CSC303: A1

Due Feb 12 at 11:50PM, EST

Be sure to include your name and student number with your assignment. All assignments are to be submitted on Markus.

You will receive 20% of the points for any (sub)problem for which you write I do not know how to answer this question. You will receive 10% if you leave a question blank. If instead you submit irrelevant or erroneous answers you will receive 0 points. You may receive partial credit for the work that is clearly on the right track.

Question 1: (10 Points) Let $A(G)$ be the adjacency matrix of a directed unweighted graph G. For a fixed set of nodes V, assume we have the sequence of graphs $G_0 = (V, E_0), G_1 = (V, E_1), \ldots G_N(V, E_N)$. Let $B(k) = A(G_0) \times A(G_1) \times \cdots \times A(G_k)$. Therefore, $B(1) = A(G_0) \times A(G_1)$.

(a) [5 points] What is the meaning of $B(1)_{ij}$? (That is, the (i, j) entry in the matrix $B(1)$). Assume the array is 1-indexed. Justify your answer.

(b) [5 points] Suppose that $V = \{v_1, v_2, \ldots v_n\}$ for $n > 2$.

   (i) Further suppose that $E_i = \{(v_{[1+(i \mod n)]}, v_{[1+(i+1 \mod n)]})\}$. What is the smallest value of k such that $B(k)_{1,n} \neq 0$? Briefly justify your answer.

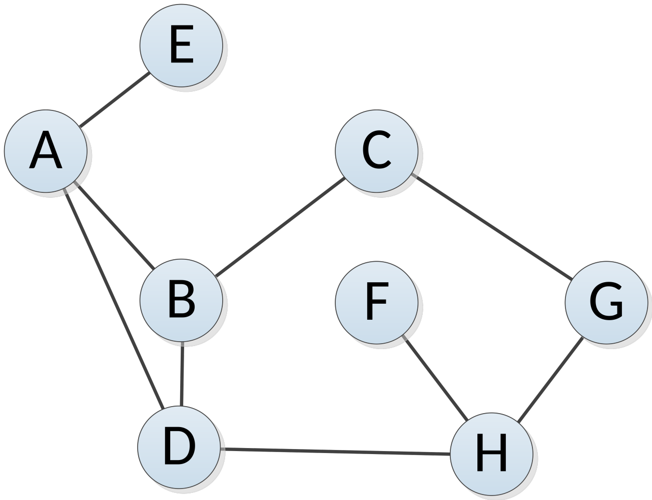
   (ii) What if $E_i = \{(v_{[1+(i \mod n)]}, v_{[1+(i+1 \mod n)]}), (v_{[1+(i \mod n)]}, v_{[1+(i+2 \mod n)]})\}$? Briefly justify your answer.

Question 2: (20 points) In class, we learned the definitions of both the clustering coefficient, and of a local bridge.

(a) [15 points] For any constant $c > 0$, is it always possible to construct a graph $G = (V, E)$ such that there exits some node $A \in V$ where A has a clustering coefficient that is less than c, and A is **not** adjacent to a local bridge (i.e. for all $(A, B) \in E$ then (A, B) is not a local bridge)? If yes, describe how to create such a graph and briefly justify why the clustering coefficient can be made to be arbitrarily low. If no, briefly justify why it is impossible.

(b) [5 points] In class we learned about membership closure and about the dispersion of an edge. How are these ideas similar? How are these ideas dissimilar? You can assume that focal closure is also occurring.

Question 3:(30 points) This question concerns the strong triadic closure property. Consider the graph below.



(a) [5 points] Suppose edge (D, H) is a strong edge. Label the remaining edges so as to maximize the number of strong edges (equivalently minimizing the number of weak edges) while satisfying the strong triadic closure property.

(b) [5 points] Briefly describe how you went about labeling the graph once the edge (D, H) was labelled as being strong.

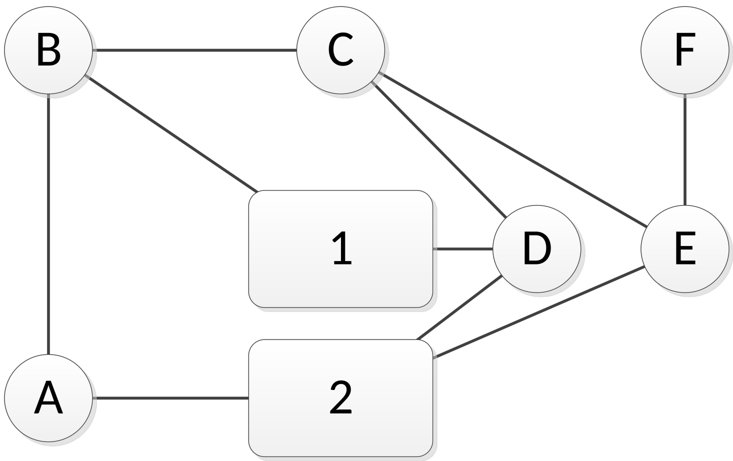
(c) [5 points] Now suppose edge (D, H) is a weak edge. Label the remaining edges so as to maximize the number of strong edges while satisfying the strong triadic closure property.

(d) [10 points] You are now told that the graph has two communities: $C_1 = \{A, E, B\}, C_2 = \{C, G\}$. Using this information, walk through the Rosenstein algorithm. Walk through each edge considered by the algorithm and write the number of STC violations caused by the edge currently under consideration. If there is a tie for which edge to consider next, then use the edge with the earliest endpoint in alphabetical order (if this leads to another tie, then tie-break by the other endpoint, again by earliest alphabetical order). What is the final set of strong edges? What is the final number of STC violations in the graph?

