CSCI4190: Introduction to Social Networks

Project Report: Task 4 – Simulating Epidemics using SIR, SIS, and SIRS Epidemic Models

CHOI Jang Hyeon
1155077214

1. Abstract

This paper aims to simulate an epidemic based on SIR, SIS, and SIRS models using a real-world dataset – a user network of Twitch, a livestreaming platform. This network of users is formed based on friendship and could represent a social network in the general society; each user is a person, and an edge represents the people they meet often.

The infection probability, number of initial adopters, the duration of infection, and the duration of removal have been experimented with to examine their effects on not only the lifespan but also the average infection rate per day of the disease in the Twitch network.

The lifespan of the disease under the SIR model and the average rate of infection under the SIS and SIRS model have increased when the duration of the infection increased. It was surprising to discover that the rate of infection converged to a positive value as time passed, meaning that the number of infected people was stable under a 500-day timeframe.

The paper additionally examines the relationship between the Twitch network structure and the spread of the disease.

2. Methodology

2.1 Dataset

The dataset used in this project is "Twitch Social Networks" available at Stanford University's SNAP (Stanford Network Analysis Project) database [7]. Each node in the network represents a user and an edge between two nodes indicates friendship between the users. There are six of these networks in the database, each differing in the language that the users use: German (DE), English (ENGB), Spanish (ES), French (FR), Portuguese (PT), and Russian (RU). Only one of these networks was used to simulate the epidemic. Hence, the EN (also referred to as ENGB) network was chosen based on its density and number of users. Other aspects of the dataset, such as node features, were not necessary for the simulations.

2.2 Software

The simulation of the epidemics is programmed with Python and the snap.py library, a Python version of Stanford's SNAP library that is originally written in C++ [5]. NodeXL, a visualization package for Microsoft Excel, was used to illustrate the ENGB network structure, and Matplotlib was used to illustrate scatterplots and line graphs [3].

2.3 Independent variables

The epidemic is simulated on the Twitch dataset by randomly selecting one or more initial adopters. These adopters are infected with the disease and each have a probability p of spreading the disease to their friend (a node that they are connected to). The infected people will stay infected for t_i days (an increment of t is equal to one day) and either return as susceptible or are removed from the network afterwards, depending on the epidemic model. Thus, the four

independent variables used to measure their effects on the lifespan and average rate of infection is the following:

- Infection probability p
- Number of initial adopters n
- Duration of infection t_i
- Duration of removal period r

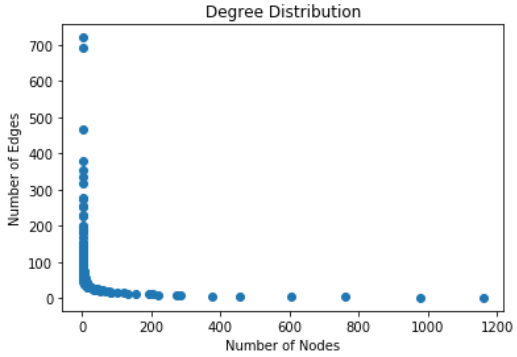
3. Statistics

Network Details	
Nodes	7,126
Edges	35,324
Self-loops	0
Connected components	1
Maximum Geodesic Distance	10
Average Geodesic Distance	3.68
Clustering Coefficient	0.1309
Modularity	0.42
Closed Triads (Triadic Closures)	29,266

There are a total of 7,126 nodes in the ENGB network and 35,324 edges that connect them into one large, single component. The geodesic distance refers to the shortest path between two nodes, and the maximum is 10 in this network. The average geodesic distance, therefore, refers to the average number of edges between two nodes. An average geodesic distance of 3.68 means that several paths exist between two nodes such that there often is a path of 3.68 edges to the other node [1].

The clustering coefficient of 0.1309 and modularity of 0.42 indicates that, although the clusters themselves are quite sparse, there are several edges that connect them to other clusters within the network. Hence, dense communities within the network are uncommon and edges are somewhat evenly connected between nodes [6].

