Crime Busting by An Iterative 2-phase Propagation Method

Team # 13855

Summary

In this article, we develop an efficient method to deal with the problem of finding out the likelihood of each person involved in a conspiracy. All the possible suspected conspirators work in the same office complex and communicate with each other, thus forming a *message network*. We design a model and develop corresponding methodology to tackle this model and finally produce a *priority list* of potential conspirators.

We assign each person(i.e. a node in message network) a real number to represent the likelihood of being involved in conspiracy. This number will be updated during the iterative propagation process.

We also assign a real number to each topic to indicate how probable this topic is involved in the conspiracy. This number is **not fixed** and will also be iteratively modified during the propagation.

In the article, we formally define these two values and develop their formulations so that we could design an algorithm that could calculate these two functions.

We do several iterations of the following 2-phase propagation process until convergence.

Each iteration consists of the following two phases:

Person Phase In this phase, we recalculate the suspiciousness of each node, based on the suspiciousness of its neighbours and messages between it and its neighbours.

Topic Phase In this phase, we recalculate the suspiciousness of each topic, based on the suspiciousness of people who talk about this topic.

We also exploit an *exponential decay* between two iterations to make the effect of messages attenuate as the distance increase.

The final suspiciousness of each person is used to produce the priority list of conspirators.

At last, we do a series of experiments to evaluate our models and analyse the results. We also discuss our model's sensitivity to the choice of the parameters and its scalability to other applications.

Contents

1 Problem Restatement

We are given a group of suspected conspirators who form a communication network by sending messages of various topics to each other. Some of the given people are known to be involved in the conspiracy while some others innocent. We are required to identify the most probable conspirators among the remaining ones.

We are also encouraged to find out the leaders of the conspiracy and talk about several other topics, such as how could the semantic and text analysis improve our methodology.

2 Problem Analysis

In order to find out the most probable conspirators, we believe that a priority list would be an ideal output. Therefore, we assign the ith person a real number S_i ($S_i \in [0,1]$) to indicate the likelihood of him involved in the conspiracy.

And we try to figure out a method to eventually determine these numbers.

We also believe that topics have their suspiciousness. Different topics suggest different probability of the speakers being conspirators. For example, a cabal who frequently talk about having a secret convention are highly suspicious.

On the other hand, we think that the people who are talking about a topic can reversely affect the suspiciousness of that topic. For instance, the conspirators may develop some argots which perhaps seem innocuous and are not considered as suspicious at first. Therefore, we make the suspiciousness of a topic changeable following the intuition that frequent mentions among conspirators can make a topic more suspicious.

We thereby develop our iterative 2-phase propagation method to solve this problem.

3 Basic Assumptions

In this section, we discuss several key assumptions we have made and rationale for making these assumptions.

Assumption 1. *We assume that the suspiciousness of a person is determined by* ***both*** *the people with whom he talks* ***and*** *the topics he talk about.*

We believe that a conspirator may be a good friend to some non-conspirator and they talks of a lot of everyday topics. So being intimate with some highly suspected conspirators does not inevitably suggest a extremely high probability of involvement in the conspiracy. We must take both these two factors into account with some elaborate formulations to determine one's suspiciousness.

Assumption 2. *We treat the influence of a message as bidirectional.*

For example, no matter one send a suspicious message to a probable conspirator or receive one from him, he becomes more suspicious. In Section 4.2, we discuss the preprocessing works we have done to tailor the data to meet this assumption.

Assumption 3. *We assume the suspiciousness of prior known conspirators and non-conspirators to be fixed as* 1 *and* 0 *respectively, and will never change during the iterative propagation process.*

Assumption 4. *We treat one message on multiple topics as several messages that each is on a single topic.*

Assumption 5. *We assume the impacts of suspiciousness imposed on a specific person by different contacts are independent.*

We explain the rationale for making this assumption in Section 4.3.2 after we present our models.

Assumption 6. *The impacts of suspiciousness imposed by someone diminish as the distance to this person in the message network becomes longer based on an exponential decay rule.*

We discuss this in further details in Section 4.3.4.

4 Models and Methodology

4.1 Definitions

Name	Definition	Value or Expression
$\mathbf{P_N}$	$\mathbf{P_N}$ is the total number of people.	83
$\mathbf{M_N}$	$\mathbf{M_N}$ is the total number of messages.	910 $^{\alpha}$
$\mathbf{T_N}$	$\mathbf{T_N}$ is the total number of topics of messages.	15
$\mathcal{P}$	$\mathcal{P}$ is the set of all people(represented by their IDs).	$\{p \vert 0 \leqslant p < \mathbf{P_N}, p \in \mathbb{N}\}$
$\mathcal{T}$	$\mathcal{T}$ is the set of all topics(represented by their IDs).	$\{t \vert 1 \leqslant t \leqslant \mathbf{T_N}, t \in \mathbb{N}\}$
p_i	p_i is a specific person whose id is i, where $i \in \mathcal{P}$	
t_i	t_i is a specific topic whose id is i, where $i \in \mathcal{T}$	
t_{xy}	t_{xy} is the topic of the message sent from p_x to p_y	
S_i	S_i is the likelihood of p_i involved in the conspiracy.	$0 \leqslant S_i \leqslant 1$
R_i	R_i is the suspiciousness of t_i.	$0 \leqslant R_i \leqslant 1$ $^{\beta}$
DR_S	DR_S is the decay rate of the propagation of S function.	
DR_R	DR_R is the decay rate of the propagation of R function.	
cDR_S	cDR_S indicates the current degree of decay of S.	$cDR_S = (DR_S)^n$ $^{\gamma}$
cDR_R	cDR_R indicates the current degree of decay of R.	$cDR_S = (DR_S)^n$

$^{\alpha}$ We will explicate this value in Section 4.2.

