COMP4880/8880 Assignment 3 - Resilience, Contagion, and Homophily

This assignment is graded out of 15, and counts towards 15% of the final grade.

You should submit your answers via Gradescope and your code via wattle. To avoid losing marks both your answers and code must be submitted before the deadline.
Register for Gradescope at http://gradescope.com using your anu e-mail and include your student ID number with sign-up. Use the entry code **MZ35V8** to sign up for COMP4880/COMP8880.

Submitting answers: Prepare answers to your homework in a single PDF file and submit it via GradeScope. If you complete the assignment in a jupyter notebook (or R notebook ... etc) submitting a pdf copy of the notebook (containing all answers) will suffice. You are responsible for clearly identifying the question being answered, and also make clear what source code was used to answer the question. Some questions require you to include source code in the answer pdf. At submission time you will need to indicate which page(s) of the PDF answer each question. An introduction to submission through gradescope is available here: https://www.youtube.com/watch?v=KMPoby5g_nE

Submitting code: Upload your code to wattle by the same submission deadline. Put all the code into a single file and upload it. A jupyter, R notebook, or zip file is acceptable.

Completing the assignment You may choose to use any programming language. We highly recommend a python notebook with the networkx package. You can use third party packages such as graph libraries, for example networkx in python, igraph in R, and so on -- this is strongly recommended.

Question 1: Resilience

This question requires you to use the unweighted undirected global flight network from Assignment 0 available here: http://seeslab.info/downloads/air-transportation-networks/ . Please read the description of the data on the page where you download the data. Node information is stored in “global-cities.dat” (node ids are the second column), and an edge list (on node ids) is stored in “global-net.dat”.

1.1. **(1 mark)** Use moments of the degree sequence to calculate the critical threshold under random node removal for the global flight network.

1.2. **(2 mark)** Simulate 20 random node removal sequences for the global flight network and calculate the average fraction of nodes that need to be removed before the largest component contains less than 2% of the number of nodes in the original graph. Please report: the fraction of nodes removed in each simulation, and the average fraction of nodes removed across all simulations.

1.3. **(1 mark)** Does the simulated result support the calculation from the degree sequence? List the main reasons why they might be different? (Short answer or bullet points are expected)

1.4. **(1 mark)** Is the condition that “the largest component contain less than 2% of nodes in the graph” a realistic condition for breakdown of the flight network? Come up with another condition you believe to be more realistic. Explain why your new condition is more realistic.

Question 2: Contagion

A computer virus is spreading through a set of computers and network hubs. We will model this as a bipartite graph with two types of nodes: computers C and hubs H. Both computers and network hubs are susceptible to this computer virus. Computers are only connected to hubs (computers do not infect other computers directly) and hubs are only connected to computers (hubs do not infect other hubs directly). Hubs may connect to multiple computers and computers may connect to multiple hubs, all connections are bi-directional. There is a different infection rate for computer to hub infection than from hub to computer infection, both are nonzero.

2.1. **(1 mark)** Write down an expression for the rate of change of the fraction of nodes infected in the network version of the SI model for this bipartite case. Make sure to define all the variables you use, and make explicit any assumptions you make. Your expression should be in matrix form. Hint: Look at the lecture slides (specifically the first part of the second lecture on contagion) and consider how you might modify the expression, in particular consider changing beta (in the lecture slides beta is the probability per unit time that infection will be transmitted between two individuals).

2.2. **(1 mark)** Use your bipartite SI model to derive an approximate expression for the probability that each node in the network (both computers and hubs) is infected at time t. This expression will be vector valued with a domain of the nonnegative real numbers. You may assume that we care only about the early time so the fraction of infected individuals is small. Also assume at t=0 there is one infected node. Tip: your solution should involve finding the dominant eigenvector of some matrix.

2.3. **(2 marks)** You have been given a bipartite graph ("computer_hub_graph.csv" on wattle, stored as an adjacency list, the format is described here: https://networkx.github.io/documentation/networkx-1.10/reference/readwrite.adjlist.html) with 30 computers and 15 hubs. Assume that the probability per unit time that infection will be transmitted from a computer to a hub is 0.05, and 0.01 from a hub to a computer. Using your answer to part 2 above plot out the average number of nodes that are infected at time t (for $t < 10$). On your plot the x-axis should be time and the y-axis should be the average number of nodes that are infected. Also determine the 5 nodes that are most likely to be infected at time t=10.

2.4. **(2 marks)** Now simulate the SI model on the network (given by "computer_hub_graph.csv" on wattle) assume that the probability per unit time that infection will be transmitted from a computer to a hub is 0.05, and 0.01 from a hub to a computer. Start the infection from a single random node. Run this simulation 1000 times (with a new random start node each time) then plot the result (for $t < 10$). On your plot the x-axis should be time and the y-axis should be the average number of nodes that are infected. Also determine the 5 nodes that are most likely to be infected at time t=10. Tip: the time until an infection spreads across an edge can be modelled as an exponential random variable.

2.5. **(1 mark)** Give at least three reasons why your answer to part 3 is not the same as your answer to part 4. (bullet points or short answer)

Question 3: Homophily

Download the **CiteSeer for Document Classification** dataset from https://linqs.soe.ucsc.edu/data (direct link to data:

https://linqs-data.soe.ucsc.edu/public/lbc/citeseer.tgz). The nodes in this graph are computer science papers and the edges are citations. Each paper belongs to one category from: Agents, AI, DB, IR, ML, and HCI. For this question you should make the graph undirected (to do this interpret each directed edge as undirected). Details of the graph are in “citeseer.cites” while node categories are in “citeseer.content”. Please review the README to understand this dataset. Note: there are some nodes in “citeseer.cites” that are not present in “citeseer.content” and thus do not have an assigned category, please remove these nodes from the graph before answering any of the questions below.

3.1. **(1 mark)** Calculate and report: the number of nodes, the number of edges, the number of nodes in each category, the number of edges wholly in each category (to do this count the number of edges where both ends are in the category).

3.2. **(1 mark)** Does the CiteSeer for Document Classification graph exhibit homophily with respect to category? Calculate a relevant measure to support your statement. Also explain what the measure computes and why it can be used to detect homophily.

3.3. **(1 mark)** Does the CiteSeer for Document Classification graph exhibit homophily with respect to node degree? Calculate a relevant measure to support your statement. Also explain what the measure computes and why it can be used to detect homophily.