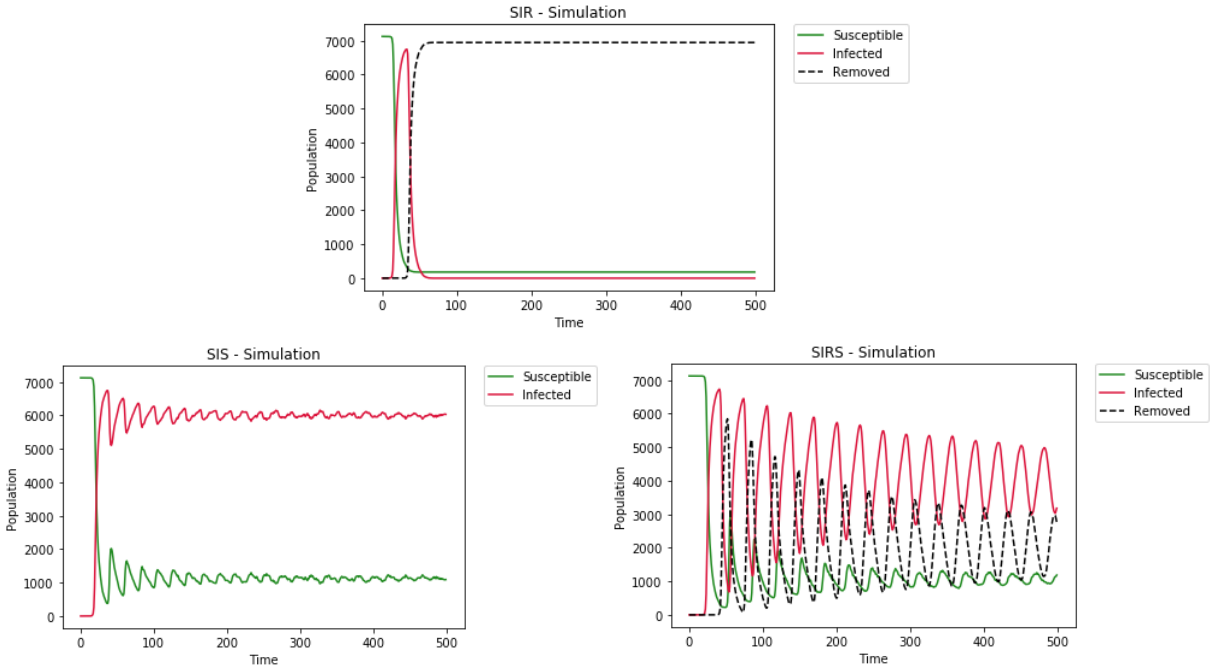
The figure below shows the number of nodes that have a certain number of edges. The distribution follows a power-law distribution in which only few of nodes have a substantial number of edges. The maximum number of edges that a node has is 720, and more than 1,000 nodes have less than 10 edges.



(A scatter plot that distributes the number of nodes against the number of edges.)

4. Results

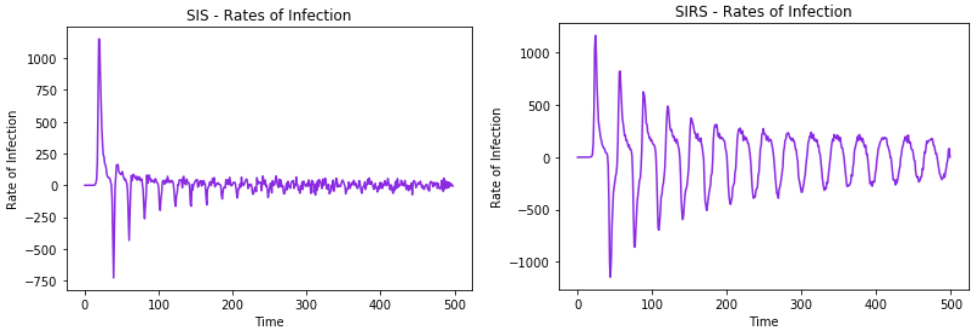
Given one initial adopter, the infection probability of $p = 0.1$, infection period of $t_i = 20$ days, and the removal period (for SIR and SIRS) of $r = 10$ days during a 500-day period, the disease persisted under the SIS and SIRS models while the disease died out under the SIR model.



(Simulations of the SIR, SIS, and SIRS models under the same conditions.)

