MIDTERM EXAMINATION
Networked Life (NETS 112)
October 1, 2013
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: _____/10

Problem 2: _____/20

Problem 3: _____/20

Problem 4: _____/10

Problem 5: _____/15

Problem 6: _____/25

TOTAL: _____/100

Problem 1 (10 points) Answer “True” or “False” for each of the following assertions.

a. The largest possible number of edges in an undirected network of N vertices grows roughly as N^2/2.
True. The largest possible number of edges is N(N-1) / 2 ~ N^2 / 2.

b. The distance between any two vertices is always less than the average-case diameter.
False. Note that the question specified "average-case" diameter.

c. Typically, in large-scale social networks, there are a reliable number of vertices with degree much higher than the average.
True. Most large-scale social networks exhibit a heavy-tail degree distribution.

d. If we remove an edge from a graph, the diameter might decrease.
False. Removing an edge can only increase the distance between two vertices.

e. In the forest fire demo, the more vertices we delete from the grid, the greater the fraction of the grid that will burn.
False. Deleting vertices turns them into parking lots, which decreases the extent of contagion.

f. In the viral spread demo, an infected vertex will always pass the infection to each of its neighbors.
False. Not always -- depends on the value of the "stickiness" parameter.

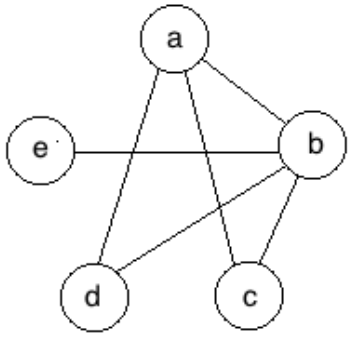
g. In the altruistic contagion model, if we add an edge between vertices A and B, the equilibrium wealths of vertices A and B will increase, and the equilibrium wealths of all other vertices will decrease.
True. The degree of A and B will increase, which increases their wealth. The sum of all degrees also increases (i.e. the denominator term), which decreases everyone else's wealth.

h. As long as short paths exist in a network, it is possible to solve the navigation or “small world” problem.
False. The existence of short paths is necessary, but not sufficient. An algorithm still has to be able to find those short paths.

i. It is impossible for a network to have 20 vertices, a maximum degree of 2, and a diameter of 3.
True. 20 > 2^3. Note that the question said "impossible."

j. In Kleinberg's model, we consider navigation to be efficient if the number of steps required is about the square root of N.
False. The number of steps needs to be about log(N) or less.

Problem 2 (20 points) Consider the network below.



Graph G

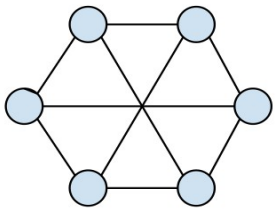
a. (5 points) Calculate the average-case diameter, which is defined as the average distance between pairs of vertices.

Sum of distances = 14
Number of pairs = 10
Diameter = 14/10 = 1.4

b. (5 points) Calculate the equilibrium wealth of each vertex in the altruistic contagion model considered in class.

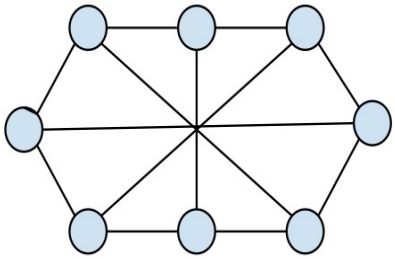
Wealth = (degree(vertex) / sum of all degrees) * total wealth
Sum of all degrees = 12
Total wealth = 5
Wealth(a) = 3*5/12
Wealth(b) = 4*5/12
Wealth(c) = Wealth(d) = 2*5/10
Wealth(e) = 1*5/12

c. (5 points) Draw a network with 6 vertices in which the maximum degree is 3 and the worst-case diameter is 2.



Partial credit was given for networks which met at least two of the requirements.

d. (5 points) Is it possible for a network to have more than 6 vertices and still have a maximum degree of 3 and worst-case diameter of 2? If so, draw such a network.



Partial credit was given for showing some work.

Many students applied the equation $N \leq \Delta^D$ to show that this should be possible ($6 \leq 3^2$), but this equation doesn't actually apply, since the question was asking about worst-case diameter. The equation tells you that an average-case diameter of 2 is possible, but this does not imply that a worst-case diameter of 2 is also possible.

Problem 3 (20 points)

a. (5 points) In the forest fire demo, what is the parameter that we varied between the different trials? What is the effect of this parameter on the extent of contagion?

