Probability adjustment for sensitive information diffusion in social network

Yucheng Lu

1 Introduction

Nowadays, more and more institutions and social networking platforms tend to prevent sensitive information such as rumors, personal information or trade secrets from spreading. In some special cases such as explosion of a new computer virus in the network or political campaign in the social network, the cascading of rumors or virus is not desirable.

Consider the following scenario in our daily life: A has same probability to share information with B and C, who are the only two neighbors of A. It means that A has the same probability to leak sensitive information or privacy to B and C. Suppose that we know in a certain period, A has certain probability to spread sensitive information to its neighbors. At the meantime, A is informed that B is a 'gossiper', which means that B has higher probability to forward information received to more people. Under these facts, one obvious solution for A to protect its privacy from spreading is to 'talk' less to B. Notwithstanding, it doesn't seem practical to limit the diffusion of information transmitted by A. To maintain the entropy A is forwarding, A should 'talk' less to B but 'talk' more to C. We can see that there are some common acknowledgements when it comes to protecting the sensitive information: 1) Limiting the diffusion of information just for protecting sensitive information is not practical. 2) In order to maintain the overall flowing and transmissions in the current network, the structure of the network should be altered as little as possible. In certain scenarios the circumstance around the infected nodes may not be known to all.

In my project, I use a strategy based on probability adjustment of information transmission to achieve this target. I design the algorithms for probability adjustment both in the background of informed network and uninformed network. My work can be separated into three parts. In the first part, I build and analyze the Dynamic Routes Model, turn the problem into a convex optimization problem and design **Algorithm I** for probability adjustment in the informed network. In the second part, I use the Multi-arm bandit to analyze the situation in the uninformed network. And I design **Algorithm II** to find the optimal solution used in the uninformed network. In the last part, I simulate and obtain some experimental results from existing data.

2 System Model

2.1 Dynamic Routes Model

Let $G = (N, E)$ be a connected network with a set of finite nodes $N = \{1,2, \dots, n\}$ and a set of links E. In this paper, we assume that the evolution of the network structure is much slower compared with the speed of information spreading, and thus can be neglected. The nodes have no difference except for sensitive level. A node is either a sensitive node or a normal node. Sensitive information can only be transmitted from a sensitive node to a normal node. In Dynamic Routes Model, information can flow from node i to node j and vice versa as long as there is an undirected path between i and j, denoted by $(i,\ j) \in$ E. I assume that G has no self-loops and no multiple links between any two nodes. Let $|S(t)|$ be the size of the infected node set (or

simply the number of infected nodes) at time t. To keep the notation simple, we will also use $S(t)$ to represent the set of infected notes at time t, $i.e.,\{i \in N|S_i(t) = 1\}$ and $\partial S(t)$ to represent the set of edges originating from $S(t)$ to $N\backslash S(t)$ at time t. We allow that the diffusion starts from a single user ($|S(0)| = 1$) or a connected initial component ($S(0) > 1$). Clearly, all the infected nodes remain connected at any time $t > 0$. Let $N(S(t)) = \{j \in N\backslash S(t) \mid \exists (i,j) \in E, i \in S(t)\}$ be the set of 'neighbors' of the infected nodes at time t, and $\partial(S(t), j) = \{(i,j) \in E \mid i \in S(t), j \in N(S(t))\}$ be the set of edges originating from $S(t)$ to the neighboring node j. For each $r \in \partial S(t)$, it has a parameter β_r, which is the probability of sensitive information on this route. I assume that $\beta_r \ll 1$ in our model because sensitive information diffusion is not frequent.

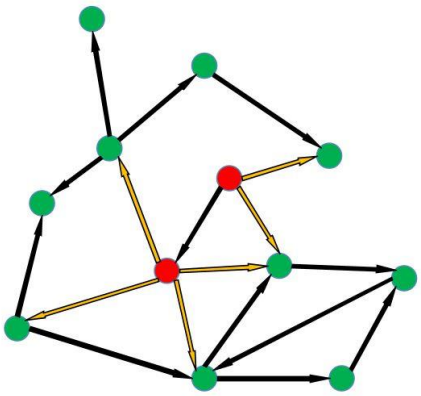

Fig 1. The red nodes form set $S(t)$, the yellow routes form set $\partial S(t)$.