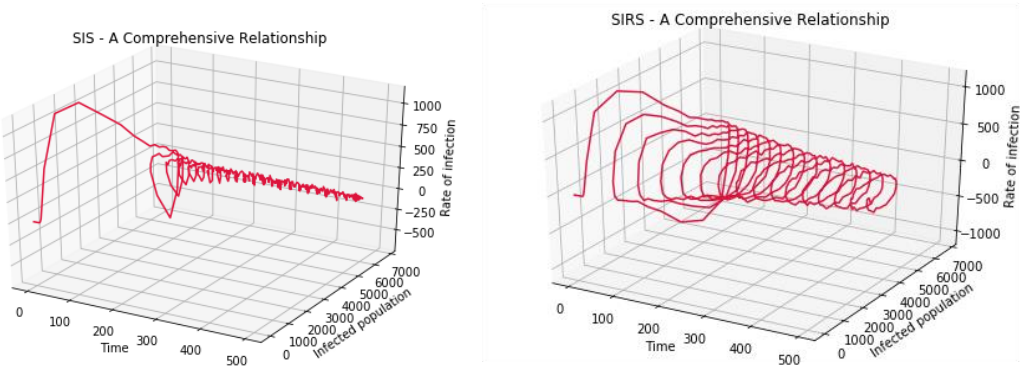
The SIS and SIRS models have persisting diseases with fluctuating numbers of infected people and, hence, were measured by their average rates of infection. On the other hand, the lifespan of the disease was measured for the SIR model instead.

If the rates of infection were illustrated from simulating the SIS and SIRS models, it would show fluctuating graphs like sound waves. Nevertheless, it can be seen in the graphs below that the rates of infection converge to a value close to 0. Experiment results in section 4.2 show what these values can be. The rate of infection can be negative if not enough people are infected with probability p before the duration of the infection ends.



(Rates of infection for SIS and SIRS models.)

Combining the number of infected people with the rate of infection throughout 500 days shows an interesting graph.



(A 3D illustration of the relationship between the number of infected people and the rate of infection over time.)

These 3D illustrations of the SIS and SIRS models show that even if the infection rate converges to a value close to 0, the number of infected people remains in a positive range over a 500-day period.

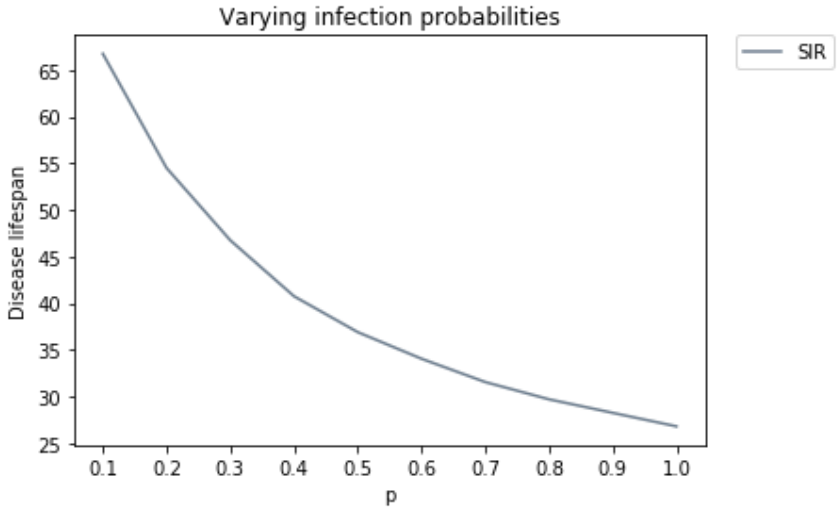
To determine the effects that the infection probability *p* has on the lifespan and the rate of infection of the disease, values from 0.1 to 1.0 were experimented incrementally for each of the three epidemic models. Also, the number of initial adopters *n* were increased by a factor of 2, and the infection period t_i was increased by increments of 10 as well. Removal periods *r* were held constant at 10 for SIR while it was incremented by 10 for the SIRS model. For higher accuracy, 20 simulations were run with a timeframe of 500 days for each time, and their values were averaged.

4.1 Disease lifespan under the SIR model

The lifespan of the disease is measured by logging the number of days that have passed until the total number of infected people is equal to 0, which is when the disease dies out. The following tables record the day that the disease dies for each modification of independent variables.