We varied a *connectivity* parameter that determines the probability that a vertex will be forest. The higher this parameter, the greater the contagion. When we increase the probability that a vertex will be forest, we increase the expected size of a vertex's connected component, and therefore increase the expected amount of the grid that will burn.

b. (5 points) In the viral epidemic demo, what are the two parameters that we varied? What is the effect of each parameter on the extent of contagion?

We varied a *connectivity* parameter that determines the probability of rewiring local connections to random long-distance connections, and we varied a *stickiness* parameter that determines the probability that an infected vertex will pass on the infection to its neighbors. The higher the stickiness parameter, the greater the contagion – this relationship is fairly straightforward. Increasing the connectivity parameter also increases the contagion, to an extent – ideally you want some, but not all, connections to be rewired. This creates the necessary mix of local and long-distance edges.

c. (5 points) What do we mean by the "tipping" or "threshold" phenomenon that was observed in these demos?

We observe a tipping phenomenon when there is some value q of a parameter such that, below this value, contagion is very limited / contained, and above this value, contagion is nearly complete.

Some students stated that the tipping point is the point at which contagion is nearly complete. This is almost, but not quite, correct. The key point is that there is an *exponential increase* at the tipping point, rather than a *linear* increase. We would *not* observe a tipping phenomenon if contagion increased steadily and eventually affected the majority of the network.

d. (5 points) Compare and contrast the viral epidemic demo with Kleinberg's model. Consider both the network formation processes, and the network dynamics (i.e. the way in which information spreads through the network). Comment on the extent to which there is a "purpose" for the participants each dynamic.

Network formation: In the viral demo, we *rewire* edges from one vertex to a *random* destination vertex. In Kleinberg's model, we *add* edges from one vertex to a destination vertex based on the *distance* between the two vertices. Both models involve creating long-distance edges, but the main differences are the 1) rewiring vs. adding, and 2) distance is a factor in Kleinberg, but not in the viral epidemic.

Network dynamics: In the viral demo, contagion spreads *probabilistically* through the network. Vertices have no "purpose" or "choice" – they *passively* spread the contagion to their neighbors. In Kleinberg's model, information is spread with the purpose of reaching a target. Each vertex *actively* chooses a neighbor (the one geographically closest to the target) to pass the message to. Here, dynamics are controlled by an algorithm, whereas in the viral demo, dynamics are purely probabilistic.

Problem 4 (10 points)

In The Tipping Point, Gladwell gives some cases where people applied the ideas of the book to real situations. Choose something (a product, a trend, an idea, a practice, etc.) that you would like to see become more widely used or adopted, and describe how the ideas in the book could be applied to make this happen. Which of the "three rules of epidemics" according to Gladwell (the law of the few, the stickiness factor, the power of context) did you use?

Most answers that demonstrated an understanding of the reading were given full credit. Answers that did not use ideas from the book, but rather from class, were given partial credit, based on how "reasonable" they were.

A surprising number of students chose the payments app Venmo (created by Penn grads) as the product they would like to see become more widely used. I've never tried it before, but these students were quite effective salesmen, and convinced me to download it. I suggest that the entire class do so, and we can leverage the law of the few to turn Venmo into the next big thing!

Problem 5 (15 points)

Consider the in-class biased voting experiments.

High level note – some of you were not present in class for these experiments, and had trouble with this question. Note that summary slides were posted on the website, and you should have been able to answer (at least parts b and c) based on those slides. Class attendance is mandatory, and you are responsible for anything that happens in class, even if you are absent.

a) (5 points) Summarize the rules and incentives or payoffs for these experiments.

The goal of the class is to converge on one of two colors. Each student is given a preference color. If a significant majority of the class chooses one color, then anyone who chose that color gets points, as follows: if the color is your preference, you receive 2 points, otherwise 1. Anyone who chooses the minority color gets 0 points. If the class does not converge, no one receives points.

b) (5 points) In one of the experiments, the network was defined by connections between students of the same gender, and then there were 5 students who had connections to everyone. How did the existence of these connectors affect the dynamics of the experiment? What was the outcome of this experiment?