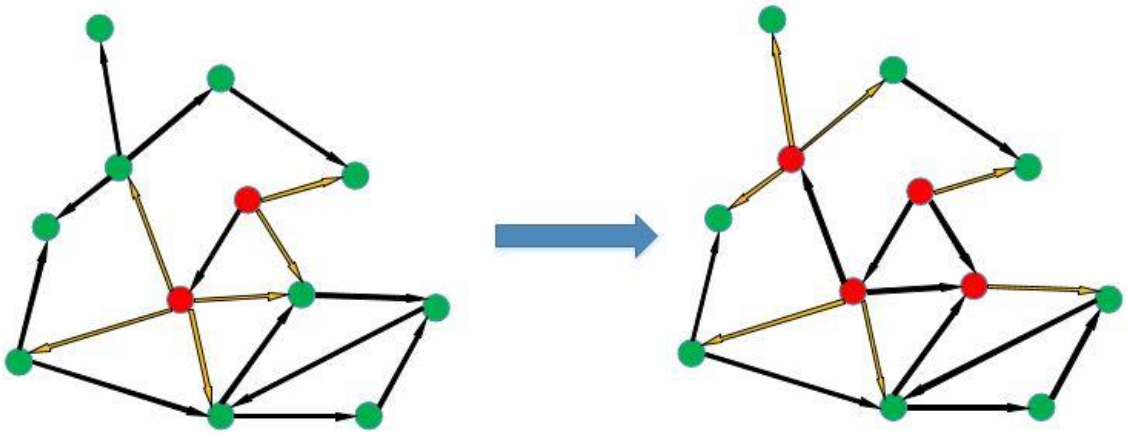

Fig 2. Possible change from $S(t)$ and $\partial S(t)$ to $S(t+1)$ and $\partial S(t+1)$

Normally, sensitive information is time-sensitive. So I assume that γ nodes will change from sensitive node to normal node during each time slot.

2.2 Informed and Uninformed Network

The Informed Network refers to network whose structure is known beforehand. For instance, the network in Fig 1. is an Informed Network. Conversely, we can define network whose structure is ambiguous as Uninformed Network. In this type of network, we can only know certain part of the network. In some cases, when it comes to monitoring the sensitive information, we can only know

the sensitive region, which means we only know which nodes are sensitive ones and their relationships.

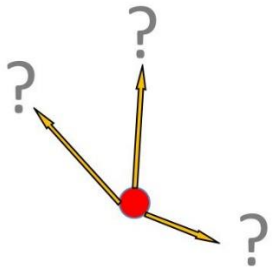

Fig 3. We cannot know what's going on with normal nodes

2.3 Multi-arm Bandit

The multi-armed bandit problem for a gambler is to decide which arm of a *K*-slot machine to pull to maximize his total reward in a series of trials. Many real-world learning and optimization problems can be modeled in this way. We can model the scenario of probability adjustment in uniformed network as multi-arm bandit. In most case, users or rumor protectors are always familiar with the sensitive region, which means they know about the sensitive nodes and the relationship among those nodes. However, the situation outside such region can be ambiguous. If there is a certain period of time many rumors are being spread such as political activity and we want to take control of the spreading of these rumors, we may want to know about the outside as much as possible and leak as little rumors as possible at the main time. Which is to say, within certain steps, we have to maximize the reward, say, minimize the infected nodes as much as possible. The case can be simplified as a multi-arm bandit problem.

