CSCI 4190 Introduction to Social Networks
Project Report
Task 5 – Simulate Epidemics

Group 18
Tsang Hing Wing 1155034860
Tsang Hing Wa 1155063930

1 Abstract

This social network analysis (SNA) project involves a simulation of an epidemic of disease (or an idea/innovation in the social network context) in a real world social network depicted by a real Google+ dataset. By using the three epidemic models: SIS, SIR and SIRS, the infectious behaviour between nodes would be traced and observed to figure out the characteristics and conditions that would lead to the simulation result. This analysis is powered by Stanford Network Analysis Platform (SNAP) and several library tools and a program was constructed to simulate the epidemics.

2 Objective

In this project, we are interested in the characteristics of different epidemic models such as SIS, SIR and SIRS and how long could an epidemic survives under different conditions. The epidemic models are widely used to simulate the infectious behaviour of diseases. In fact, the spread of diseases is just like the spread of ideas and innovations in a social network, since both diseases and ideas would be passed person to person, and across similar kinds of human networks. Ultimately, we would like to find out what are the factors controlling the speed and duration of information spreading in a social network and what are the major differences in behaviours between the three epidemic models.[1]

3 Methodology

3.1 Data Source

A Google+ dataset is selected to be the platform of conducting epidemic simulations. Google+ is a social media platform owned by Google Inc. which allows users to create virtual groups called "circles" to organize their friends, share information within the circles or engage in any other social activities. The dataset is available in the SNAP Datasets of Stanford Large Network Dataset Collection (http://snap.stanford.edu/data) free of charge. The entire Google+ dataset consists of 132 sub-networks. For simplicity,

[1] For simplicity and consistency, a "disease" in this report would be generally referring to an idea or innovation spreading in a social network, unless otherwise specified.

one sub-network dataset has been chosen as our research target (file name: 100518419853963396365).

3.2 Tool

SNAP 3.0 for C++ developed by the Standford University was used to conduct the social network analysis, which would then be imported to Microsoft Visual Studio installed on Windows 10 machines for coding and testing. The network dataset would be imported to Visual Studio as a graph structure in SNAP for further manipulations. In addition, third party packages were deployed in complement with SNAP to facilitate the analysis. A list of the tools is as follows:

- Gnuplot (http://www.gnuplot.info/)
  - a graphing and plotting package, needed for plotting structural properties of networks (e.g., degree distribution);
- NodeXL (http://nodexl.codeplex.com/)
  - a graphical front-end that integrates network analysis and SNAP into Microsoft Office and Excel.

3.3 Experiment Design

In this project, we would reproduce the branching process by assuming each node in the dataset is a person who is susceptible to a disease, and each edge in the network represents a contact between two nodes. Some people in the network would be randomly chosen to be infected by a new disease (i.e. pathogens). In each contact with these pathogens, there would be a probability (p) that the disease would be infected to the other side of contact who is susceptible. We define 1 wave of infectious behaviours (or recovery if possible) as 1 time step. The epidemic would stop at time t when at that time there is no infected person left in the network.

We would design a series of test cases as the input parameters to the epidemic models, and record the running results for each test case. Since the main objective of this project is to observe how long the epidemic would progress until it stops infecting anyone and what are the deterministic factors behind, we would record the time taken (i.e. t) for each run of simulation to stop. For the most accuracy, we would run the simulation for 10 to 100 times, depending on the possibility, and take the average of recorded time as the result.

In this study, we would mainly focus on the following factors that are suspected to have impact on the epidemic duration, which are:

1. the infection probability (contagiousness);
2. the number of initial adopters (severity);
3. the length of infectious period;
4. the network structure.

When measuring the effect of these factors, we would hold the other parameters in the model constant. For example, when the focus is infection probability, we would run the simulations under different probabilities, but holding the same number of initial adopters and infection period among all runs.

