**MIDTERM EXAMINATION**
**Networked Life (NETS 112)**
**October 22, 2015**
**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. You may also make annotations directly on any diagrams given.***

**Name: Paul Erdos**

**Penn ID: 112**

**Problem 1: _____/10**

**Problem 2: _____/15**

**Problem 3: _____/10**

**Problem 4: _____/10**

**Problem 5: _____/20**

**Problem 6: _____/20**

**Problem 7: _____/15**

**TOTAL: _____/100**

**Problem 1 (10 points)** Clearly answer "True" or "False" for each of the following assertions. (Rohan)

a. In one of the course readings, we saw evidence that the diameter of the Facebook graph has increased dramatically with the number of users.

False: The diameter of the Facebook graph has actually decreased slightly over time as each new user tends to create a large number of friendships that increase the connectivity of the graph.

b. In the mathematical collaboration network, there are vertices that do not lie in the giant component.

True: Consider, for example, two authors who write a paper together and then never write a paper again - they will not lie in the giant component.

c. In most real-world, large-scale social networks, the number of edges actually present grows more rapidly than the number of vertices.

True: The number of possible edges grows as a function of n^2. Furthermore, if actual edges represent relationships, each new person added to a graph will almost certainly create more than one new edge. More likely is that each new vertex will create many relationships, which means that the number of actual edges will naturally grow faster than the number of vertices.

d. If a network has a clustering coefficient much higher than the overall edge density, there must be distinct communities present.

True: The clustering coefficient captures the connectivity of different communities - if it is significantly higher than the background edge density (as it is in all real social networks), then there must be communities present.

e. In the giant component demo studied in class, the giant component emerges suddenly when the average degree is about the square root of the population size.

False: The giant component emerges at average degree = 1 (p = 1/n). If you didn't remember this, consider a real world example: the Facebook network has about 10^9 users - if this was true, the giant component would not emerge until the average person had 10^4.5 (about 32,000) friends.

f. The diameter of the giant component must always be finite.

True: Nodes in the same connected component must by definition have a path between them. This means that every pairwise shortest path is finite, because path lengths are only defined as infinite when no path exists.

g. In "Six Degrees", it is argued that clustering of connectivity appears in large social networks, but not in biological or physical networks.

False: Clustering naturally occurs in most large scale networks.

h. The property of a network not containing any cycles is a monotone property.

False: Consider a tree with exactly n-1 edges and no cycles. If we add one edge to this network, we will immediately form a cycle. In fact, any connected component with n vertices and >= n edges must contain a cycle. If we can add edges and 'break' some property of the network, the property is not monotone.

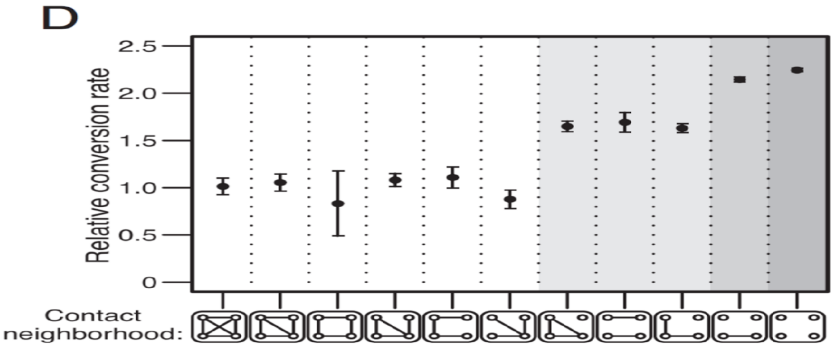
i. The only properties known to have a tipping point or threshold behavior in the Erdos-Renyi model are giant component and small diameter.

False: Any monotone property has a tipping point in the Erdos-Renyi model for network formation. We can think up as many monotone properties as we would like, but consider for example the property of a network containing a cycle of length 5.

j. The smaller components in the squash network were geographically diverse.

This question was sufficiently ambiguous that I accepted both answers. Within the smaller components, the vertices were not geographically diverse. Across the smaller components, the vertices were geographically diverse.

**Problem 2 (15 points)** The diagram below is taken from one of the assigned readings. Precisely describe the experiment in the article. Clearly explain what the x and y axes are showing or measuring, and discuss the result that the diagram is summarizing and why it is interesting.
(Chris)

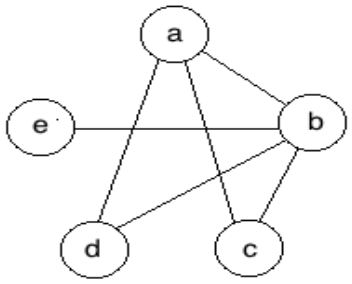



This diagram is taken from the article "Structural Diversity in Social Contagion." The goal of the experiment is to analyze the growth of Facebook. Facebook recruits new users by emailing them and showing that some of their real world friends are already using Facebook. The authors find that recruitment success is tightly controlled by the number of connected components in an individual's contact neighborhood (his friends in the email), rather than by the actual size of the neighborhood.