Varying probabilities of infection

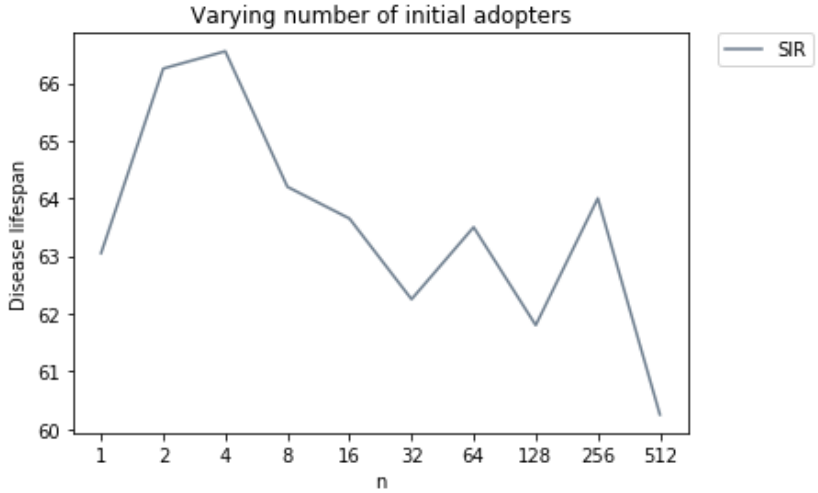
p	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
n	1	1	1	1	1	1	1	1	1	1
t_i	20	20	20	20	20	20	20	20	20	20
SIR	66.75	54.5	46.75	40.75	36.9	34.05	31.55	29.7	28.25	26.9



The lifespan of the disease decreases as it become more contagious. Under the SIR model, infected people are removed after the duration of the disease. Therefore, a more contagious disease will remove more people from the network sooner and reduce its lifetime.

Varying numbers of initial adopters

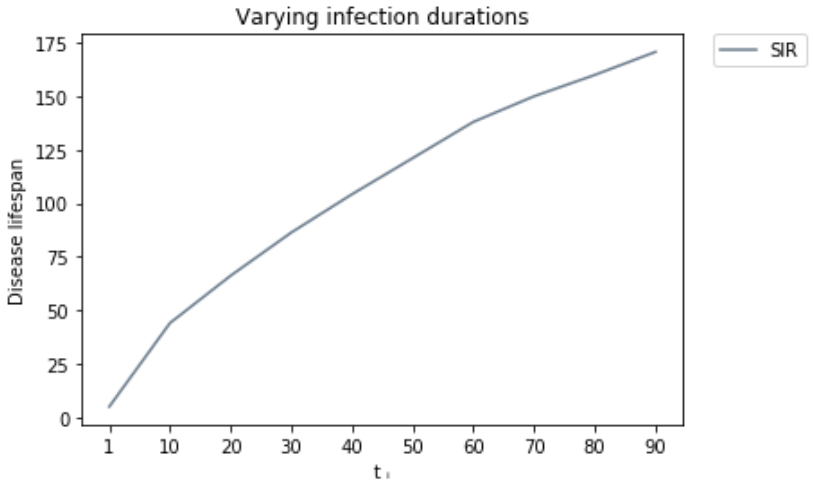
p	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
n	**1**	**2**	**4**	**8**	**16**	**32**	**64**	**128**	**256**	**512**
t_i	20	20	20	20	20	20	20	20	20	20
SIR	63.05	66.25	66.55	64.2	63.65	62.25	63.5	61.8	64.0	60.25



The lifespan of the disease shows a declining trend as the number of initial adopters increased. This is intuitive since a greater number of initial adopters have increased the speed of spreading the disease and caused an sooner removal of people from the network.

Varying infection durations

p	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
n	1	1	1	1	1	1	1	1	1	1
ti	**1**	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**
SIR	5.1	44.1	66.2	86.35	104.2	121.05	137.95	149.9	159.85	170.6



Increasing the infection duration has elongated the disease lifespan. The reason is that the disease has remained in the network longer as the infection duration increased. Hence, infected nodes were removed at much later stages, resulting with a more stubborn disease.

The duration of removal does not affect the lifespan of the disease under the SIR model because removed nodes do not return to a susceptible state. Therefore, varying removal durations were not experimented with the SIR model.