For consistency and simplicity, the following assumptions hold in this project:

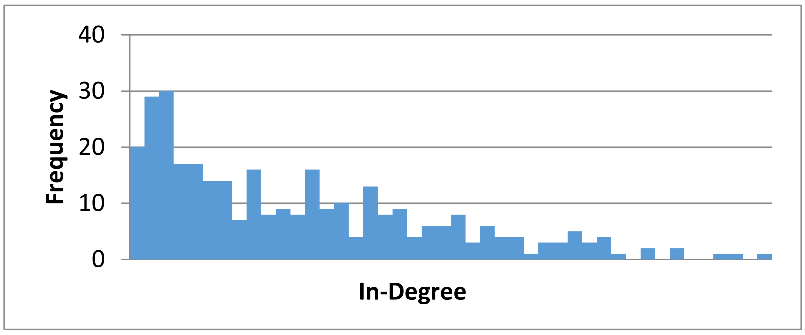
- there is no such concept of contact period that specifies the timing of contact between two nodes. Contact would exist permanently as long as there is an edge;
- the infection probability, infectious period and recovery period would not change in each run of simulation;
- recovery probability is equal to 1 (for SIR and SIRS model only);
- the graph of network would not change during the epidemics;
- there is no upper limit on the epidemic duration, number of infections and repeated infections (for SIRS only).

4 Data Statistics

Table 1 below shows some network statistics generated by the "netstat" example of the SNAP toolkit on our targeted graph.

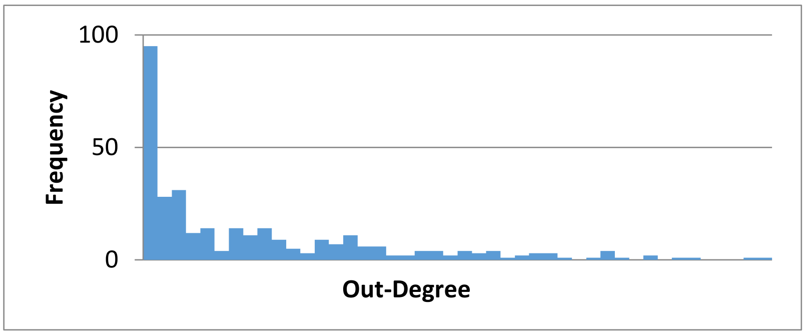
100518419853963396365.edges:	Directed	Self Edges:	0
Nodes:	326	BiDir Edges:	4234
Edges:	10297	Closed triangles:	104181
Zero Deg Nodes:	0	Open triangles:	363625
Zero InDeg Nodes:	2	Frac. of closed triads:	0.222701
Zero OutDeg Nodes:	65	Connected component size:	1.000000
NonZero In-Out Deg Nodes:	259	Strong conn. comp. size:	0.794479
Unique directed edges:	10297	Approx. full diameter:	4
Unique undirected edges:	8180	90% effective diameter:	2.360603

Table 1. Network statistics generated by "netstat" in SNAP



Minimum In-Degree	0
Maximum In-Degree	118
Average In-Degree	31.586
Median In-Degree	24.000

Table 2. In-degree frequency distribution and summary



Minimum Out-Degree	0
Maximum Out-Degree	175
Average Out-Degree	30.712
Median Out-Degree	15.000

Table 3. Out-degree frequency distribution and summary

Given the (weakly) connected component size equals 1 and there is no node with zero node degrees, the graph itself is one huge connected component where all nodes are connected to each other by at least one outgoing or incoming edge. The diameter is small, meaning the network members are relatively close to each other. One thing that is worth to mention is that among the 326 nodes, around 20% of nodes has zero out-degree, which implies they were only included in others' circles, but not to connect themselves with others.

5 Result

5.1 Effect of infection probability

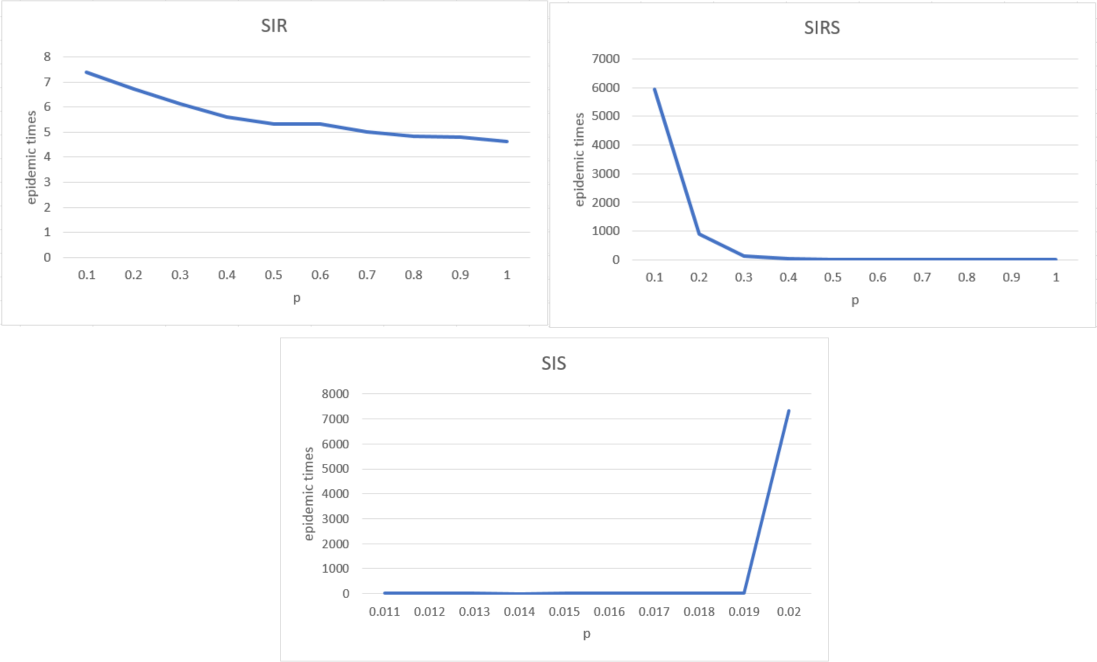
In this part, we are interested in how the infection probability would affect the duration of epidemic. We have defined 10 levels of probability for each epidemic model to find out the relationship. To observe the sole effect of infection probability on t, we set the following conditions for all runs:

