Machine Learning Solution to Criminal Pri- ority Machine learning is carefully selected by us to play the key role in the en- tire 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, reason- ing 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 close- ness centrality to determine the center of the suspicious network from diﬀerent aspects: ▶Degree centrality [Freeman, 1979] indicates activeness of a member, i.e. the member who tends to have more links to its surroundings. As explained in [Xu & Chen, 2003], degree centrality is not quite reliable to indicate the team leader in a criminal network. For a graph G(V, E), the normalized

Team # 14531 Page 5 of 19 degree centrality of node i is as follows: CD(i) = ∑|V | j=1 ν(i, j) |V | −1 , i ̸= j (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 [Freeman, 1979] 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 oth- er two members transmit information between themselves. The normalized betweenness centrality is deﬁned as: CB(i) = ∑|V | j=1 ∑|V | k<j ωj,k(i) |V | −1 , k ̸= i (2) where ωj,k(i) indicates whether the shortest path between node j and node k passes through node i. ▶Closeness centrality [Sabidussi, 1966] 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: Cc(i) = ∑|V | j=1 ρ(i, j) −Ccmin Ccmax −Ccmin , i ̸= j (3) where ρ(i, j) is the length of the shortest path connecting nodes i and j. Ccmin and Ccmax 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 possession rate of its confederacy among all its

Team # 14531 Page 6 of 19 neighbors, which illustrates his or her compactness 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. 3 0 1 0 0 0 0.25 0.5 0.75 0.25 0.5 0.75 Known non-conspirators Known conspirators Harvey Elsie Alex Yao Ulf Paul Jean Chris Paige Derlene Gard Ellin Tran Este Jia Figure 2: Possession rate of neighboring accomplices distribution • Number of current non-suspicious messages from the known con- spirators Table 2.1 is the topics mentioned between known conspirators.1 It is obvi- ous that a known conspirator rarely talks about irrelevant topics, i.e. topics irrelevant to their conspiracy, with his or her accomplices even though some unknown topics appear among them, which accounts for a very small propor- tion. 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. 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 ob- tained by solving an optimization problem related to training set by gradient ascent algorithm. 1Topic No. 16 in the raw data is regarded as wrong and thus discarded.

Team # 14531 Page 7 of 19 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) 2.2.1 Logistic regression We consider a training set of size m: {(x(1), y(1)), (x(2), y(2)), · · · , (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; θ) = hθ(x) = 1 1 + e−θT x, (4) where θ ∈Rn is the parameter vector. Then, under the framework of generalized linear model, the maximum a poste- riori (MAP) estimate of the parameter θ is given by min θ J(θ), (5) where the cost function is given by J(θ) = 1 m m ∑ i=1 [−y(i) log(hθ(x)(i)) −(1 −y(i)) log(1 −hθ(x(i)))] + λ 2m n ∑ j=1 θ2 j, (6) with λ being the regularization parameter. 2.2.2 Gradient descent The cost function J(θ) 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 θ := θ −α∇θJ(θ), (7) where α is a small positive constant.

Team # 14531 Page 8 of 19 2.2.3 Leave-one-out cross validation As we are only informed of the correct classiﬁcation of 15 nodes, at a given round, we only use 14 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 15 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 Sj = {(x(l), y(l))|l = 1, 2, · · · , j −1, j + 1, · · · , 15}. (8) Using this training set, parameter vector θ(j) is obtained, and the corresponding hypothesis is tested on both Sj and the left-out (x(j), y(j)), arriving at this round's training error εSj and C-V error εj respectively. Hence, by averaging them over j, the training error is given by εS = 1 15 15 ∑ j=1 εSj, (9) and the cross validation error is given by ε = 1 15 15 ∑ j=1 εj. (10) 2.2.4 Selecting regularization parameter The regularization parameter λ (λ > 0) is selected optimally as to minimize the cross validation error, i.e. λ = arg min λ>0 ε. (11) 2.3 Results By training the logistic regression with our leave-one-out cross validation s- trategy, λ is optimally set to 1.9 and the overall C-V error ε = 0.27 (training error εS = 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. 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 (⋆).

Team # 14531 Page 9 of 19 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) Figure 3 illustrates the probability of criminal involvement estimated by hθ(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. 10 20 30 40 50 60 70 80 0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1 Rank in the priority list Probability of being criminal All the members Gretchen (manager) Jerome (manager) Dolores (manager) Figure 3: Probability of conspiracy versus the corresponding rank in priority list 2As 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

Team # 14531 Page 10 of 19 2.4 Semantic model enhancement When talking about enhancement of the accuracy, we should reduce the diﬀer- ence the performance with those who could deal with it more ﬂexibly and exactly: human being. A very important way we humans solve cases is analysis the com- munication contents including messages and records. However, just as aforemen- tioned, our ability so quite limited when handling huge amount of information that we have to utilize machine to help us. Therefore, semantical 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 topis (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. Conspiratorial Message 1 Conspiratorial Message 2 Conspiratorial Message x suspicious word 3 suspicious word n suspicious word 1 suspicious word 2 Conspiratorial Dictionary New conspiratorial Message New non-conspiratorial Message Undetermined Message 1 Undetermined Message 2 Undetermined Message 3 Undetermined Message 4 Undetermined Message m Legend: Figure 4: Framework of topic semantic diﬀusion 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 case during the detection period. Detectors' reasoning plays a indispensable role through out the entire process. Therefore, we draw the problem of topic semantic

Team # 14531 Page 11 of 19 diﬀusion into a topic similarity measurement task based on expert dictionary. As seen in Figure 4, a conspiratorial dictionary is ﬁrstly constructed from the conspir- atorial messages about known suspicious topics. Resource allocation mechanism on bipartite network outperforms in extracting the hidden information of networks [Zhou, 2007], which is exploited by us to unfold the similarity among diﬀerent top- ics 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 = d1, d2...dn and T = t1, t2...tm. 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(tl) = n ∑ i=1 ailf(di) D(di) . (12) Equation 12 expresses the calculation of the resource held by t(l) after the ﬁrst step : f(tl). D(di) indicates the degree of the node di and ail is deﬁnrd as follows: ail = {1, ditl ∈E 0, otherwise. (13) Intuitive explanation of step 1 is to arrange the resource averagely by degree of ti from T to D if ti 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 ′(ti) = m ∑ l=1 ailf(dl) D(dl) = m ∑ l=1 ail D(dl) n ∑ j=1 ajif(tj) D(tj) . (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 2.4 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. 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 2.4, 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

Team # 14531 Page 12 of 19 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) 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 con- struct a bipartite graph with the topics, we will also get similarity among topics based on speaker who transmit them. However, the determination of the relation- ship 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 com- mitting conspiracy, however, the likelihood does not proportionally indicate the leadership inside the network, for the identiﬁcation of leaders is further complicat- ed by its topology. Thus we adopt LeaderRank, a node ranking algorithm closely related to net- work 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 robust- ness against random noise, LeaderRank is also appropriate for addressing criminal network problems, which usually suﬀer from incompleteness and incorrectness.

Team # 14531 Page 13 of 19 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 tradi- tional algorithms including HITS and PageRank [L¨u et al., 2011]. This method is mathematically equivalent to random walk mechanism on the directed network with adaptive probability, leading to a parameter-free algorithm readily applica- ble 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 homeostasis 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 ν(t) according to the voting from its neighbors: νi(t + 1) = |V |+1 ∑ j=1 µij Dout(j)νi(t) (15) Where µij is a binary indicator with value 1 if node i points to j and 0 otherwise. Dout(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 νi(Tc)+νgn(Tc)/|V |, where νgn(Tc) 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 GTS connected by suspicious topics only, so as to minimize the coupling of the company's hierarchical structure to the conspiracy relations. Supposing Tij denotes the set of topics mentioned by messages from node i to node j, and TS denotes the set of known suspicious topics (TS = {7, 11, 13}). Then GTS is the maximum subgraph of the original graph G, whereas Tij ⊆TS, for all (i, j) ⊆ETS (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 GTS has to be reversed to be compatible with LeaderRank's original design. The reversed sub-network induced by suspicious topics is denoted by G′ TS.

Team # 14531 Page 14 of 19 3.4 Results By running LeaderRank on G′ TS , 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 organiza- tion. Name LeaderRank score Yao 2.67 Alex 2.21 Paul 1.92 Elsie 1.62 Table 4: Partial results of LeaderRank on G′ TS 3.5 Empirical support Empirical analysis of criminal networks has found that a leader of a criminal or- ganization tends to carefully balancing his or her degree centrality and betweenness centrality. It has been proposed that the leader usually maintains a high between- ness centrality but a relatively low degree centrality, for enhancing eﬃciency and meanwhile ensuring safety [Morselli, 2010]. 0.00 0.01 0.02 0.03 0.04 0.05 Betweenness centrality 8 10 12 14 16 18 20 Degree centrality Known conspirators High conspiracy prob. Yao (inferred leader) Figure 5: The joint distribution of betweenness centrality and degree centrality

Team # 14531 Page 15 of 19 And our inference that Yao is the leader is thus empirically supported. Figure 5 illustrates the joint distribution of betweenness centrality CB and degree centrality (Din+Dout) 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 reallo- cation 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 [Xu, 2005], the method for inferring criminality or con- spiracy 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 con- nected 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 [Sen, 1968]. Given two priority lists {pk} = {p1, p2, · · · , pn} and {qk} = {q1, q2, · · · , qn} (for example, p2 = 5 means node 2 is ranked 5th by {pk} list), then (i, j), i ̸= j is said to be a concordant pair if their rankings agree in two lists, i.e. pi > pj, qi > qj or pi < pj, qi < qj; (i, j) is said to be a discordant pair if their rankings disagree in two lists.