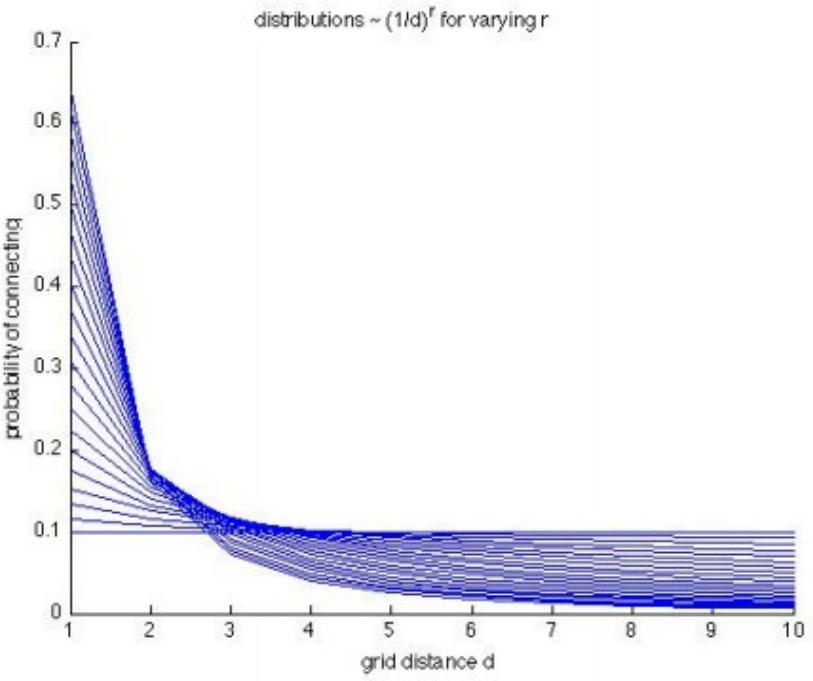
The connectors had a position of "power" or "influence" because they could communicate with everyone, and therefore had access to the most information. They facilitated the transfer of information between the two gender components, and therefore helped the class reach a consensus. Four of the five connectors had the same preference color, and managed to convince the class to converge to that color.

c) (5 points) In another one of the experiments, the network was defined by connections between students who were born in adjacent months. Furthermore, color preferences were chosen so that a student's preference was different from that of his neighbors. How did this network structure and preference rule affect the dynamics of the experiment? What was the outcome of the experiment?

The set-up of this experiment – namely, the lack of connectors / long-distance edges – made it very difficult for different parts of the network to communicate with each other. Information remained local, contained in "month" clusters, rather than spreading. In the end, most students simply chose their preference color, and consensus was not reached.

Problem 6 (25 points)

The image below is taken from the slides and illustrates the effect of *r* on the networks generated by the Kleinberg model.



a. (4 points) What is the main point that this graph is making about the effect of *r*?

Each line represent a different value of *r*. The flat line at 0.1 represents $r = 0$. As *r* increases, the lines get successively steeper. This illustrates the fact that, for small values of *r*, distance does not make a large difference in the probability of connection. We are about equally likely to make a long-distance connection as a local connection. But at large values of *r*, distance makes a big difference, and we are much less likely to make long-distance connection as compared to a local connection.

Many students thought that the graph was illustrating the fact that $r = 2$ is the only value that permits efficient navigation. Although this is a true fact, it is *not* illustrated by the graph. The graph is not plotting anything related to the efficiency of navigation.

b. (6 points) If *r* is too large, why might efficient navigation be difficult?

At large values of r, we are more likely to add local edges, rather than long-distance edges, to the network. The lack of long-distance edges makes it difficult to make "large hops" across the network. Indeed, short paths between vertices may not even exist.

c. (6 points) What about if *r* is too small?

At small values of r, we are equally likely to add long-distance edges as local edges. So there are a fair number of long-distance connections, and short paths do exist in the network. However, it may be hard to find these short paths using only local information. For instance, maybe we take a big hop towards the target, and end up in its general vicinity. But the short path requires us now to take a big hop *away*

from the target, in order to get back to it. Since we only have local information, we don't know this. So instead, we take lots of small steps to get the rest of the way there, and this is inefficient.

Suppose we are generating a network in Kleinberg's model, and we are about to add a long distance edge to vertex u. The distance between vertex u and vertex v is 5, and the distance between vertex u and vertex w is 10. Let p denote the probability that we add the u-v edge, and let q denote the probability that we add the u-w edge.

d. (3 points) If $r = 0$, what is numerical value of the ratio p/q?

p ~ (1/5)^0
q ~ (1/10)^0
p/q = 1

e. (3 points) What is the ratio if $r = 1$?

p ~ (1/5)^1 = 1/5
q ~ (1/10)^1 = 1/10
p/q = 2

e. (3 points) What is the ratio if $r = 2$?

p ~ (1/5)^1 = 1/25
q ~ (1/10)^1 = 1/100
p/q = 4

Networked Life
CSE 112
Prof. Michael Kearns
Midterm Examination
March 1, 2007

NAME: _____

PENN ID: _____

Exam Score:

Problem 1:_____/10
Problem 2:_____/15
Problem 3:_____/12
Problem 4:_____/12
Problem 5:_____/15
Problem 6:_____/15
Problem 7:_____/9
Problem 8:_____/12

TOTAL:_____/100

1. (10 points) Answer "True" or "False" to each of the following assertions.
One point each, no partial credit

a) The maximum number of edges possible in an undirected network of N vertices is N^2.

