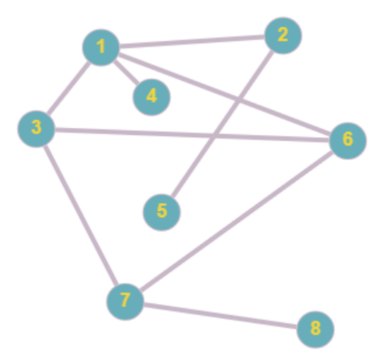
Rumor spread in social networks
STAT 471 Project Group Report
by Shuxin Qiao, Hanbo Gong, Jianfeng Wang

Introduction

Objective
In this project, we investigated the spread of rumors through a modified Voter Model simulation. In particular, we are interested in how using different values of parameters (strength of relations, self-confidence, etc) affects how the process behaves.
By developing the model, we desired to understand how rumors spread over time and space, and possible ways to control them.

An Extension of the Voter Model
While the standard voter model is defined such that nodes are connected to 4 surrounding nodes, we modified the model so that any node may be in relation to any number of nodes. This change models the real world much more accurately because the connections between people are generally not based on neighbors to their geographical N,W,S, and E.

In addition, people may affect the opinions of others to different degrees. For example, professors can affect many students and their impact on others is greater than those who do not participate in social activities frequently. To simulate reality better, our modified voter model is designed to randomly generate a relation matrix based on various parameters. These changes make our model much more versatile than the original voter model with its artificial restriction of 4 neighbors to a node's geographical N,W,S, and E.
The graph to the right is an example of our model with nodes being connected to varying numbers of nodes.



A graph of possible network

Mathematical Framework and Equations

Representation of neighbours

- Matrix Form is used. Row and column order are the same mappings to objects (node) in order.
- Size *n x n* where *n* is the number of population.
- Node (i,j) in the matrix is the weighted influence value of i^{th} person to j^{th} person.

- Diagonal Node (i, i) is the self-confidence value that i[th] person has.
- Note that Node (i,j) generally ≠ Node (j, i) because the degree to which the i-th person affects the j-th person is often different from the degree to which the j-th person affects the i-th person (ie. parents and their children). The relation matrix, therefore, is generally **asymmetric**

$$\begin{pmatrix} 5.8 & 1.3 & 2.9 & 10.1 & 0 & 5.4 & 0 & 0 \\ 4.7 & 7.6 & 0 & 0 & 3.9 & 0 & 0 & 0 \\ 2.9 & 0 & 9.8 & 0 & 0 & 1.2 & 8.2 & 0 \\ 7.7 & 0 & 0 & 10.7 & 0 & 0 & 0 & 0 \\ 0 & 16.4 & 0 & 0 & 9.8 & 0 & 0 & 0 \\ 3.8 & 0 & 14.2 & 0 & 0 & 2.7 & 11.9 & 0 \\ 0 & 0 & 0.5 & 0 & 0 & 6.7 & 3.3 & 5.4 \\ 0 & 0 & 0 & 0 & 0 & 0 & 3.8 & 10.3 \end{pmatrix}$$

Example of a relation matrix

Representation of nodes

Assume the population is *n*. In our model, a person can either believe in the rumour (node = 1) or not believe in the rumour (node = 0).
Define X to be the 2-state {0,1} continuous-time Markov process with some *transition rates* $(\lambda_i)_{0\to 1}$ and $(\lambda_i)_{1\to 0}$. At time $t \geq 0, X_t(i)$ is the state of node i for $i \in \{1, 2, 3, ...n\}$. Then define X_t as follows:

$X_t = \langle X_t(1), X_t(2), ...X_t(n)\rangle^\intercal$ where $X_t(i)$ is either 0 or 1 for $i \in \{1, 2, 3, ...n\}$

Transition Rates

We defined the transition rates $(\lambda_i)_{0\to 1}$ and $(\lambda_i)_{1\to 0}$ as:

$$(\lambda_i)_{0\to 1} = \frac{1}{\sum_{j=1}^n B_{ij}} \sum_{j=1}^n B_{ij} X_t(j)$$
$$(\lambda_i)_{1\to 0} = \frac{1}{\sum_{j=1}^n B_{ij}} \sum_{j=1}^n B_{ij} (1 - X_t(j))$$

