CSCI4190 Project Report

2018 – 2019 Spring term

LIU, Muzi 1155077104
ZHU, Leyan 1155077028

5/4/19

CSCI 4190

Table of Contents

Abstract

In this report, we will present the works we have completed for the course project of course "CSCI4190 Social Networks". Specifically, we simulate several typical epidemic models, and then discuss about the influences of the variable factors on how widespread the epidemic is. Based on the simulation results we obtain, we show some conclusion, as well as our understandings of the epidemic models.

The report is organized as the following ways: In section 1, we specify the data and the methodologies we use to conduct simulations. In section 2, we introduce the epidemic models we simulated, the simulation settings we design, and present simulation results for each model. The analysis of the results and further discuss about the influencing factors of epidemic models are also included in this section. Finally, we summarize the whole report and declare our task allocation in the last two sections.

1 Task specification

1.1 Data statistics and analyzing tools

The data set that we build simulations on top of is a who-trust-whom online social network of a general consumer review site Epinions.com, whose key features are summarized in Table 1: Data statistics [1]:

Nodes	75879
Edges	508837
Nodes in largest WCC	75877 (1.000)
Edges in largest WCC	508836 (1.000)
Nodes in largest SCC	32223 (0.425)
Edges in largest SCC	443506 (0.872)
Average clustering coefficient	0.1378
Number of triangles	1624481
Fraction of closed triangles	0.0229
Diameter (longest shortest path)	14
90-percentile effective diameter	5

Table 1: Data statistics

The data is interpreted as a directed graph, where the influences from one node to another are also directed. For simplicity, we refer node n_i as the *neighbor* of node n_j if there is an edge from n_j to n_i, which does not necessarily imply that n_j is a neighbor of n_i.

The tool we use to program and analysis is **S**tanford **N**etwork **A**nalysis **P**latform (**SNAP**), which is a general-purpose network analysis and graph mining library. It is also the library recommended by TAs in tutorials [2].

1.2 Task selection

The task set that we select is the fourth one: “Simulate epidemics”. We simulate SIR, SIS and SIRS models as required, and analyze the effect of network structure from the facet of initial adopters’ selection.

1.3 Methodologies

To reasonably design the simulation, we mainly use the method of variable control to limit the number of changed factors and thus focus on the effects made by one factor only. The default settings in simulations are introduced in Section Simulation Results and Analysis.

2 Simulation Results and Analysis

In this project, we simulate SIR, SIS and SIRS models respectively. Although they are different models, they are variations of the same basic idea, and thus have some similar factors. To control the variables for later analysis, we set the default values for each factor here. In later report, all the factors that are not assigned to a value specifically can be assumed as the default values.

In all the models, the default values are as follows: contagion probability $p = 0.2$; size of initial adopter $= 10$ and the initial adopters are randomly selected; time interval of remaining in infectious state $t_l = 2$; time interval of remaining in remove state $t_r = 2$.

2.1 SIR Model

In SIR model, each node has three states: susceptible (S), infectious (I), and removed (R). Initially, all the nodes are in the susceptible (S) state, except for the initial adopters. The initial adopters are in the infectious (I) state. Then in each time step, the nodes in I-state will turn their neighbors with S-state into I-state with a contagion probability p. I-state nodes only persist for t_l time steps. After t_l steps, those nodes enter the removed (R) state and does not affect other nodes any more.

From the simulation, we abstract some graphs to describe and disclose some patterns of the development of epidemics.

2.1.1 Evolution of the States of Nodes

The below Figure 1: Evolution of the states of nodes describes the evolution of the number of nodes in each state in a typical simulation. The horizontal axis is the time step, and the vertical axis is the number of nodes. Each vertical bar represents a capture of the network in a time step, and the lengths of sub-bars with different colors are the number of nodes with the corresponding state. Thus, the total length of each bar is the total number of nodes in the network, which is 75,879.

From Figure 1, we can observe that the susceptible nodes decrease in number continuously, while the number of removed nodes increases continuously. The number of nodes in I-state increases first and decreases to zero at the end. Finally, a part of the nodes are in removed state and another part of them are in susceptible state. The whole network is in a Nash Equilibrium and does not change.

However, besides the general evolution pattern we observe, we also want to know that whether the epidemics will be affected if we tune the factors.

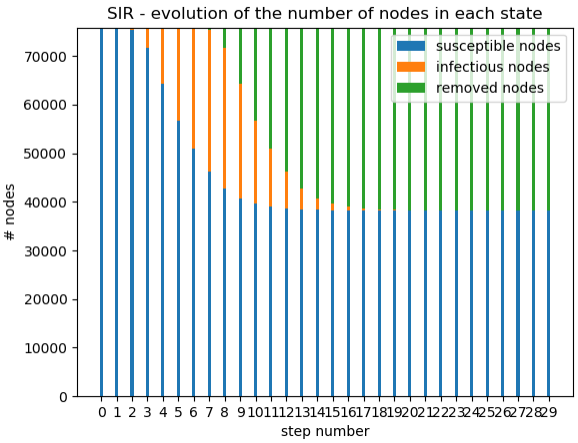


Figure 1: Evolution of the states of nodes

2.1.2 Contagion Probability vs. Disease Spreading

To find the relations between contagion probability and the spreading of epidemic, we plot the time step it takes to reach the end of the simulation, where all the nodes are either susceptible or removed; and also plot the number of nodes that are finally removed with the change of contagion probability.