False

b) In Kleinberg's model for navigation in social networks, a value of r=2 is the only value that permits rapid navigation.

True

c) Corporate portals are an example of web pages that might be part of the component IN.

False

d) A hub is a page that points to a lot of other good hubs.

False

e) A chromatic number greater than 4 is an example of a monotone graph property.

True

f) A clustering coefficient greater than 0.4 is an example of a monotone graph property.

False

g) It is more "prestigious" to have a high Erdos number than a low one.

False

h) In the contagion or "random walk" model of economic exchange, each vertex's wealth will depend solely on their clustering coefficient.

False

i) When using a log-log plot, if the data appears linear, it is normally distributed.

False

j) The α-model of network formation was developed to provide a better explanation of naturally observed degree distributions than the Erdos-Renyi model.

False

2. (15 points) Consider the version of the Erdos-Renyi in which at each step, we choose two vertices that are not already connected uniformly at random, and then add the edge between them. Suppose that at some point during this process, there are K distinct connected components of size $C_1 > C_2 > C_3 > \ldots > C_K$.

5 points for each part; partial credit as described.

a) Explain why it is difficult for two very large components to coexist in this process.

By definition of a component, two different components have no edges between them. But if they are both very large, the number of possible (or missing) edges between them must be very large (on the order of the product of the sizes of the two components). Thus, it quite likely that a randomly chosen missing edge will connect the two components into one. This argument can be made mathematically precise, but full credit for saying something close to this line of thought.

b) What can you say about the relative likelihood of the growth of the different components? What general type of process discussed in class does this remind you of?

Similar to the reasoning in part a), the larger a component is, the more "missing edges" it has to the rest of the network. Thus larger components are more likely to grow than smaller components. This is obviously reminiscent of "rich get richer" processes generally, and of preferential attachment more specifically. Full credit for both saying that large components are advantaged in growth, and mentioning one of these two processes; 2 points for only describing the growth properties or only mentioning one of the two processes.

c) Based on your answers above and discussions in class, what do you think the distribution of component sizes looks like in this Erdos-Renyi process, and how does it contrast with the degree distribution?

The distribution of component sizes will be heavy-tailed, and specifically a power law, while the degree distribution we know from class to be sharply peaked (Poisson). Full credit for saying both of these; 1 point for only recalling the degree distribution correctly.

3. (12 points) Consider very large networks in which both the degree of any vertex and the network diameter are relatively small.

a) Discuss, drawing as much as possible on course material and readings, why we would be interested in such networks.

Relevant materials and ideas:

- Obviously we are interested in large networks because so many of the ones of interest are enormous. Examples include social networks, the Internet, the web, the human brain, and so on.
- We are interested in small-diameter networks because of the many large-scale networks that have been documented to have rather small diameter. Relevant course citations include the six degrees of Travers and Milgram, the many small-diameter networks measured by Watts (Kevin Bacon graph, C. Elegans, North America power grid), and a number of others.
- We are interested in bounded-degree networks because of the belief that many networks, especially social networks, do have some fundamental limits on the degree --- i.e. you can only have so many friends because time is limited. Relevant course concepts include Gladwell's Magic Number 150 and the cortex ratio experiments of Dunbar.

2 points each for giving reasonable justifications (not necessarily those above) for each of these three parts.

b) Are there any mathematical limitations to arbitrarily large networks of small degree and small diameter? If your answer is "no", briefly explain why not. If your answer is "yes", give specific values for the network size, diameter and degree that cannot be simultaneously achieved, and explain why.

Yes, there are clear limitations. For instance, it is impossible to have a network with 100 vertices in which every vertex has degree at most 2, yet the network has diameter 5. The most efficient arrangement would be a 100-cycle, and clearly on average a pair of vertices is much more than 5 steps apart. 2 points for answering Yes; 4 additional points for clearly describing any scenario that cannot be achieved.

4. (12 points) For each of the networks shown below, indicate which vertex (A or B) has the **higher** PageRank value.

2 points each, no partial credit

a) Circle A or B

B

b) Circle A or B

B

c) Circle A or B

A

d) Circle A or B

A

e) Circle A or B

A

f) Circle A or B

B

5. (15 points) This question refers to the following class readings:

"*An Experimental Study of the Small-World Problem*".Travers, Milgram.
"*Navigation in a Small World*". Kleinberg.
"*Identity and Search in Social Networks*". Watts, Dodds, Newman.

Write a brief essay in which you describe what phenomenon Kleinberg and Watts, Dodds and Newman are attempting to explain *above and beyond* the findings of Travers and Milgram. Be sure to not only describe this phenomenon, but to indicate why its empirical presence requires explanation (that is, why might it be difficult to achieve). Finally, briefly contrast the different models and answers proposed by Kleinberg and Watts, Dodds and Newman.