We can use the same notation mentioned in 2.1 and model the uninformed network as a multi-arm bandit model. Here set $\partial S(t)$ is the target paths we want to make adjustment to. Let A be the set of 'arms' we can choose from. We assume that the initial transmission probability of each path is β, and the upper bound and lower bound probability of each path are β_{min} and β_{max}, respectively. (The assumption of upper bound and lower bound originates from the previous fact that we have to change the structure of network as little as possible. In the Uninformed Network, however, we cannot do calculation due to lack of data. As a result, an upper bound and a lower bound is provided beforehand.) Each time, I select two elements belonging to $\partial S(t)$ and adjust their probabilities (detailed procedure will be discussed in the later chapter). The reason I choose two elements is because for fixed n, of all the $C_n^k, (k = 2, 3, \ldots, n-1)$, C_n^2 is the smallest, through which I can minimize the size of A.

3 Dynamics analysis

3.1 Informed network

In order to illustrate the impact of probability adjustment on the informed network, I target $|\partial S(t)|$ as analyzing object. From the assumption from 2.1, we can know that at time t, the probability for the destination node of j ($j \in \partial S(t)$) to get involved into the sensitive region is $1-(1-\beta_j)^{|\partial(S(t),j)|}$, we can obtain the expectation of variable quantity in set $\partial S(t)$, which is:

$$\Delta_1 \triangleq E[|\partial S(t+1)| - |\partial S(t)|] = \sum_{j \in \partial S(t)} [1-(1-\beta_j)^{|\partial(S(t),j)|}](d_j - |\partial(S(t),j)|)$$

We can simplify the right side of the equality by Taylor Expansion Formula, and we can get:

$$\Delta_1 = \sum_{j\in\partial S(t)} \beta_j |\partial(S(t),j)|(d_j - |\partial(S(t),j)|)$$

Assuming the network subjects to Power-Law distribution, thus we can rewrite $|\partial(S(t),j)|$ as:

$$|\partial(S(t),j)| = \sum_{i\in S(t)} \frac{d_i d_j}{\sum_1^N d_m}$$

And we can get:

$$\Delta_1 = \sum_{j\in\partial S(t)} \beta_j \sum_{i\in S(t)} \frac{d_i d_j}{\sum_1^N d_m}\left(d_j - \sum_{i\in S(t)} \frac{d_i d_j}{\sum_1^N d_m}\right)$$

Recall from 2.1 that there are γ nodes moving from $\mathrm{S(t)}$ to $N\backslash \mathrm{S(t)}$, we can write this part of change by

$$\Delta_2 = \gamma(\sum_{i\in S(t)} \frac{d_i}{\sum_1^N d_m} - \sum_{i\in N\backslash S(t)} \frac{d_i}{\sum_1^N d_m})D$$

Where D is the average degree in the network. Let $\Delta_1 = \Delta_2$, and we can get the following result:

$$\sum_{j=1}^{|\partial S(t)|} \beta_j d_j^2 = \frac{\sum_{i\in S(t)} \frac{d_i}{\sum_1^N d_m}\left(1 - \sum_{i\in S(t)} \frac{d_i}{\sum_1^N d_m}\right)}{(\sum_{i\in S(t)} \frac{d_i}{\sum_1^N d_m} - \sum_{i\in N\backslash S(t)} \frac{d_i}{\sum_1^N d_m})D}\,\gamma$$

Make $C \triangleq \frac{\sum_{i\in S(t)} \frac{d_i}{\sum_1^N d_m}\left(1 - \sum_{i\in S(t)} \frac{d_i}{\sum_1^N d_m}\right)}{(\sum_{i\in S(t)} \frac{d_i}{\sum_1^N d_m} - \sum_{i\in N\backslash S(t)} \frac{d_i}{\sum_1^N d_m})D}$, which is an bounded constant, the equality can be simplified as $\sum_{j=1}^{|\partial S(t)|} \beta_j d_j^2 = C\gamma$. If γ is increased, say, due to environmental interference. We can rewrite the equality to a differential form:

$$\sum_{j=1}^{|\partial S(t)|} (\beta_j + \Delta\beta_j) d_j^2 = C(\gamma + \Delta\gamma)$$