- 2 initial adopters;
- infectious period[2] = 1;
- recovery period[3] = 3; (for SIRS model only)

The data below shows the effect of different infection probability (p) on the duration of epidemic (t) under different epidemic models. The t values are taken by averaging the results of 100 runs of simulation for each p level.

p	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**
SIR	7.38	6.73	6.12	5.61	5.33	5.32	5.01	4.83	4.79	4.64
SIRS	5935	885	127.6	35.56	24.45	12.92	14.69	7.01	8.16	4.76

p	**0.011**	**0.012**	**0.013**	**0.014**	**0.015**	**0.016**	**0.017**	**0.018**	**0.019**	**0.02**
SIS[4]	3.34	3.48	3.45	3.02	3.52	3.36	4.05	3.45	3.41	7349



Graph 1. Effect of infection probability

[2] The infectious period refers to the time period that a node remains infectious to other susceptible nodes after it has been infected by the disease.

[3] The recovery period refers to the time period that a node has recovered from a disease and gain immunity, such that during this period, it would not get infected, nor does it transmit the disease to other nodes.

[4] Due to the exceptionally long running time of SIS model, the probability scale is adjusted accordingly.

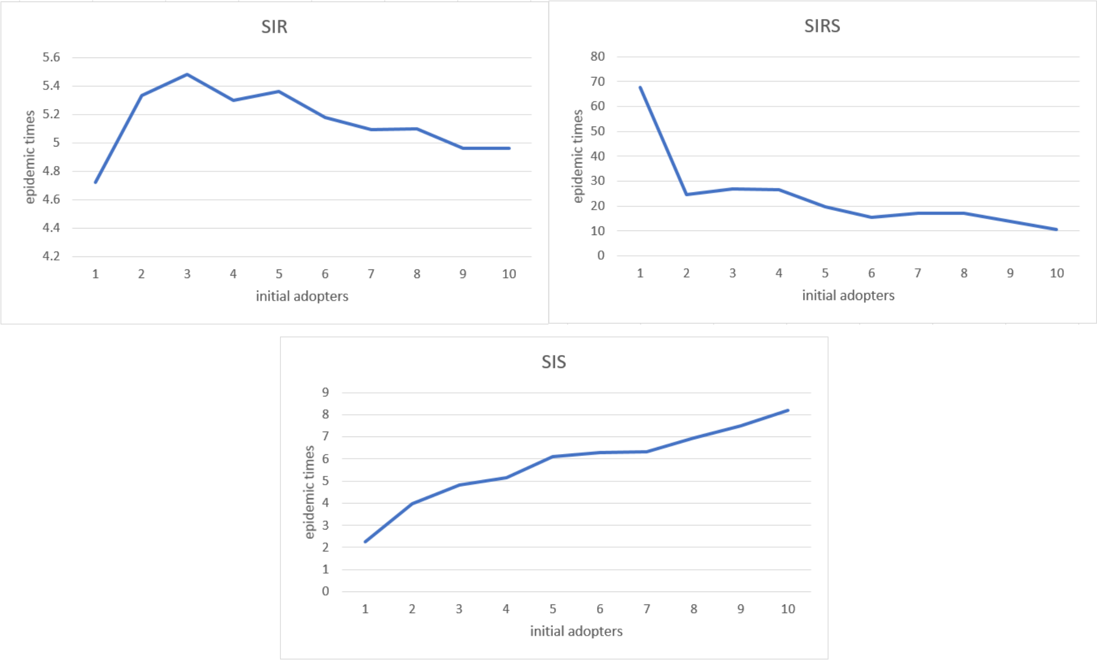
5.2 *Effect of number of initial adopters*

