CS 7280: Network Science

Fall 2022

Assignment-4

Learning Objectives

The objective of this assignment is to experiment with the concepts we covered in Module-4 about network epidemics, and see how the theoretical results that were derived in class compare to simulation results.

Delivery

Please submit your Jupyter notebook **Assignment4-YOURLASTNAME.ipynb.** Ensure all graphs are **generated** and **appropriately labeled** in your notebook.

Note

First review the Epidemics on Network python module available here. You can download this module and run the examples given in the documentation to become familiar with it.

Install the library using: pip install EoN.

This homework, especially the parts 2 and 3 might take several minutes to run. Be aware of this and plan to complete it accordingly.

Dataset

The file fludata.txt has the list of students, teachers and staff at a school. The interaction between them was measured based on the proximity of the sensors that people were carrying on them. The data file has three columns, the first two columns are the IDs of the people and the third column is the number of interactions. Construct an undirected graph from that text file using functions in networkX module. For the purpose of this assignment, we consider only the **unweighted** graph (i.e., you can ignore the third column).

Part-1 (40 Points)

Consider a pathogen with transmission rate of 0.01 and recovery rate of 0.5. Suppose that an outbreak started at node 325 (“patient-0”).

Run the SIS model given the above parameters 10 times and answer the following questions. You can use the function *eon.fast_SIS* for efficient simulation.

A) With the previous parameters, plot the number of infected individuals over time and the number of susceptible individuals over time (i.e. create a line plot with 10 lines for each). What do you observe? Comment on your observations.

B) Remember that we modeled the initial exponential increase of the number of infected nodes as $I(t) \approx I_0 e^{t/\tau}$, where τ is a time constant. Note that here $I_0 = 1$ as only one node was infected initially. Now, using a *single* simulation, let us try to fit an exponent to the curve of the number of infections, say for $I(t) \leq 100$ (the exponential region of the outbreak, where the *number of infected* is less than or equal to 100) and print the estimated time constant τ. Plot the simulated number of infected nodes and the exponential fit you estimated (i.e. line plot with two lines). Finally, examine quantitatively how close the fit is using R-squared error.
Hint: You may find the function scipy.optimize.curve_fit to be helpful for finding the exponent.

C) Repeat the above estimation several times with the same configuration as part 1B (~25x) running *fast_SIS* and fitting τ each time. Show the distribution of the exponent τ by plotting a box plot. Then, compute the expected value of τ using the three theoretical derivations we covered in class: **(1)** random distribution shown in the Lesson 9 Canvas lecture *"SIS Model"* , **(2)** arbitrary distribution shown in the Lesson 9 Canvas lecture *"Summary of SI, SIS, SIR Models with Arbitrary Degree Distribution",* and **(3)** arbitrary distribution from the textbook found in *Ch. 10, Equation 10.21.* Plot the theoretical calculations as dots on the box plot. In text, compare these theoretical estimates to the empirical distribution and indicate which of the two derivations matches your simulation results most accurately.
Hint: When running your simulations, you may want to discard the outbreaks that died out stochastically – *check whether the number of infected nodes at the last time step is 0 and ignore them*. *When doing these calculations, be careful regarding the definition of moments and the underlying assumptions for some formulas.*

D) Similarly, compute the percentage of the population that remains infected at the endemic state, plot its distribution across several simulation runs, and compare it with the beta theoretical value we derived in class (in the lesson 9 module, "*SIS Model"*). Choose the most appropriate plot to compare these values.
Hint: In a single simulation, the number of infected individuals at endemic state$I(t)$can be approximated by averaging the values of $I(t)$ across the time interval where $I(t)$ starts to fluctuate in a range rather than showing significant increasing/decreasing trend.
Hint: You can exclude the cases that died out stochastically from your plot.

Part-2 (25 Points)

Next, let us vary the transmission rate and see how it affects the spread of infection. Since we know that only the ratio of the transmission rate and the recovery rate matters, let us keep the recovery rate constant at 0.5 and vary only the transmission rate.

A) Vary the transmission rate and plot the **number of infected individuals** at the endemic state against the transmission rate. Make sure to cover a wide range of values so that your plot covers from the "0 endemic state" case to the "100% endemic state" case. For each value of the transmission rate, compute the average of several simulations (~10) to avoid outliers.

B) Based on the equations derived in lecture, find the minimum theoretical values of the transmission rate for an epidemic to occur. Plot the curve of the expected values at each value of beta, along with your theoretical minimum values, against your simulated results from part 2A, and analyze how they compare.

Part-3 (30 Points)

Now, let us see how the choice of "patient-0" affects the spread of an outbreak. Consider every node of the network as patient-0, and run the SIS model using the parameters in Part 1 to compute τ. Run the simulation with each node in the simulation as patient-0. **Hint**: You can skip cases where the infection quickly diminishes to 0.

A) Plot a scatter plot between the value of τ that corresponds to each node, and different centrality metrics of that node: degree centrality, closeness centrality, betweenness centrality, and eigenvector centrality. Remember to use the **unweighted** centrality metrics.
B) Compute the Pearson correlation coefficient between each centrality metric and τ, along with a confidence level for that correlation.
C) Rank these centrality metrics based on Pearson's correlation coefficient, and determine which of these metrics can be a better predictor of how fast an outbreak will spread from the initial node. Comment on your observations.

Note: You may get different results between different runs as the simulation is random. Grading will be based on correctness of the implementation as well as how well you justify your observations.

Part-4: Knowledge Question (5 Points)

Answer the following food for thought question from Lesson 10 – Submodularity of Objective Function:

Prove that a non-negative linear combination of a set of submodular functions is also a submodular function.

Tip: Make sure you understand the definition of linearity.