4.2 Average rates of infection under the SIS and SIRS models

To solve for the average infection rate, the infection rate per day needs to be calculated first. The infection rate equals to the following:

$$\frac{Number\ of\ infected\ people\ on\ day\ (t+1)\ -\ Number\ of\ infected\ people\ on\ day\ t}{day\ (t+1)\ -\ day\ t} = number\ of\ infected\ people\ on\ day\ \mathrm{t}+1$$

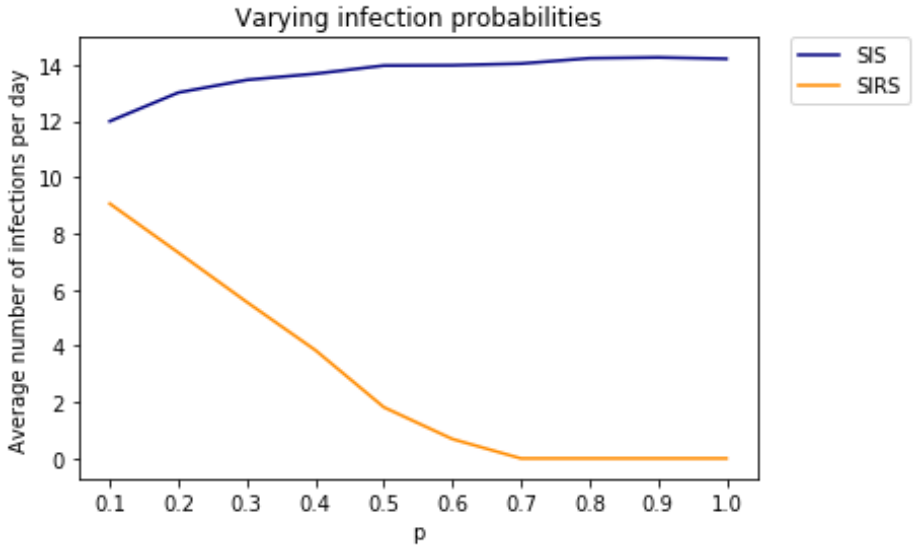
The number of infected people per day in each of the 500 days is then summed as the total number of infected people and divided by the total number of days. This will give the average number of infected people per day:

$$\frac{Total\ number\ of\ infected\ people}{Total\ number\ of\ days\ =\ 500} = average\ number\ of\ infected\ people\ per\ day$$

The rate of infection, or number of infected people per day, measures the speed at which the disease is being spread or, in other words, how many new infections there are per day in the entire network. The following tables show how varying the infection probability, number of initial adopters, infection duration, and removal duration affect the average number of newly infected people per day.

Varying probabilities of infection

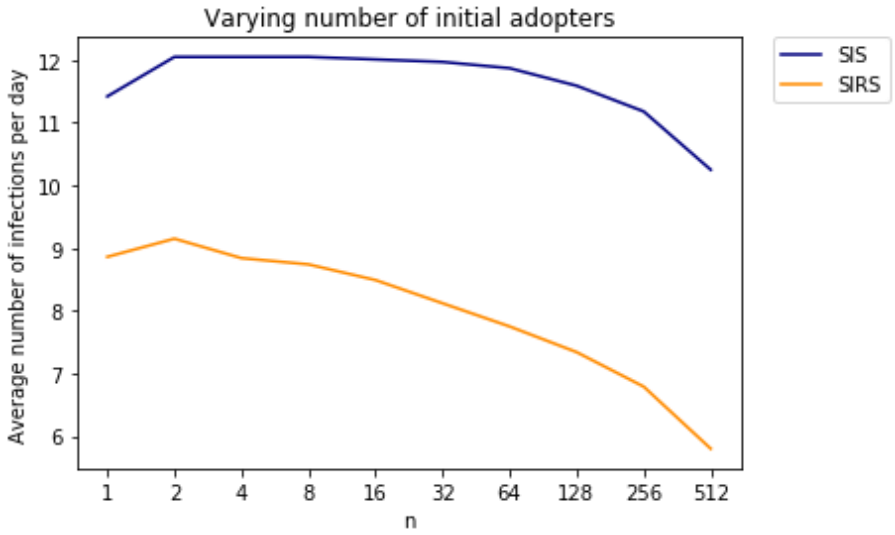
p	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
n	1	1	1	1	1	1	1	1	1	1
t_i	20	20	20	20	20	20	20	20	20	20
r	10	10	10	10	10	10	10	10	10	10
SIS	11.99	13.01	13.46	13.68	13.97	13.98	14.04	14.23	14.26	14.21
SIRS	9.06	7.32	5.56	3.85	1.82	0.69	0.0	0.0	0.0	0.0



From the results of simulating the two epidemic models on varying probabilities of infection, it was interesting to find out that the average number of cases slightly increased for the SIS model but decreased for the SIRS model. One reason behind the phenomenon with the SIRS model could be that, the quicker the disease spread, more nodes were removed from the graph at a time. Thus, the disease reaches and spends all of its infection period at dead-end nodes before any of the removed node could reenter the network for the disease to persist to. In the SIRS model, the disease started to die out when $p \geq 0.7$; when there are no more newly infected people per day, then it means that that disease has died out.