Team # 14531 Page 16 of 19 Name Dolores Crystal Jerome Sherri Neal Rank (previous) 1 2 3 4 5 Prob. of conspiracy (previous) 0.555 0.508 0.388 0.316 0.299 Rank (new) 2 1 6 9 3 Prob. of conspiracy (new) 0.621 0.629 0.405 0.393 0.504 Name Christina Jerome William Dwight Beth Rank (previous) 6 7 8 9 10 Prob. of conspiracy (previous) 0.555 0.508 0.388 0.316 0.299 Rank (new) 14 5 4 12 16 Prob. of conspiracy (new) 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 τ = (number of concordant pairs) −(number of discordant pairs) 1 2n(n −1) . (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 τ = 0.86, justifying the robustness of the machine learning ap- proach. 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 prob- ability, as indicated by the distribution of dots skewed from the diagonal line. The augmented probability is compatible with the new information that expands both the set of suspicious topics and known conspirators.

Team # 14531 Page 17 of 19 0 0.05 0.1 0.15 0.2 0.25 0.3 0.35 0.4 0.45 0.5 0.8 0.82 0.84 0.86 0.88 0.9 0.92 0.94 0.96 0.98 1 Probability of wrong classification Kendall's tau Conspirators Non−conspirators Figure 6: The expected Kendall's tau declines as misclassiﬁcation probability in- creases 0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1 0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1 Probability of being a conspirator (Requirement 1) Probability of being a conspirator (Requirement 2 ) Training set Unknown nodes Chris Gretchen (manager) Jerome (manager) Dolores (manager) Figure 7: Criminal probabilities before and after the change of conditions 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.

Team # 14531 Page 18 of 19 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 cen- trality in topology, conspiratorial level of connectivity and communication contents. Experimental results hits about 73% correct prediction rate in the basis 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 en- hancement lead us to some deeper discursion 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 from aspects on the topology property of a criminal leader. Finally, the high value of Kendall's tau illustrates the nonsensitive property of the model. The weakness of our model is mainly about features. Diﬀerent networks may not share the same features. Thereby the features of particular network may become meaningless when type of network changes. Furthermore, large amounts of information might seriously limit the capability of the resource allocation strategy because the construction of bipartite graph itself becomes a problem when facing with too much noise information. 5.2 Further discussion When taking about the good portability of a model on diﬀerent models, we could focus on two aspects: the similarity of two networks and diﬀerence. Some common pattern appeared in diﬀerent kinds of networks,including biological network, are the small-world property, power-law degree distributions, network transitivity and community structure [Girvan, 2001]. Either topology or transmission properties shared by diﬀerent networks could help to build a model with good portability. However, on the other hand, the fundamental diﬀerence in mechanism of in- formation transmission decides the distinguishing models: people ﬁnd out clues through evidence or relative materials to break down the invisibility of a criminal network based on reasoning and deduction. Even though share some similar prop- erties, individuals in other networks like some biological networks communicate with each other under some relatively ﬁxed principles or unchangeable pattern for division of work. The co-occurrence-based approaches always fail to characterize biological interactions [Chen & Sharp, 2004], where particular dynamics function of analytical model might perform better like utilizing revised infectious disease model to predict disfunction or diseased biological components.

Team # 14531 Page 19 of 19 References [Baker & Faulkner, 1993] Baker, W. E. & Faulkner, R. R. (1993). The Social Organization of Conspiracy: Illegal Networks in the Heavy Electrical Equipment Industry. [Chen & Sharp, 2004] Chen, H. & Sharp, B. M. (2004). Content-rich biological network constructed by mining PubMed abstracts.. BMC bioinformatics 5, 147. [Freeman, 1979] Freeman, L. (1979). Centrality in social networks conceptual clar- iﬁcation. Social networks. [Girvan, 2001] Girvan, M. (2001). Community structure in social and biological networks. PNAS. [Krebs, 2002] Krebs, V. E. (2002). Mapping Networks of Terrorist Cells. 24(3), 43–52. [L¨u et al., 2011] L¨u, L., Zhang, Y.-C., Yeung, C. H., & Zhou, T. (2011). Leaders in social networks, the Delicious case.. PloS one 6(6), e21202. [Morselli, 2010] Morselli, C. (2010). Assessing Vulnerable and Strategic Positions in a Criminal Network. Journal of Contemporary Criminal Justice 26(4), 382– 392. [Sabidussi, 1966] Sabidussi, G. (1966). The centrality index of a graph. Psychome- trika. [Sen, 1968] Sen, K. P. (1968). Estimates of the regression coeﬃcient based on Kendall's tau. Journal of the American Statistical Association. [Wheat, 2007] Wheat, C. (2007). Algorithmic Complexity and Structural Models of Social Networks. Management , 1–38. [Xu, 2005] Xu, J. (2005). Criminal network analysis and visualization. Communi- cations of the ACM 48(6). [Xu & Chen, 2003] Xu, J. & Chen, H. (2003). Untangling Criminal Networks : A Case Study. World Trade , 232–248. [Zhou, 2007] Zhou, T. (2007). Bipartite network projection and personal recom- mendation. Physical Review E 76(4), 1–7.

|  |  |
| --- | --- |
| 0 | 1 |
|  | Team Control Number |
| For office use only | For office use only |
|  |  |
| T1 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | F1 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  14531 |
| T2 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | F2 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| T3 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | F3 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
|  | Problem Chosen |
| T4 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | F4 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
|  | C |
|  | 2012 Mathematical Contest in Modeling (MCM) Summary Sheet |
|  | (Attach a copy of this page to each copy of your solution paper.) |
|  | Type a summary of your results on this page. Do not include |
|  | the name of your school, advisor, or team members on this page. |
|  | Finding Conspirators in the Network: |
|  | Machine Learning with Resource-allocation Dynamics |
|  | Problem Clarification: A conspiracy network is embedded into a network of |
| KEY CONCEPT |  |
|  | 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, |
| M | non-criminals and suspicious topics, we are fundamentally asked to estimate the |
| achine learning |  |
|  | probability of criminals involvement for the not identified individuals , and to clarify |
| L | the leader of conspirators. Besides, relevant discussions are suggested. |
| ogistic regression | 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 |
| S | machine learning.(ii) A conspirator is reluctant to mention topics related to crime |
| emantic diffusion |  |
|  | when talking with an outsider.(iii) Conspirators tend not to talk about irrelevant |
| B | topics frequently with each other. (iv) The leader of conspiracy tries to minimize risk |
| ipartite graph | by restricting direct contacts. (v) A non-conspirator has no  idea of who are |
|  | conspirators, thus treating conspirator and non-conspirators equally. |
| R | Model Design and Justification：The probability of conspiracy for an unidentified |
| esource-allocation |  |
|  | node is modeled as a sigmoid function in terms of a linear combination of the node’s |
| dynamics | features (logistic regression), whereas features are formulated from local topological |
|  | measures and the node’s semantic messaging patterns. Parameters of this model |
| K | are trained by using a subset of identified conspirators and non-conspirators. The |
| endall’s tau coefficient |  |
|  | performance of the model is enhanced by discovering potential relationship of |
|  | similarities among topics via topic-word bipartite dynamics. Resource-allocation |
| K | dynamics are performed to identify the leader of the conspirators, which win |
| EY TECHNIQUES |  |
|  | theoretical evidence in criminal network research. |
|  | Results and Sensitivity Analysis：(i)The accuracy of the machine learning |
| G | scheme is measured by its performance on leave-one-out cross validation. Basic |
| radient descent |  |
|  | solution gets 73% prediction accuracy and semantic enhanced solution win 87% |
| R | correct rate. (ii)The insensitivity of priority conspirator list is manifested by analyzing |
| evised Leader Rank | Kendall’s  tau. This argument  is 0.86  illustrates high stability of  the model |
|  | performance.(iii)The leader we predicted tends to be Yao and the top three in |
| B | priority list are Dolores, Crystal and Jerome (known conspirators excluded). |
| ipartite graph |  |
|  | Strengths and Weaknesses Discussion： |
| transmission | The combination of both the topology properties and semantic affinity among |
|  | individuals leads to a good performance. The time complexity is linear in the whole |
|  | process in mining of semantic potential information, which is suitable with large |
|  | amounts of data. |
|  | However, when facing with large amounts of data, our model prefer obtaining |
|  | assistance from semantic network analysis to form the expert dictionary. Such |
|  | features might also meaningless when change a new network background. |