In this part, we are interested in how the number of initial adopters would affect the duration of epidemic. In each epidemic model, we will select a number of nodes, ranging from 1 to 10, to get infected by the disease. To observe the sole effect of initial adopters on t, we set the following conditions for all runs:

- infection probability = 0.5[5];
- infectious period = 1;
- recovery period = 3; (for SIRS model only)

The data below shows the effect of different number of initial adopters (ini_a) on the duration of epidemic (t) under different epidemic models. The t values are taken by averaging the results of 100 runs of simulation for each ini_a level.

ini_a	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
SIR	4.72	5.33	5.48	5.3	5.36	5.18	5.09	5.1	4.96	4.96
SIRS	67.5	24.45	26.97	26.65	19.84	15.55	17.12	17.02	13.77	10.51
SIS	2.26	3.99	4.81	5.14	6.11	6.27	6.33	6.93	7.5	8.21



Graph 2. Effect of number of initial adopters

[5] Due to the exceptionally long running time of SIS model, the infection probability used is set to 0.015 instead of 0.5.

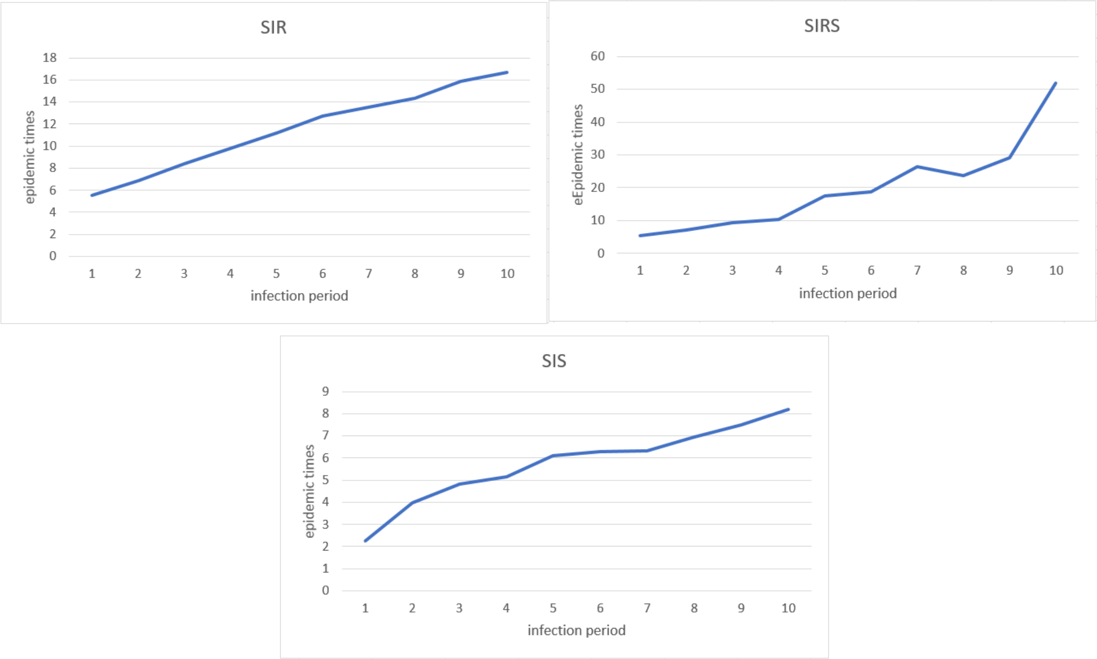
5.3 Effect of length of infectious period

In this part, we are interested in how the length of infection period would affect the duration of epidemic. In each epidemic model, we will apply different infectious period to control the exposure of nodes to disease. To observe the sole effect of infectious period on t, we set the following conditions for all runs:

- 2 initial adopters;
- infection probability = 0.5[6];
- recovery period = 5[7]; (for SIRS model only)

The data below shows the effect of different lengths of infection period (t_I) on the duration of epidemic (t) under different epidemic models. The t values are taken by averaging the results of 100 runs of simulation for each t_I level.

