MIDTERM EXAMINATION
Networked Life (MKSE 112)
October 18, 2012
Prof. Michael Kearns

This is a closed-book exam. You should have no material on your desk other than the exam itself and a pencil or pen. If you run out of room on a page, you may use the back, but be sure to indicate you have done so.

Name: ____

Penn ID: ____

Problem 1: ________/15

Problem 2: ________/15

Problem 3: ________/10

Problem 4: ________/15

Problem 5: ________/15

Problem 6: ________/20

Problem 7: ________/10

TOTAL: _________/100

Problem 1 (15 points) For each of the following statements, simply write "TRUE" or "FALSE".

a. There always exist networks with arbitrarily large population, maximum degree 3, and diameter 6.

FALSE

b. The Erdos-Renyi model of network formation will generate high clustering coefficients if the edge density is large enough.

FALSE

c. Preferential Attachment will always generate connected networks.

TRUE

d. The current average-case diameter of the Facebook social graph is less than 3.

FALSE

e. All else being equal, adding long-distance edges to a network will increase contagion.

TRUE

f. Having a clustering coefficient of at least 0.25 is a monotone property.

TRUE

g. Having a clustering coefficient of at least 0.25 has a tipping or threshold point.

TRUE

h. There are networks in which every vertex has the same degree.

TRUE

i. The sum of the degrees in a network must always equal the number of edges.

FALSE

j. "The Tipping Point" describes how Hush Puppies spread virally due to an email marketing campaign.

FALSE

k. For the clustering coefficient to be large, the overall edge density must also be large.

FALSE

l. C. Elegans is the name of the author of a famous paper on electricity networks.

FALSE

m. Adding random long-distance edges to a grid network will reduce the diameter.

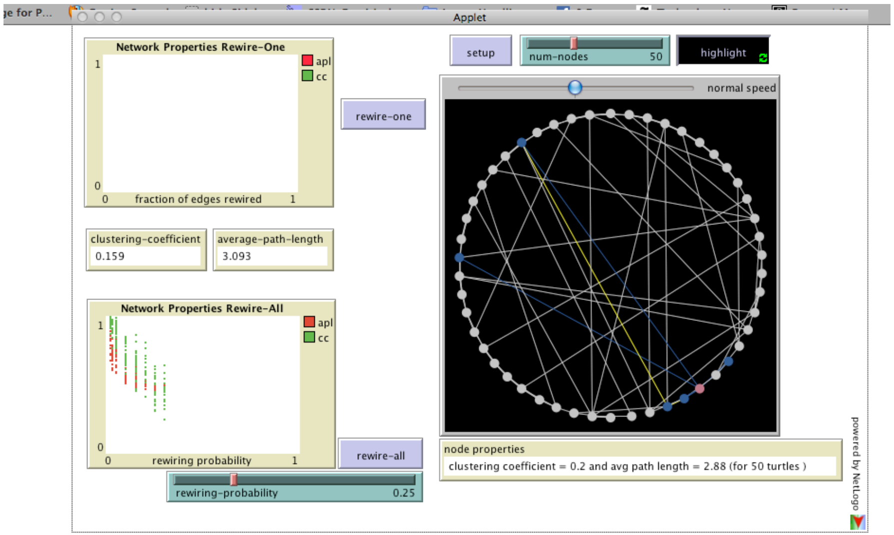
TRUE

n. The Erdos Number of most mathematicians is in the range of 5 to 10.

FALSE

o. The distribution of U.S. city sizes obeys a normal or Gaussian distribution.

FALSE



Problem 2 (15 points) The screenshot above is from one of the demos we examined in class.

(a) As precisely as possible, describe the mathematical model of network formation underlying this demo, including a description of how the rewiring probability affects the networks generated.

For this part, you should literally describe the model: you start with a cycle with connections two hops in each direction, and the rewiring parameter determines the probability with which each original edge is replaced with a random edge. Nothing else is required for this problem. If you describe the overall purpose of the demo but not the formation model, 0 credit.

(b) What are the red and green dots measuring?