|  |  |  |  |
| --- | --- | --- | --- |
| 0 | 1 | 2 | 3 |
| Finding Conspirators in the Network: Machine Learning with Resource-allocation Dynamics |  |  |  |
| KEY CONCEPT |  |  | 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 are fundamentally asked to estimate the probability of criminals involvement for the not identified individuals , and to clarify the leader of conspirators. Besides, relevant discussions are suggested. 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 bipartite dynamics. Resource-allocation dynamics are performed to identify the leader of the conspirators, which win theoretical evidence in criminal network research. Results and Sensitivity Analysis：(i)The accuracy of the machine learning scheme is measured by its performance on leave-one-out cross validation. Basic solution gets 73% prediction accuracy and semantic enhanced solution win 87% correct rate. (ii)The insensitivity of priority conspirator list is manifested by analyzing Kendall’s tau. This argument is 0.86 illustrates high stability of the model performance.(iii)The leader we predicted tends to be Yao and the top three in priority list are Dolores, Crystal and Jerome (known conspirators excluded). Strengths and Weaknesses Discussion： The combination of both the topology properties and semantic affinity among individuals leads to a good performance. The time complexity is linear in the whole process in mining of semantic potential information, which is suitable with large amounts of data. However, when facing with large amounts of data, our model prefer obtaining assistance from semantic network analysis to form the expert dictionary. Such features might also meaningless when change a new network background. |
|  |  |  |  |
|  | Machine learning |  |  |
|  |  |  |  |
|  | Logistic regression |  |  |
|  |  |  |  |
|  | Semantic diffusion |  |  |
|  |  |  |  |
|  | Bipartite graph |  |  |
|  |  |  |  |
|  | Resource-allocation |  |  |
|  | dynamics |  |  |
|  |  |  |  |
|  | Kendall’s tau coefficient |  |  |
|  |  |  |  |
| KEY TECHNIQUES |  |  |  |
|  |  |  |  |
|  | Gradient descent |  |  |
|  |  |  |  |
|  | Revised Leader Rank |  |  |
|  |  |  |  |
|  | Bipartite graph |  |  |
|  | transmission |  |  |
|  |  |  |  |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 |
| (cid:84)(cid:101)(cid:97)(cid:109) (cid:35) (cid:49)(cid:52)(cid:53)(cid:51)(cid:49) |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | (cid:80)(cid:97)(cid:103)(cid:101) (cid:49) (cid:111)(cid:102) (cid:49)(cid:57) |  |  |  |
|  |  |  |  |  |  | (cid:67)(cid:111)(cid:110)(cid:116)(cid:101)(cid:110)(cid:116)(cid:115) |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| (cid:49) | (cid:73)(cid:110)(cid:116)(cid:114)(cid:111)(cid:100)(cid:117)(cid:99)(cid:116)(cid:105)(cid:111)(cid:110) | (cid:46) (cid:46) (cid:46) (cid:46) (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:46) |  | (cid:46) |  | (cid:46) |  | (cid:46) |  | (cid:46) |  | (cid:46) |  | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:50) |
| (cid:50) |  | (cid:65) (cid:77)(cid:97)(cid:99)(cid:104)(cid:105)(cid:110)(cid:101) (cid:76)(cid:101)(cid:97)(cid:114)(cid:110)(cid:105)(cid:110)(cid:103) (cid:83)(cid:111)(cid:108)(cid:117)(cid:116)(cid:105)(cid:111)(cid:110) (cid:116)(cid:111) (cid:67)(cid:114)(cid:105)(cid:109)(cid:105)(cid:110)(cid:97)(cid:108) (cid:80)(cid:114)(cid:105)(cid:111)(cid:114)(cid:105)(cid:116)(cid:121) |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:52) |
|  | 2.1 | Feature formulation . |  | . | . | . | . | . | . | . | . | . | . | . |  | . |  | . |  | . |  | . |  | . | . | . | . | . | . | . | . | . | . | . | 4 |
|  | 2.2 | Methods . . . . . . | . | . | . | . | . | . | . | . | . | . | . | . |  | . |  | . |  | . |  | . |  | . | . | . | . | . | . | . | . | . | . | . | 6 |
|  |  | 2.2.1 | Logistic regression |  |  |  | . | . | . | . | . | . | . | . |  | . |  | . |  | . |  | . |  | . | . | . | . | . | . | . | . | . | . | . | 7 |
|  |  | 2.2.2 | Gradient descent |  |  | . | . | . | . | . | . | . | . | . |  | . |  | . |  | . |  | . |  | . | . | . | . | . | . | . | . | . | . | . | 7 |
|  |  | 2.2.3 | Leave-one-out cross validation |  |  |  |  |  |  |  |  |  | . | . |  | . |  | . |  | . |  | . |  | . | . | . | . | . | . | . | . | . | . | . | 8 |
|  |  | 2.2.4 | Selecting regularization parameter . |  |  |  |  |  |  |  |  |  |  |  |  |  |  | . |  | . |  | . |  | . | . | . | . | . | . | . | . | . | . | . | 8 |
|  | 2.3 | Results . . . . . . . | . | . | . | . | . | . | . | . | . | . | . | . |  | . |  | . |  | . |  | . |  | . | . | . | . | . | . | . | . | . | . | . | 8 |
|  | 2.4 |  | Semantic model enhancement |  |  |  |  |  | . | . | . | . | . | . |  | . |  | . |  | . |  | . |  | . | . | . | . | . | . | . | . | . | . | . | 10 |
| (cid:51) |  | (cid:73)(cid:100)(cid:101)(cid:110)(cid:116)(cid:105)(cid:102)(cid:121)(cid:105)(cid:110)(cid:103) (cid:116)(cid:104)(cid:101) (cid:108)(cid:101)(cid:97)(cid:100)(cid:101)(cid:114) (cid:111)(cid:102) (cid:116)(cid:104)(cid:101) (cid:99)(cid:111)(cid:110)(cid:115)(cid:112)(cid:105)(cid:114)(cid:97)(cid:99)(cid:121) (cid:46) |  |  |  |  |  |  |  |  |  |  |  |  |  |  | (cid:46) |  | (cid:46) |  | (cid:46) |  | (cid:46) |  | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:49)(cid:50) |
|  | 3.1 | LeaderRank . . . . | . | . | . | . | . | . | . | . | . | . | . | . |  | . |  | . |  | . |  | . |  | . | . | . | . | . | . | . | . | . | . | . | 13 |
|  | 3.2 |  | Suspicious topic sub-network extraction . |  |  |  |  |  |  |  |  |  |  |  |  | . |  | . |  | . |  | . |  | . | . | . | . | . | . | . | . | . | . | . | 13 |
|  | 3.3 | Edge reverse . . . . | . | . | . | . | . | . | . | . | . | . | . | . |  | . |  | . |  | . |  | . |  | . | . | . | . | . | . | . | . | . | . | . | 13 |
|  | 3.4 | Results . . . . . . . | . | . | . | . | . | . | . | . | . | . | . | . |  | . |  | . |  | . |  | . |  | . | . | . | . | . | . | . | . | . | . | . | 14 |
|  | 3.5 | Empirical support . | . | . | . | . | . | . | . | . | . | . | . | . |  | . |  | . |  | . |  | . |  | . | . | . | . | . | . | . | . | . | . | . | 14 |
|  | 3.6 | Discussion . . . . . | . | . | . | . | . | . | . | . | . | . | . | . |  | . |  | . |  | . |  | . |  | . | . | . | . | . | . | . | . | . | . | . | 15 |
| (cid:52) |  | (cid:69)(cid:118)(cid:97)(cid:108)(cid:117)(cid:97)(cid:116)(cid:105)(cid:110)(cid:103) (cid:116)(cid:104)(cid:101) (cid:77)(cid:111)(cid:100)(cid:101)(cid:108) (cid:46) |  | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:46) |  | (cid:46) |  | (cid:46) |  | (cid:46) |  | (cid:46) |  | (cid:46) |  | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:49)(cid:53) |
|  | 4.1 | Sensitivity analysis | . | . | . | . | . | . | . | . | . | . | . | . |  | . |  | . |  | . |  | . |  | . | . | . | . | . | . | . | . | . | . | . | 15 |
|  |  | 4.1.1 Priority list | . | . | . | . | . | . | . | . | . | . | . | . |  | . |  | . |  | . |  | . |  | . | . | . | . | . | . | . | . | . | . | . | 15 |
|  |  | 4.1.2 | Probability inﬂation |  |  |  |  | . | . | . | . | . | . | . |  | . |  | . |  | . |  | . |  | . | . | . | . | . | . | . | . | . | . | . | 16 |
| (cid:53) |  | (cid:68)(cid:105)(cid:115)(cid:99)(cid:117)(cid:115)(cid:115)(cid:105)(cid:111)(cid:110) (cid:97)(cid:110)(cid:100) (cid:67)(cid:111)(cid:110)(cid:99)(cid:108)(cid:117)(cid:115)(cid:105)(cid:111)(cid:110) |  |  |  |  | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:46) |  | (cid:46) |  | (cid:46) |  | (cid:46) |  | (cid:46) |  | (cid:46) |  | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:46) | (cid:49)(cid:56) |
|  | 5.1 | Summary . . . . . | . | . | . | . | . | . | . | . | . | . | . | . |  | . |  | . |  | . |  | . |  | . | . | . | . | . | . | . | . | . | . | . | 18 |
|  | 5.2 | Further discussion . | . | . | . | . | . | . | . | . | . | . | . | . |  | . |  | . |  | . |  | . |  | . | . | . | . | . | . | . | . | . | . | . | 18 |

|  |
| --- |
| 0 |
| (cid:84)(cid:101)(cid:97)(cid:109) (cid:35) (cid:49)(cid:52)(cid:53)(cid:51)(cid:49) (cid:80)(cid:97)(cid:103)(cid:101) (cid:50) (cid:111)(cid:102) (cid:49)(cid:57) |
| (cid:49) (cid:73)(cid:110)(cid:116)(cid:114)(cid:111)(cid:100)(cid:117)(cid:99)(cid:116)(cid:105)(cid:111)(cid:110) |
| As illustrated in Figure 1, criminals and conspirators tend to form organization- |
| al 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 [Baker & Faulkner, 1993]. |
| 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 d- |
| iﬀerent 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 [Krebs, 2002]. |
| 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 |

|  |
| --- |
| 0 |
| (cid:84)(cid:101)(cid:97)(cid:109) (cid:35) (cid:49)(cid:52)(cid:53)(cid:51)(cid:49) (cid:80)(cid:97)(cid:103)(cid:101) (cid:51) (cid:111)(cid:102) (cid:49)(cid:57) |
| 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 [Krebs, 2002]. 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 |
| [Morselli, 2010]. |
| 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 intro- |
| duced, which try to discover the hidden topological patterns by partitioning the |
| big network into small closely connected cliques [Xu, 2005]. Despite the light they |
| shed upon the internal structures of criminal networks, these methods still suﬀer |
| from intimidating complexity with large databases [Wheat, 2007]. |
| In this paper, we carefully combine the local-feature-based methods with ap- |
| proaches 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 compu- |
| tationally 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 sen- |
| sitive 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. |