t_I	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
SIR	5.55	6.84	8.41	9.78	11.17	12.75	13.52	14.34	15.88	16.67
SIRS	5.46	7.13	9.28	10.35	17.43	18.63	26.37	23.57	29.2	51.84
SIS	2.26	3.99	4.81	5.14	6.11	6.27	6.33	6.93	7.5	8.21



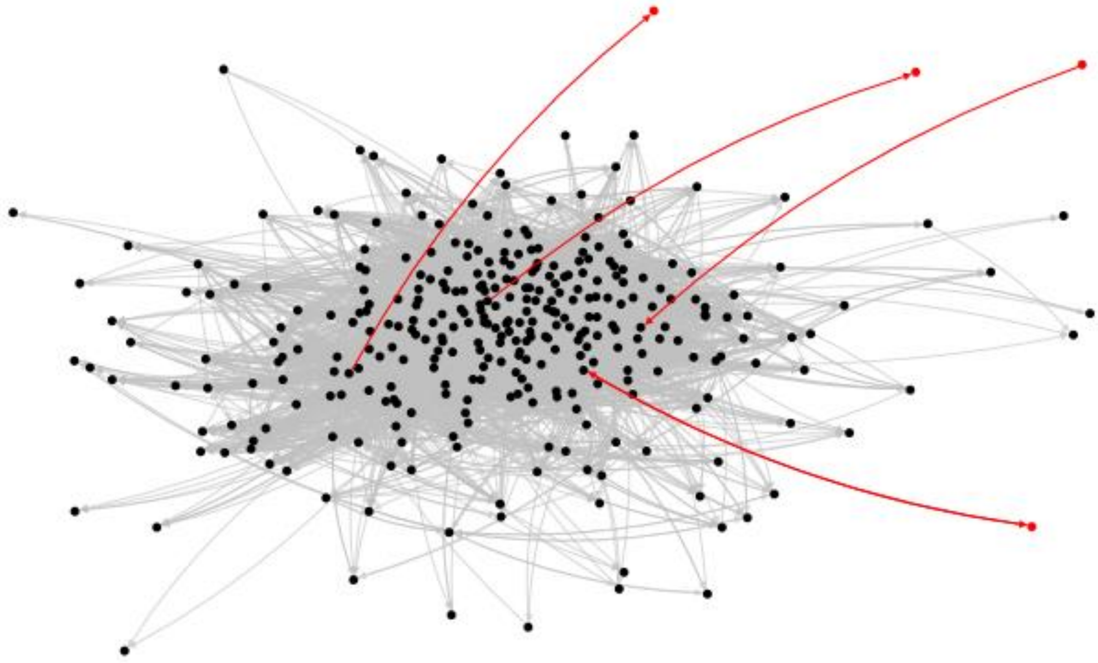
Graph 3. Effect of number of initial adopters

[6] Due to the exceptionally long running time of SIS model, the infection probability used is set to 0.015 instead of 0.5.

[7] Recovery period is set to 5 since the running time for SIRS would be exceptionally long when t_I increases if the recovery period is 3.

5.4 *Effect of network structure*

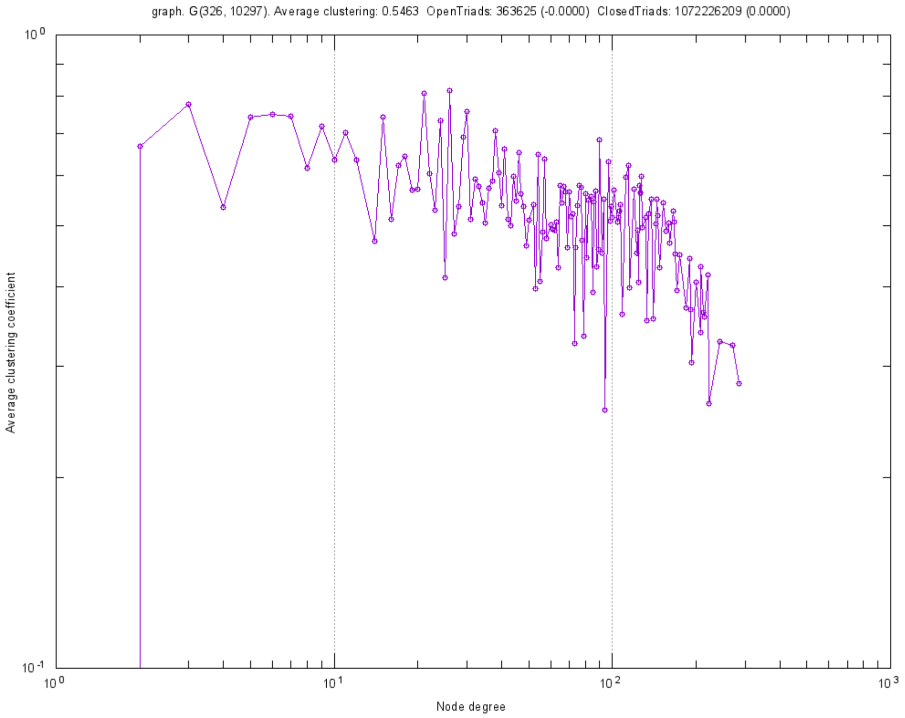
In this part, we are interested in the network structure and how it may potentially affect the duration of epidemic. The figure below shows the visualized graph of our network dataset, which is laid out by using the Fruchterman-Reingold algorithm. The edges in red are the 4 bridges identified by the "concomp" example of SNAP in the graph. Each of these four bridges are connected to a single, distant node, and the bridge is also the only edge that the single node is connected by.