Figure 2 shows the result we obtained from simulations. The horizontal axis is the contagion probability for each simulation. The yellow bars represent the number of nodes removed in the last time step, and the blue line shows the change of required time step to reach Nash equilibrium, where there is no infectious node. We can see a stable increase of removed nodes as we enlarge the contagion probability p. The pattern of the number of removed nodes is expected, as the larger the p is, the more nodes are infected in each time step and finally be removed.

However, the results of required time step show a more complex situation. We hypothesis that the pattern in the figure is caused by the different effect of p indifferent situations. For example, when $p < 0.4$, the infection is effectively accelerated with p increases and thus speeds up the whole process; while when $0.4 < p < 0.6$, the enlarge of the number of infected nodes slowdown the process, and overwhelms the effect of infection acceleration; but when $p > 0.6$, the acceleration takes advantage again and thus causes the second decrease of required time step.

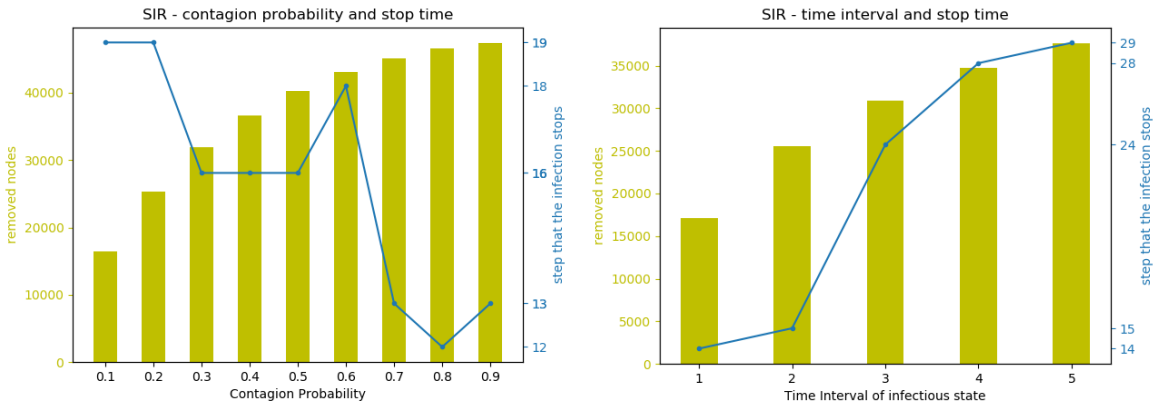


Figure 2: required time step and the number of removed nodes with different contagion probability

2.1.3 Time Interval t_l vs. Disease Spreading

Despite of the contagion probability, another factor that we are interested in is the time interval t_l of a node staying in the I-state. Following Figure 3 summarizes what we find in the simulations. Similar to Figure 2, we plot the final removed nodes and the ending time steps against t_l. Both the number of removed nodes and the ending time step show a clear increasing pattern. It is natural to infer that with the growth of t_l, the time for an infectious node being removed is postponed, thus it takes more time for the network to enter the equilibrium, where all the infectious nodes should be converted to removed nodes. Also, the longer the nodes remain infectious, the more chances their neighbors getting infected as well. That explains why the number of removed nodes increases monotonously.

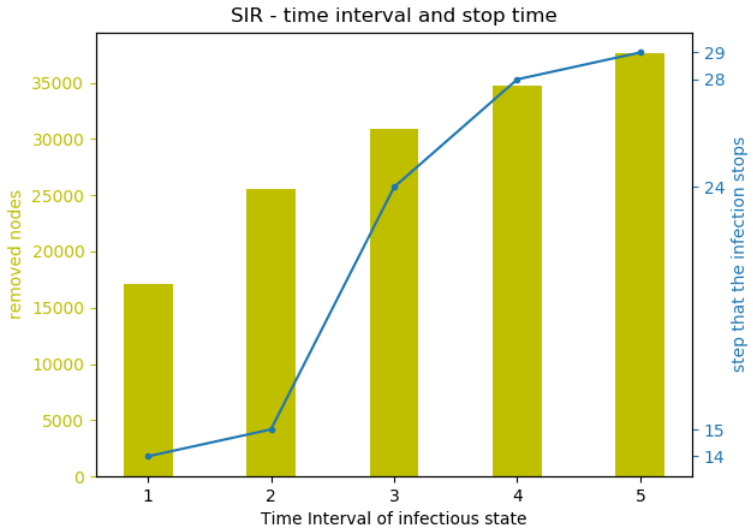


Figure 3: required time step and the number of removed nodes with different time intervals