Which is:

$$\sum_{j=1}^{|\partial S(t)|} \Delta\beta_j d_j^2 = C\Delta\gamma$$

Recall from the previous chapters that we have some limitations of protections. First of all, $\Delta\beta_j$ is obviously bounded. It can be expressed as:

$$\left|\Delta\beta_j - \beta_c\right| \le 0\ , j \in \partial S(t)$$

Where β_c is a constant. The output entropy should stay as a constant, which means that overall $\Delta\beta_j$ should be 0. It can be expressed as:

$$\sum_{j=1}^{|\partial S(t)|} \Delta\beta_j = 0$$

Also, as stated before, we should make change to the network as little as possible. Here I denote change as δ, and $\delta = \sum_{j=1}^{|\partial S(t)|} \Delta\beta_j^2$. Here I don't use ABS function because it is not differentiable.

As a result, I change the problem above into:

Minimize:

$$\delta = \sum_{j=1}^{|\partial S(t)|} \Delta\beta_j^2$$

Subject to:

$$\sum_{j=1}^{|\partial S(t)|} \Delta\beta_j d_j^2 = C\Delta\gamma$$

$$\sum_{j=1}^{|\partial S(t)|} \Delta\beta_j = 0$$

$$\left|\Delta\beta_j - \beta_c\right| \leq 0 \, , j \in \partial S(t)$$

which is a classical convex optimization problem. Due to the number of variables, I use the iteration algorithm to solve this problem. And **Algorithm I** here can be used to solve this problem:

Algorithm I

Initiate $\lambda_0,\ \Delta\beta_0$

$k \leftarrow 0$

$d_0 \leftarrow argmin\{\nabla\delta(\Delta\beta_0)^T d\}$

While

  If $bounded$ or $d_k = 0$

    Break

  Else

    $k \leftarrow k + 1$

    $d_k \leftarrow argmin\{\nabla\delta(\Delta\beta_{k-1})^T d\}$

    $\lambda_k \leftarrow argmin\{\delta(\Delta\beta_{k-1} + \lambda_{k-1} d_{k-1})\}$

    $\Delta\beta_k \leftarrow \Delta\beta_{k-1} + \lambda_k d_k$

End

3.2 Uninformed network

As discussed in 2.3, in given time, our target is to minimize the nodes receiving sensitive information under the background of uninformed network. First we have to initiate set A. This is done by adjusting transmission probability of two routes selected from $\partial S(t)$. Considering of our task is time-sensitive, I combine $\varepsilon - greedy$ and $Upper\ Confidence\ Bound$ strategy and design **Algorithm II** as follows:

Algorithm II

$A \leftarrow \{\}$

$R \leftarrow \mathbf{0}^{|A|}$

$T \leftarrow time$

For $i \in \partial S(t)$

  For $j \in \partial S(t) \backslash \{i\} \cap \{k \mid p(k) < p(i)\}$

If $\frac{p(i)+p(j)}{2} > \frac{\beta_{min}+\beta_{max}}{2}$

$p'(i) \leftarrow \beta_{max}$

$p'(j) \leftarrow p(i)+p(j)-\beta_{max}$

action $\leftarrow \{p'(i), p'(j)\}$

Else

$p'(i) \leftarrow p(i)+p(j)-\beta_{min}$

$p'(j) \leftarrow \beta_{min}$

action $\leftarrow \{p'(i), p'(j)\}$

$A \leftarrow$ action

End

End

For t $\leftarrow$ 1 to T

If $random(0,1) < \varepsilon$

$action \leftarrow random(A)$

Else

$action \leftarrow argmax_a[Q_t(a) + c\sqrt{\frac{\ln(t)}{N_t(a)}}]$

$R \leftarrow bandit(action)$

$Q_{t+1} \leftarrow (1-\alpha)^n Q_1 + \sum_{i=1}^{t} \alpha(1-\alpha)^{t-i} R_t$