Varying number of initial adopters

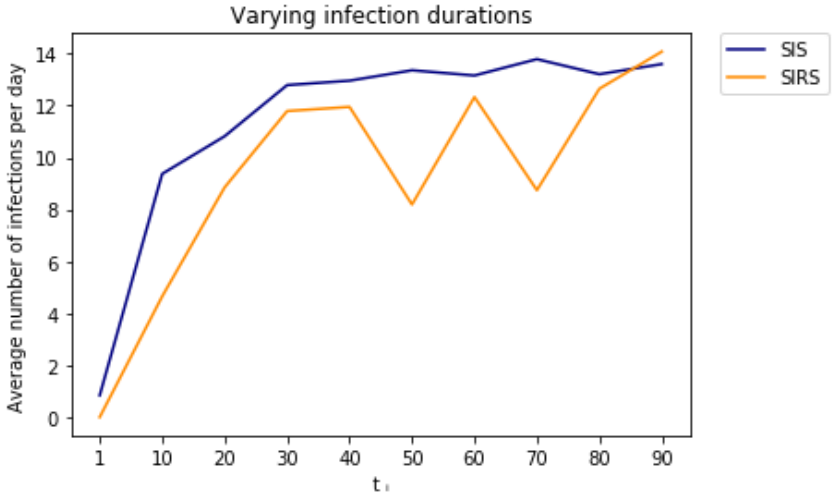
p	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
n	**1**	**2**	**4**	**8**	**16**	**32**	**64**	**128**	**256**	**512**
t_i	20	20	20	20	20	20	20	20	20	20
r	10	10	10	10	10	10	10	10	10	10
SIS	11.42	12.05	12.05	12.05	12.01	11.97	11.87	11.59	11.18	10.25
SIRS	8.86	9.15	8.84	8.74	8.49	8.12	7.75	7.34	6.79	5.80



The different number of initial adopters had delicate effects on the average infection rate. Under the SIS model, the average infection rate seemed to have briefly increased and plateaued before decreasing. The SIRS model also showed a similar behavior. The reason could be that as the disease could spread faster with more initial adopters, there were lesser and lesser susceptible nodes to infect each day, especially when there are 512 initial adopters and are given 500 days to infect.

Varying infection durations

p	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
n	1	1	1	1	1	1	1	1	1	1
t_i	**1**	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**
r	10	10	10	10	10	10	10	10	10	10
SIS	0.84	9.37	10.82	12.79	12.96	13.36	13.16	13.79	13.21	13.60
SIRS	0.0	4.65	8.85	11.79	11.95	8.19	12.33	8.74	12.65	14.08

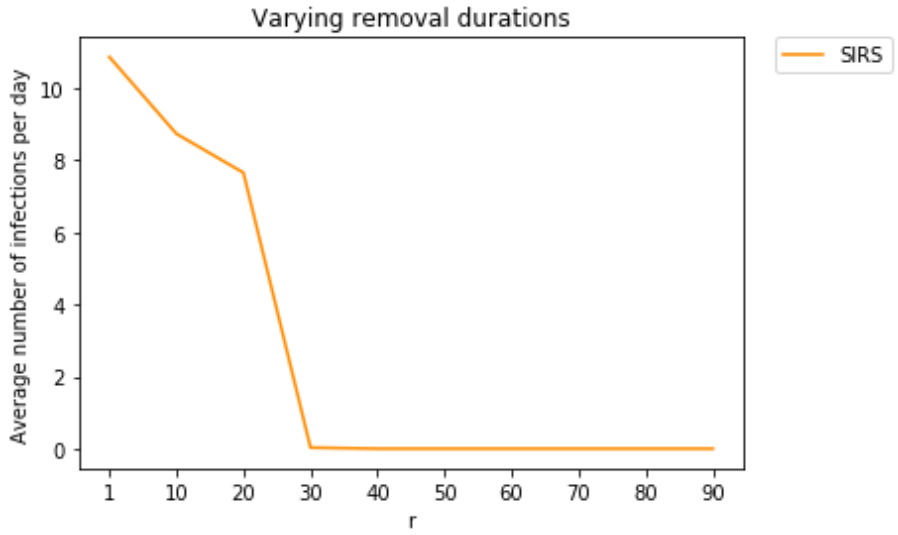


Increasing durations of infection showed a gradual escalation in the average rate of infection for both models. The SIRS model resulted with more unstable values as t_i increased.