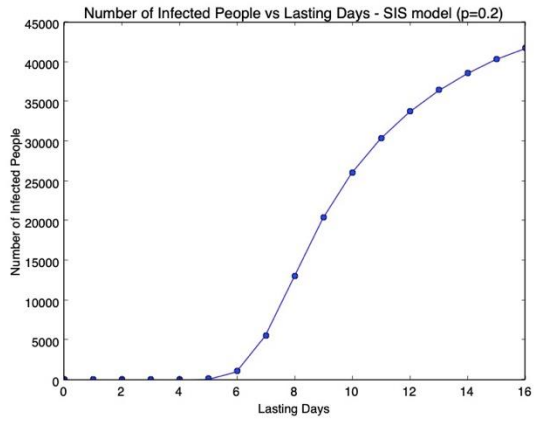
2.2 SIS Model

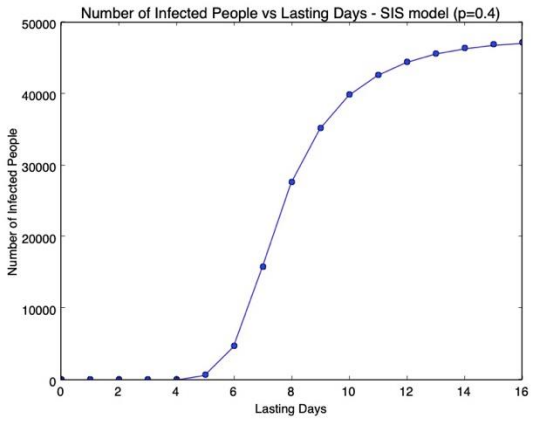
In this section, we look into the SIS model performed on the network structure [3]. For SIS epidemic model, every node only has two states, different from SIR model, namely Susceptible and Infectious. It does not have any removed state, so the node cycling back to the Susceptible state is ready to catch the disease again. The mechanism is very similar to that of SIR model. If we take a snapshot of the whole network at any time, we can see that some nodes are in the I(Infectious) state and the others are in the S(Susceptible) state. Those in I state remain infectious for t_l steps. After t_l steps, it is no longer infectious, and it returns to the S state. For those in S state, each of them has a probability p of passing the disease to each of its susceptible neighbors. We seek to evaluate the spread of disease based on different parameters, such as the probability p to pass the disease and the time t_l that a node remains infectious.

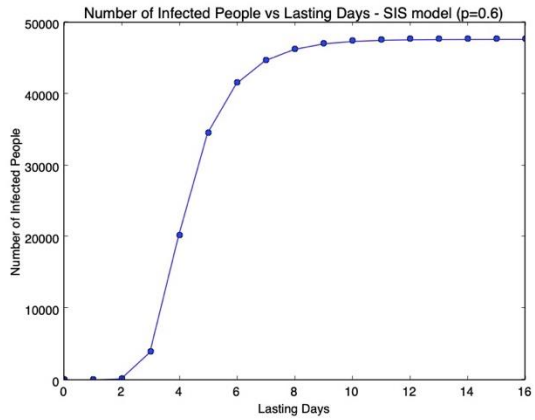
2.2.1 Number of Infected People vs. Lasting Days of the Disease

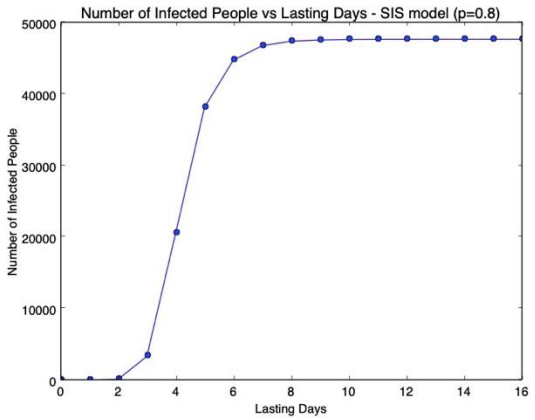
First, we evaluated the number of people that have once been infectious before according to the lasting days of the disease. According to the definition, intuitively, the number must keep increasing till a limit where the disease stops passing (in other words, all people in the network are in the susceptible state). Since SIS model can run for an extremely long time as it cycles through the nodes

potentially multiple times. For example, for a node v, it first passes the disease to its neighbor node u, then it returns back to the susceptible state. Then for some time later, it is probable that its neighbor u transmits the disease back to v, so on and so forth. In this scenario, the disease can hardly stops spreading. Therefore, we set the lasting days to range from 1 to 16 and analyzed the spread speed of disease based on SIS model. The results are illustrated in the following diagrams.





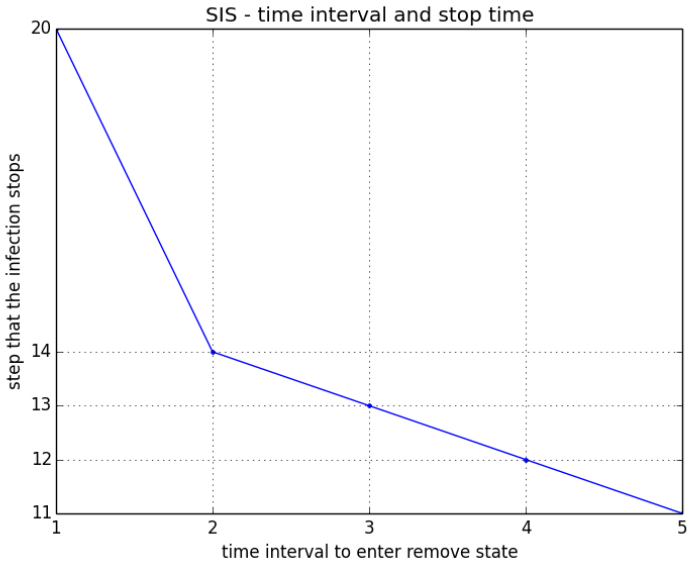




We set different probability p (i.e. 0.2, 0.4, 0.6, 0.8 respectively) to analyze the difference. It is shown that larger contagion probability p results in faster transmission of the disease. For probability that is relatively small (0.2 and 0.4), the disease starts to spread quickly at the 6th day, while for probability that is relatively large (0.6 and 0.8), the disease starts to spread quickly at only the 2nd or 3rd day. And we can see that, same as our intuition, all these four graphs increase dramatically and quickly, showing that SIS model spreads fast with only small probability like 0.2. Another finding from the graphs is that we can see that they all converge to some number between 40000 and 50000. In this interval, it starts to spread rather slowly. We think that it is probably because these nodes are in strongly connected components so that the disease will mainly spread within this SCC, and for nodes who are more isolated or weekly connected to other nodes, it is more difficult for the disease to be passed to them. More details will be demonstrated in the Section 2.4 to analyze the relationship between SCC and the spread of disease.