$^{\beta}$ We will give detailed definitions and formulations of S_i and R_i in the following sections.

$^{\gamma}$ n is the # of current iterations.

Table 1: Model Definitions

4.2 Preprocessing

Before we present our models, we would like to address the preprocessing works we have done to the data.

First, we find that there are two self-towards messages, which are sent by p_3 and p_{30} to themselves. We expurgate these two messages since we believe that it makes no sense to send a message to oneself. Also, we discover a message whose topic is 18, which is invalid, so we expurgate this message as well. Still, there are different employees with same names in the office complex who are distinguished primarily by node IDs, are renamed with a suffix though. For instance, p_4 Gretchen 1 and p_{32} Gretchen 2.

We build a graph model using the provided data, in which people are vertices and messages are edges.

In addition, since we have assumed that a message has bidirectional influence (in Section 3), we have done the following modifications to the graph. For each edge in the graph (i.e. a message), we create another message of the same topic whose sender and receiver are swapped, which is a common trick to turn a directed graph into an undirected one. **For this reason, in the remaining part of this article, every time we mention receiving a message, we mean both sending and receiving a message.**

When a message contains more than one topics, we split it into several messages for the convenience of manipulation.

These preliminary works could answer why M equals 910 rather than 400.

4.3 Methodology

4.3.1 Overview

As we described in the abstract, our method exploits the propagation of suspiciousness in the message network to determine how likely a person is involved in the conspiracy.

Our model can be better understood based on the following intuition. **Suspiciousness can be passed along with the message.** i.e. You become more suspicious when you received a message from a suspected person. On the other hand, **the topic of the message can affect the probability of transmission of suspiciousness.** Talking with someone about a dubious topic would increase the likelihood of transmitting his suspiciousness to you.

Following this intuition, we defined S_i for each person p_i and R_i for each topic t_i and calculate them in turn during an iterative propagation process.

Definition 1 (Person Suspiciousness Function S). *S_i is the likelihood of person p_i to be involved in the conspiracy.*

Definition 2 (Topic Suspiciousness Function R). *R_i is defined as the suspiciousness transmission rate of topic t_i.*

We give more specific definitions and formulate the expressions of these two functions in Section 4.3.2 and 4.3.3.

We do several iterations to calculate the value of S and R. Each iteration consists of a *P phase* and a *T phase*. In each phase, we use the value of S and R calculated in the previous *P&T phases* to recalculate the new S or R values.

We believe that a direct contact has a far more powerful influence than an indirect contact, we exploit an *exponential decay* as the number of iteration grows to make the propagation decrescendo.

Algorithm 1 shows an overview of our algorithm.

Algorithm 1 Top view of the whole methodology

```
Initialization:
  Build the graph and assign appropriate initial values to $S$ and $R$.
Iterations:
  for $i = 1 \to MAX\_ITER\_NUM$ do
    PropagateP($cDR_S$); { P phase with the current decay degree of $S$}
    PropagateT($cDR_R$); { T phase with the current decay degree of $R$}

    $cDR_S \leftarrow cDR_S * DR_S$;
    $cDR_R \leftarrow cDR_R * DR_R$;
  end for
  return $S$;
```

4.3.2 Formulations of S Function

According to the aforementioned discussions and Assumption 1, we try to formulate S function for a person p_i based on the impacts imposed by all p_i's neighbours.

For a specific neighbour of p_i: p_j, we define the impact imposed on p_i by p_j as $C_{ji} = S_j * R_{t_{ji}}$.

For two distinctive neighbour of p_i: p_j and p_k, we believe that under most circumstances, the presence of C_{ki} has limited effect on C_{ji}. Therefore, we assume C_{ji} and C_{ki} are independent.

So for a person p_i:

$$\begin{aligned} S_i &= \bigcup_{(j,i)\in\mathbf{M}} C_{ji} \\ &= C_{ji} + \bigcup_{(k,i)\in\mathbf{M}-\{\mathbf{j}\}} C_{ki} - C_{ji} * \bigcup_{(k,i)\in\mathbf{M}-\{\mathbf{j}\}} C_{ki} \\ &= \ldots \quad \ldots \end{aligned} \tag{1}$$

We thus convert the calculation of arbitrary union of all p_i's neighbours into one of its sub-problem of calculating $\bigcup_{(k,i)\in\mathbf{M}-\{\mathbf{j}\}} C_{ki}$.

We could first calculate $\bigcup\limits_{(k,i)\in\mathbf{M}-\{\mathbf{j}\}} C_{ki}$, then use it to calculate $\bigcup\limits_{(j,i)\in\mathbf{M}} C_{ji}$. And we could apply the method in Euqation 1 recursively to convert the calculation of $\bigcup\limits_{(k,i)\in\mathbf{M}-\{\mathbf{j}\}} C_{ki}$ into solving its sub-problem until the number of neighbours decreases to 1, in which case it simply equals C_{ki}.

We design the following Algorithm 2 to calculate S_i iteratively based on this method.

