CSC303: A1

Due Feb 16th at 11:59PM

Include your name and student number with your assignment, **on the last page**. 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." If instead you submit irrelevant, erroneous, or blank answers then you will receive 0 points. You may receive partial credit for the work that is clearly "on the right track."

A latex file of this assignment is available on Quercus.

Note: Yed graph editor is a free, relatively simply, multiplatform, graph editor `https://www.yworks.com/products/yed`

Question 1: (15 points) Let $A(G)$ be the adjacency matrix of a simple undirected graph $G = (V, E)$. Let $B(G) = A(G) + I$. Let $C^{(k)} = B(G)^k$, i.e.:

- $C^{(1)} = B(G)^1 = B(G)$
- $C^{(2)} = B(G)^2 = B(G) \times B(G)$
- $C^{(3)} = B(G)^3 = B(G) \times B(G) \times B(G)$

(a) [10 points] What holds in G *if and only if* $c_{ij}^{(k)} > 0$? (That is, whether the (i, j) entry in the matrix $C^{(k)}$ is strictly positive). Justify your answer; be rigorous and make sure your proof shows both directions of the iff. [HINT: you may want to start with $B(G)^2$]

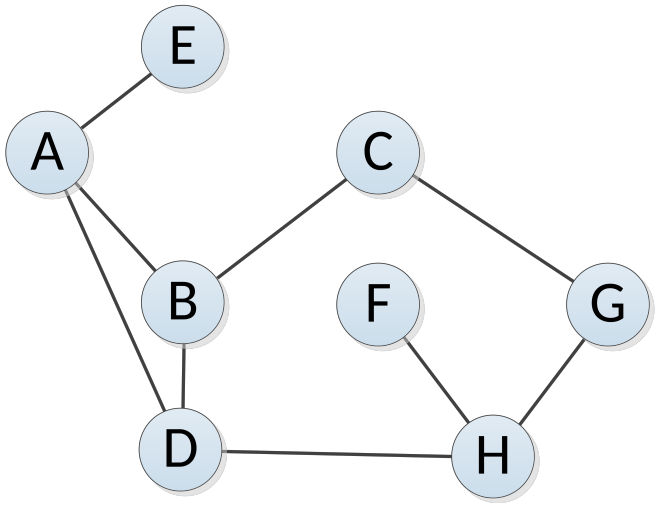
(b) [5 points] Given all evaluations of $B(G)^k$, how can you use this to determine if a graph has more than one connected component, in $O(|V|)$ time? What values of k does your approach require? Note that you do not need to return the components, only determine whether or not there is only a single connected component. Briefly justify.

Question 2:(20 points) You are given an edge weighted graph $G(V, E, w)$, where all edges are labelled as strong or weak. This graph does not obey strong triadic closure.

(a) [15 points] One of your classmates claims that after they added edges to G so that it obeyed STC, the clustering coefficient of one of the nodes has decreased. Is this possible? If no, provide a proof. If yes, provide an example and explain how it meets the conditions. Note that as this node's clustering coefficient has allegedly decreased, it could never have been undefined.

(b) [5 points] After thinking for some time, another classmate claims that if a node has a clustering coefficient of zero, then all edges involving the node must be local bridges. Is this correct? If no, provide a counterexample. If yes, provide a proof. Note: a clustering coefficient of zero is *not* undefined.

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 Rozenshtein 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:(20 points) Assume we have have a node-weighted undirected graph $G = (V, E, w)$ with N nodes (i.e., $N = |V|$). In this graph, all nodes $v \in V$ are either red ($w(v) = R$) or blue ($w(v) = B$). There are N_R red nodes, and N_B blue nodes. Therefore $N = N_R + N_B$. The graph contains K edges (i.e., $K = |E|$).

(a) [5 points] Assume that we create the multiset E_K, containing K possible edges created by selecting endpoints from V, uniformly at random. For simplicity, you are allowed to form self-loops (e.g., $e = (v_i, v_i)$) and you are allowed to include the same edge more than once (e.g., edges $e, e' \in E_K$ such that $e \neq e'$ and both edges have the same endpoints; i.e. E_K is a multiset).

Let the random variable X be the number of edges in E_K (*including* any duplicate edges) that are between two nodes of different colour. What is the probability that $X = k$ for $0 \leq k \leq K$? (HINT: you've seen this distribution in your prerequisite statistics course! If you are rusty, then feel free to consult your notes/textbook from your past statistics class.)