2.2.2 Complete Days vs. Infectious Interval

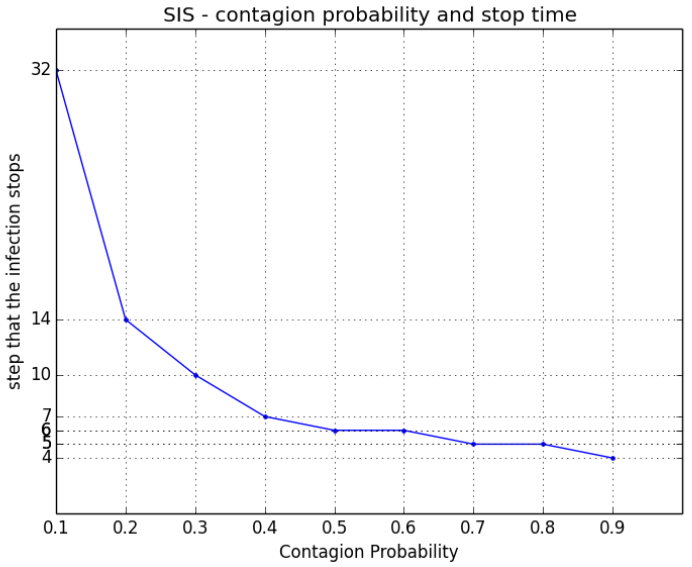
We set the complete condition to be the number of people that have experienced the disease reaching half of the total number of nodes in the network. Then we looked into the number of complete days with respect to the infectious interval. We set the infectious interval to range from 1 to 5 and set the contagion probability to be 0.2.



The diagram shows that the complete days decrease with the infectious interval. Since the longer the infectious interval is, the larger the opportunity is for a node to transmit the disease to its neighbors. Therefore, longer infectious interval results in the faster spread of the disease.

2.2.3 Complete Days vs. Contagion Probability

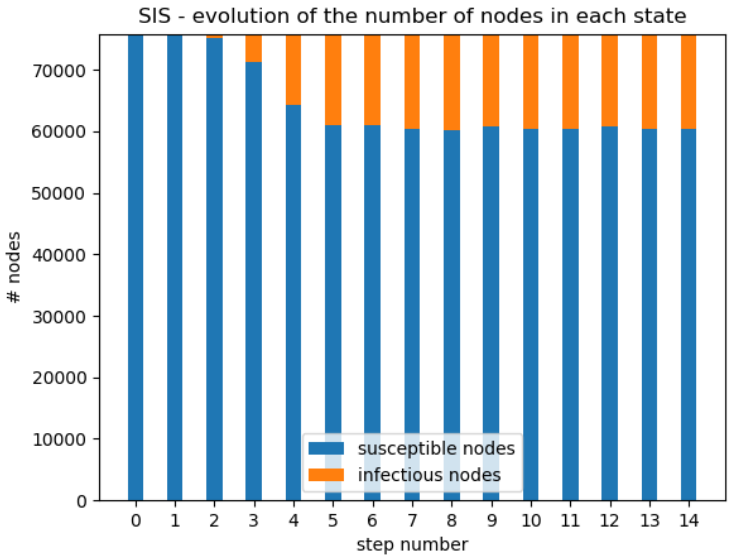
The complete condition is the same as that in Section 2.2.2. We looked into the number of complete days with respect to the contagion probability. We set the contagion probability to range from 0.1 to 0.9 with the step of 0.1 and set the infectious interval to be 2.



The diagram shows that the complete days decrease with the contagion probability. Similar to Section 2.2.2, larger contagion probability gives a node more chances to spread the disease to its neighbors, thus resulting in faster transmission of the disease. And we can see that from probability 0.1 to 0.2, there is a dramatic drop of the number of complete days. It shows that for SIS model, a little larger contagion probability can cause a huge increase of speed to the spread of disease.

2.2.4 Evolution of the State of Nodes

The complete condition is the same as that in Section 2.2.2. We looked into the state of all nodes in the network with time goes on. We set the contagion probability to be 0.2 and the infectious interval to be 2.



This diagram shows that after spreading the disease to more people for several initial steps, the number of people in S/I state almost remains the same. It implies that In SIS model, the spread of disease can hardly stop since the number of people who are in infectious state remains to be that large, which is caused by the disease keeps cycling back to people from their neighbors.

2.2.5 Critical Value of Contagion Probability

In SIS model, there is a special property that it has a particular critical value of the contagion probability p (depending on the network structure). When the contagion probability get across this value from a smaller value to a larger value, the network will shift from one that dies out quickly to one that persists for a very long time.

In this section, we tried to find the critical value for this particular network structure. Since if the disease persists for a long time, the program will keep running, so we set the threshold to be 50. That is, when the disease still persists after 50 rounds, we let it exit the loop and consider it as the situation that the disease persists for a long time. And we run 10 times for each probability to avoid occasional cases.