(e) [5 points] Are any of the edges bridges, or local bridges? If so, list these edges and their spans.

Question 4:(35 points) Consider the following graph. It is a social-affiliation network of students on day 0.

Each day, triadic, focal, or membership closures can occur. Given an open triangle ($e_{a,b}, e_{a,c}$) then the probability that a new connection ($e_{b,c}$) forms due to this triangle is 0.5 for triadic closure, 0.2 for membership closure, and 0.1 for focal closure. If a missing edge can be created by multiple closures, then assume that each potential closure works independently.

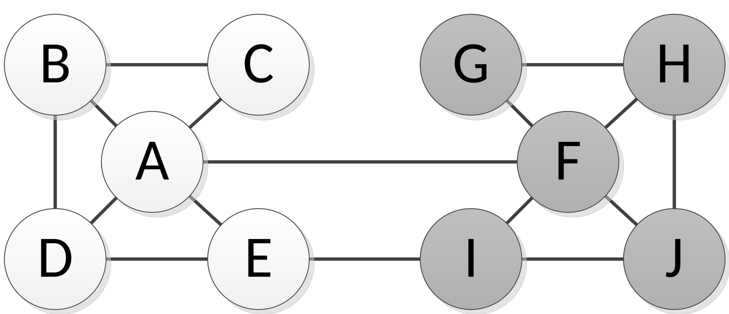


(a) [15 points] For each edge that could be created during the day, list the type(s) of closure(s) that would produce this edge (i.e. triadic, focal, or membership closure) and the probability that it occurs.

(b) [15 points] What is the probability that the nodes A and F are friends on day 2? Justify your answer.

(c) [5 points] You decide to start a new club, and are aiming to make everyone members – however, at the start of day 0 you can enroll only one person into your new club. Assuming you are very lucky, how quickly could you enroll everyone into your club? Who would you have to enroll on day 0 to achieve this? Briefly justify your answer.

Question 5:(25 points) Consider the following social graph in which white nodes correspond to tennis players, and grey nodes correspond to golf players.



- [5 points] Is there evidence of homophily among golf and tennis players? Give a detailed *quantitative* justification for your conclusion.
- [20 points] Compute the betweenness of the edges (A, F) and (E, I). Justify your answer.

Question 6: (20 points)

The following question requires you to use the NetLogo software package. I strongly recommend running it on a teach.cs machine (i.e. a CDF machine) with the command `netlogo`. Please ask TAs this week if you are having trouble with Netlogo.

If you are having troubles running netlogo remotely, then you can locally install the same version of netlogo from `http://ccl.northwestern.edu/netlogo/4.1.3/`. Despite the warning, it seems to work with OpenJDK 11.

Start Netlogo and load the Segregation model from the SampleModels/SocialScience Library. This implements a version of the Schelling model discussed in class. Note that there is a slight difference, instead of X agents desiring at least n of their neighbours to also be X, in this variant X agents desire at least $n\%$ of their non-empty neighbours to also be X. This has no significant impact on the observed trends.

We would like you to run *three* simulations of the Segregation model setting the parameters as follows: consider two different numbers of agents, 1200 and 2400; and consider five settings of the threshold variable (or "% similar-wanted" as it is called in the software), 20%, 30%, 50%, 70%, and 80%. Notice that you have ten combinations of settings, and must run three simulations for each. (You can set the speed faster to ensure each simulation proceeds quickly, or slower if you want to watch the patterns emerge).

For each simulation, record the final "% Similar" once the simulation converges (when all agents are happy) and the number of rounds of movement, or "Ticks" required. For each of the ten combinations of settings, report:

(i) the average (over the three simulations) of "% Similar" value and the "Ticks" value at convergence in the table provided;

(ii) the minimum value observed over the three simulations; and

(iii) the maximum value.

Please hand in the table on the final page of the assignment with these values to make marking easier.

On the basis of your observations, draw some qualitative conclusions about the impact of the number of agents and the similarilty threshold on the final degree of population homogeniety and the time taken for the Schelling model to converge. Provide possible explanations for these observed patterns. What is the general trend as the desired % similarity increases? Are there any violations of this trend, and if so, why?

NOTE: It is likely that for the setting $N = 2400$ and $t = 80\%$, the simulation will not terminate. For any setting that does not terminate, indicate for how long it ran (the tick counter is at the top of the display), and what conclusions, if any, can be observed from the plots provided by netlogo. When the desired % similarity is high, you may wish to increase the simulation speed.

	$N = 1200$		$N = 2400$	
	%-Sim	Ticks	%-Sim	Ticks
$t = 20\%$	Avg. Min. Max.	Avg. Min. Max.	Avg. Min. Max.	Avg. Min. Max.
$t = 30\%$	Avg. Min. Max.	Avg. Min. Max.	Avg. Min. Max.	Avg. Min. Max.
$t = 50\%$	Avg. Min. Max.	Avg. Min. Max.	Avg. Min. Max.	Avg. Min. Max.
$t = 70\%$	Avg. Min. Max.	Avg. Min. Max.	Avg. Min. Max.	Avg. Min. Max.
$t = 80\%$	Avg. Min. Max.	Avg. Min. Max.	Avg. Min. Max.	Avg. Min. Max.