Here, the rate is determined by a weighted average of the "neighbors" $Y_t(j)$ values. Note that for the rate at node i, we *need* to use B_{ij} values instead of B_{ji} values because we are concerned with to what degree the nodes that i is in relation to, are influencing node i.
The sum $\sum^n_{\substack{j=1 \\ j:B_{ij}>0}} B_{ij} X_t(j)$ is overall the "neighbors" of node i that are multiplied by their respective weights which we divide by the sum of the weights of the "neighbors" of node I ($\sum_{j=1}^n B_{ij}$)

Generator

The process X solve the *(L,v)-martingale problem* if X_0 has distribution *v* and

$$f(X_t) - f(X_0) - \int_0^t Lf(X_u)du$$

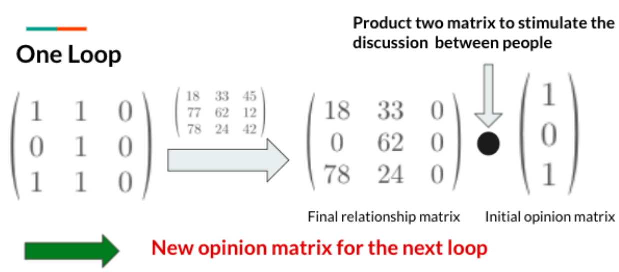
is a martingale for each $f \in D$ with respect to $\sigma(X_s, s \leq t)$

Simulation

Parameters setting

Our simulation can change a series of parameters including :

1. **Population size** - the expected number of people who participate in the spread of the rumor
2. **Loop time** - how many times will the simulation run.
3. **Expected believers size** - the number of believers before information interactions
   The personal opinion about the rumor can be considered as a Bernoulli single trial, therefore the total number of believers in the population can be considered as a Binomial variable. The model will randomly generate *the initial opinion matrix* of nx1 size.
4. **Expected relation size** - the expected number of people any individual can affect.
   The relation between any two people also can be considered as Bernoulli single trial and the total number of people any individual can affect should follow Binomial distribution. After setting this expected relation size, *the basic relationship matrix* can be randomly generated.
5. **The exact value for self-confidence or expected value (uniform distribution)**
   The confidence level can measure how confident a person is, which is highly related to a person's ability to influence others. Confident people will communicate their opinions to others and even try to persuade others, however people who are not confident rarely express their opinions and affect others weakly. As a result, it is not accurate to only use binary relations to describe how people interact with each other. We introduce the confidence level matrix into our model, which is randomly generated following a *uniform distribution*. After inputting the expected value for the uniform distribution, the model will randomly generate each diagonal in the range of [0,2E(x)] to create *the confidence level matrix.*
6. **Binary relation (1/0) or float value relation (uniform distribution)**
   In the previous steps, the basic relationship matrix (Binomial) and the confidence level matrix (uniform) have been generated. We combine them by floating the corresponding number from the confidence level matrix to the replace "1" in the basic relationship matrix and obtain the final
   relationship matrix.



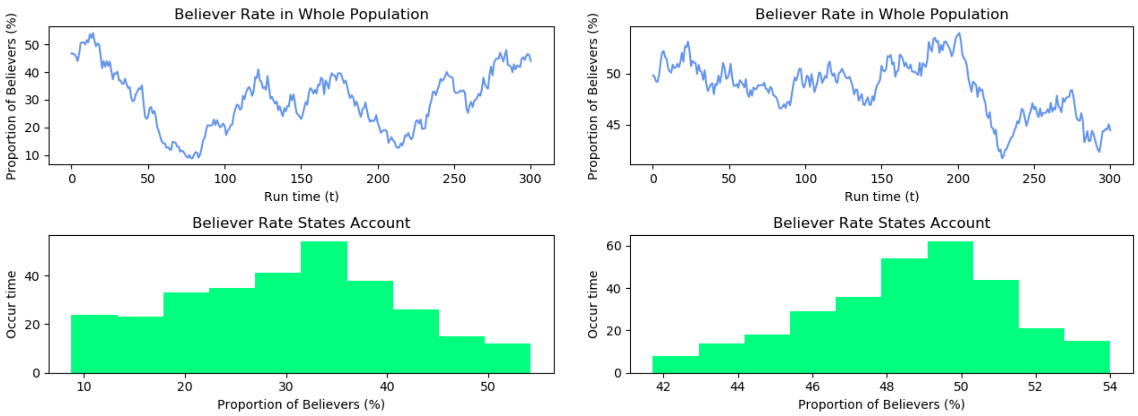
After setting all of the parameters above, we product ***the final relationship matrix*** with ***the initial opinion matrix*** to simulate the one loop of the spread of rumor and then the opinion matrix will be updated, which will be applied to the next interaction. The complete process is shown in Figure.