Graph 4. Network layout

As mentioned in Section 4, the entire graph itself is a weakly connected component, in which all components are connected by at least one path regardless of the edge direction. According to the result of "concomp", the graph would also become one strongly connected component if all the nodes with zero in-degree or out-degree are excluded, since they are the only components (consist of single node) that do not have bidirectional edges, and hence not possible to access the rest of the graph from these components.

Also, in the graph below, we can observe an inverse relationship between node degrees and average clustering coefficient in the network. This could be explained by the fact that when a person has more friends or followers, the likelihood that their friends or followers would know each other decreases. The average clustering coefficient is 0.5463, meaning on average more than half of all pairs of a node's friends would be friends themselves, resulting in the large number of closed triangles in the graph.



Graph 5. Average clustering coefficient vs Node degree

6 Discussion

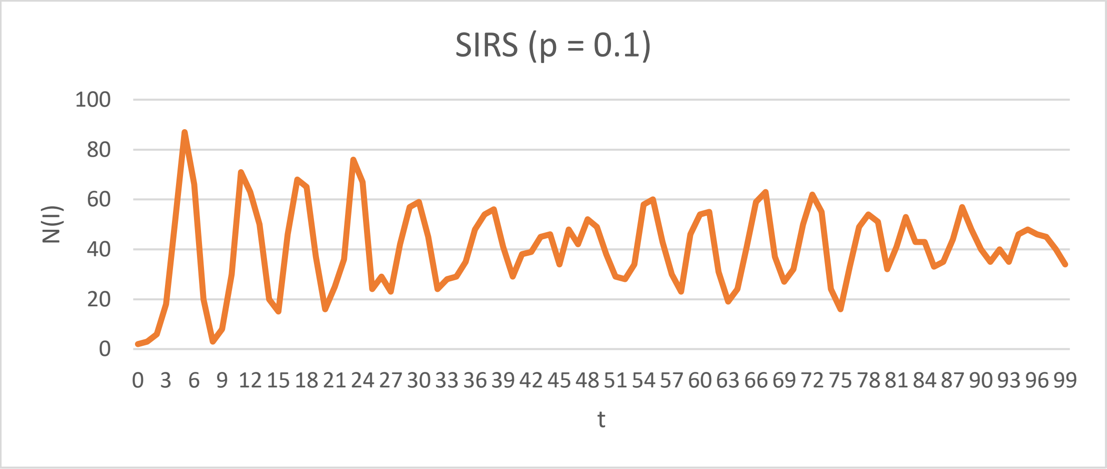
6.1 Factors affecting the duration of epidemic