This table illustrates the corresponding time for the disease to stop, question mark in the table means that it runs over 50 rounds.

p	1	2	3	4	5	6	7	8	9	10
0.01	1	2	6	1	1	1	1	7	3	2
0.02	?	?	?	2	1	?	1	?	?	1
0.03	4	?	?	1	2	?	?	?	1	?

We can see that when probability is 0.01, for all 10 times the disease terminates fast in steps less than 10. And when probability is a bit larger, (0.02 and 0.03), it stops fast in only around a half of altogether 10 times. So we guess that the critical value is somewhere between 0.01 and 0.02. And to get more accurate, we also ran the test for probability = 0.015.

p	1	2	3	4	5	6	7	8	9	10
0.015	?	1	1	1	1	1	1	1	1	1

It clearly shows that the disease terminates fast in most of the test. And sometimes it may still spread for a long time due to the network structure. For example, the initial adopters have many neighbors so that it can transmit down the disease. And the details to analyze the effect of initial adopters will be illustrated in Section 2.4.2.

2.3 SIRS Model

The SIRS model is a combination of the above-mentioned SIR and SIS models. In SIRS model, a node experiences four states iteratively: susceptible (S) state – infectious (I) state – removed (R) state – susceptible (S) state. In R-state, the node will not be infected by other nodes. R-state persists for another time interval t_r.

2.3.1 Evolution of the States of Nodes

For SIRS model, we plot the evolution of the proportion of nodes in each state as well. Figure 4 shows the pattern of a simulation for 50 timesteps. We can observe that the network oscillates back and forth since timestep 10. During 10 to 50 timesteps, the number of infectious nodes basically remain unchanged, while the number of susceptible nodes and removed nodes are oscillating. The oscillation may be a result of the model mechanism: the amount of newly infected nodes is close to the amount of newly removed nodes, thus the susceptible nodes decreases, and the removed nodes increases in size.

Similarly, we also analyze the influence of several factors in SIRS and try to get a deeper insight of SIRS model.

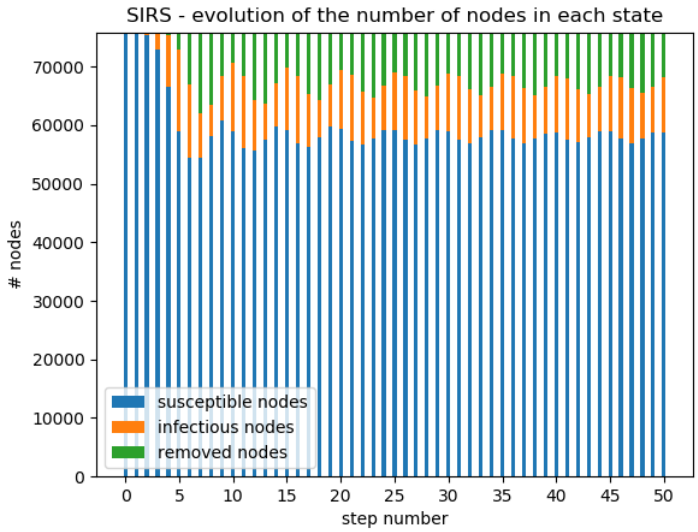


Figure 4: Node evolution

2.3.2 Contagion Probability p vs. Disease Spreading

As in SIRS mode, the network will never reach an equilibrium, it is not appropriate to measure the level of spread speed by the number of timesteps that is required to reach an end situation; or to measure the level of widespread by the number of removed nodes at the end. Instead, we simulate for 50 timesteps, which is enough for the network to reach the oscillation and take an average of the proportion of infectious nodes in the last 30 timesteps. Figure 5 shows the changes of the average proportion of infectious nodes with growing contagion probability.

Surprisingly, for different p, the proportion of infectious nodes remains unchanged. Note that in Figure 5, the maximum proportion is 0.118, while the minimum is 0.117, which only variates in 0.001. As the variations are small enough to omit, we may infer that the contagion probability p does not have much effect to the level of widespread for epidemics in SIRS models.

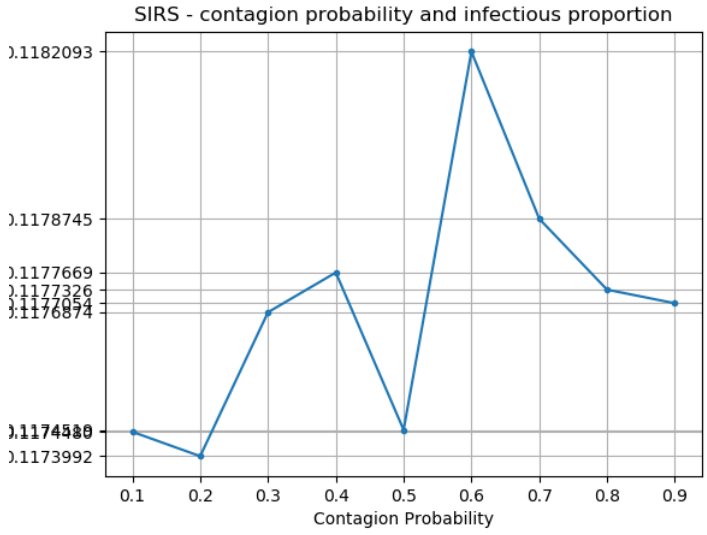


Figure 5: contagion probability vs. proportion of infected nodes

As the contagion probability has no effect to the level of widespread of epidemics, a question that we naturally arise is that what kind of factors will affect it then? Another important adjustable factor in SIRS model is the time intervals t_l and t_r. The analysis of these two factors is presented below.