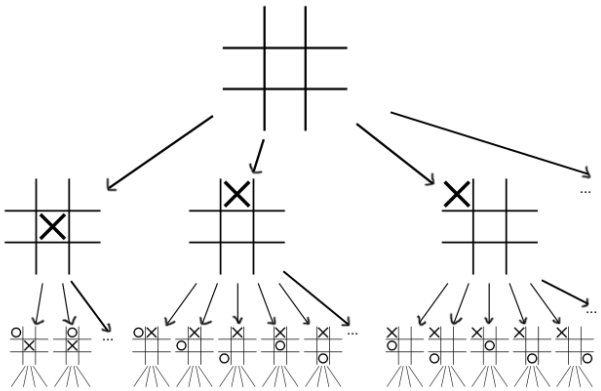
(b) [5 points] Assume that the observed number of cross-community edges in G is very probable under the distribution of our random variable X. In other words, if k_G is the number of cross-group edges in G, then assume that $P(X = k_G)$ is high. Would this be evidence for homophily with respect to the colour of nodes?

(c) [5 points] Now assume that G be a clique (i.e., a graph containing all possible edges). Let $p := N_R/N$ and $q := N_B/N$ (i.e., p is the proportion of nodes that are red, and q is the proportion of nodes that are blue). What is the proportion of cross-community edges in G with respect to p and q? Briefly justify your answer.

(d) [5 points] Your answer for the proportion of cross-community edges in a clique should not equal $2pq$. Is your answer evidence against homophily in G? What are some of the problems with our proposed homophily test? You should have at least 2 problems. Explain [Note that there are many correct ways to answer this question].

Question 5:(5 points) In this question, we will define a type of game tree. For a given game with deterministic moves and a fixed starting state, we will create the game tree graph by creating one node for every possible valid state in the game (i.e., any state reachable by a sequence of valid moves starting from the starting state), and drawing a directed edge between a pair of nodes A and B if there is a valid move in the game that transitions the game from state A to state B.

For the game tic-tac-toe, a subgraph of the full game tree is drawn below. You can assume that the X player will always go first. We will consider the game state to consist of only the contents of the board at a given point in time (i.e., we consider the tic-tac-toe state to consist of only the contents of each of the 9 spaces on the board).

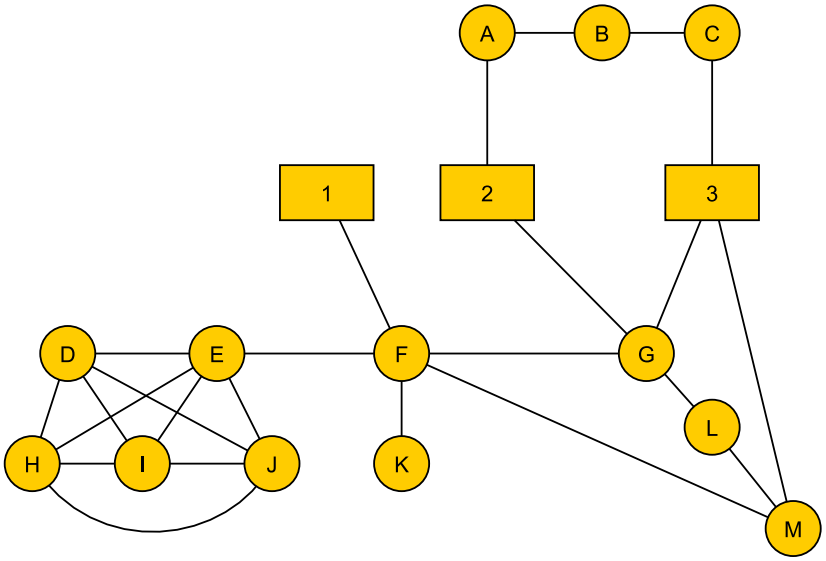


(a) [5 points] Under the previous definition, is the full tic-tac-toe game tree a rooted tree? Justify. State any necessary assumptions you make.

Question 6: (30 Points) Considering the following social-affiliation graph. This graph represents the friendships and affiliations of a small group, on day 0. Assume that each day, t, triadic, focal, or membership closures can occur – assume that the graph does not change until day $t+1$ at which point all relationships are updated at once.

Assume that triadic closures occur with probability 0.5, membership closures with probability 0.4, and focal closures with probability 0.2.

If a missing edge can be created by multiple closures, then assume that each potential closure works independently.



(a) [10 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. You only need to list edges that directly involve L or 2 (i.e., L or 2 is one of the endpoints), or are potentially caused by L or 2 (i.e., L and 2 are not an endpoint of the edge of interest, but is part of a closure that could create the edge of interest).