When t_i is too small, it did not provide enough time for the disease to spread with probability $p = 0.1$, causing the infection rates to be very low.

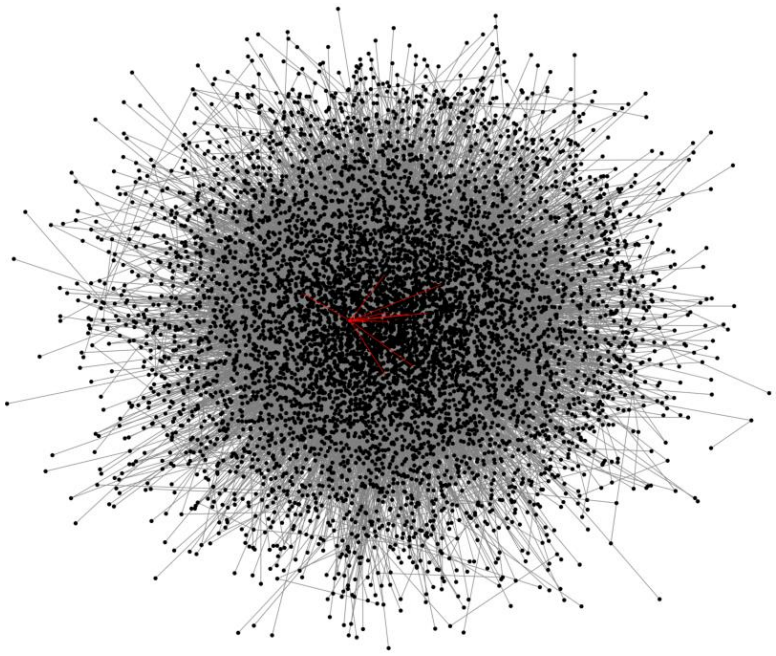
Varying removal durations

p	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
n	1	1	1	1	1	1	1	1	1	1
t_i	20	20	20	20	20	20	20	20	20	20
r	**1**	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**
SIRS	10.86	8.73	7.65	0.03	0.0	0.0	0.0	0.0	0.0	0.0



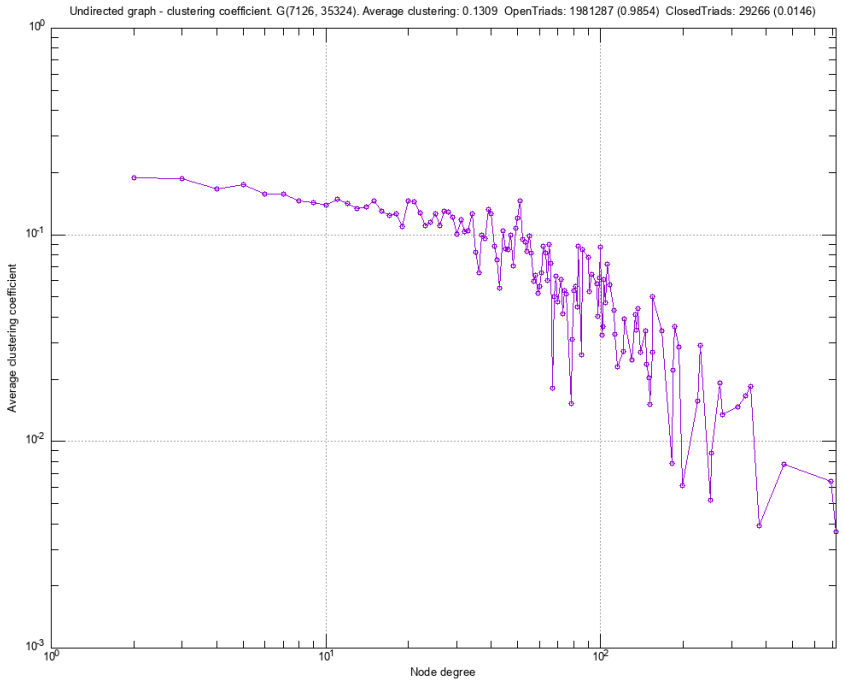
Increasing the removal duration *r* has led the disease to die out because the results show that the number of infections per day is 0 at certain values of *r*. The specific intervals in which the disease persists is when $r \leq t_i$ and the disease essentially dies out when $r > t_i$. In other words, if the infected nodes remain infected longer than nodes remaining removed, the disease is able to persist at a positive rate. However, if the nodes remain removed for longer periods of time, the disease will eventually reach a dead-end node and die out.