2.3.3 Time Interval t_l vs. Disease Spreading

Figure 6 presents how the average proportion of infectious nodes changes with the time interval of nodes staying in S-state. It clearly shows a linear relationship between the two factors. As the infectious nodes stay in I-state for a long time, the contact between them to susceptible nodes will also increase. Thus for each infectious node, they are able to infect more neighbors, causing a larger proportion of infected nodes in the scale of the whole network.

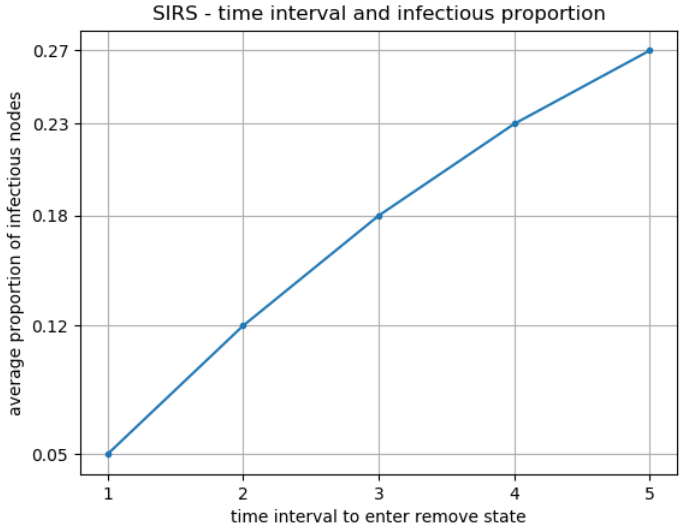


Figure 6: time interval between I-state and R-state vs. proportion of infected nodes

2.3.4 Time Interval t_r vs. Disease Spreading

Although the changes of the proportion of infectious nodes and t_l are in direct proportion, it is in inverse proportion with t_r as Figure 7 shows. We infer that the reason of this inverse proportional pattern is because the nodes in R-state will not be infected, and thus when t_r becomes larger, it becomes harder to infect the node in long term.

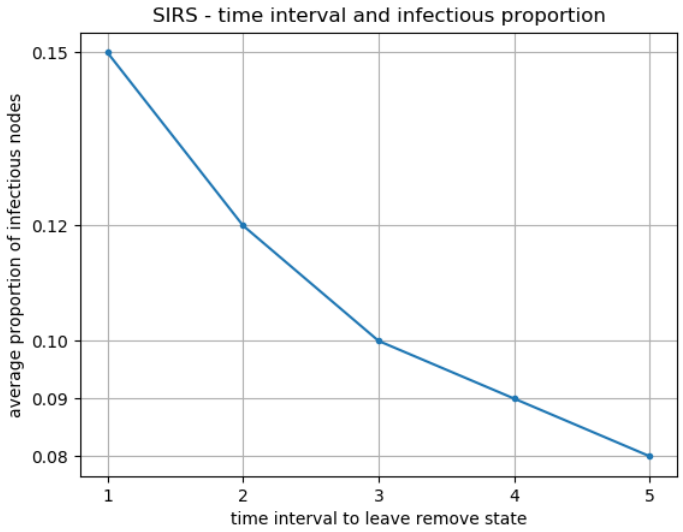


Figure 7: time interval between R-state and S-state vs. proportion of infected nodes

Compared with contagion probability p, both t_l and t_r are influential to the infected node proportion, as the proportion varies largely with the time intervals.

2.4 Epidemics and Network Structure

In this section, we will analyze the performance of epidemics with respect to different network structures such as the average out degree of nodes and the strongly connected components in the network.

2.4.1 Basic Reproductive number R_0

There are two possibilities for a disease in the branching process model. One is the disease reaches a wave where it infects no one, thus dying out after a finite number of steps. The other one is the diseases continues to infect people in every wave, proceeding infinitely through the contact network.

From what was introduced in lectures of epidemics, we know that an important factor to determine whether the disease can die out quickly or persist infinitely is the basic reproductive number.

Specifically, the basic reproductive number is defined as:

$$R_0 = k \cdot p$$

, where k is the number of new people everyone meets, p is the contagion probability.

So that the basic reproductive number is just the expected number of new cases of the disease caused by a single individual.

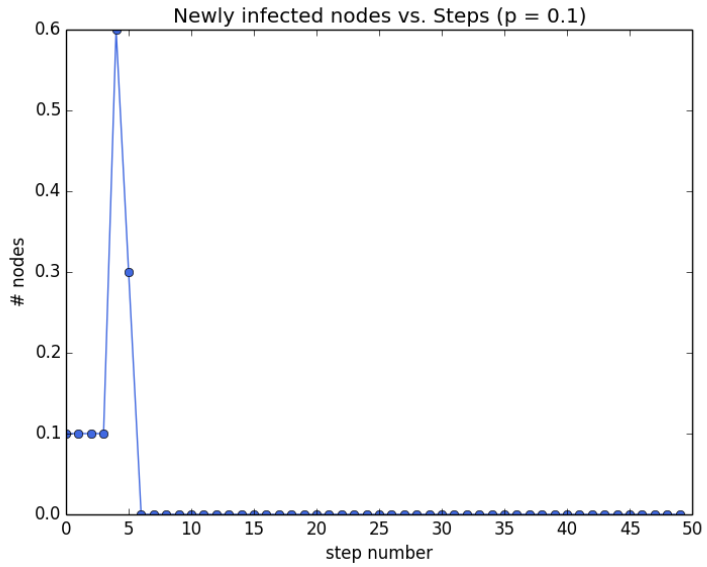
It is believed that if $R_0 < 1$, then with probability 1, the disease dies out after a finite number of waves. And if $R_0 > 1$, then with probability greater than 0, the disease persists by infecting at least one person in each wave. In this section, we tried to verify the statement by experiment.