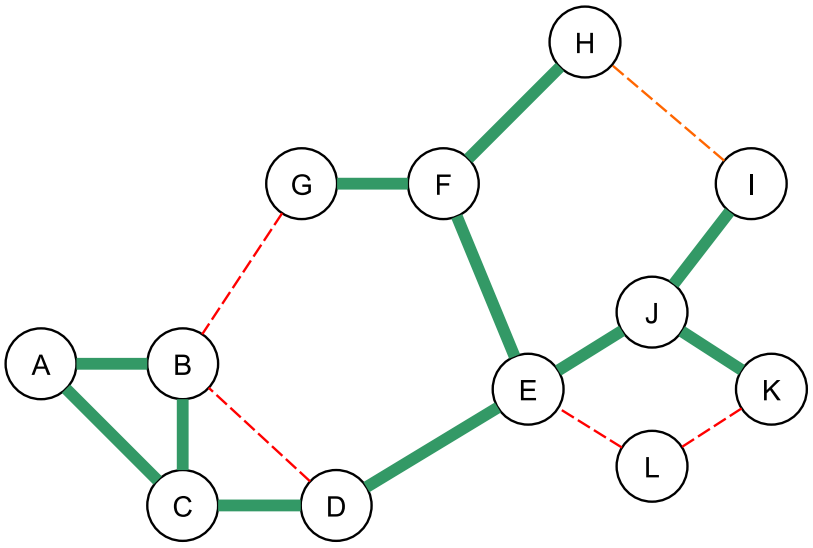
(b) [5 points] In class and tutorial, we discussed how studies suggest that influence is a factor in mutable factors such as obesity. In principle, we could apply a social-affilation network such as the above to model obesity (obesity would be an affiliation), but this is problematic as the model does not allow for the dissolution of edges – i.e., obese nodes can never get back into shape. To resolve this, we propose adding a probability, p, with which membership in the obesity affiliation can decay every timestep (for simplicity, you can assume that you cannot create and dissolve and edge in the same timestep). What new limitation does this change cause? How could we fix this? [This question can have several answers]

(c) [5 points] Which edges in the graph may form in a single timestep, are between two people, and could not be predicted using only the social network (as opposed to the full social-affiliation network)? Briefly justify why these are the only possible such edges.

(d) [5 points] Give an example of a node in this network with bonding capital, and an example of a node in this network with bridging capital. Briefly justify.

(e) [5 points] Over a sufficiently long period of time, what will this model converge to? (you do not need to justify this answer). Suggest a way to modify the model so that the behaviour is less unrealistic.

Question 7: (10 Points) Consider the following signed network where a solid green edges denotes friendship (an edge labeled +) and a red dashed edges represents enemies (an edge labeled −).
Specify the minimum number of edge signs (if any) that need to be changed so that the network can be completed to a strongly balanced network. Provide an explanation as to how the initial or modified network can be completed into a strongly balanced one.



Question 8: (15 points)

The following question requires you to use the NetLogo 6.2.2 software package. I strongly recommend running it on a teaching lab machine with the command `netlogo`. Please ask TAs this week if you are having trouble with Netlogo.

Start Netlogo and load the Segregation model. Go to File, Models Library, and select SampleModels/SocialScience/Segregation. 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. If they have no none-empty neighbours, they are also satisfied. This has no significant impact on the observed trends.

You can remotely access the teach.cs machines by following any of these instructions: `https://www.teach.cs.toronto.edu//using_cdf/x2go.html`, `https://www.teach.cs.toronto.edu//using_cdf/rdp.html`, or `https://www.teach.cs.toronto.edu//faq.html#HOWTO9`

Note that the simulation *technically* can be run in a web browser but this is very slow! I do not suggest it, as the runs will likely take an unreasonably long time to converge.

In the model, all individuals are blue or orange. Happy individuals are represented by squares, and unhappy individuals are represented by X's.

We would like you to run *three* simulations of the Segregation model setting the parameters as follows: consider two different densities, 70% and 95%; and consider four settings of the threshold variable (or "% similar-wanted" as it is called in the software), 25%, 50%, and 75%. Notice that you have six 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 six 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 five 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.

NOTE: For any setting where the model does not converge within 5000 ticks (the tick counter is at the top of the display), indicate for how long it ran, and what conclusions, if any, can be observed from the plots provided by netlogo. When the desired % similarity is high, you will want to increase the simulation speed.

	density = 70%		density = 95%	
	%-Sim	Ticks	%-Sim	Ticks
$t = 25\%$	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.
$t = 75\%$	Avg. Min. Max.	Avg. Min. Max.	Avg. Min. Max.	Avg. Min. Max.

END OF ASSIGNMENT 1

If you are typsetting the assignment using the provided LaTeX, then please write your name and student number below.

NAME: Your name should go here, on the last page.
STUDENT NUMBER: You student number should go here, on the last page.