Algorithm 2 Algorithm for calculating S_i

```
Initialization:
  $S$ and $R$ values are those produced in the last P phase and T phase
  Set each $C_{ji} = S_j * R_{t_{ji}}$
  $Result \leftarrow 0$ {The reason for using the notation Result instead of $S_i$ is to avoid confusion with $S_i$ calcualated in last P Phase}
Iterations:
  for all $p_j \in p_i$'s neighbours do
    $Result = Result + C_{ji} - Result * C_{ji}$ {Calculate $S_i$ iteratively using the method in Equation 1}
  end for
  return $(1 - cRD_S) * S_i + cRD_S * Result$ {We do not use Result directly as $S_i$, becasue of the exponential decay, explained in Section 4.3.4}
```

In the *P phase* of one iteration, for each p_i which is neither known conspirator nor known innocent, we use Algorithm 2 to recalculate S_i to propagate the suspiciousness to its neighbours.

The initial value of S is set as follows: All the known conspirators have $S = 1$ while known non-conspirators have $S = 0$. And other ordinary people have $S = 0.5$.

4.3.3 Formulations of R Function

As we mentioned before, we believe that topics are also of different suspiciousness, which are related to the suspiciousness of those who frequently talks about these topics.

In our model, we define the suspiciousness asscociated with a topic to be the likelihood of the suspiciousness to be passed along the message of that topic.

We interpret this definition by defining R as follows: Suppose p_i has only one neighbour p_j, which is a known conpirator. We define $R_{t_{ji}}$ as the probability of p_i to be a conspirator.

This definition can be understood this way: All the suspiciousness of p_i comes from p_j since p_j is the only neighbour of p_i. And because p_j is a known conspirator, where $S_j = 1$, S_i can reflect the likelihood of p_j's suspiciousness to be transmitted to p_i.

Now we present the formulation of R. Suppose (p_j, p_i) is a message from p_j to

p_i of topic t_x. According to Equation 1 in Section 4.3.2, we have

$$\begin{aligned} S_i &= \bigcup_{(k,i)\in\mathbf{M}} C_{ki} \\ &= \bigcup_{(k,i)\in\mathbf{M}} S_k * R_{t_{ki}} \\ &= S_j * R_x + \bigcup_{(k,i)\in\mathbf{M}-\{\mathbf{j}\}} C_{ki} - S_j * R_x * \bigcup_{(k,i)\in\mathbf{M}-\{\mathbf{j}\}} C_{ki} \end{aligned} \tag{2}$$

So we have

$$R_x = \frac{S_i - \bigcup_{(k,i)\in\mathbf{M}-\{\mathbf{j}\}} C_{ki}}{S_j * (1 - \bigcup_{(k,i)\in\mathbf{M}-\{\mathbf{j}\}} C_{ki})} \tag{3}$$

We could use Algorithm 2 described in Section 4.3.2 to calculate $\bigcup_{(k,i)\in\mathbf{M}-\{\mathbf{j}\}} C_{ki}$.

Since there may be many messages sharing a common topic t_x, using Equation 3 will yield one R_x for each message, producting several R_x. We then use the arithmatic average to modify R_x under the rule of the exponential decay.

Algorithm 3 shows how to calculate R_x at the *T phase* at a specific iteration.

Algorithm 3 Algorithm for calculating R_x

```
Initialization:
  $S$ and $R$ values are those produced in the previous P phase and T phase
  $Result \leftarrow 0$
  $Count \leftarrow 0$ {Count is the number of messages on $t_x$}
Iterations:
  for all edges $(p_j, p_i)$ where $t_{ji} = x$ do
    $Temp \leftarrow \bigcup_{(k,i)\in\mathbf{M}-\{\mathbf{j}\}} C_{ki}$ {Using the method in Algorithm 2}
    $Result \leftarrow Result + \frac{S_i - Temp}{S_j*(1-Temp)}$ {Equation 3}
    $Count \leftarrow Count + 1$
  end for
  $Result \leftarrow Result/Count$ {Taking average over all messages}

  return $(1 - cRD_R) * R_x + cRD_R * Result$ {Exponential decay, explained in Section 4.3.4}
```

The initial value of R is set as follows: All the known suspicious topics have $R = 0.7$ while all other topics have $R = 0.05$.

In Section 6.2, we discuss some more elaborate allocations of the initial value of R.

4.3.4 Exponential Decay

As we mentioned, we exploit an exponential decay in the calculation of S and R to make the near people have a stronger influence than far ones.

To better understand this, consider a simple scenario: a chain. If the message network is a chain with the beginning person a known conspirator and all others unknown, the suspiciousness of the first person would keep transmitting to all other people. All the S_i on this chain would converge to 1 as the number of iteraions approaches to the infinity.

So we do not simply use the S and R calculated in each iteration to replace those calculated in the previous iteration. Instead, we use them to modify the previous results based on the exponential decay rule. This is the reason for the expression at the end of both Algorithm 2 and 3.

$$(1 - cRD_S) * S_i + cRD_S * Result$$
$$(1 - cRD_R) * R_x + cRD_R * Result$$

cRD_S and cRD_R shrink to a constant proportion compared with the previous iteration.

$$cDR_S = (DR_S)^n$$
$$cDR_R = (DR_R)^n$$

where n is the number of the current iteration.

DR_S and DR_R are set to 0.5 in the basic experiments, and we will discuss our model's sensitivity to different settings of DR_S and DR_R in Section 6.2.