- Whereas Travers and Milgram emphasized that large social networks have small diameter, the later works focused on the fact that people could actually *find* the short paths in a distributed fashion, using only very local information about the overall network topology.
- The reason this requires explanation is that sometimes simple "local" algorithms can fail to find the short paths even though they exist. An example are networks in which one must travel "away" from the target destination in order to travel the shortest path. For instance, we all know that the fastest way to travel to a geographic location is not to always try to move in its direction at all times --- we might need to go south to the airport (a "long distance link") in order to reach a northern destination most quickly.
- Kleinberg's model permits navigation solely through grid coordinates --- his local algorithms always forward messages to the neighbor whose grid address is closest to that of the target. Watts et al. have a richer model of social navigation that allows navigation via multiple "dimensions" --- geography, profession, religion, etc. Mathematically, Kleinberg's result shows a "knife's edge" requiring very particular conditions for efficient navigation, while Watts et al. have a more "robust" model that permits fast navigation under a wider range of circumstances.

5 points each for showing a clear grasp of each of the above 3 issues.

6. (15 points) This problem refers to the definition of the clustering coefficient discussed in class.

5 points each.

a) Briefly give the definition of the clustering coefficient of an individual vertex, and of an entire network.

The clustering coefficient of a vertex v is computed by first taking the neighbors of v. Letting the number of neighbors be k, we then count how many edges there are between the neighbors of v (ignoring v itself, which is used only to find its neighbors). If r is the number of such edges, the clustering coefficient of v is then r/[k(k-1)/2], which is simply the fraction of possible edges appearing among v's neighbors.

The clustering coefficient of the entire network is simply the average of the clustering coefficients of all the vertices.

b) Briefly describe the criterion used to determine whether a network is highly clustered or not.

Let c be the clustering coefficient of the network. We then compute the fraction of all possible edges present in the entire network --- i.e. if the entire network has E edges and N vertices, we compute p = E/[N(N-1)/2]. We then compare c and p. If c is much larger than p we consider the network to be highly clustered; if, on the other hand, c is close to p then we do not consider it so.

c) For each of the following network formation models, briefly say whether they are highly clustered or not: Erdos-Renyi, the α-model, preferential attachment.

Only the α-model shows high clustering (at only at certain values of α, though it's not required to say this). Neither Erdos-Renyi nor preferential attachment show high clustering.

7. (9 points) For each of the network formation models below, write all of the following properties that the model exhibits: small diameter, giant component, high clustering, heavy-tailed degree distribution. If you think a model has none of these properties, write “none”.

2 points off for each missing/incorrect answer (min 0, max 9)

a) Preferential Attachment

small diameter
giant component (by definition PA always generates only a single component)
heavy-tailed degree distribution

b) Erdos-Renyi with p = 1/(2N)

none, p is too small

c) Erdos-Renyi with p = 150/N

giant component

small diameter is ambiguous, so will accept either answer

8. (12 points) The image on the last page of this exam was discussed in class and was generated from one of the February 16 human-subject consensus experiments. (Notice that it is rotated; be sure to read it with the numbers 5, 10,…, 35 going down the vertical axis.)

a) Briefly but precisely describe what this figure is showing, describing what the vertical and horizontal coordinates are illustrating.

The figure is showing the progression of a consensus experiment. For each of the 36 players, there is a row of colored bars; these are arranged on the vertical axis. The horizontal axis represents the elapsed time in the experiment, up to the maximum of 180 seconds. The color at row i at time t is the color currently chosen by player i at that moment of the consensus experiment.

4 points for showing a clear understanding of the diagram.

b) Write a brief analysis of the image, pointing out interesting examples of both individual and collective behavior. What was the final outcome of this experiment? Feel free to make annotations on the image and refer to them in your analysis.

2 points for indicating this is a failed consensus experiment
3 points for at least one clear example/discussion of individual behavior
3 points for at least one clear example/discussion of collective behavior

MIDTERM EXAMINATION
Networked Life
CIS 112
Spring 2008
March 6, 2008
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.

Name: _____

Penn ID: _____

Problem 1: _____/20

Problem 2: _____/15

Problem 3: _____/15

Problem 4: _____/10

Problem 5: _____/15

Problem 6: _____/10

Problem 7: _____/15

TOTAL: _____/100

Problem 1 (20 points) Indicate whether each of the follow statements is true or false.