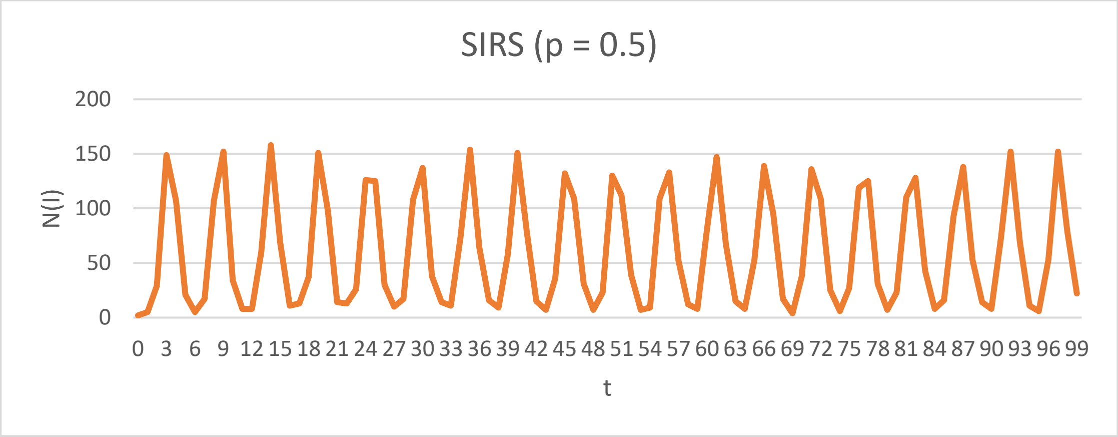
6.1.1 Infection probability

According to our simulation results, there exists a negative relationship between infection probability and epidemic duration. In the SIR epidemic model, as the infection probability increases from 0.1 to 1, the infection period gradually decreases from 7.38 to 4.64, which implies that, the higher the infection probability, the lower the epidemic duration. This may be explained by the nature of SIR epidemic. An SIR epidemic on a finite graph would spread out within a bounded group of nodes. When infected nodes recover, they will be removed from the group permanently. As a result, the number of potentially infectious nodes drops and results in the epidemic ending after a relatively small number of steps. In other words, if infection probability is high, the epidemic spreads rapidly and reaches a wave where it infects no one after a few number of steps.

Moreover, in the SIRS epidemic model, the result is similar to that of SIR model. The epidemic duration has a negative relationship with the contagion probability, except that the epidemic duration is much longer than that of SIR. For example, assume contagion probability is 0.5, the SIR model takes 5.33 steps to stop while the SIRS model comes to an end after 24.45 steps. It is because nodes in recovery states will become susceptible again after tR steps (i.e. the recovery period of a node). This feature allows a refill to the supply of susceptible nodes and hence the epidemic requires longer time to terminate.

Besides, the reason why SIRS would experience a much sharper decrease of duration as p increases could also be seen from the figures below. A model like SIRS which allows temporary immunity could result in oscillations of affected individuals over time caused by synchronization of disease outbreaks. When p = 0.1, the number of infected nodes fluctuates relatively more randomly as time evolves, while in p = 0.5, the number of infected nodes has formed a steady oscillation with larger fluctuations in each cycle. Infection would be more likely to stop at the time when the number of infected nodes are so small, such that the basic reproductive number ($R_0 = pk$) is smaller and may fall below the threshold of 1. In the graph of p = 0.5, the infected node numbers drop to the trough more frequently compared to the graph of p = 0.1. Therefore, as p increases, the chances that R_0 may fall below 1 will increase, hence the epidemic would not persist for very long.





Graph 6. The trends of infected numbers under SIRS model with different infection probabilities

On the other hand, in the SIS epidemic model, the changes in contagion probability makes nearly no difference in epidemic duration until it reaches a critical value (0.02 in our dataset). As the contagion probability increases from 0.01 to 0.019, the epidemic duration just varies slightly. However, when the contagion probability is 0.02, the epidemic duration increases drastically from 3.41 to 7349. The network shifts from one that dies out quickly to one that persists for a very long time. This is due to the nature of SIS model where infected nodes cycle back to susceptible state and create an unbounded supply of nodes. The epidemic is likely to cycle through the nodes multiple times.

6.1.2 Number of initial adopters

The number of initial adopters of the disease is also an important factor of epidemic duration. In SIR and SIRS model, other factors being constant, there is a general downward trend of epidemic duration when number of initial adopters increases, with some fluctuations in between. In these two models, more initial adopters mean that the epidemic spreads faster at the beginning (infect more nodes) and hence more nodes will be removed from consideration at one particular wave. As a result, it is easier to reach a wave where it infects no one, given that the infection probability is lower than 1. The sharp decrease of epidemic duration in SIRS model is contributed by the similar oscillation property that is explained in the last part, except that the larger infection probability is replaced as more initial adopters. Nevertheless, there is one exception in SIR model: when there is only one initial adopter, the average epidemic duration is the lowest. Since there is only one initial adopter, the epidemic will either stop in 1 step or spread for a finite number of steps, which lead to the smaller average value.