The x-axis shows the connectivity of the contact neighborhood of the recruited individual. The far left side represents a fully-connected contact neighborhood and the far right side represents a completely disconnected contact neighborhood. The y-axis represents the probability that the recruit joins Facebook.

The diagram shows that lower connectivity (greater number of connected components) among the friends in the email leads to higher recruitment rate, indicating that potential users are swayed by structural diversity.

**Problem 3 (10 points)** Let S be some set of vertices in a graph or network. Then the *subgraph induced by S* is the graph obtained by paying attention only to the vertices in S and the edges between them, and ignoring all other vertices and edges. For example, in the graph shown below, the subgraph induced by S = {a,b,c} is the triangle between those three vertices, and the subgraph induced by S = {c,d,e} consists of three isolated vertices.
(Chris)



Graph G

Now consider the specific 5-vertex graph shown above; let's call it H. Consider the following property of a graph G: "G contains H as an induced subgraph". This means that there exist 5 vertices in G whose induced subgraph looks exactly like H above.

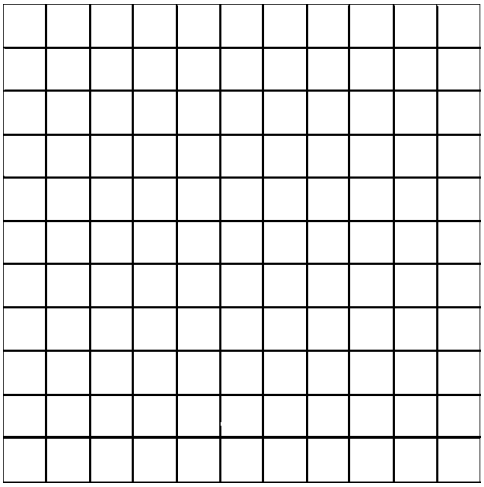
a. Is the property of containing H as an induced subgraph a monotone property? Why or why not?

The property of containing H is NOT monotone. Suppose we have the property (H is an induced subgraph) and we add one more edge. If the new edge is between any two of the vertices in H, then H will change into a different graph, and we will no longer have H as an induced subgraph.

b. Consider a graph G over N vertices that is generated according to the Erdos-Renyi model. If N is very large and we add enough edges, do you expect that at some point G will contain H as an induced subgraph? Why or why not?

If N is very large and we add enough edges, G will have a VERY large number of induced subgraphs. Specifically, any combination of 5 vertices in G form a subgraph, so the number of subgraphs is on the order of N^5. Because we have so many possible subgraphs to choose from, there is a high probability that at least one of them will look like H.

**Problem 4 (10 points)** Consider the 12 by 12 grid graph shown below, where there is a vertex at every corner or intersection point. Consider the process, discussed in class, of routing a message from vertex A to vertex B by always forwarding to the neighbor whose grid address is nearest to the destination (ties are broken arbitrarily).
(Rohan)



a. Exactly how many hops or steps will it take to route the message from A to B? Are there many possible paths the message might take or only one?

It will take exactly 11 hops to get from point A to point B. At each step, the message will be forwarded to a neighbor that is closer (by grid distance) to the target. If two of a node's neighbors are the same distance from the target it will be indifferent to where it forwards the message. Here the tie will be broken randomly (for example, A is initially indifferent as to whether it forwards the message up or to the right) so there are many possible paths.

b. Draw in new a "long-distance" edge added to the grid such that the shortest-path distance from A to B becomes as small as possible, but that the answers to part a. above are unchanged.

If we add an edge from the node directly below A or directly to the left of A that connects directly to B, we will reduce the shortest path between A and B from 11 to 2. However, a navigation algorithm that only has local information would never forward the message away from the target, and the nodes below and to the left of A are 12 hops away from B by grid distance. Thus, the message will still be forwarded along one of the 11 hop paths, and the answer to part a will remain unchanged.

**Problem 5 (20 points)** This problem refers to the assigned reading "Can Cascades be Predicted?", by Cheng et al.. which describes an attempt to predict whether a given piece of content posted on Facebook will "go viral".

a. (10 points) The article begins by discussing a technical difficulty with simply predicting whether a piece of content will reach a given number of reshares. What is this technical difficulty, and how do the authors propose getting around it?
(Rohan)

Virtually all posts on Facebook do not go viral, so a predictive model that simply guesses that every post will reach a very low number of reshares will be right 99.99%+ of the time. To get around this difficulty, the authors restrict their study to posts that reach some threshold of k reshares, and try to predict if each of these posts will eventually reach the median number of reshares f(k) for all posts that also reached k reshares. This controls for the fact that most posts don't go viral, allows the authors to study the life of a viral post and helps normalize the heavy tail distribution of reshares across posts.

b. (10 points) Briefly but clearly describe the approach the authors take to their problem, and summarize their main findings. Topics for discussion might include the performance the authors achieve and how it compares to the baseline, the various categories of "features" they introduce and what they measure, and the relative values these features seem to have and how it changes as the cascade grows.
(Chris)

The authors analyze one month's worth of Facebook photo reshare data by considering the predictive value of many different features (content, origins, network structure (structural) , time between reshares (temporal), etc.). They find that the temporal and structural features are key predictors of cascade size, while the origins of the content become less important as the cascade progresses. They also find that initial breadth (through a broadcast) rather than depth in a cascade is a better indicator of larger cascades.