|  |
| --- |
| 0 |
| (cid:84)(cid:101)(cid:97)(cid:109) (cid:35) (cid:49)(cid:52)(cid:53)(cid:51)(cid:49) (cid:80)(cid:97)(cid:103)(cid:101) (cid:52) (cid:111)(cid:102) (cid:49)(cid:57) |
| (cid:50) (cid:65) (cid:77)(cid:97)(cid:99)(cid:104)(cid:105)(cid:110)(cid:101) (cid:76)(cid:101)(cid:97)(cid:114)(cid:110)(cid:105)(cid:110)(cid:103) (cid:83)(cid:111)(cid:108)(cid:117)(cid:116)(cid:105)(cid:111)(cid:110) (cid:116)(cid:111) (cid:67)(cid:114)(cid:105)(cid:109)(cid:105)(cid:110)(cid:97)(cid:108) (cid:80)(cid:114)(cid:105)(cid:45) |
| (cid:111)(cid:114)(cid:105)(cid:116)(cid:121) |
| Machine learning is carefully selected by us to play the key role in the en- |
| tire 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, reason- |
| ing 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. |
| (cid:50)(cid:46)(cid:49) (cid:70)(cid:101)(cid:97)(cid:116)(cid:117)(cid:114)(cid:101) (cid:102)(cid:111)(cid:114)(cid:109)(cid:117)(cid:108)(cid:97)(cid:116)(cid:105)(cid:111)(cid:110) |
| • (cid:67)(cid:101)(cid:110)(cid:116)(cid:114)(cid:97)(cid:108)(cid:105)(cid:116)(cid:121) |
| We exploit three types of centrality including degree, betweenness and close- |
| ness centrality to determine the center of the suspicious network from diﬀerent |
| aspects: |
| (cid:73) (cid:68)(cid:101)(cid:103)(cid:114)(cid:101)(cid:101) (cid:99)(cid:101)(cid:110)(cid:116)(cid:114)(cid:97)(cid:108)(cid:105)(cid:116)(cid:121) [Freeman, 1979] indicates activeness of a member, i.e. |
| the member who tends to have more links to its surroundings. As explained |
| in [Xu & Chen, 2003], degree centrality is not quite reliable to indicate the |
| team leader in a criminal network. For a graph G(V, E), the normalized |

|  |
| --- |
| 0 |
| (cid:84)(cid:101)(cid:97)(cid:109) (cid:35) (cid:49)(cid:52)(cid:53)(cid:51)(cid:49) (cid:80)(cid:97)(cid:103)(cid:101) (cid:53) (cid:111)(cid:102) (cid:49)(cid:57) |
| degree centrality of node i is as follows: |
| ∑|V | |
| j(cid:61)(cid:49) ν(i, j) |
| , i ̸= j (1) CD(i) = |
| |V | − 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. |
| (cid:73) (cid:66)(cid:101)(cid:116)(cid:119)(cid:101)(cid:101)(cid:110)(cid:110)(cid:101)(cid:115)(cid:115) (cid:99)(cid:101)(cid:110)(cid:116)(cid:114)(cid:97)(cid:108)(cid:105)(cid:116)(cid:121) [Freeman, 1979] 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 oth- |
| er two members transmit information between themselves. The normalized |
| betweenness centrality is deﬁned as: |
| ∑|V | ∑|V | |
| k<j ωj,k(i) j(cid:61)(cid:49) |
| , k ̸= i (2) CB(i) = |
| |V | − 1 |
| where ωj,k(i) indicates whether the shortest path between node j and node |
| k passes through node i. |
| (cid:73) (cid:67)(cid:108)(cid:111)(cid:115)(cid:101)(cid:110)(cid:101)(cid:115)(cid:115) (cid:99)(cid:101)(cid:110)(cid:116)(cid:114)(cid:97)(cid:108)(cid:105)(cid:116)(cid:121) [Sabidussi, 1966] 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: |
| ∑|V | |
| j(cid:61)(cid:49) ρ(i, j) − Ccmin |
| , i ̸= j (3) Cc(i) = |
| Ccmax − Ccmin |
| where ρ(i, j) is the length of the shortest path connecting nodes i and j. |
| Ccmin and Ccmax are the minimum and maximum lengths of the shortest |
| paths respectively. |
| • (cid:78)(cid:117)(cid:109)(cid:98)(cid:101)(cid:114) (cid:111)(cid:102) (cid:107)(cid:110)(cid:111)(cid:119)(cid:110) (cid:110)(cid:101)(cid:105)(cid:103)(cid:104)(cid:98)(cid:111)(cid:114)(cid:105)(cid:110)(cid:103) (cid:99)(cid:111)(cid:110)(cid:115)(cid:112)(cid:105)(cid:114)(cid:97)(cid:116)(cid:111)(cid:114)(cid:115) |
| 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 |
| its confederacy among all its in Figure 2, we calculate the possession rate of |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|  |  |  |  | (cid:48) |  |  |  |  |
|  | (cid:100) |  |  |  |  |  |  |  |
|  |  |  | (cid:53) |  | (cid:101) |  |  |  |
|  |  |  |  |  |  |  |  | (cid:110) |
|  |  |  |  | (cid:53) |  |  |  |  |
|  |  |  |  |  |  | (cid:121) |  |  |
|  | (cid:110) |  |  |  |  |  |  |  |
|  |  | (cid:111) |  |  |  |  |  |  |
|  |  |  |  | (cid:53) |  |  |  |  |
|  |  |  |  |  |  | (cid:120) |  |  |
|  |  | (cid:102) |  |  |  |  |  |  |
| (cid:48) |  |  |  | (cid:49) |  | (cid:53) | (cid:53) | (cid:53) |
|  |  |  | (cid:51) |  |  |  |  |  |
|  |  |  |  |  |  |  |  | (cid:101) |
|  |  |  |  |  |  | (cid:110) |  |  |
|  | (cid:115) | (cid:108) |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  | (cid:97) |
|  |  |  |  | (cid:101) |  |  |  |  |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|  |  |  |  | (cid:48) |  |  |  |  |  |
|  | (cid:100) |  |  |  |  |  |  |  |  |
|  |  |  | (cid:53) |  | (cid:101) |  |  |  |  |
|  |  |  |  |  |  |  |  |  | (cid:110) |
|  |  |  |  | (cid:53) |  |  |  |  |  |
|  |  |  |  |  |  | (cid:121) |  |  |  |
|  | (cid:110) |  |  |  |  |  |  |  |  |
|  |  | (cid:111) |  |  |  |  |  |  |  |
|  |  |  |  | (cid:53) |  |  |  |  |  |
|  |  |  |  |  |  | (cid:120) |  |  |  |
|  |  | (cid:102) |  |  |  |  |  |  |  |
| (cid:48) |  |  |  | (cid:49) |  | (cid:53) | (cid:53) | (cid:53) | (cid:48) |
|  |  |  | (cid:51) |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  | (cid:101) |  |
|  |  |  |  |  |  | (cid:110) |  |  |  |
|  | (cid:115) | (cid:108) |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  | (cid:97) |  |
|  |  |  |  | (cid:101) |  |  |  |  |  |
|  | (cid:101) |  |  |  |  |  |  |  |  |
| (cid:115) |  |  |  |  |  |  |  |  |  |
|  |  |  | (cid:48) |  |  |  |  |  |  |
| (cid:115) |  |  |  |  |  |  |  |  |  |

|  |  |
| --- | --- |
| 0 | 1 |
| neighbors, which illustrates his or her compactness 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. |
| (cid:48) |  |
| (cid:100) |  |
| (cid:53) (cid:101) |  |
| (cid:110) |  |
| (cid:53) |  |
| (cid:121) |  |
| (cid:110) |  |
| (cid:111) |  |
| (cid:53) |  |
| (cid:120) |  |
| (cid:102) |  |
| (cid:49) (cid:48) (cid:53) (cid:53) (cid:53) (cid:48) |  |
| (cid:51) |  |
| (cid:101) |  |
| (cid:110) |  |
| (cid:115) (cid:108) |  |
| (cid:97) |  |
| (cid:101) |  |
| (cid:101) |  |
| (cid:115) |  |
| (cid:48) |  |
| (cid:115) |  |

|  |
| --- |
| 0 |
| (cid:84)(cid:101)(cid:97)(cid:109) (cid:35) (cid:49)(cid:52)(cid:53)(cid:51)(cid:49) (cid:80)(cid:97)(cid:103)(cid:101) (cid:54) (cid:111)(cid:102) (cid:49)(cid:57) |
| neighbors, which illustrates his or her compactness 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. |
| (cid:48) |
| (cid:100) |
| (cid:53) (cid:101) |
| (cid:110) |
| (cid:53) |
| (cid:121) |
| (cid:110) |
| (cid:111) |
| (cid:53) |
| (cid:120) |
| (cid:102) |
| (cid:49) (cid:48) (cid:53) (cid:53) (cid:53) (cid:48) |
| (cid:51) |
| (cid:101) |
| (cid:110) |
| (cid:115) (cid:108) |
| (cid:97) |
| (cid:101) |
| (cid:101) |
| (cid:115) |
| (cid:48) |
| (cid:115) |
| Figure 2: Possession rate of neighboring accomplices distribution |
| • (cid:78)(cid:117)(cid:109)(cid:98)(cid:101)(cid:114) (cid:111)(cid:102) (cid:99)(cid:117)(cid:114)(cid:114)(cid:101)(cid:110)(cid:116) (cid:110)(cid:111)(cid:110)(cid:45)(cid:115)(cid:117)(cid:115)(cid:112)(cid:105)(cid:99)(cid:105)(cid:111)(cid:117)(cid:115) (cid:109)(cid:101)(cid:115)(cid:115)(cid:97)(cid:103)(cid:101)(cid:115) (cid:102)(cid:114)(cid:111)(cid:109) (cid:116)(cid:104)(cid:101) (cid:107)(cid:110)(cid:111)(cid:119)(cid:110) (cid:99)(cid:111)(cid:110)(cid:45) |
| (cid:115)(cid:112)(cid:105)(cid:114)(cid:97)(cid:116)(cid:111)(cid:114)(cid:115) |
| Table 2.1 is the topics mentioned between known conspirators.(cid:49) It is obvi- |
| ous that a known conspirator rarely talks about irrelevant topics, i.e. topics |
| irrelevant to their conspiracy, with his or her accomplices even though some |
| unknown topics appear among them, which accounts for a very small propor- |
| tion. 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. |
| (cid:50)(cid:46)(cid:50) (cid:77)(cid:101)(cid:116)(cid:104)(cid:111)(cid:100)(cid:115) |
| 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 ob- |
| tained by solving an optimization problem related to training set by gradient ascent |
| algorithm. |