Conseqently, as the iterations proceed, the previous results play an increasingly important role while the impacts of current propagation iteration diminishes until convergence.

5 Results and Analyses

In this secion, we discuss the basic results of our models applied to Requirement 1 and the corresponding analyses. We put the discussion of Requirement 2, 3 and 4 into Section 6.

5.1 Requirement 1

5.1.1 Main Task

In this basic requirement, there are 8 known conspirators [1] and 8 known non-conspirators. There are also 3 known suspicious topics.

In this basic experiment, the initial values of different variables and constants are set as follows. All the known conspirators have $S = 1$ while known non-conspirators have $S = 0$. And other ordinary people have $S = 0.5$. All the known suspicious topics have $R = 0.7$ while all other topics have $R = 0.05$.

When applying our model to this scenario, a priority list of all 83 people is produced, which is shown in Table 2.

And the R value for all topics are shown in Table 3.

[1] Because there are two people named Elsie

ID	S value	ID	S value	ID	S value	ID	S value
49	1	38	0.793437	12	0.640781	70	0.224238
18	1	50	0.790764	60	0.61854	77	0.217232
21	1	30	0.76857	35	0.611702	73	0.215225
67	1	32	0.766605	45	0.608586	76	0.204149
43	1	6	0.760001	14	0.607512	53	0.203252
7	1	41	0.754968	39	0.551087	55	0.197634
37	1	44	0.746949	69	0.528206	75	0.18895
54	1	40	0.724757	1	0.519437	52	0.183797
81	0.850742	20	0.720356	26	0.475945	58	0.179185
10	0.847161	8	0.718942	51	0.444996	59	0.176167
17	0.838916	33	0.714468	72	0.430927	63	0.163895
13	0.829731	31	0.711505	82	0.409267	61	0.16224
3	0.819499	24	0.701129	25	0.39586	2	0
4	0.816793	19	0.695165	80	0.388173	78	0
28	0.809526	11	0.694822	79	0.366544	0	0
15	0.806096	27	0.686166	56	0.364273	74	0
16	0.804627	29	0.685091	57	0.352322	68	0
34	0.802074	46	0.679116	23	0.287067	48	0
36	0.798483	42	0.668063	71	0.27554	65	0
22	0.796896	9	0.667589	62	0.255381	64	0
47	0.793939	5	0.664651	66	0.239405		

Table 2: Priority list for Requirement 1

ID	R value	ID	R value	ID	R value	ID	R value
1	0.114903	5	0.0474539	9	0.133353	13	0.591751
2	0.16544	6	0.0839097	10	0.102462	14	0.13817
3	0.203057	7	0.569282	11	0.549412	15	0.153815
4	0.0863024	8	0.108256	12	0.0928846		

Table 3: Suspiciousness of topics(R values) for Requirement 1

We find out that among all the three suspicious topics, t_{13} has a slightly larger R value than the other two. And as the data describes, t_{13} is believed as the key in the conspiracy, which validates our models and methodology.

In addition, among all other topics, t_2 and t_3 have relatively high suspiciousness. When we put t_2 under scrutiny, we find that the data show some of the message traffic of t_2 contain Spanish words. And based on our model, t_2 is relatively frequently talked about among suspicious people. So it may be the case that t_2 contains some form of argots or jargons which need to be further investigated. Therefore, our model can help to find out potentially suspicious topics which could direct the investigation of ICM officers.

5.1.2 Senior Managers

The three senior managers of the company stated in the problem are Gretchen, Jerome and Delores, but there is not anyone whose name is Delores in the name list and we found there is a name called Dolores instead. Furthermore we found two Gretchen and two Jerome in the name list. In this way, we analyze the five people, Dolores, two Jerome and two Gretchen whose IDs are 4, 10, 16, 32 and 34.

First, we found that these peoples S values are as shown in Table 4.

Name & ID	S value
Gretchen No.4	0.816793
Dolores No.10	0.847161
Jerome No.16	0.804627
Gretchen No.32	0.766605
Jerome No.34	0.802074

Table 4: S values for Senior Managers

We can see that the suspiciousness of Dolores No.10, Jerome No.16 , Jerome No.34 and Gretchen No. 4 are relatively high. But we cannot confirm that anyone of them is conspirator, because none of their suspiciousness is conspicuously high. As a result, we must further analyze these three people.

We analyze the messages talked by the people mentioned above which are conspiratorial. We define the conspiratorial messages talked with person whose priority is comparatively high as marked messages. We took the highest 20 people (including known conspirators) as highly suspected people. We count the conspiratorial messages, marked messages and total messages sent or received by the five people mentioned above. The data is displayed in Table 5.

Name & ID	Marked Msgs	Conspiratorial Msgs	Total Msgs
Gretchen No.4	2	5	16
Dolores No.10	3	6	17
Jerome No.16	1	5	10
Gretchen No.32	0	3	30
Jerome No.34	0	4	24

Table 5: Messages of Senior Managers

When analyzing these data, we find out that although Gretchen No.32 and Jerome No.34 are in the upper third in the whole priority list, they never discuss conspiratorial topics with highly suspected conspirators.

The reason of the fact that they have relatively high suspiciousness is that even some non-conspiratorial message sent or received by a person will slightly increase the suspiciousness of that person in our model.