(a) Recall the “contagion” model of economic exchange discussed in class: Given an undirected connected graph, each vertex begins with an equal amount of currency. At each step, every vertex divides its current wealth equally among its neighbors. As this process is repeatedly infinitely, wealth will be distributed uniformly across the vertices.
FALSE

(b) The frequency of English words appearing in the New York Times over the past 20 years is well-approximated by a power law distribution.
TRUE

(c) "Having a diameter greater than 6" is a monotone property of a graph.
FALSE

(d) The Alpha Model with a large value for the parameter alpha corresponds to Watt's “Solaria” world.
TRUE

(e) In Kleinberg's model of navigation in social networks there is a parameter r governing the distribution of “long-distance” connections. Rapid search is not possible in this model when r = 2 because the long-distance connections will actually not travel very far.
FALSE

(f) The maximum number of edges that an undirected graph with N vertices can have is N(N+1)/2.
FALSE

(g) The PageRank of a web page can be interpreted as the probability that a certain kind of web surfer will visit that page.
TRUE

(h) All monotone graph properties exhibit tipping behavior in the Erdos-Renyi model.
TRUE

(i) “Organic” search results refer to those ranked according to objective criteria that cannot be easily manipulated.
FALSE

(j) In matters of spatial distribution, it is relatively straightforward to infer individual preferences from collective behavior.

FALSE

Problem 2 (15 points)

(a) Explain in words the primary differences between a normal or Poisson distribution, and a power law distribution.

Correct comments: Normal/Poisson sharply peaked around mean, has exponential decay away from mean, unlikely to draw values far from mean, etc; power law has long tails, slow decay away from mean, likely to draw values many times larger than the mean, etc.

(b) Describe the standard way of testing whether a set of sampled data points are better fit by a normal or Poisson distribution, or by a power law distribution. Feel free to illustrate your description with diagrams.

Plot y-axis equal to log of the value of the quantity in question (e.g. degree), x-axis equal to log of the number of observations, frequency, probability, etc., with that value. Then power laws will appear (nearly) linear with negative slope equal to the power, while Normal/Poisson will have high curvature away from a line. Plots showing these two cases would be appropriate.

(c) Name a few naturally occurring data sources that empirically seem to obey power law distributions.

Many examples from class: degrees in a social network, North American city sizes, distances traveled by dollar bills, file sizes on a computer, etc.

Problem 3 (15 points) Briefly but precisely describe the main definitions and ideas behind Kleinberg's "Hubs and Authorities" algorithm and the PageRank algorithm. Describe conditions or examples for which you think the two algorithms would disagree on the importance of a page.

Here either pseudo-code for both algorithms as was given in class, or the English descriptions that led to them are satisfactory. E.g., A good authority is a page which is pointed to by a lot of good hubs, and a good hub is a page that points to a lot of good authorities. A page with high PageRank is one which is linked from pages with high PageRank. Both types of answers are equally acceptable.

A page that has many outlinks but no inlinks would be a good candidate for a page that would have high hub weight but low PageRank. Similarly, a page linked from a few such hubs but no other pages would have high authority weight but low PageRank. Other differences can arise due to the fact that in PageRank, a page "p" gives 1/N(p) of its rank to each of the pages it links to, whereas in Hubs & Authorities, a hub p gives h(p) to each of the pages it links to. In dividing by N(p), PageRank discounts transmitted rank from

pages that have many links, whereas Hubs and Authorities does not. This could certainly cause discrepancies.

Problem 4 (10 points) Consider any of the network formation models we discussed in which both "local" and "long-distance" edges are present. Describe such a model as precisely as you can, and discuss what aspects of the real world such a model is intended to capture. Discuss which of the following properties networks generated by the model will and will not have: heavy-tailed degree distribution, small diameter, and high clustering coefficient.

We expect you to describe a model like Kleinberg's, where you start off with a grid (or a line or a cycle) and then add either uniformly random long-distance edges, or long-distance edges added according to a power-law distribution; also fine if instead of adding long-distance edges we instead "rewire" grid edges with fixed probability (as in the early epidemic demo).

For these models, the properties entailed would be small diameter and high clustering, but not heavy-tailed degrees.

5 points for clearly describing the model; 2 point for describing real world aspects the model captures; 3 points for listing the correct properties. Points will be deducted for confusing/conflating models; if you list multiple models, points will be deducted for those that do not meet the criteria specified in the problem.

Unacceptable answers here include the alpha model and Erdoes-Renyi model. However, in case you list one of these unacceptable models but give a reasonable argument that there will indeed be a mixture of distances about these models, a maximum of 5 points can be given. (A reasonable argument must clearly indicate that because in both models, there is a significant probability p with which both long distance connections and local connections are equally likely to be added.)

Models like preferential attachment are simply wrong and 0 point is given. (In the preferential attachment model, many nodes have degree only one!)