|  |  |
| --- | --- |
| 0 | 1 |
| (cid:84)(cid:101)(cid:97)(cid:109) (cid:35) (cid:49)(cid:52)(cid:53)(cid:51)(cid:49) | (cid:80)(cid:97)(cid:103)(cid:101) (cid:55) (cid:111)(cid:102) (cid:49)(cid:57) |
| Jean Alex Elsie Poul | Ulf Yao Harvey |
| Jean (cid:49)(cid:49)⋆ | 8 14 |
| Alex 1 (cid:49)(cid:51)⋆ | (cid:49)(cid:49)⋆ 3,(cid:55)⋆ |
| Elsie (cid:49)(cid:49)⋆ | (cid:49)(cid:51)⋆ |
| Poul (cid:49)(cid:49)⋆ (cid:55)⋆ | (cid:55)⋆ 4 |
| Ulf (cid:55)⋆, (cid:49)(cid:49)⋆, (cid:49)(cid:51)⋆ | (cid:49)(cid:51)⋆ |
| Yao (cid:49)(cid:51)⋆ (cid:55)⋆, (cid:49)(cid:49)⋆, (cid:49)(cid:51)⋆ (cid:55)⋆,9 | (cid:49)(cid:51)⋆ 2, (cid:55)⋆ |
| Harvey | (cid:49)(cid:51)⋆ |
| Table 1: Topics among known conspirators ( known conspiratorial topics are those |  |
| with star and highlighted in blue) |  |
| (cid:50)(cid:46)(cid:50)(cid:46)(cid:49) (cid:76)(cid:111)(cid:103)(cid:105)(cid:115)(cid:116)(cid:105)(cid:99) (cid:114)(cid:101)(cid:103)(cid:114)(cid:101)(cid:115)(cid:115)(cid:105)(cid:111)(cid:110) |  |
| We consider a training set of size m: | {(x(cid:40)(cid:49)(cid:41), y(cid:40)(cid:49)(cid:41)), (x(cid:40)(cid:50)(cid:41), y(cid:40)(cid:50)(cid:41)), · · · , (x(cid:40)m(cid:41), y(cid:40)m(cid:41))}, |
| where x(cid:40)i(cid:41) is an n-dimensional feature vector, and y(cid:40)i(cid:41) | indicates the classiﬁcation |
| of the agent, i.e. | y(cid:40)i(cid:41) = 1 for conspirators and y(cid:40)i(cid:41) = 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 |  |
| 1 |  |
| P (y = 1|x; θ) = hθ(x) = | , (4) |
| 1 + e−θ(cid:84) x |  |
| where θ ∈ (cid:82)n is the parameter vector. |  |
|  | Then, under the framework of generalized linear model, the (cid:109)(cid:97)(cid:120)(cid:105)(cid:109)(cid:117)(cid:109) (cid:97) (cid:112)(cid:111)(cid:115)(cid:116)(cid:101)(cid:45) |
| (cid:114)(cid:105)(cid:111)(cid:114)(cid:105) (MAP) estimate of the parameter θ is given by |  |
| min J(θ), | (5) |
| θ |  |
| where the cost function is given by |  |
| ∑ | ∑ |
|  | λ |
| 1 m J(θ) = [−y(cid:40)i(cid:41) log(hθ(x)(cid:40)i(cid:41)) − (1 − y(cid:40)i(cid:41)) log(1 − hθ(x(cid:40)i(cid:41)))] + | θ(cid:50) (6) |
|  | j , |
|  | 2m |
| i(cid:61)(cid:49) | j(cid:61)(cid:49) |
| with λ being the regularization parameter. |  |
| (cid:50)(cid:46)(cid:50)(cid:46)(cid:50) (cid:71)(cid:114)(cid:97)(cid:100)(cid:105)(cid:101)(cid:110)(cid:116) (cid:100)(cid:101)(cid:115)(cid:99)(cid:101)(cid:110)(cid:116) |  |
|  | The cost function J(θ) 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 |  |
| θ := θ − α∇θJ(θ), | (7) |
| where α is a small positive constant. |  |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|  | Jean | Alex | Elsie | Poul | Ulf | Yao | Harvey |
| Jean |  | (cid:49)(cid:49)⋆ |  |  | 8 |  | 14 |
| Alex |  |  | 1 | (cid:49)(cid:51)⋆ | (cid:49)(cid:49)⋆ | 3,(cid:55)⋆ |  |
| Elsie |  | (cid:49)(cid:49)⋆ |  |  | (cid:49)(cid:51)⋆ |  |  |
| Poul | (cid:49)(cid:49)⋆ |  | (cid:55)⋆ |  | (cid:55)⋆ |  | 4 |
| Ulf |  | (cid:55)⋆, (cid:49)(cid:49)⋆, (cid:49)(cid:51)⋆ |  |  |  | (cid:49)(cid:51)⋆ |  |
| Yao | (cid:49)(cid:51)⋆ | (cid:55)⋆, (cid:49)(cid:49)⋆, (cid:49)(cid:51)⋆ | (cid:55)⋆,9 |  | (cid:49)(cid:51)⋆ |  | 2, (cid:55)⋆ |
| Harvey |  |  |  |  |  | (cid:49)(cid:51)⋆ |  |

|  |
| --- |
| 0 |
| (cid:84)(cid:101)(cid:97)(cid:109) (cid:35) (cid:49)(cid:52)(cid:53)(cid:51)(cid:49) (cid:80)(cid:97)(cid:103)(cid:101) (cid:56) (cid:111)(cid:102) (cid:49)(cid:57) |
| (cid:50)(cid:46)(cid:50)(cid:46)(cid:51) (cid:76)(cid:101)(cid:97)(cid:118)(cid:101)(cid:45)(cid:111)(cid:110)(cid:101)(cid:45)(cid:111)(cid:117)(cid:116) (cid:99)(cid:114)(cid:111)(cid:115)(cid:115) (cid:118)(cid:97)(cid:108)(cid:105)(cid:100)(cid:97)(cid:116)(cid:105)(cid:111)(cid:110) |
| As we are only informed of the correct classiﬁcation of 15 nodes, at a given |
| round, we only use 14 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 15 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(cid:40)j(cid:41), y(cid:40)j(cid:41)) is left out, and the |
| training set is given by |
| , j − 1, j + 1, · · · , 15}. (8) Sj = {(x(cid:40)l(cid:41), y(cid:40)l(cid:41))|l = 1, 2, · · · |
| Using this training set, parameter vector θ(cid:40)j(cid:41) is obtained, and the corresponding |
| hypothesis is tested on both Sj and the left-out (x(cid:40)j(cid:41), y(cid:40)j(cid:41)), arriving at this round’s |
| respectively. training error εS(cid:106) and C-V error εj |
| Hence, by averaging them over j, the training error is given by |
| (cid:49)(cid:53)∑ |
| 1 1 (9) εS = εS(cid:106) , |
| 5 |
| j(cid:61)(cid:49) |
| and the cross validation error is given by |
| (cid:49)(cid:53)∑ |
| 1 |
| (10) ε = εj. |
| 15 |
| j(cid:61)(cid:49) |
| (cid:50)(cid:46)(cid:50)(cid:46)(cid:52) (cid:83)(cid:101)(cid:108)(cid:101)(cid:99)(cid:116)(cid:105)(cid:110)(cid:103) (cid:114)(cid:101)(cid:103)(cid:117)(cid:108)(cid:97)(cid:114)(cid:105)(cid:122)(cid:97)(cid:116)(cid:105)(cid:111)(cid:110) (cid:112)(cid:97)(cid:114)(cid:97)(cid:109)(cid:101)(cid:116)(cid:101)(cid:114) |
| The regularization parameter λ (λ > 0) is selected optimally as to minimize |
| the cross validation error, i.e. |
| λ = arg min ε. (11) |
| λ>(cid:48) |
| (cid:50)(cid:46)(cid:51) (cid:82)(cid:101)(cid:115)(cid:117)(cid:108)(cid:116)(cid:115) |
| By training the logistic regression with our leave-one-out cross validation s- |
| trategy, λ is optimally set to 1.9 and the overall C-V error ε = 0.27 (training |
| error εS = 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. |
| 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 (⋆). |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 0 | 1 | 2 | 3 | 4 | 5 |
| (cid:84)(cid:101)(cid:97)(cid:109) (cid:35) (cid:49)(cid:52)(cid:53)(cid:51)(cid:49) |  |  |  |  | (cid:80)(cid:97)(cid:103)(cid:101) (cid:57) (cid:111)(cid:102) (cid:49)(cid:57) |
| 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 |

|  |  |
| --- | --- |
| 0 | 1 |
|  | (cid:52) |
| (cid:108) | (cid:51) |

|  |  |  |  |
| --- | --- | --- | --- |
| 0 | 1 | 2 | 3 |
|  |  | (cid:49) |  |
|  | (cid:57) |  |  |
|  | (cid:56) |  | (cid:115) |
|  |  |  | (cid:41) |
|  | (cid:55) |  | (cid:41) |
|  |  |  | (cid:41) |
|  | (cid:54) |  |  |
|  | (cid:53) |  |  |
|  | (cid:52) |  |  |
| (cid:108) | (cid:51) |  |  |
|  | (cid:50) |  |  |
|  | (cid:49) |  |  |