Furthermore, we can explain the large total number of messages by the fact that they are senior managers. Because they are senior managers, they have to contact with lots of people. So considering that they may need to contact with all kinds of people and the limited number of marked and conspiratorial messages, we opine that their suspiciousness are not very high.

On the other hand, we notice that Dolores No.10 has small number of total messages but the largest number of marked and conspiratorial messages among the five. In other words, the ratio of suspicious messages is very high of Dolores No.10.

Therefore, we can conclude that Dolores No.10 is probably a conspirator.

5.1.3 Conspiracy Leaders

To determine the leader(s) of the conspiracy, we think that the person who contact with more conspirators using conspiratorial messages is more likely to be the leader.

Therefore, we count the number of probable conspirators a person contacts with using conspiratorial messages. The same as above, we consider the people who are top 20 in the priority list as probable conspirators. The statistics of the top 20 people in the priority list is shown in Table 6.

ID	#	ID	#	ID	#	ID	#
3	3	15	1	22	2	43	7
4	1	16	1	28	2	49	3
7	4	17	3	34	0	54	4
10	2	18	4	37	2	67	6
13	1	21	6	38	2	81	1

Table 6: # of Suspicious contacts of highly suspected conspirators on conspiratorial topics

From the Table 6, we can find that p_{21}, p_{43} and p_{67} have relatively large number of probable conspirators contacted on conspiratorial topics. So we can draw a conclusion that p_{21} Alex, p_{43} Paul and p_{67} Yao are probably the leaders of conspirators.

6 Evaluation and Further Discussions

6.1 Requirement 3: Impacts of NLP developments on our model

The development of Natural Language Processing will not only improve the performance of our current model, but also could help us to build more powerful models which is far beyond our expectation using the information that we are provided now.

The precision of topic extraction is a key constraint to the performance of our model. Take the case EZ for example, in which topic 4 is summarized as *George's stress*. If we are only given the information that Harry, Dave and George are involved in this topic (which is the case in our current ICM case), we would think of Harry as highly dubious since he shares a topic with and basically only with those known conspirators.

However, if we are informed of the content of all the messages, we would discover that we are totally misled. The fact is Harry has no idea where George's stress comes from and Dave wants George to appease Harry, which may on the quite opposite prove the innocence of Harry.

So some important information are lost during the topic extraction and classification. As the topic discovery develops by enhancing or replacing the currently used topic models such as *Latent Dirichlet Allocation* and *probabilistic Latent Semantic Indexing*, our model's performance would enhance.

Our model will also benefit a lot from the development of the textual attitude/emotion extraction and analysis. The development in all these techniques will produce fine-grained topics which may be able to identify the nuances between Harry and Dave when they are talking about George's stress.

With these informations, we may add relationships to the topics to make them not isolated but interactive.

Regardless of all these impacts based on the evolutionary changes of NLP, which is unlikely to be achieved in the recent future, introducing pragmatic NLP techniques to the data will also ameliorate our current model.

A simple example is that we could use text analysis to each topic to assign a suspiciousness value to each topic, rather than simply marking some topics as suspicious. We do this (by human efforts rather than NLP techniques) in our experiments, which is discussed in Section 6.2.2.

6.2 Model Sensitivity

In this section, we discuss our model's sensitivity to the prior knowledges, the settings of parameters and initial values of S and R functions with a series of experiments.

6.2.1 Requirement 2: Changing Prior Knowledge

Priority list change If Chris has gone rogue and t_1 is connected to the conspiracy as well, there would be a considerable change of the sequence in the priority list with several major leap on the rank of the most suspected, and a subtle variation of the most dubious topics measured by the R value of each topic. The result is shown in Table 13, which is put into the Appendix for the compactness of the article due to its large volume.

Analysis As our assumption indicated, the whole communication network model sees one more affirmative conspirator as one more evil propagation source with the S value of 1, and one more message delivering conspiracy as several more dangerous links initiated by high R value. Reasonably, people who have closer relationship with this new evil via more doubtable links should gain more attention in that chances are gauge on them may change tremendously or even flip over. On the other side, judgment on people who have few immediate contacts with Chris and talk less about topic 1 may remain a lot more similar.

Comparing the result, it is rather obvious that some node with a rank of top 20 50 in the previous priority list and an S value between 0.52 0.73 which was hard to discriminate them from conspirators have a much higher S value and move up at least 7 places on the priority list. Among these nodes are p_{32}, p_{45},

p_{14}, p_{69}, p_{25}, p_{31} and p_{20}. Basically, these nodes are mostly among top 30 now due to their abnormal message or their intimate connection with Chris, which can be seen in Table 7 and 8. This can also be verified by the illustration in Figure 1 that nodes mentioned above are either interlocutors of t_1 (yellow links) or contacts of Chris (red links).