Problem 5 (15 points) Consider the classic paper "An Experimental Study of the Small World Problem" by Travers and Milgram. Briefly present and defend any three criticisms of the methodology or findings of this paper.

The following critiques are among the legitimate answers:

1. Small overall sample size --- very few initial parties even contacted by T&M

2. Even smaller completion rate --- only 64 chains completed

3. Bias introduced by failing to account for chains that never complete

4. Conflation of existence of short paths and their discovery by subjects from only local info, which is not pointed out until Kleinberg

5. Lack of analysis of how people decided who to forward the letter to, an issue analyzed in the Dodds et al study

6. Sample biased towards families of higher income.

7. Lack of incentives for subjects to actually try his/her best to find the shortest path.

8. The conclusion that connectors played a major role is probability an artifact of the small sample size of the experiment.

You need to list 3 critiques, 5 points for each correct critique.

Problem 6 (10 points) Consider a search engine which offers both "organic" and "sponsored" search results in response to user queries, and suppose the sponsored search results are listed in the order of the bid-per-click made by advertisers on the search term.

(a) Describe ways in which a person or organization might try to "game" the organic search results --- that is, manipulate the ranking algorithm used by the search engine in order to make their page appear higher in the organic results. You are free to draw on the class discussion of the common elements in organic search employed by search engines such as Google and Yahoo!.

For either (a) or (b), you receive 5 points if you list at least two correct methods. You receive 2.5 points if you list only one correct method. And beyond that, 1 point is deducted for each wrong answer.

For part (a), possible answers include:

1. Gaming PageRank by arranging to have lots of pages point to yours, possibly by creating those pages yourself

2. Adding hidden terms via tags in the html

3. Having terms on the page that don't have anything to do with its content but are just popular search terms

4. Typo squatting --- having your page contain common typos so that your page will appear in response to searches on them (this applied to (b) below as well)

5. Adding a lot of links from one's page to other important but irrelevant web sites

6. Hire a search engine optimization company (SEO) for you.

7. Trading links with some reputable sites

(b) Describe ways in which a person or organization might try to "game" the sponsored search results --- that is, to cause their page to appear higher in the sponsored results, to cause their advertising costs to be lower, or to cause the advertising costs of their competitors to be higher.

Possible answers include:

1. Bid jamming --- making your bid as close to the one above you without exceeding it

2. Clicking on competitors ads in order to cost them advertising budget

3. Typo-squatting, or put false description to lure customers
4. Bid search queries totally unrelated to their business in order to increase exposure.

5. Hire a search engine marketing (SEM) company for you.

Problem 7 (15 points) This problem considers the assigned paper "Graph Structure in the Web" by Broder et al.

(a) Briefly but precisely name and describe the five regions of the web identified in the paper.

Strongly Connected Component (SCC): group of web pages such that for any two pages A and B in the SCC, there exists a directed path of hyperlinks from A to B.

IN: pages for which there exists a directed path of hyperlinks leading to the SCC but not vice versa.

OUT: pages that can be found by following a directed path of links from the SCC but not vice versa.

TENDRILS: pages which can not reach the SCC or be reached from the SCC via directed hyperlink paths. These are pages for which a directed path leads to OUT, or for which a directed path leads from in IN.

DISCONNECTED: pages for which there are no paths to or from any of the pages in the weakly connected component (WCC) = SCC + IN + OUT + TENDRILS.

(b) Which of the five regions does a page belong to the very moment it is created, and why?

The newly created page presumably has no hyperlinks pointing to it since it is brand new. Therefore it cannot be in SCC or OUT. However, any of the other three regions are

possible depending on whether it has no outbound hyperlinks (DISCONNECTED), outbound hyperlinks pointing to IN or the SCC (IN), or outbound hyperlinks point to OUT (TENDRILS).

(c) In what ways is the creator or author of a web page in control of which of the five regions their page falls? In what ways are they not in control?

A web page author controls what pages her page links to but not what pages link to her page. If no pages in the SCC link to her page, the author can ensure her page belongs to IN by linking to pages in the SCC or to pages in IN. If pages in the SCC link to her page, the author can ensure her page belongs to OUT by not linking to pages in SCC or IN.

Similarly, if her page links to a page in IN or SCC, she cannot control whether her page is linked to. Thus she cannot control whether her page ends up in IN or SCC. And if her page does not link to pages in IN or SCC, she cannot control whether her page lands in DISCONNECTED, OUT or even TENDRILS in the case that a page in IN links to her page.

MIDTERM EXAMINATION
Networked Life (CIS 112)
March 4, 2010
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.