|  |  |
| --- | --- |
| 0 | 1 |
| (cid:48) |  |
|  | (cid:116) |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|  |  |  |  |  |  |  |  |  | (cid:32) |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  | (cid:65)(cid:108)(cid:108)(cid:32)(cid:116)(cid:104)(cid:101)(cid:32)(cid:109)(cid:101)(cid:109)(cid:98)(cid:101)(cid:114)(cid:115) |  |  |  |  |
|  |  |  |  |  | (cid:71)(cid:114)(cid:101)(cid:116)(cid:99)(cid:104)(cid:101)(cid:110)(cid:32)(cid:40)(cid:109)(cid:97)(cid:110)(cid:97)(cid:103)(cid:101)(cid:114)(cid:41) (cid:74)(cid:101)(cid:114)(cid:111)(cid:109)(cid:101)(cid:32)(cid:40)(cid:109)(cid:97)(cid:110)(cid:97)(cid:103)(cid:101)(cid:114)(cid:41) |  |  |  |  |
|  |  |  |  |  | (cid:68)(cid:111)(cid:108)(cid:111)(cid:114)(cid:101)(cid:115)(cid:32)(cid:40)(cid:109)(cid:97)(cid:110)(cid:97)(cid:103)(cid:101)(cid:114)(cid:41) |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|  |  |  |  |  | (cid:115) |  |  |  |  |  |
|  |  | (cid:49) |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  | (cid:50) |  |  |
|  |  |  |  |  | (cid:49) |  |  |  |  |  |
|  |  |  |  | (cid:115) |  |  |  |  |  |  |
|  |  |  |  | (cid:50) |  |  |  |  |  |  |
|  | (cid:50) |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  | (cid:51) |
|  |  |  |  |  |  | (cid:115) |  |  |  |  |
|  |  |  |  |  |  | (cid:51) |  |  |  |  |
|  |  |  |  |  |  |  |  |  | (cid:52) |  |
|  |  |  | (cid:120) |  |  |  | (cid:115) |  |  |  |
|  |  |  |  |  |  |  | (cid:110) |  |  |  |
| (cid:58) |  |  |  |  |  |  |  |  |  |  |

|  |  |
| --- | --- |
| 0 | 1 |
|  | (cid:120) |
| (cid:58) |  |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|  |  |  |  |  |  |  |  |  |  |  | (cid:49) |  |  |  |  |
|  |  |  |  |  |  |  | (cid:115) |  |  |  |  |  |  |  |  |
|  |  | (cid:49) |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  | (cid:50) |  |  |  |
|  |  |  |  |  |  |  | (cid:49) |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  | (cid:115) |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  | (cid:50) |  |  |  |  |  |  |  |  |  |
|  | (cid:50) |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | (cid:51) |
|  |  |  |  |  |  |  |  | (cid:115) |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  | (cid:51) |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  | (cid:52) |  |  |
|  |  |  | (cid:120) |  |  |  |  |  | (cid:115) |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  | (cid:110) |  |  |  |  |  |  |
| (cid:58) |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  | (cid:101) |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  | (cid:108) |  |  |  |  |  |
|  |  |  |  | (cid:101) |  |  |  |  |  |  |  |  |  | (cid:109) |  |
|  |  |  |  |  |  |  |  |  |  | (cid:121) |  |  |  |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|  | (cid:120) |  |  | (cid:115) |  |  |
|  |  |  |  | (cid:110) |  |  |
| (cid:58) |  |  |  |  |  |  |
|  |  |  | (cid:101) |  |  |  |
|  |  |  |  |  | (cid:108) |  |
|  |  | (cid:101) |  |  |  | (cid:109) |
|  |  |  |  |  | (cid:121) |  |

|  |
| --- |
| 0 |
| (cid:84)(cid:101)(cid:97)(cid:109) (cid:35) (cid:49)(cid:52)(cid:53)(cid:51)(cid:49) (cid:80)(cid:97)(cid:103)(cid:101) (cid:49)(cid:48) (cid:111)(cid:102) (cid:49)(cid:57) |
| (cid:50)(cid:46)(cid:52) (cid:83)(cid:101)(cid:109)(cid:97)(cid:110)(cid:116)(cid:105)(cid:99) (cid:109)(cid:111)(cid:100)(cid:101)(cid:108) (cid:101)(cid:110)(cid:104)(cid:97)(cid:110)(cid:99)(cid:101)(cid:109)(cid:101)(cid:110)(cid:116) |
| When talking about enhancement of the accuracy, we should reduce the diﬀer- |
| ence the performance with those who could deal with it more ﬂexibly and exactly: |
| human being. A very important way we humans solve cases is analysis the com- |
| munication contents including messages and records. However, just as aforemen- |
| tioned, our ability so quite limited when handling huge amount of information that |
| we have to utilize machine to help us. Therefore, semantical 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 topis (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. |
| (cid:49) |
| (cid:115) |
| (cid:49) |
| (cid:50) |
| (cid:49) |
| (cid:115) |
| (cid:50) |
| (cid:50) |
| (cid:51) |
| (cid:115) |
| (cid:51) |
| (cid:52) |
| (cid:115) (cid:120) |
| (cid:110) |
| (cid:58) |
| (cid:101) |
| (cid:108) |
| (cid:109) (cid:101) |
| (cid:121) |
| Figure 4: Framework of topic semantic diﬀusion |
| 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 |
| case during the detection period. Detectors’ reasoning plays a indispensable role |
| through out the entire process. Therefore, we draw the problem of topic semantic |

|  |
| --- |
| 0 |
|  |
|  |

|  |
| --- |
| 0 |
| (cid:84)(cid:101)(cid:97)(cid:109) (cid:35) (cid:49)(cid:52)(cid:53)(cid:51)(cid:49) (cid:80)(cid:97)(cid:103)(cid:101) (cid:49)(cid:49) (cid:111)(cid:102) (cid:49)(cid:57) |
| diﬀusion into a topic similarity measurement task based on expert dictionary. As |
| seen in Figure 4, a conspiratorial dictionary is ﬁrstly constructed from the conspir- |
| atorial messages about known suspicious topics. Resource allocation mechanism |
| on bipartite network outperforms in extracting the hidden information of networks |
| [Zhou, 2007], which is exploited by us to unfold the similarity among diﬀerent top- |
| ics 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(cid:49), d(cid:50)...dn and T = t(cid:49), t(cid:50)...tm. 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: |
| ∑ |
| ailf (di) |
| . (12) f (tl) = |
| D(di) |
| i(cid:61)(cid:49) |
| Equation 12 expresses the calculation of the resource held by t(l) after the ﬁrst |
| is deﬁnrd as follows: step : f (tl). D(di) indicates the degree of the node di and ail |
| { |
| 1, ditl ∈ E |
| (13) ail = |
| 0, otherwise. |
| Intuitive explanation of step 1 is to arrange the resource averagely by degree of ti |
| from T to D if ti 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 : |
| ∑ ∑ ∑ |
|  |
| ′ ailf (dl) ail ajif (tj) |
| f = . (14) (ti) = |
| D(dl) D(dl) D(tj) |
| j(cid:61)(cid:49) l(cid:61)(cid:49) l(cid:61)(cid:49) |
| 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 2.4 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. |
| 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 2.4, 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 |

|  |
| --- |
| 0 |
| (cid:84)(cid:101)(cid:97)(cid:109) (cid:35) (cid:49)(cid:52)(cid:53)(cid:51)(cid:49) (cid:80)(cid:97)(cid:103)(cid:101) (cid:49)(cid:50) (cid:111)(cid:102) (cid:49)(cid:57) |
| Rank Topic Number Similarity to known suspicious topics |
| 1 (cid:49)(cid:49)⋆ 0.750 |
| 2 (cid:55)⋆ 0.667 |
| 3 (cid:49)(cid:51)⋆ 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) |
| 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 con- |
| struct a bipartite graph with the topics, we will also get similarity among topics |
| based on speaker who transmit them. However, the determination of the relation- |
| ship 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). |
| (cid:51) (cid:73)(cid:100)(cid:101)(cid:110)(cid:116)(cid:105)(cid:102)(cid:121)(cid:105)(cid:110)(cid:103) (cid:116)(cid:104)(cid:101) (cid:108)(cid:101)(cid:97)(cid:100)(cid:101)(cid:114) (cid:111)(cid:102) (cid:116)(cid:104)(cid:101) (cid:99)(cid:111)(cid:110)(cid:115)(cid:112)(cid:105)(cid:114)(cid:97)(cid:99)(cid:121) |
| Our machine learning scheme tries to estimate the likelihood of a node com- |
| mitting conspiracy, however, the likelihood does not proportionally indicate the |
| leadership inside the network, for the identiﬁcation of leaders is further complicat- |
| ed by its topology. |
| Thus we adopt LeaderRank, a node ranking algorithm closely related to net- |
| work 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 robust- |
| ness against random noise, LeaderRank is also appropriate for addressing criminal |
| network problems, which usually suﬀer from incompleteness and incorrectness. |

|  |  |  |
| --- | --- | --- |
| 0 | 1 | 2 |
| Rank | Topic Number | Similarity to known suspicious topics |
| 1 | (cid:49)(cid:49)⋆ | 0.750 |
| 2 | (cid:55)⋆ | 0.667 |
| 3 | (cid:49)(cid:51)⋆ | 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 |