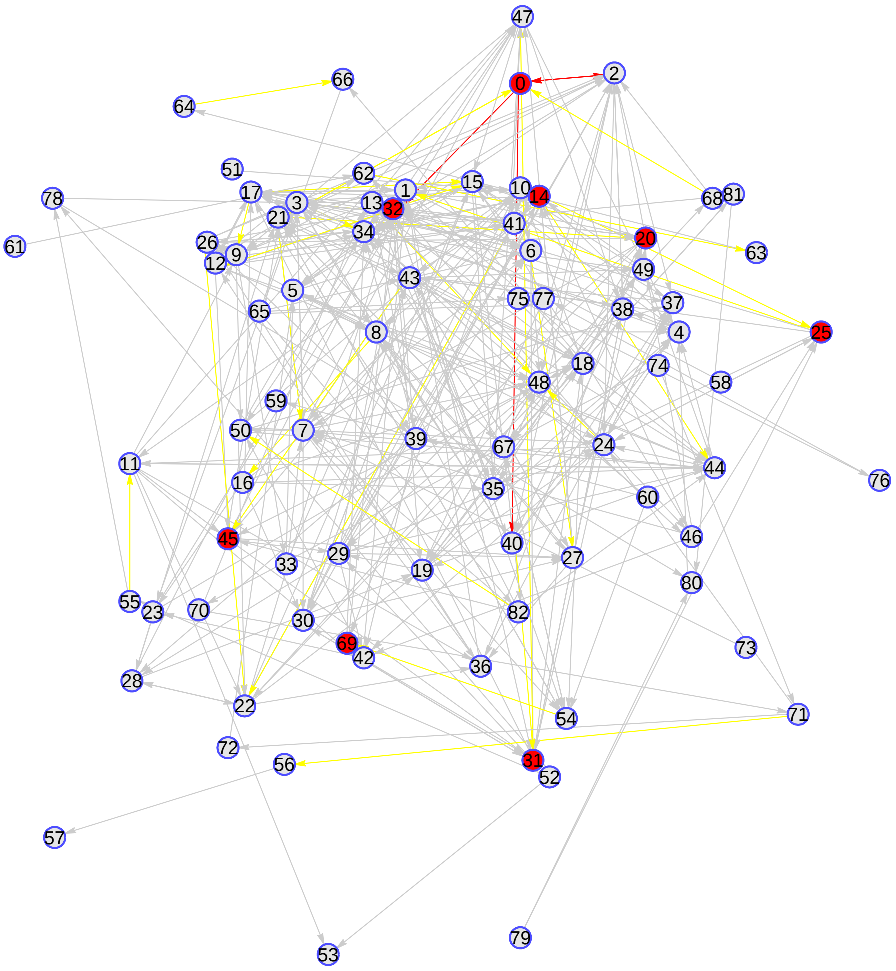

Figure 1: All connections with Chris(Node 0) and topic 1

Interlocutor No.	Identity	Number of messages	Topics
32	Unknown	3	3,6,9
21	conspirator	3	1
2	non-conspirator	6	2,9,14,15
68	non-conspirator	1	1

Table 7: Chris's message record

To give a specific illustration, we present Gretchen 2, p_{32}, who grabs our attention because it epitomizes the effect of the new information. p_{32} has heard from Chris 3 times and each time with a different topic [Table 7], and it has also been involved in up to three messages about t_1 [Table 8]. These facts are definitely the reason that its rank jumps from 17th to 6th and its likelihood of being a conspirator soars from 76.66% to 83.46% [Table 13].

Accordingly, those nodes which are far from Chris and have less involved in t_1, almost remain the same places in the priority list [Table 13]. Among them are

From	To	From	To	From	To
0	21	17	15	40	31
0	21	21	0	41	22
3	32	21	7	43	45
3	34	21	20	54	69
8	16	24	48	55	11
10	27	25	1	62	63
13	48	26	22	64	66
14	25	31	47	68	0
14	44	32	12	71	56
17	9	32	15	82	50

Table 8: All messages about topic 1

extremely-suspected conspirators like p_3, p_{16} and p_{34} and unsuspected employees like p_{53}, p_{58} and p_{59}. There are also another explanation counts for this result which is that new information changes are extremely outnumbered when judging these people considering their abundant message records.

6.2.2 Initial Value for R function

We try to set different initial values of R function for different topics, based on their textual characteristics.

Topic No.	R value	Topic No.	R value	Topic No.	R value
1	0.6	6	0.2	11	1
2	0.4	7	1	12	0.4
3	0.1	8	0.3	13	1
4	0.2	9	0.2	14	0
5	0.5	10	0.3	15	0

Table 9: R value initiation after topic analysis

These R values in Table 9 is given based on a somehow casual deduction. For instance, a Spanish atmosphere in this company is observed due to its frequent usage in communication like massages about t_2, t_7, t_{10}, t_{12}, in which t_7 is affirmatively connected to the conspiracy. Nevertheless, the contents of these topics are also dubious. Such as choosing a best restaurant for lunch, inviting a certain group of people for ski trip, these topics are all about an isolated activity with restricted participants, which might be a conspiracy meeting. Those close relationships between topics and their suspicious details might increase their R value.

As shown in Table 10, an initiation of R value can cause considerable effect on the priority list, which could reflect our model's sensitivity to the settings of initial values of R function. Also, we may expect that a better *Semantic and Text Analyses* can bring about more precise and discreet result by an accurate initiation of R as we mentioned in Section 6.1.