The red dots measure, for each random network generated at a given rewiring, the average-case diameter or "path length", and the green dots the clustering coefficients.

(c) What is the primary point the demo is illustrating?

That while both quantities are decreasing with rewiring, diameter is falling much faster, so there is a "sweet spot" where high clustering and low diameter occur simultaneously. They must get this last point for full credit.

(a) For full credit these three points are required :

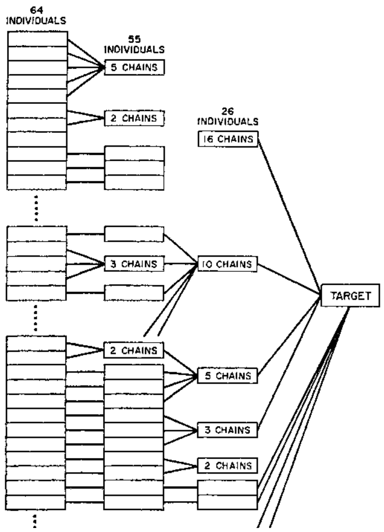
- *Specifying formation - 1 point*
- *Specifying starting configuration - 2 points*
- *Discussing Rewiring parameter - 2*

(b)

- *Specifying what red dots measure- 2.5*
- *Specifying what green dots measure - 2.5*

(c) For full credit all three observations should be specified:

- *Both quantities decrease - 1*
- *Greater decrease for diameter - 2*
- *There is a sweet spot of high clustering and low diameter - 2*



Problem 3 (10 points) The image above is taken from one of the assigned readings and was discussed in one of the lectures.

(a) Briefly but precisely describe the topic of the paper.

This is from Travers and Milgram; you should describe the chain-letter experiment and navigation in social networks.

(b) Briefly but precisely describe the primary point this particular diagram is making. Clearly circle that part of the diagram that emphasizes this point most strongly.

The diagram is showing the different paths to the target that the 64 completed chains took. The main point is that there was a single individual who was the penultimate step for 16 of the 64 chains, indicating that this person had a special role in the navigation process and is perhaps a hub/connecter/high-degree individual. The circle should be around the rectangle that says "16 chains".

(a)

- *Identifying Travers and Milgram experiment - 0.5*
- *Describing the chain letter experiment - 1.5*
- *Describing the import in terms of navigation in networks/degrees of separation/small world hypothesis -3*

(b)

- *Identifying the pattern of convergence - 2*
- *Significance of the penultimate link and identifying connectors- 2*
- *Specifying the hub correctly (encircling the rectangle with 16 chains) - 1 (Partial credit if the entire or part of the penultimate stage has been encircled but mostly no credit if only the rectangle with 5 or 10 chains has been pointed out)*

Problem 4 (15 points) Recall the tennis-ball exercises we performed on the first day of class, and again a few weeks into the semester.

(a) Briefly describe the nature of this exercise, i.e. what we actually did.

There were a couple of variants, but broadly they all involved picking a source and target individual, asking them to say something about themselves (e.g. hometown and some hobbies), then asking everyone who knew them to stand up and throw the ball to one of them etc., and seeing how long it took to get to the target.

(b) Briefly describe what the exercise is meant to investigate, and relate it to subsequent readings in the course.

Obviously the main point is navigation in networks --- finding short paths from only local information and local forwarding in the network. You should at least mention Travers and Milgram, and ideally also the Columbia paper and possible "The Tipping Point".

(c) Briefly describe how we computed shortest paths in the exercise, and why doing it the same way in the real world would be difficult.

Here you should recall that after doing the navigation, Professor then had the original source stand up, followed by everyone they knew, followed by everyone who knew someone standing up, etc. with everyone STAYING up until the target stood. You should NOT confuse this with the navigation exercise itself. The reason this would be difficult in the real world is that we are essentially doing a "multicast" here, and doing it on the scale of something like FB would be hard

(a) Any variation of the exercise discussed in class got full credit. Points were taken off for answers that only said we were trying to simulate navigation in networks or trying to find the shortest path but did not mention how we went about doing it. The ball-passing exercise needed to be described.