|  |
| --- |
| 0 |
| (cid:84)(cid:101)(cid:97)(cid:109) (cid:35) (cid:49)(cid:52)(cid:53)(cid:51)(cid:49) (cid:80)(cid:97)(cid:103)(cid:101) (cid:49)(cid:51) (cid:111)(cid:102) (cid:49)(cid:57) |
| (cid:51)(cid:46)(cid:49) (cid:76)(cid:101)(cid:97)(cid:100)(cid:101)(cid:114)(cid:82)(cid:97)(cid:110)(cid:107) |
| LeaderRank algorithm is a state-of-art achievement on node ranking, which is |
| more tolerant of noisy data and robust against manipulations than more tradi- |
| tional algorithms including HITS and PageRank [L¨u (cid:101)(cid:116) (cid:97)(cid:108)(cid:46), 2011]. This method |
| is mathematically equivalent to random walk mechanism on the directed network |
| with adaptive probability, leading to a parameter-free algorithm readily applica- |
| ble 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 homeostasis 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 ν(t) according to the voting from its neighbors: |
| |V |(cid:43)(cid:49)∑ |
| µij |
| (15) νi(t + 1) = νi(t) |
| Dout(j) |
| j(cid:61)(cid:49) |
| is a binary indicator with value 1 if node i points to j and 0 otherwise. Where µij |
| the out-degree of node j. The fraction of above two arguments Dout(j) denotes |
| 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 νi(Tc)+νgn(Tc)/|V |, |
| where νgn(Tc) is the score of the ground node at steady state. |
| (cid:51)(cid:46)(cid:50) (cid:83)(cid:117)(cid:115)(cid:112)(cid:105)(cid:99)(cid:105)(cid:111)(cid:117)(cid:115) (cid:116)(cid:111)(cid:112)(cid:105)(cid:99) (cid:115)(cid:117)(cid:98)(cid:45)(cid:110)(cid:101)(cid:116)(cid:119)(cid:111)(cid:114)(cid:107) (cid:101)(cid:120)(cid:116)(cid:114)(cid:97)(cid:99)(cid:116)(cid:105)(cid:111)(cid:110) |
| As the criminal network in embedded in a network of company employees, we |
| extract the sub-network GT(cid:83) connected by suspicious topics only, so as to minimize |
| the coupling of the company’s hierarchical structure to the conspiracy relations. |
| Supposing Tij denotes the set of topics mentioned by messages from node i to |
| node j, and TS denotes the set of known suspicious topics (TS = {7, 11, 13}). Then |
| is the maximum subgraph of the original graph G, whereas GT(cid:83) |
| (16) Tij ⊆ TS, for all (i, j) ⊆ ET(cid:83) |
| (cid:51)(cid:46)(cid:51) (cid:69)(cid:100)(cid:103)(cid:101) (cid:114)(cid:101)(cid:118)(cid:101)(cid:114)(cid:115)(cid:101) |
| 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 GT(cid:83) 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(cid:83) |

|  |
| --- |
| 0 |
| (cid:84)(cid:101)(cid:97)(cid:109) (cid:35) (cid:49)(cid:52)(cid:53)(cid:51)(cid:49) (cid:80)(cid:97)(cid:103)(cid:101) (cid:49)(cid:52) (cid:111)(cid:102) (cid:49)(cid:57) |
| (cid:51)(cid:46)(cid:52) (cid:82)(cid:101)(cid:115)(cid:117)(cid:108)(cid:116)(cid:115) |
| By running LeaderRank on G′ , a ranking score is assigned to every node in |
| T(cid:83) |
| 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 organiza- |
| tion. |
| Name LeaderRank score |
| Yao 2.67 |
| Alex 2.21 |
| Paul 1.92 |
| Elsie 1.62 |
| Table 4: Partial results of LeaderRank on G′ |
| T(cid:83) |
| (cid:51)(cid:46)(cid:53) (cid:69)(cid:109)(cid:112)(cid:105)(cid:114)(cid:105)(cid:99)(cid:97)(cid:108) (cid:115)(cid:117)(cid:112)(cid:112)(cid:111)(cid:114)(cid:116) |
| Empirical analysis of criminal networks has found that a leader of a criminal or- |
| ganization tends to carefully balancing his or her degree centrality and betweenness |
| centrality. It has been proposed that the leader usually maintains a high between- |
| ness centrality but a relatively low degree centrality, for enhancing eﬃciency and |
| meanwhile ensuring safety [Morselli, 2010]. |

|  |  |  |
| --- | --- | --- |
| 0 | 1 | 2 |
|  |  | Known conspirat ors |
|  | 20 |  |
|  |  | High conspiracy prob. |
|  |  | Yao (inferred leader) |
|  | 18 |  |
|  | 16 |  |
| Degreecentrality | 14 |  |
|  | 12 |  |
|  | 10 |  |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| Know |  |  | rs |  |  |  |  |
|  | Know | n conspirato | rs |  |  |  |  |
|  | High Yao( | conspiracy p inferred lead | rob. er) |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

|  |
| --- |
| 0 |
| (cid:84)(cid:101)(cid:97)(cid:109) (cid:35) (cid:49)(cid:52)(cid:53)(cid:51)(cid:49) (cid:80)(cid:97)(cid:103)(cid:101) (cid:49)(cid:53) (cid:111)(cid:102) (cid:49)(cid:57) |
| And our inference that Yao is the leader is thus empirically supported. Figure 5 |
| illustrates the joint distribution of betweenness centrality CB and degree centrality |
| (Din +Dout) 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. |
| (cid:51)(cid:46)(cid:54) (cid:68)(cid:105)(cid:115)(cid:99)(cid:117)(cid:115)(cid:115)(cid:105)(cid:111)(cid:110) |
| 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 reallo- |
| cation dynamics on the network, is generally more computationally inexpensive |
| compared to traditional methods like block-modeling. Our scheme accommodates |
| large databases with higher eﬃciency. |
| (cid:52) (cid:69)(cid:118)(cid:97)(cid:108)(cid:117)(cid:97)(cid:116)(cid:105)(cid:110)(cid:103) (cid:116)(cid:104)(cid:101) (cid:77)(cid:111)(cid:100)(cid:101)(cid:108) |
| (cid:52)(cid:46)(cid:49) (cid:83)(cid:101)(cid:110)(cid:115)(cid:105)(cid:116)(cid:105)(cid:118)(cid:105)(cid:116)(cid:121) (cid:97)(cid:110)(cid:97)(cid:108)(cid:121)(cid:115)(cid:105)(cid:115) |
| Considering the usual incompleteness, imprecision and even inconsistency with |
| criminal social networks [Xu, 2005], the method for inferring criminality or con- |
| spiracy 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 con- |
| nected to criminal activity, and Chris, who was considered innocent before, has |
| now proved guilty. |
| (cid:52)(cid:46)(cid:49)(cid:46)(cid:49) (cid:80)(cid:114)(cid:105)(cid:111)(cid:114)(cid:105)(cid:116)(cid:121) (cid:108)(cid:105)(cid:115)(cid:116) |
| 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 (cid:75)(cid:101)(cid:110)(cid:100)(cid:97)(cid:108)(cid:108)(cid:39)(cid:115) |
| (cid:116)(cid:97)(cid:117) coeﬃcient τ [Sen, 1968]. Given two priority lists {pk} = {p(cid:49), p(cid:50), · · · , pn} and |
| {qk} = {q(cid:49), q(cid:50), · · · , qn} (for example, p(cid:50) = 5 means node 2 is ranked 5th by {pk} |
| list), then (i, j), i ̸= j is said to be a concordant pair if their rankings agree in two |
| lists, i.e. pi > pj, qi > qj or pi < pj, qi < qj; (i, j) is said to be a discordant pair if |
| their rankings disagree in two lists. |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 0 | 1 | 2 | 3 | 4 | 5 |
| (cid:84)(cid:101)(cid:97)(cid:109) (cid:35) (cid:49)(cid:52)(cid:53)(cid:51)(cid:49) |  |  |  |  | (cid:80)(cid:97)(cid:103)(cid:101) (cid:49)(cid:54) (cid:111)(cid:102) (cid:49)(cid:57) |
| Name | Dolores | Crystal | Jerome | Sherri | Neal |
| Rank (previous) | 1 | 2 | 3 | 4 | 5 |
| Prob. of conspiracy (previous) | 0.555 | 0.508 | 0.388 | 0.316 | 0.299 |
| Rank (new) | 2 | 1 | 6 | 9 | 3 |
| Prob. of conspiracy (new) | 0.621 | 0.629 | 0.405 | 0.393 | 0.504 |
| Name | Christina | Jerome | William | Dwight | Beth |
| Rank (previous) | 6 | 7 | 8 | 9 | 10 |
| Prob. of conspiracy (previous) | 0.555 | 0.508 | 0.388 | 0.316 | 0.299 |
| Rank (new) | 14 | 5 | 4 | 12 | 16 |
| Prob. of conspiracy (new) | 0.335 | 0.407 | 0.409 | 0.352 | 0.334 |

|  |
| --- |
| 0 |
| Rank (new) 14 5 4 12 16 |
| Prob. of conspiracy (new) 0.335 0.407 0.409 0.352 0.334 |
| Table 5: Change with top 10 in the priority list (known conspirators excluded) |
| Then (cid:75)(cid:101)(cid:110)(cid:100)(cid:97)(cid:108)(cid:108)(cid:39)(cid:115) (cid:116)(cid:97)(cid:117) is deﬁned as |
| (number of concordant pairs) − (number of discordant pairs) |
| τ = . (17) |
| (cid:49) (cid:50) n(n − 1) |
| τ lies in the range of [−1, 1], whereas 1 for perfect ranking agreement, −1 for utter |
| disagreement. |
| The (cid:75)(cid:101)(cid:110)(cid:100)(cid:97)(cid:108)(cid:108)(cid:39)(cid:115) (cid:116)(cid:97)(cid:117) between two priority lists obtained in Requirement 1 and |
| Requirement 2 is τ = 0.86, justifying the robustness of the machine learning ap- |
| proach. |
| 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. |
| (cid:52)(cid:46)(cid:49)(cid:46)(cid:50) (cid:80)(cid:114)(cid:111)(cid:98)(cid:97)(cid:98)(cid:105)(cid:108)(cid:105)(cid:116)(cid:121) (cid:105)(cid:110)(cid:13)(cid:97)(cid:116)(cid:105)(cid:111)(cid:110) |
| 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 prob- |
| ability, as indicated by the distribution of dots skewed from the diagonal line. The |
| augmented probability is compatible with the new information that expands both |
| the set of suspicious topics and known conspirators. |