The authors achieve strong performance (nearly 80% accuracy) in predicting whether a cascade will continue to grow in the future. Furthermore, the authors find that the growth of a cascade becomes more predictable as more of its reshares are observed.

**Problem 6 (20 points)**

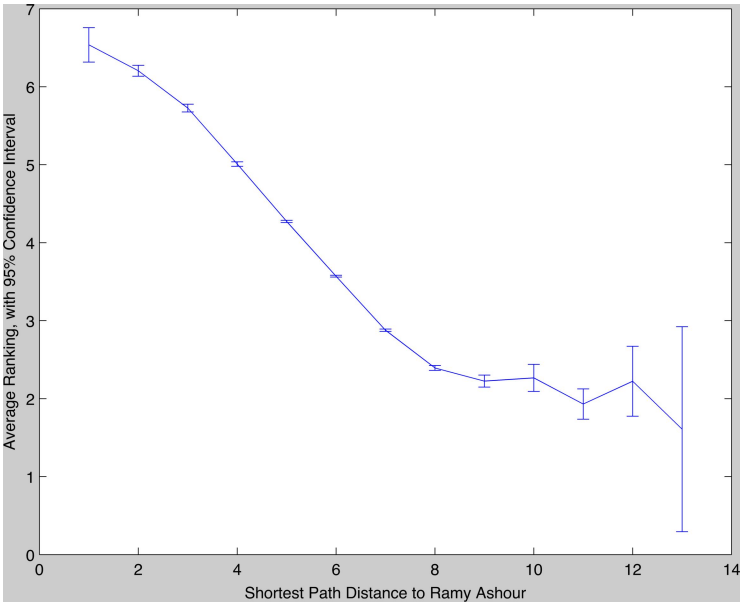
In class and the readings, we examined three different network formation models that will all yield a clustering coefficient higher than the overall edge density. Briefly but as precisely as you can, describe each model, and for each, say whether you think the model generates networks with clear "community" structure or not, and why.
(Rohan)

The first model that we studied was the alpha model. In the alpha model we examine each pair of vertices, and if they don't share any common neighbors we connect them with some background probability p. However, if they do share some fraction of common neighbors x, we connect them with the probability p + (x/N)^a, where a is a parameter that we can adjust. For any fixed a and N, the probability of connecting two arbitrary vertices increases as a function of x, the number of common neighbors they share. This introduces a bias towards connecting friends of friends, and thus creates a clustering coefficient higher than the background edge density. For smaller values of a (< 1) this bias is amplified, whereas larger values of a will do the opposite. For a = 1, the bias toward connecting friends of friends increases as a linear function of the number of neighbors two vertices share.

The second model that we studied was the (rewired) ring model. Here we have a ring where each vertex is connected to its immediate clockwise and counterclockwise neighbors, and also two its neighbors two hops away. Here each node has four neighbors, and those neighbors are themselves connected by 3 edges. There are four choose 2 = 6 possible edges, so each node has a clustering coefficient of .5. Because the network is perfectly symmetrical, the network clustering coefficient is also .5 (the average of each node's individual clustering coefficient). Because each node has a degree of 4, there are (4/2)*n = 2n total edges in this network. As always, there are n choose 2 = n(n-1)/2 possible edges, so the background edge density is 2n / (n(n-1)/2) = ~ 4/n. Clearly as n goes to infinity, 4/n goes to 0, but the clustering coefficient will remain constant at .5.

The third model that we studied (in lecture) is the community or coloring model. Here we partition the network into k distinct categories (thought of as colors or communities) and then run a modified Erdos-Renyi on the graph. When we examine two nodes of the same 'color', we connect them with probability p. When we examine two nodes of different 'colors', we connect them with probability q. Assuming p is significantly larger than q, this model will create highly clustered graphs where each of the k communities is densely intraconnected but sparsely interconnected.

**Problem 7 (15 points)** The image below is reproduced from one of the assigned articles, and was also discussed in lecture. Briefly but precisely describe exactly what the x and y axes are measuring, and what point the diagram is making.
(Chris)




This diagram is from Dr. Kearns's paper about the network of registered squash players in the US. The vertices in the network are squash players and two players are connected by an edge if they have played in a registered match against one another. This is reminiscent of the network of coauthorships among mathematicians.

The x-axis represents the shortest path from a squash player to Ramy Ashour, and the y-axis represents the average ranking of the squash players. The rankings are determined by the governing body, with a higher ranking indicating a better player.

The diagram is making the point that "Ashour number" is negatively correlated with the quality of the player. That is, on average, players that are closer to Ashour in the squash network tend to have higher rankings. This is not surprising, but it is still interesting to see because it confirms something that we might expect to be true about the network.