(b) For full credit, the answer needs to mention all of the below in some form or shape:

- *Navigation in networks*
- *Finding shortest path*
- *Using only local information/local forwarding*
- *Related readings - Travers and Milgram, Columbia paper, Tipping point*

(c) For full credit the answer needs to:

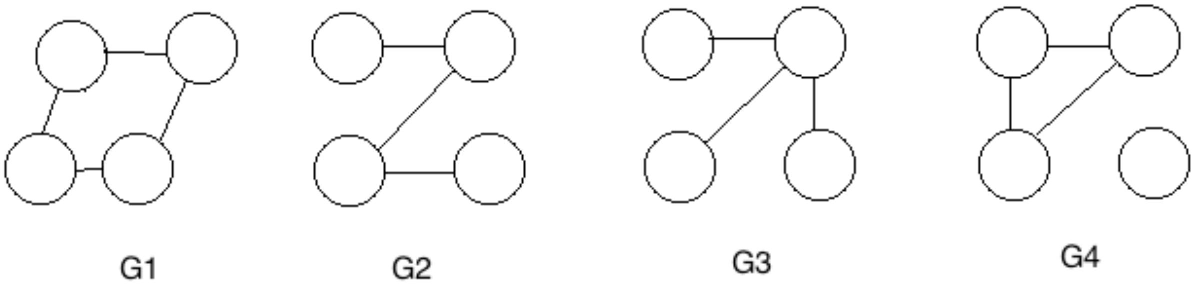
- *Mention that everyone who stood up had to remain standing till the target was reached (Points have been taken off if this wasn't clear)*
- *Bring up the question of scale in real networks in some form (Just saying it is not feasible or 'everyone can't be asked to stand up together' is not sufficient).*

.

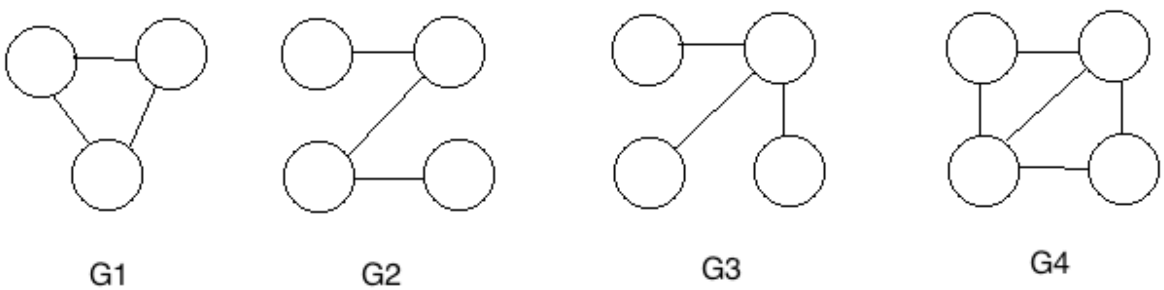
Problem 5 (15 points)

There is no need for any calculations here, and grades are based on whether you get the answers right. (i.e. there is no partial credit for this question.) Each part has 5 points, and the answer to part a,b and c is G4, G1, G4 respectively.

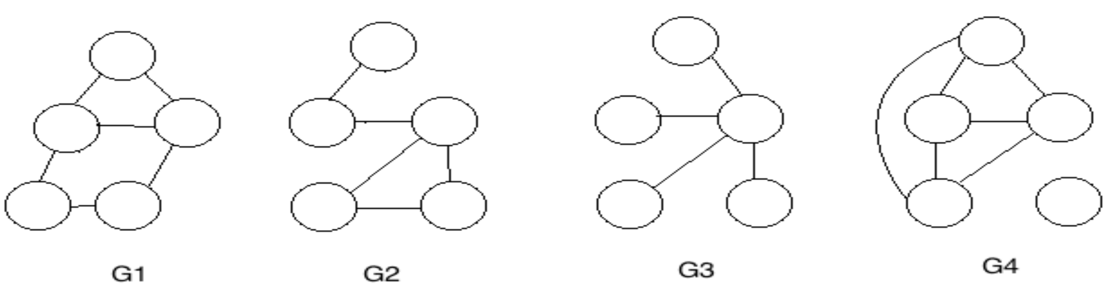
(a) Which of the following networks has the largest clustering coefficient?