End

4 Experiment and Analysis

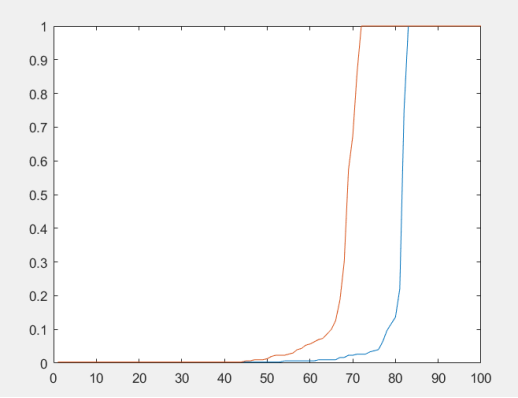

Fig 4. Comparison between using the strategy and not using the strategy

We can see that the main effects of transmission probability adjustment on a given graph is delaying the diffusion of sensitive information.

Results of uninformed network are as follows:

For graph with Power-Law distribution:

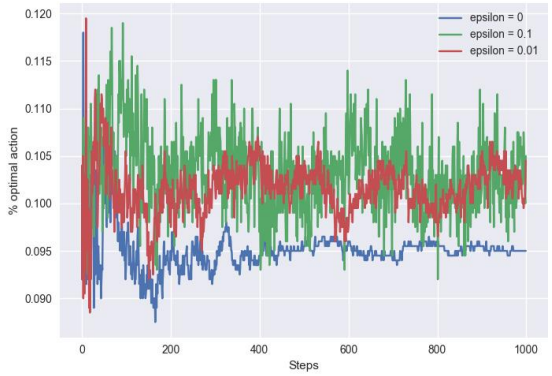

Fig 5. Result of optimal action rate using $\varepsilon - greedy$ method

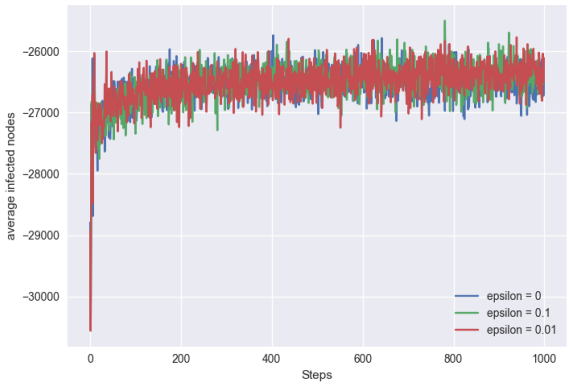

Fig 6. Result of average infected nodes using $\varepsilon - greedy$ method

For graph with uniform distribution:

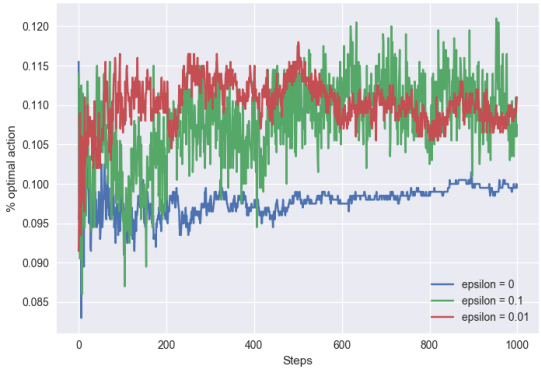

Fig 7. Result of optimal action rate using $\varepsilon - greedy$ method

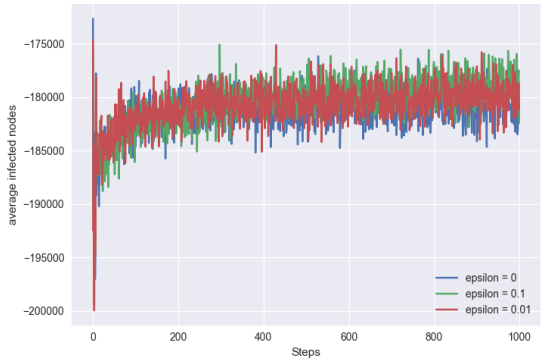

Fig 8. Result of average infected nodes using $\varepsilon - greedy$ method

5 Conclusion

(To be continued…About to finish before week 15.)