ID	S	Rank Change	ID	S	Rank Change	ID	S	Rank Change
7	1.0000	0	20	0.8459	-1	56	0.5691	-7
18	1.0000	0	41	0.8440	1	51	0.5585	9
21	1.0000	0	44	0.8426	-1	79	0.5518	-4
37	1.0000	0	14	0.8422	4	71	0.5351	10
43	1.0000	0	6	0.8420	2	23	0.5123	2
49	1.0000	0	11	0.8402	6	57	0.4927	-3
54	1.0000	0	33	0.8386	-2	70	0.4357	-3
67	1.0000	0	8	0.8373	5	66	0.4313	0
10	0.8556	9	9	0.8349	2	55	0.4193	0
17	0.8556	6	27	0.8326	-9	77	0.3808	-4
81	0.8555	3	19	0.8315	5	63	0.3443	-6
3	0.8551	1	5	0.8283	5	61	0.2836	1
13	0.8551	4	29	0.8272	-7	58	0.2772	-3
15	0.8545	-3	45	0.8257	-4	59	0.2754	-2
34	0.8543	-3	1	0.8251	3	76	0.2716	1
28	0.8543	-7	46	0.8203	-2	52	0.2237	-1
50	0.8541	-7	42	0.8103	4	75	0.1948	2
38	0.8527	1	25	0.8070	1	53	0.1921	-1
36	0.8525	-4	35	0.8068	-4	73	0.1836	-1
22	0.8525	1	12	0.8000	5	0	0.0000	0
30	0.8523	-1	26	0.7946	3	2	0.0000	0
32	0.8519	3	60	0.7923	-2	48	0.0000	0
47	0.8516	1	69	0.7827	7	64	0.0000	0
4	0.8512	-1	39	0.7228	-15	65	0.0000	0
40	0.8500	2	82	0.6642	3	68	0.0000	0
16	0.8494	6	72	0.6196	3	74	0.0000	0
24	0.8490	-1	80	0.6123	-1	78	0.0000	0
31	0.8480	-6	62	0.5934	-5			

Table 10: Priority list after topic analysis

6.2.3 Decay rate

As we said in Section 4.3.4, we use the decay rate (represented by DR_S and DR_R) to control the degree of propagation. In specific, DR_S controls the propagation of people's suspiciousness S while DR_R controls the propagation of topics' suspiciousness R.

So we could achieve a variety of different goals by use flexible settings. Since displaying the priority list is too space consuming, we just do two experiments as illustration.

First we set them both to be 1 (i.e. diable the exponential decay) to test the effect of introducing such decay. The resulting priority list is shown in Table 11.

The results turn out to be coordinate with our postulation in Section 4.3.4. A majority of nodes have a tendency to have a S function converged to 1.

Then we set DR_R to be 0, which means the suspiciousness of topics will never change during the iterations. This shows the results under the condition that we only take those topics believed suspicious in the problem statement as dubious. The results are shown in Table 12.

ID	S value	ID	S value	ID	S value	ID	S value	ID	S value
40	1	49	1	13	1	25	1	57	0.976686
26	1	54	1	14	1	81	1	51	0.974359
27	1	67	1	15	1	23	1	73	0.958333
28	1	69	1	17	1	71	0.999999	61	0.931401
30	1	77	1	22	1	33	0.999998	58	0.907174
31	1	80	1	18	1	72	0.999965	59	0.906998
32	1	82	1	19	1	62	0.999909	63	0.872648
34	1	24	1	20	1	76	0.999849	74	0
35	1	1	1	21	1	70	0.999839	0	0
37	1	3	1	29	1	60	0.999805	68	0
38	1	4	1	5	1	53	0.999273	78	0
39	1	6	1	50	1	52	0.998788	48	0
41	1	7	1	36	1	75	0.998326	2	0
43	1	8	1	16	1	66	0.997768	65	0
44	1	10	1	42	1	79	0.997184	64	0
45	1	11	1	46	1	55	0.996628		
47	1	12	1	9	1	56	0.994773		

Table 11: Priority list when setting $DR_S = DR_R = 1$

ID	S value	ID	S value	ID	S value	ID	S value	ID	S value
67	1	15	0.803964	19	0.682531	1	0.422953	53	0.176714
43	1	38	0.801247	46	0.680683	80	0.399671	75	0.168908
21	1	50	0.801028	29	0.673473	79	0.381148	52	0.168527
54	1	22	0.798566	27	0.667362	26	0.377577	58	0.165943
37	1	47	0.797955	42	0.664048	56	0.373485	59	0.164828
7	1	34	0.79138	9	0.663387	57	0.366902	63	0.157118
18	1	30	0.772823	60	0.652621	25	0.325067	61	0.157118
49	1	6	0.763542	5	0.652462	82	0.308527	68	0
81	0.853917	41	0.754253	24	0.64319	23	0.240106	74	0
10	0.84965	32	0.74663	12	0.587289	71	0.218424	65	0
17	0.839955	33	0.736335	14	0.574714	62	0.204772	64	0
13	0.834948	44	0.713927	35	0.569191	77	0.196454	78	0
4	0.819865	40	0.709913	45	0.561554	70	0.192799	48	0
28	0.819123	8	0.708059	39	0.534129	66	0.19245	2	0
3	0.817	20	0.702925	69	0.526705	76	0.180432	0	0
16	0.811792	31	0.688978	51	0.478053	73	0.179955		
36	0.803966	11	0.68853	72	0.445297	55	0.176714		

Table 12: Priority list when setting $DR_R = 0$

6.3 Requirement 4: Model Scalability and Other Applications

Our model is based on the property that some particular feature possessed by some nodes may propagate through some edges. In our case, for instance, suspiciousness can propagate through communications. As a result, our model accords with the kind of networks in which features may transmit through edges and edges' probability to transmit is also taken into account. Furthermore, we find that many pragmatic problems accord with this kind of network.

We illustrate the scalability of our model by applying our model to the problem of the spread of viral disease within human or other population.