(b) Which of the following networks has the largest edge density?



(c) Which of the following networks has the largest clustering coefficient *relative to its edge density?*



Problem 6 (20 points) Suppose we think of the edges in a network as representing communication links, and imagine there is an adversarial party who might destroy or take over certain vertices in the network. For instance, the network might be the Internet and a terrorist organization might be able to compromise certain computers in the network. Consider two types of attack: one in which the adversary is able to *choose* a small number of vertices to destroy, and one in which the adversary is only able to destroy a small number of *randomly* chosen vertices.

(a) Suppose the network is generated according to the Erdos-Renyi model. Discuss the vulnerability of the network's global connectivity to each type of attack.

Remember Professor Kearns said during the exam this is only about connectivity (really, connected components), not contagion of any kind. For part (a), the expected answer is that there is no real difference between the two kinds of attack, and that (assuming there is some minimal edge density in the first place), the network should remain connected even if a few vertices are knocked out or deleted. More generally, as long as there is a giant component, that component should remain connected. Since most/all degrees in E-R are roughly equal, there shouldn't be vertices that are more vulnerable or cause more damage than others, so random and worst-case attacks should be about the same. If your discussion is along these lines, you are given full credit.

(b) Suppose the network is generated according to Preferential Attachment. Discuss the vulnerability of the network's global connectivity to each type of attack.

In contrast, here we have PA, which first of all, at least as described in class, generates trees, which are of course vulnerable to being disconnected. But even if you don't say/see that point, the main point is that in PA there is a big difference between the two types of attack due to the different degree distributions. Because the heavy-tailed degree distribution , a random or typical vertex will have very low degree, so the random attack will leave most of the network still connected. But a worst-case attack would target the high-degree vertices or hubs, which will shatter the network into many small components.

Each part has 10 points. We considered 5 points for discussing each type of attack. You must have provided both the correct answer (2 points for each attack per model) and the proper discussion of it (3 points for each attack per model).

Some of you have just compared the attacks with each other and it is not clear how devastating each of them alone is, so 1-2 points are taken from you depending on the rest of your answer.

If your argument is not sound you are given partial credit; however, if it is completely wrong or irrelevant, there is not partial credit. If you have described what the E-R or PA models and their properties are, and this information is relevant to the correct answer and shows your understanding of the problem/models, you are given partial credit (1-3 points depending on the context).

Note that if you said anything that shows you did not understand the models or problem, some points are taken from you (1-4 points depending on the severity of the mistake). Also this can affect your answer to other parts of the question and lower your grade even more.

Problem 7 (10 points) The following equation is taken directly from one of the lectures:

$$R(p) = \sum_{q \in POINTS(p)} R(q) / out(q)$$

Succinctly and precisely describe what this equation is for. Be sure to say what p, q, R(p), POINTS(p) and out(q) are, and what the name of the equation is.

The intended answer is as follows:
This is the PageRank equation. (2 points)
p and q are vertices (or websites), (1 points)
R(p) and R(q) are their ranks, (1 points)
out(q) is the outdegree of vertex q, (1 points)
and POINTS(p) are the vertices pointing to p. (1 points)
The equations says that each page pointing to p gives p its "share" R(q)/out(q) of q's ranks, and summing over all such q determines R(p). (4 points)

Some of you forgot to mention some of the variables, and so you did not get the corresponding grade. Some have provided examples of page rank calculation. If the examples illustrates the point, partial credit is given.

The most common reason that many of you did not get the full credit for this question is that instead of describing what the equation is for, a bunch of information about the pagerank algorithm was given, including how to run it, how it converges, who invented it, etc. You are given partial credit for that (0-3 depending on how relevant the information is and what the context --i.e. rest of your answer-- is.)

Note that again, if you have said anything that shows you do not understand the algorithm/equation, depending on the severity of your mistake, 1-2 points are taken from you. Also people who did not get the equation right, often lost the 4 points of description automatically.