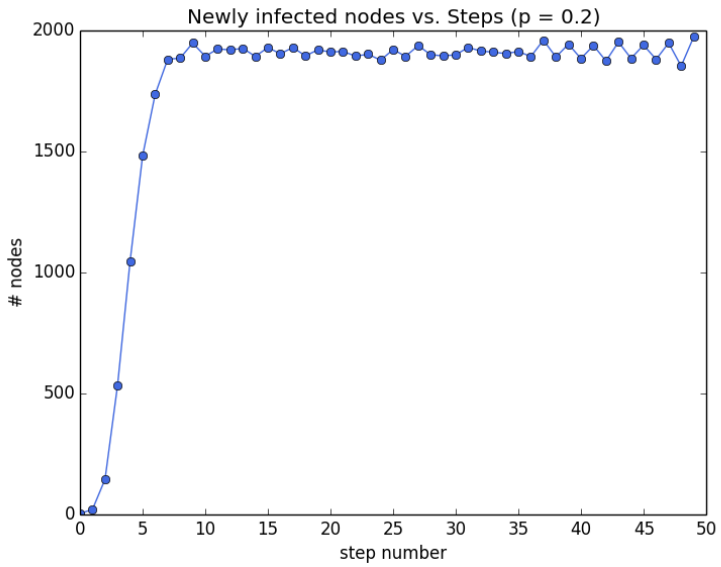
2.4.1.1 Average degree k

In the specific dataset we are using, the parameter k is actually the average out-degree of each node. We have calculated the average out-degree and it's around 6.7.

2.4.1.2 Epidemics and R_0

Since for the SIR model, no matter what contagion probability we choose, it will finally have no newly infected nodes since if a node has entered the removed state, then it is immune to the disease, i.e. it will never be infected again. Therefore, we conducted the experiment based on SIS model. Since the average out-degree of the network is around 6.7, in order to look into the situation when $R_0 > 1$ and $R_0 < 1$, we selected the contagion probability 0.1 and 0.2 respectively to satisfy these two conditions. And we looked into the number of newly infected nodes in each round. And we totally run the program for 10 times, each with 50 rounds. Finally, we take the average of results of these 10 experiments to generate a diagram. Also, we set the infectious interval = 1 and the number of initial adopters = 1 to simulate the branching process.





We can see that when p = 0.1, $R_0 < 1$, the corresponding number of newly infected people quickly decreases to 0, which means the disease stops spreading. While when p = 0.2, $R_0 > 1$, the corresponding number of newly infected people first keeps increasing and then it almost remains the same by oscillating in a small range of values, which means the disease will persist for a really long time.

2.4.2 Selection of initial adopters

In the above simulations all the initial adopters are randomly selected from the whole network. However, during the simulations, we observe that the results may vary even if we simulate the same model with the same settings. We hypothesis that the random selection of initial adopters is the reason of this variation. Thus, we conduct further simulations to figure out the effect of the topological features of the initial adopter.

2.4.2.1 Set size of initial adopters

One naïve factor is the number of the initial adopters. We conduct simulations on top of the SIR model for simplicity, as the it enters an equilibrium at the end.

Figure 8 shows the relation between the number of initial adopters and the required step to reach equilibrium; as well as the relation between the number of initial adopters and the number of nodes in R-state in equilibrium. The figure does not . We find that

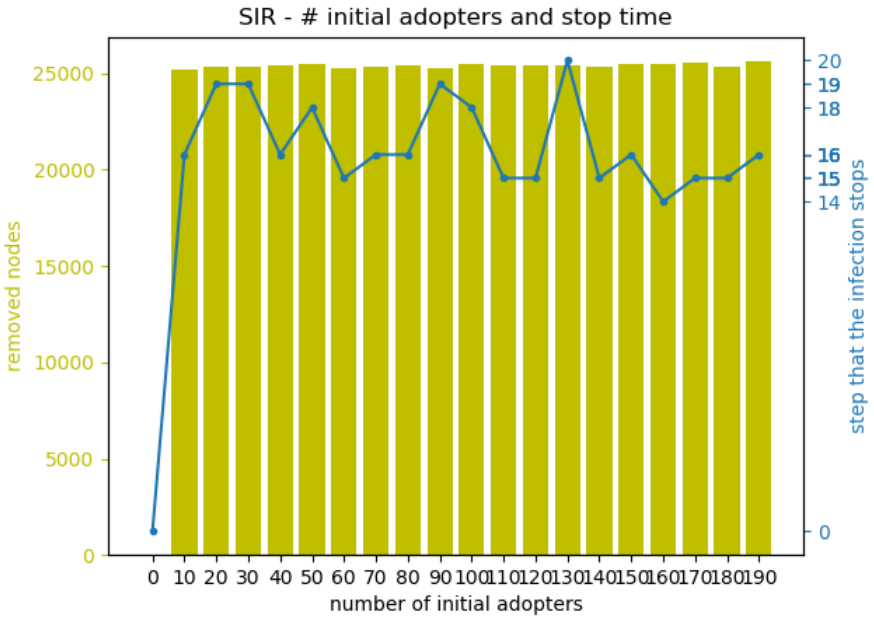


Figure 8: influence of the number of initial adopters

2.4.2.2 According to node degree

Another possible factor of the set of initial nodes that may affect the spreading process is the out-degree of the initial adopters. We find the 10 nodes with the largest degree in the network and the other 10 nodes with the smallest degree in the network, and conduct simulations where the initial

adopters are the two sets respectively. The two figures below are the simulations where the initial adopters are the set of 10 nodes with largest degree (left subgraph) and with smallest degree (right subgraph) respectively.