Running models under different settings

For simplification, use default settings for settings not mentioned.

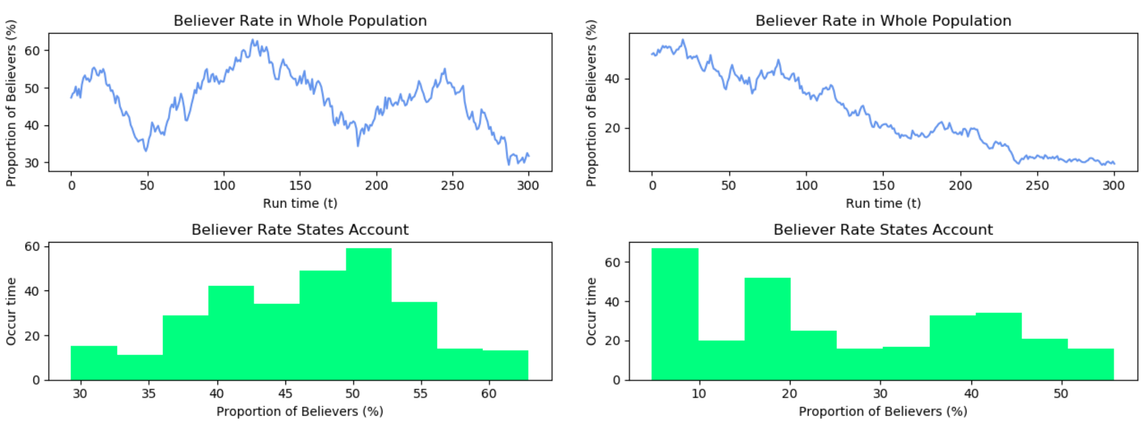
Default setting: Population size: 1000; Relation size: ~half of sample size; Believer size: ~half of sample size; Self-confidence: all 1; Relation value: binary (0/1)

Comparison 1: Population size: 500 VS. 5000 in 300 runs.



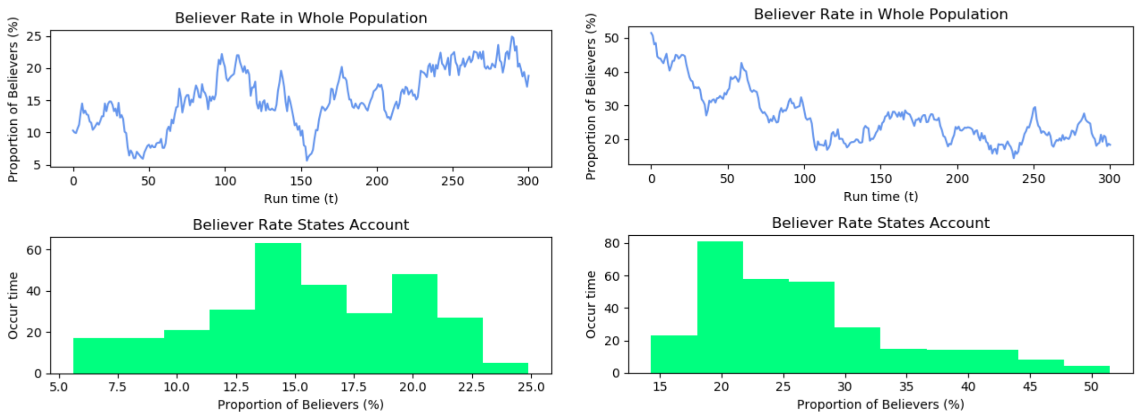
From the histogram X-axis, we can see the range of states simulation goes with 5000 population size is smaller than the left graph. Even the potential curve shows less spread. This could give evidence that the larger population size is more consistent under the same condition or conversely much harder to change.

Comparison 2: Relation size: ~1/10 sample size VS. ~1/2 in 300 runs.



From line plots, we can see the more relations each person has, the more fluctuation the simulation will have. That could give a hint that giving fewer relationships for each one will cause the reduction of income information which can help the convergence to a single state result.

Comparison 3: Believer size: ~1/10 sample size VS. ~1/2 in 300 runs.