In this virus instance, we ignored the dynamics of viral growth within individuals. In other words, we assume that the ability of a individual to transmit virus is constant. For each member p_i of the population of n individuals, we could use S_i in our model to represent his virion level. And $R_{ij} * S_i$ is the expected rate of transmission infectious particles from individual p_i to individual p_j. In this instance, the transmission rate will also declines exponentially with distance between individuals, which is fairly reasonable. As a result, the distribution of the final size of this epidemic can be estimated by our model, for the various initial patterns of infection.

To sum up, our model can solve the problem of prioritizing and categorizing the infected nodes in this network as well as other problems having the same property we presented at the beginning of this section.

6.4 Model Drawbacks

One drawback for our model is that the presence of a neighbour can never reduce one's suspiciousness.

A communication with a person with a quite low suspiciousness will also increase one's suspiciousness, although very slightly.

Talking with a person with low S value on a topic with a low R value may induce an increase in suspiciousness of at most $S * R$, which is typically comparable to 10^{-2}, but having many communications still may result in a relatively high suspiciousness.

Appendices

Appendix A Tables

Node	S(Old)	Rank(Old)	S(New)	Rank(New)	Rank Variation	S Variation
0	0	-	1	0	-	1
45	0.608586	38	0.76127	24	-14	0.152684
14	0.607512	39	0.745367	27	-12	0.137855
55	0.197634	61	0.414291	50	-11	0.216657
69	0.528206	41	0.721606	30	-11	0.1934
63	0.163895	66	0.345939	55	-11	0.182044
32	0.766605	17	0.834614	6	-11	0.068009
25	0.39586	47	0.637846	39	-8	0.241986
31	0.711505	25	0.795509	17	-8	0.084004
20	0.720356	22	0.803063	15	-7	0.082707
71	0.27554	53	0.439828	48	-5	0.164288
56	0.364273	50	0.49602	46	-4	0.131747
9	0.667589	33	0.735734	29	-4	0.068145
22	0.796896	12	0.829145	8	-4	0.032249
62	0.255381	54	0.409737	51	-3	0.154356
12	0.640781	35	0.717535	32	-3	0.076754
27	0.686166	29	0.755073	26	-3	0.068907
15	0.806096	8	0.835188	5	-3	0.029092
82	0.409267	46	0.562015	44	-2	0.152748
1	0.519437	42	0.619194	40	-2	0.099757
8	0.718942	23	0.767619	22	-1	0.048677
47	0.793939	13	0.815351	12	-1	0.021412
3	0.819499	5	0.840111	4	-1	0.020612
34	0.802074	10	0.822496	9	-1	0.020422
50	0.790764	15	0.808414	14	-1	0.01765
26	0.475945	43	0.605413	43	0	0.129468
40	0.724757	21	0.774412	21	0	0.049655
11	0.694822	28	0.738048	28	0	0.043226
41	0.754968	19	0.792661	19	0	0.037693
44	0.746949	20	0.783491	20	0	0.036542
17	0.838916	3	0.848474	3	0	0.009558
52	0.183797	63	0.189941	63	0	0.006144
10	0.847161	2	0.850208	2	0	0.003047
61	0.16224	67	0.16395	67	0	0.00171
81	0.850742	1	0.850546	1	0	-0.000196
16	0.804627	9	0.822188	10	1	0.017561
53	0.203252	60	0.203409	61	1	0.000157
58	0.179185	64	0.177722	65	1	-0.001463
59	0.176167	65	0.172444	66	1	-0.003723
57	0.352322	51	0.382491	53	2	0.030169
70	0.224238	56	0.225577	58	2	0.001339
75	0.18895	62	0.187607	64	2	-0.001343
66	0.239405	55	0.235887	57	2	-0.003518
73	0.215225	58	0.211614	60	2	-0.003611
77	0.217232	57	0.212775	59	2	-0.004457
13	0.829731	4	0.831232	7	3	0.001501
51	0.444996	44	0.444483	47	3	-0.000513
76	0.204149	59	0.202278	62	3	-0.001871
38	0.793437	14	0.79481	18	4	0.001373
72	0.430927	45	0.429841	49	4	-0.001086
5	0.664651	34	0.663021	38	4	-0.00163
23	0.287067	52	0.284276	56	4	-0.002791
80	0.388173	48	0.384105	52	4	-0.004068
39	0.551087	40	0.553689	45	5	0.002602
79	0.366544	49	0.369136	54	5	0.002592
4	0.816793	6	0.818802	11	5	0.002009
46	0.679116	31	0.680476	36	5	0.00136
29	0.685091	30	0.686281	35	5	0.00119
42	0.668063	32	0.669088	37	5	0.001025
36	0.798483	11	0.799232	16	5	0.000749
35	0.611702	37	0.611383	42	5	-0.000319
60	0.61854	36	0.61792	41	5	-0.00062
28	0.809526	7	0.811838	13	6	0.002312
33	0.714468	24	0.717777	31	7	0.003309
24	0.701129	26	0.703047	33	7	0.001918
19	0.695165	27	0.69616	34	7	0.000995
6	0.760001	18	0.760596	25	7	0.000595
30	0.76857	16	0.766879	23	7	-0.001691

[α] Variation = New value - Old value

[β] The priority list ranks all nodes with unknown identity. Nodes with an S value of 0 or 1 are not included.

Table 13: Priority list Comparison