On the contrary, there exists a positive relationship between the number of initial adopters and epidemic duration in the SIS model. The epidemic terminates after 2.26 steps and 8.21 steps respectively in the case of 1 initial adopter and 10 initial adopters. The phenomenon is opposite to that of SIR and SIRS model because there is no concept of “removed” or “recovered” in SIS model. A smaller number of initial adopter implies smaller number of infected neighbouring nodes and results in shorter epidemic duration.

6.1.3 Length of Infectious period

In terms of the infectious period, we discovered it has a positive relationship with the epidemic duration in all three models, which could be intuitive since the longer the infectious period, the more likely that a susceptible individual would expose to the disease, no matter it had caught the disease or not. In the graph of SIR and SIS, the trends are relatively linear when comparing to that of SIRS. Note that in this setting we have changed the recovery period to 5 instead of 3 for SIRS model since the epidemic could go very far if the recovery period is only 3. This somehow illustrates the property that SIRS, like SIS, could proceed very far under certain circumstances and conditions, and it appears that there is a threshold that if exceeded, the epidemic would persist for very long time, with possible synchronization as mentioned before.

6.1.4 Network Structure

In general, it is very difficult to analyse the epidemic behaviour in terms of the network structure itself, since the interaction between the network composition and the epidemic model is highly complex and there is not yet a solid understanding on how a specific type of structure would affect the disease dynamics. One structural phenomenon that is identified by the researchers is that long-range links in a SIRS epidemic model could facilitate the synchronization of disease outbreak. In our dataset, it is observed that the network is closely connected with a small diameter and an above-average clustering coefficient, meaning some nodes in the graph that are supposed to be distant from each other (since they have few links) could access each other in a relatively small number of steps, which implies the existence of long-range links (or weak ties) to connect these nodes. As such, it could explain why the simulation under SIRS could be carried forward for so long with obvious oscillations in the number of infected.

6.2 Differences and similarities between epidemic models

The epidemic models introduced in this project: SIR, SIS, SIRS, have their own specific characteristics and usage. We could not directly compare the disease dynamics across these models, but there exist some similarities and differences between these models.

First, the crucial difference between these models lies on the characteristics of recovery and immunity. In SIS, there is no immunity gained after infecting the disease, whereas in SIR and SIRS, the nodes will experience a certain period of immunity. This major difference imposes a huge influence on how the disease would spread and persist.

Second, these models are used to represent different kinds of diseases or social phenomenon in the real world. For example, SIR is usually used to describe the disease that would only be infected once in their lifetime, and after that people would gain lifelong immunity to that. Also, SIRS could also be used to represent social phenomenon like fashion, in which old fashions that were once popular in past decades in the past may become current hits again. Choosing a correct epidemic model when investigating into different subjects is extremely important to obtain a correct prediction of epidemic dynamics.

Thirdly, in SIS and SIRS, "knife-edge" results could be observed in which at some particular levels of parameters, the epidemic will undergo a rapid switch from one that quickly finishes to one that could persist for a long time. In terms of the infection probability, the critical value for SIS model is close to 0.02 according to our simulation results. And this value is largely dependent on the structure of the network and hence it is not a definite number for all SIS epidemics. The same situation could be observed in SIRS model when the infection probability falls below 0.1.

7 Conclusion

To conclude, after a complete simulation involving thousands of trials, we found that all the numeral factors under investigation, including the infection probability, number of initial adopters and length of infectious period, would have significant relationship with the epidemic durations. From the structural view, the closely and densely populated dataset also provides evidence and support to the above relationships.

8 Reference

Easley, D., & Kleinberg, J. (2009). Networks, Crowds, and Markets: Reasoning about a Highly Connected World. doi:10.1017/cbo9780511761942

Finding community structure in very large networks. (n.d.). Retrieved from https://arxiv.org/abs/cond-mat/0408187

Gnuplot Homepage. (n.d.). Retrieved from http://www.gnuplot.info/

Network Weaving: Network Weaving 101. (n.d.). Retrieved from http://www.networkweaving.com/blog/2006/06/network-weaving-101.html

NodeXL: Network Overview, Discovery and Exploration for Excel - Home. (n.d.). Retrieved from http://nodexl.codeplex.com/

SNAP: Stanford Network Analysis Project. (n.d.). Retrieved from http://snap.stanford.edu

What are strongly and weakly connected components? (n.d.). Retrieved from https://www.quora.com/What-are-strongly-and-weakly-connected-components