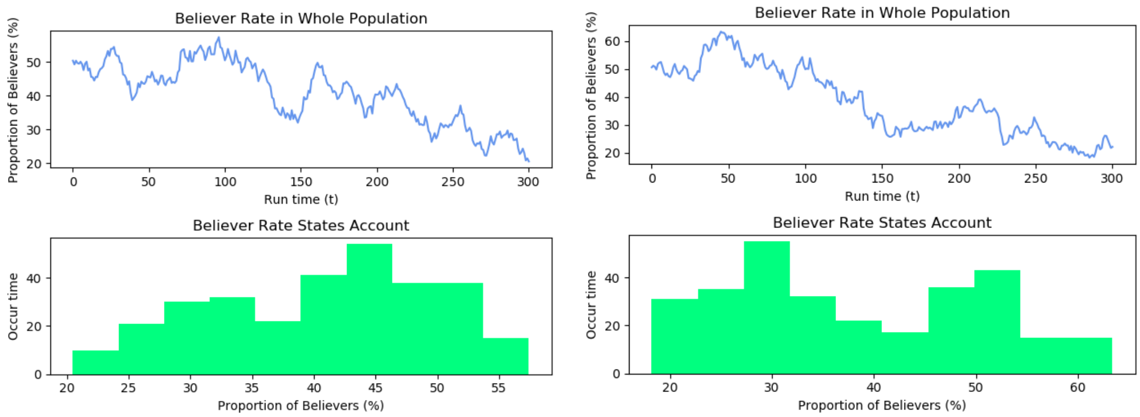
From line plots, the direct conclusion seems to be that the smaller believer percentage has more fluctuation but that is the illusion caused by different scales. According to the histogram X-axis, the larger believer size still has a larger range of fluctuation.

Comparison 4: Self-Confidence: uniform(0,200) VS. all 1 VS. all 100 in 300 runs.

There seems no significant difference between uniform(0,200) self-confidence and all 1's line plot. But there are some differences in their histogram. We need more further investigation for that certain question.

Exact 100 values for all nodes have a quiet straight line converge to result with no believer. This happens once again in the next few simulations. So it is probably not extremely rare to see and that helps us to see that high exact value self-confidence could reduce the fluctuation and speed up the convergence.

Comparison 5: Self-confidence: uniform(0,200); relation value: uniform(0,10) VS. binary in 300 runs.



To further investigate the uniform high self-confidence setting, we set the relation value also to uniform(0,10) and compare it with binary values. The results as shown are still not significantly different, but we could still see the convergence intention for those two simulations. We could also see the fluctuation is more frequent and fast in uniform relation value and even periodic pattern. That could be interpreted as the randomization characteristic of Uniform Distribution.

<u>Conclusion & Application</u>

From Section Simulation, we can conclude that parameters difference can cause different convergence speeds and fluctuation. That helps us to answer the question: what kind of properties of sample groups cause what kind of trends of rumour spread.

Population size is no doubt one of the major properties affecting the spread. Given the limitation of our simulation, only one length spread in a relation network, it takes a long time for one's opinion to spread over the space. Such a latency creates a "social circle" for each individual. Only short distance relations are affected short-distance. This "circle" may divide the population into many subgroups overlapping with each other.

Relation size is another important property. It correlates with the population size mentioned before. It helps to define the "social circle" by editing the number of its links. To some extent, we may think of it as a measure of subgroups' size. It is clear to see that fewer relations make interactions inefficient and that directly slows down the convergence speed.

Believer size is more about the initial distribution of population. It is supposed not to affect the spread since we have given random distributions to reduce the possible effects it may cause. But it can restrict the fluctuation scale in some extreme cases, such as 90% believers at the beginning. Even if we have a strong solution to converge it down to Zero Believers, we may reach the upper limit due to large fluctuations.

Self-confidence and random relation value are key points in our modified model. In comparison 4, we could see that uniformly distributed self-confidence values, even with a high limitation, do not show any significant difference from all 1 values. But a high all 100 values show a very quick convergence. That brought an interesting and a bit weird guess: considering human beings have a very large variance for self-confidence, we could take the middle part as an approximate uniform distribution. If doing so, that could mean the opinion of each individual is controlled by others rather than himself in the view of the whole population.

After having this a bit shocking possibility, we can intuitively think the overall event (like official statements) could bring a significant effect. This also takes our model into a more wide space. Except for the rumors in North Campus, we can apply this model to an even larger population like the city of Edmonton or the Province of Alberta. Rather than just rumors, this could be extended to other abstract ideas.