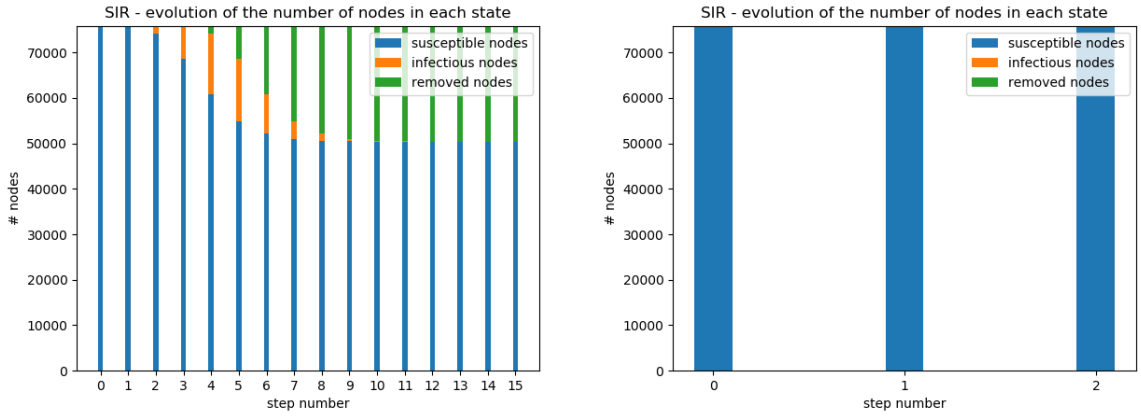


Figure 9: evolution of simulations with largest- (left) and smallest- (right) degree initial adopters

It's clear that the simulation with the adopters with large degree has more nodes infected, while the simulation with the adopters with small degree has only a small number of nodes infected, which is almost unobservable in the figure. The results indicate that the nodes with larger degree are more influential, which conforms to our common sense.

However, if we compare Figure 9 with Figure 1, we observe that the randomly selected initial adopters can actually infect more nodes than the nodes with largest degrees, as the number of removed nodes occupies almost half of the population, while the largest-degree initial adopters only cause one third of the population removed at the end. This may because the largest nodes are highly clustered in real world data, thus they are in the same strongly connected component. The largest-degree initial adopters may have a large number of common neighbors, which limits the size of the union of their neighbor sets and thus limits the total number of nodes that initial adopters have contact with.

3 Conclusion

For SIR model, higher contagion probability will result in more people infected during the spread period of the disease. And higher contagion probability as well as longer infectious interval will cause faster speed of disease transmission. Also, it is difficult for it to infect all people except that the contagion probability is 1.

For SIS model, same as SIR, higher contagion probability will result in more people infected during the spread period of the disease. And higher contagion probability as well as longer infectious interval will cause faster spread of the disease. When the contagion probability is larger than some value, the number of infectious nodes in each round will almost remain the same, making the disease persist for a long time. And there exists a critical contagion probability value for SIS model depending on different network structures. When contagion probability is smaller than this particular value, the disease dies out quickly. While when contagion probability is larger than this value, the disease persists infinitely.

For SIRS model, as time goes on, the number of infectious nodes almost remain unchanged while the number of susceptible and removed nodes are oscillating. And the contagion probability has little influence on the spread of disease. But for infectious interval and removed interval, the first has positive correlation with the spread of disease while the second has negative correlation with the spread of disease.

Regarding the influence of network structure on the spread of epidemics, basic reproductive number R_0 plays an important role. When $R_0 < 1$, the disease dies out in the end. When $R_0 > 1$, the disease persists infinitely. So, in real world, if we want to control a disease, we can reduce R_0 by either quarantining people (reduce quantity k) or by encouraging good living habits to reduce the probability p.

In terms of initial adopters of the disease, in general, initial adopters with large degree has more nodes infected while initial adopters with small degree has only a small number of nodes infected.

4 Task Allocation

Contribution of each member:

LIU Muzi:

- Simulations and analysis of SIR and SIRS models;
- Simulations and analysis about selection of initial adopters;
- Simulation design of above-mentioned tasks

ZHU Leyan:

- Simulations and analysis of SIS model;
- Simulations and analysis about basic reproductive number R_0;
- Simulation design of above-mentioned tasks

References

[1] “SNAP: Network datasets: Epinions social network,” [Online]. Available: http://snap.stanford.edu/data/soc-Epinions1.html.

[2] “SNAP: Stanford Network Analysis Project,” [Online]. Available: http://snap.stanford.edu/index.html.

[3] D. Easley and J. Kleinberg, Networks, Crowds, and Markets: Reasoning about a highly connected world, Cambridge Books, 2012.