4.3 Network structure



(NodeXL visualization of the ENGB network.)

The ENGB network has a density of 0.00139. The density of a network is the ratio of the number of connected edges to the maximum possible number of connected edges [8]. The network is not very interconnected but is still evenly so because of its modularity of 0.42. The pandemic would have spread faster if the network had more edges between nodes. An example node has been selected in the center and colored as red with its neighbors and the number of its edges shows how sparse the network is.



This figure illustrates the clustering coefficient values for nodes with certain degrees and the average clustering, number of open triads and closed triads are also included in the title. The inverse relationship of the graph indicates that the clustering coefficient decreases the more edges the node has. Furthermore, since the average clustering coefficient is equal to 0.1309 and the number of open triads (if there are three nodes, one pair of nodes is not connected) is at least nine times greater than the number of closed triads (nodes that form a triadic closure), the number of triadic closures is very low [2]. In other words, the neighbors of a node are rarely connected to each other and that the maximum infection rate could have been higher if the neighbors were more interconnected.

5. Conclusion

There were several interesting findings from simulating SIR, SIS, and SIRS epidemic models under different conditions on a real-world dataset. Experimenting with the infection probability, number of initial adopters, infection period, and removal period showed instances in which the lifespan of the disease increased in the SIR model whereas it died out in the SIRS model.

In the case of the SIR model, the lifespan of the pandemic increased proportionally to the duration of the infection. Intuitively, the longer a person stayed infected, the longer the disease remained in the network.

On the other hand, the SIS and SIRS models had persisting diseases with a positive infection rate in the 500-day period. However, the disease under the SIRS died out when $p \geq 0.7$. In terms of varying removal durations, the disease died out when $r > t_i$ and persisted when $r \leq t_i$.

The ENGB Twitch social network is highly complex, with 7,126 nodes and 35,324 edges. It is challenging to make substantial changes to the social network structure to witness any meaningful effects under the epidemic models, but a network like ENGB-Twitch must have been large enough for the disease to travel the network and leave enough time for removed nodes to turn susceptible again before infecting them in the SIS and SIRS models and allow the disease to persist.

6. References

1. Han, J., Kamber, M., & Pei, J. (2012). Advanced cluster analysis. Data Mining, 497-541. doi:10.1016/b978-0-12-381479-1.00011-3
2. Huang, H., Dong, Y., Tang, J., Yang, H., Chawla, N. V., & Fu, X. (2018). Will Triadic Closure Strengthen Ties in Social Networks? ACM Transactions on Knowledge Discovery from Data, 12(3), 30th ser.
3. NodeXL: Your social network analysis tool for social media. (n.d.). Retrieved May 01, 2021, from https://www.smrfoundation.org/nodexl/
4. Rozemberczki, B., Allen, C., & Sarkar, R. (2021). Multi-Scale Attributed Node Embedding. Journal of Complex Networks, 1-20.
5. Snap.py – SNAP for Python: Stanford network analysis platform. (n.d.). Retrieved May 01, 2021, from https://snap.stanford.edu/snappy/index.html
6. Tao, Y., & Rapp, B. (2019). The effects of lesion and treatment-related recovery on functional network modularity in post-stroke dysgraphia. NeuroImage: Clinical, 23(101865). doi:10.1016/j.nicl.2019.101865
7. Twitch social networks. (n.d.). Retrieved May 01, 2021, from https://snap.stanford.edu/data/twitch-social-networks.html
8. What is network density - and how do you calculate it? (2013, June 7). Retrieved May 01, 2021, from https://www.the-vital-edge.com/what-is-network-density/