Further Thinking & Potential Improvement

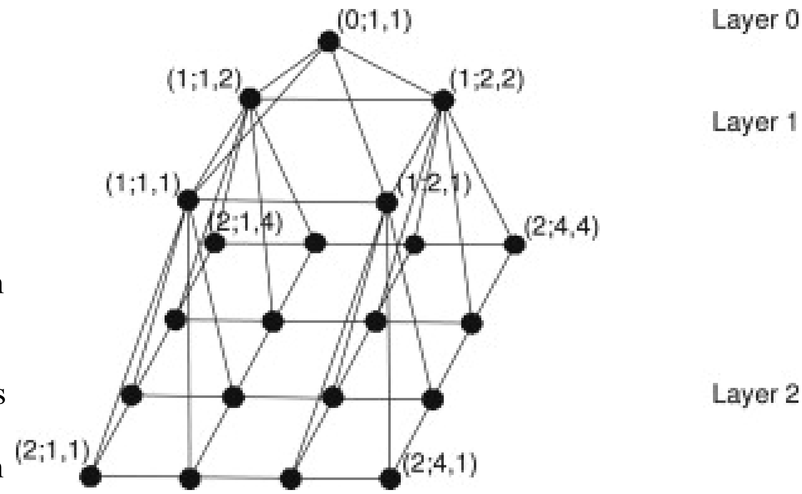
1. **Potential improvement for the confidence level matrix**

   Since the personal confidence level is different, instead of a binary relationship, our model is designed to randomly generate the confidence level matrix simply under uniform distribution with a range of [0, 2E(x)] to represent each person has individual confidence level, which may be strongly related to their impact on others. As a result, to generate a relatively accurate confidence level matrix, the value of E(x) we input is important. It should be admitted that confidence level is closely related to personality, which means that if possible, we should analyze the personality of a sample from the target population before setting the value of the expected confidence level. Such analysis can be conducted by doing a survey or questionnaire to score individual confidence levels in the sample and then their scores can be summarized to get the mean confidence score to estimate the average confidence level for the whole population. We can input this mean value as the parameter in our model to randomly generate the confidence level matrix for the target population, which may be better fitting the real situation.

2. **How does *Hidden Network Structure* affect the spread of rumors in the population?**

   In our current model, we simply assume the relationship between two people is Bernoulli single trial and the capability of an individual to affect others follows Binomial distribution. However, in the real social network, the influence between people is much more complicated and the following conditions should be considered:
   - The influence between people is not equivalent so the relationship matrix is not bound to be symmetric.
   - The influence is directional and **some strong effects only exist in one direction.** For example, in real life, the ruling class has an absolute right to express their opinions, which can impose an extremely strong impact on their advocates. Such hidden structure of the social network (Pyramidal) can change the “regular” way

of the rumor spread because if the authority declares their opinions, the percentage of believers will suddenly change, even rapidly approaching 1 or 0.



As shown in the figure, the influence exists strongly in one-way from up to down layer by layer, but the opinions from lower layers will rarely affect the upper state.
Thus, for our current model, it is hard to simulate the spread of rumors in a population with such a hidden structure. However, it may be achieved by modifying the basic relationship matrix. We can float certain raws or columns with all zero (no effect from lower layer) and all one (absolute effect from up to down) to make the randomly generated matrix also fit such specific situations.

3. **Is the algorithm a Markov Chain?**

Since the opinion matrix will be updated after every loop, the next state of the spread of rumor only depends on the result of this loop, which means it can be considered as a Markov Chain. However, it is hard to identify the exact state after each interaction and more simulations should be conducted to identify the characteristics for such Markov chain.

4. **Simulations Vs. Real-life**

Our investigation also reminds us that it is really important to have **the ability to think independently anytime and anywhere**. As our simulations show if people can be confident about their opinions and struggle with others during the discussion, the persistent fluctuations can be observed in the percentage of believers, which means that it is more difficult to make all of the people believe the rumor. And independent thinking will give us more confidence in personal opinions.
People indeed get into depression even hurt because of the spread of rumors. So, we should think critically and dependently about the information we received and carefully express our own opinions in communications rather than easily believing everything from others and passing them on. The crazy spread of rumor also can kill people just like viruses. This is also the importance of education. Education helps people to be able to distinguish rumors according to their knowledge and also taught people to express personal opinions in the right way to snuggle against those rumors.