|  |  |
| --- | --- |
| 0 | 1 |
|  | (cid:50) |
| (cid:41) |  |
|  | (cid:49) |

|  |  |  |  |
| --- | --- | --- | --- |
| 0 | 1 | 2 | 3 |
|  |  | (cid:49) |  |
|  | (cid:57) |  |  |
|  | (cid:56) |  |  |
|  | (cid:55) |  |  |
|  | (cid:54) |  |  |
|  | (cid:53) |  |  |
|  | (cid:52) |  |  |
|  |  |  | (cid:116) |
|  |  |  | (cid:115) |
|  | (cid:51) |  | (cid:115) |
|  |  |  | (cid:41) |
|  | (cid:50) |  | (cid:41) |
|  |  |  | (cid:41) |
| (cid:41) |  |  |  |
|  | (cid:49) |  |  |

|  |  |
| --- | --- |
| 0 | 1 |
| (cid:50) |  |
|  | (cid:41) |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| (cid:84)(cid:101)(cid:97)(cid:109) (cid:35) (cid:49)(cid:52)(cid:53)(cid:51)(cid:49) |  |  |  |  |  |  |  |  |  |  |  |  |  | (cid:80)(cid:97)(cid:103)(cid:101) (cid:49)(cid:55) (cid:111)(cid:102) (cid:49)(cid:57) |
|  |  | 1 |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  | 0.98 |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  | Conspirators |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  | Non−conspirators |  |  |  |  |
|  |  | 0.96 |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  | 0.94 |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  | 0.92 |  |  |  |  |  |  |  |  |  |  |  |  |
|  | Kendall’s tau | 0.9 |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  | 0.88 |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  | 0.86 |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  | 0.84 |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  | 0.82 |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  | 0.8 |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  | 0 | 0.05 | 0.1 | 0.15 | 0.2 | 0.25 | 0.3 | 0.35 |  | 0.4 | 0.45 | 0.5 |  |
|  |  |  |  |  |  | Probability of wrong classification |  |  |  |  |  |  |  |  |
| Figure 6: The expected Kendall’s tau declines as misclassiﬁcation probability in- |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| creases |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| (cid:49) |  |  |  |  |  |  |  |  |  |  |  |  |  | (cid:32) |
| (cid:57) |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| (cid:56) |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| (cid:55) |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| (cid:54) |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| (cid:53) |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| (cid:52) |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  | (cid:116) |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  | (cid:115) |  |  |  |  |
| (cid:51) |  |  |  |  |  |  |  |  |  | (cid:115) |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  | (cid:41) |  |  |  |  |
| (cid:50) |  |  |  |  |  |  |  |  |  | (cid:41) |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  | (cid:41) |  |  |  |  |
| (cid:41) |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| (cid:49) |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| (cid:48) | (cid:32) |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | (cid:48) | (cid:49) | (cid:50) |  | (cid:51) | (cid:52) | (cid:53) | (cid:54) | (cid:55) |  | (cid:56) |  | (cid:57) | (cid:49) |
|  |  |  |  | (cid:41) |  |  |  |  |  |  |  |  |  |  |
|  | Figure 7: Criminal probabilities before and after the change of conditions |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | 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. |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  | Cons Non− | pirators conspira | tors |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  | (cid:84)(cid:114)(cid:97)(cid:105)(cid:110)(cid:105)(cid:110)(cid:103)(cid:32)(cid:115)(cid:101)(cid:116) (cid:85)(cid:110)(cid:107)(cid:110)(cid:111)(cid:119)(cid:110)(cid:32)(cid:110)(cid:111)(cid:100)(cid:101)(cid:115) |  |  |  |
|  |  |  |  |  |  |  |  | (cid:67)(cid:104)(cid:114)(cid:105)(cid:115) (cid:71)(cid:114)(cid:101)(cid:116)(cid:99)(cid:104)(cid:101)(cid:110)(cid:32)(cid:40)(cid:109)(cid:97)(cid:110)(cid:97)(cid:103)(cid:101)(cid:114)(cid:41) |  |  |  |
|  |  |  |  |  |  |  |  | (cid:74)(cid:101)(cid:114)(cid:111)(cid:109)(cid:101)(cid:32)(cid:40)(cid:109)(cid:97)(cid:110)(cid:97)(cid:103)(cid:101)(cid:114)(cid:41) (cid:68)(cid:111)(cid:108)(cid:111)(cid:114)(cid:101)(cid:115)(cid:32)(cid:40)(cid:109)(cid:97)(cid:110)(cid:97)(cid:103)(cid:101)(cid:114)(cid:41) |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |

|  |
| --- |
| 0 |
| (cid:84)(cid:101)(cid:97)(cid:109) (cid:35) (cid:49)(cid:52)(cid:53)(cid:51)(cid:49) (cid:80)(cid:97)(cid:103)(cid:101) (cid:49)(cid:56) (cid:111)(cid:102) (cid:49)(cid:57) |
| (cid:53) (cid:68)(cid:105)(cid:115)(cid:99)(cid:117)(cid:115)(cid:115)(cid:105)(cid:111)(cid:110) (cid:97)(cid:110)(cid:100) (cid:67)(cid:111)(cid:110)(cid:99)(cid:108)(cid:117)(cid:115)(cid:105)(cid:111)(cid:110) |
| (cid:53)(cid:46)(cid:49) (cid:83)(cid:117)(cid:109)(cid:109)(cid:97)(cid:114)(cid:121) |
| 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 cen- |
| trality in topology, conspiratorial level of connectivity and communication contents. |
| Experimental results hits about 73% correct prediction rate in the basis 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 en- |
| hancement lead us to some deeper discursion 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 from aspects on |
| the topology property of a criminal leader. |
| Finally, the high value of (cid:75)(cid:101)(cid:110)(cid:100)(cid:97)(cid:108)(cid:108)(cid:39)(cid:115) (cid:116)(cid:97)(cid:117) illustrates the nonsensitive property of |
| the model. The weakness of our model is mainly about features. Diﬀerent networks |
| may not share the same features. Thereby the features of particular network may |
| become meaningless when type of network changes. Furthermore, large amounts of |
| information might seriously limit the capability of the resource allocation strategy |
| because the construction of bipartite graph itself becomes a problem when facing |
| with too much noise information. |
| (cid:53)(cid:46)(cid:50) (cid:70)(cid:117)(cid:114)(cid:116)(cid:104)(cid:101)(cid:114) (cid:100)(cid:105)(cid:115)(cid:99)(cid:117)(cid:115)(cid:115)(cid:105)(cid:111)(cid:110) |
| When taking about the good portability of a model on diﬀerent models, we could |
| focus on two aspects: the similarity of two networks and diﬀerence. Some common |
| pattern appeared in diﬀerent kinds of networks,including biological network, are |
| the small-world property, power-law degree distributions, network transitivity and |
| community structure [Girvan, 2001]. Either topology or transmission properties |
| shared by diﬀerent networks could help to build a model with good portability. |
| However, on the other hand, the fundamental diﬀerence in mechanism of in- |
| formation transmission decides the distinguishing models: people ﬁnd out clues |
| through evidence or relative materials to break down the invisibility of a criminal |
| network based on reasoning and deduction. Even though share some similar prop- |
| erties, individuals in other networks like some biological networks communicate |
| with each other under some relatively ﬁxed principles or unchangeable pattern for |
| division of work. The co-occurrence-based approaches always fail to characterize |
| function biological interactions [Chen & Sharp, 2004], where particular dynamics |
| of analytical model might perform better like utilizing revised infectious disease |
| model to predict disfunction or diseased biological components. |

|  |
| --- |
| 0 |
| (cid:84)(cid:101)(cid:97)(cid:109) (cid:35) (cid:49)(cid:52)(cid:53)(cid:51)(cid:49) (cid:80)(cid:97)(cid:103)(cid:101) (cid:49)(cid:57) (cid:111)(cid:102) (cid:49)(cid:57) |
| (cid:82)(cid:101)(cid:102)(cid:101)(cid:114)(cid:101)(cid:110)(cid:99)(cid:101)(cid:115) |
| [Baker & Faulkner, 1993] Baker, W. E. & Faulkner, R. R. (1993). The Social |
| Organization of Conspiracy: Illegal Networks in the Heavy Electrical Equipment |
| Industry. |
| [Chen & Sharp, 2004] Chen, H. & Sharp, B. M. (2004). Content-rich biological |
| network constructed by mining PubMed abstracts.. BMC bioinformatics (cid:53), 147. |
| [Freeman, 1979] Freeman, L. (1979). Centrality in social networks conceptual clar- |
| iﬁcation. Social networks. |
| [Girvan, 2001] Girvan, M. (2001). Community structure in social and biological |
| networks. PNAS. |
| [Krebs, 2002] Krebs, V. E. (2002). Mapping Networks of Terrorist Cells. (cid:50)(cid:52)(3), |
| 43–52. |
| [L¨u (cid:101)(cid:116) (cid:97)(cid:108)(cid:46), 2011] L¨u, L., Zhang, Y.-C., Yeung, C. H., & Zhou, T. (2011). Leaders |
| in social networks, the Delicious case.. PloS one (cid:54)(6), e21202. |
| [Morselli, 2010] Morselli, C. (2010). Assessing Vulnerable and Strategic Positions |
| in a Criminal Network. Journal of Contemporary Criminal Justice (cid:50)(cid:54)(4), 382– |
| 392. |
| [Sabidussi, 1966] Sabidussi, G. (1966). The centrality index of a graph. Psychome- |
| trika. |
| [Sen, 1968] Sen, K. P. (1968). Estimates of the regression coeﬃcient based on |
| Kendall’s tau. Journal of the American Statistical Association. |
| [Wheat, 2007] Wheat, C. (2007). Algorithmic Complexity and Structural Models |
| of Social Networks. Management , 1–38. |
| [Xu, 2005] Xu, J. (2005). Criminal network analysis and visualization. Communi- |
| cations of the ACM (cid:52)(cid:56)(6). |
| [Xu & Chen, 2003] Xu, J. & Chen, H. (2003). Untangling Criminal Networks : A |
| Case Study. World Trade , 232–248. |
| [Zhou, 2007] Zhou, T. (2007). Bipartite network projection and personal recom- |
| mendation. Physical Review E (cid:55)(cid:54)(4), 1–7. |