Name: _____

Penn ID: _____

Problem 1: _____/10

Problem 2: _____/20

Problem 3: _____/10

Problem 4: _____/10

Problem 5: _____/10

Problem 6: _____/10

Problem 7: _____/10

Problem 8: _____/20

TOTAL: _____/100

Problem 1 (10 points) For each of the following statements, simply write "TRUE" or "FALSE"

a. The Preferential Attachment network formation model explains all of the "universal" structural properties we examined in class.

b. The paper "Graph Structure in the Web" divides the pages on the Web into 7 distinct categories.

c. There is both mathematical and neuroscience evidence for the notion that there are limits to how many friendships we can maintain.

d. In controlled experiments in routing or navigation in social networks, it appears that people use geographic information mainly towards the very end of a chain.

e. If you have K friends or neighbors in a social network, the number of possible friendships among your friends grows roughly like the square root of K.

f. The PageRank algorithm can be viewed as spreading influence to the pages a particular web page points (hyperlinks) to.

g. In Kleinberg's "Hubs and Authorities" algorithm, a web page consisting of only hyperlinks to informative pages on mountain biking might obtain high authority weight for that topic.

h. "Connected" authors Christakis and Fowler are computer scientists.

i. The clustering coefficient measures how close "similar" vertices are in a network.

j. In Gladwell's terminology, a "maven" in a social network has high degree.

Problem 2 (20 points) In class we discussed the paper "The Scaling Laws of Human Travel", which makes use of data from the web site www.wheresgeorge.com.

a. Briefly describe the service provided by this web site.

b. Briefly describe the reasons for the paper's authors' interest in the data from the site.

c. Briefly describe the main finding of the paper that was discussed in class.

d. Briefly describe the connection drawn in lecture between this finding and the theoretical results of Kleinberg on navigation in social networks.

Problem 3 (10 points)

a. Let P be a monotone property of networks as defined in class. As precisely and succinctly as possible, give the definition of what it means for P to have a tipping point in the Erdos-Renyi model of network formation.

b. Name three specific monotone properties that have tipping points in the Erdos-Renyi model, and name them in the order (first to last) in which they would first appear in a network generated by Erdos-Renyi.

Problem 4 (10 points) Name two structural properties of social networks that frequently occur simultaneously, yet appear to be "in tension" with each other, in the sense that it is not obvious there should be simple mathematical models for network formation that can produce these two properties together. Then briefly describe a model we have studied or read about that indeed can do so.

Problem 5 (10 points) The assigned recent *Wired* magazine article “How Google’s Algorithm Rules the Web” discusses at length the many “signals” that inform Google’s algorithm --- perhaps individually tiny, but collectively important, contextual cues that Google uses to determine page relevancy for a given query. Briefly describe three such signals mentioned in the article, and suggest why they might be helpful in web search.

Problem 6 (10 points) In class we discussed “Rich Get Richer” processes, and the fact that they often lead to heavy-tailed distributions of whatever quantity is being allocated. The Preferential Attachment network formation model is one example of such a process, where connectivity (degree distribution) is being allocated. Give two other examples of quantities (not necessarily having to do with networks) approximately obeying a heavy-tailed distribution, and for each one briefly describe a natural “Rich Get Richer” process that might explain it.

Problem 7 (10 points)

a. Draw a connected network with exactly 10 vertices, exactly 9 edges, and the smallest diameter possible. Compute the clustering coefficient of this network.

b. Draw a connected network with exactly 10 vertices, exactly 9 edges, and the largest diameter possible. Compute the clustering coefficient of this network.

c. Draw a network (which may have multiple connected components) with exactly 10 vertices, exactly 9 edges, and the largest clustering coefficient possible.

Problem 8 (20 points) The book “Connected” and the associated articles discussed in class describe the research methodology and findings of authors Christakis and Fowler.

a) Briefly discuss the primary data source that is the basis for much of the authors’ research, and describe some of the properties of it that are different from what one might get from online social networks such as Facebook.

b) Much of the authors’ work meticulously establishes that certain behaviors, mental and physical states exhibit contagion in social networks. Name three things that are the subject of such contagion studies by Christakis and Fowler.

c) Briefly but carefully describe what the authors mean by “Three Degrees of Influence”.

MIDTERM EXAMINATION
Networked Life (MKSE 112)
October 27, 2011
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: _____/10

Problem 2: _____/10

Problem 3: _____/15

Problem 4: _____/10

Problem 5: _____/15

Problem 6: _____/15

Problem 7: _____/15

Problem 8: _____/10

TOTAL: _____/100

Problem 1 (10 points) For each of the following statements, simply write "TRUE" or "FALSE".

a. The Preferential Attachment network formation model was introduced in order to explain high clustering coefficients in social networks.

b. The smallest number of edges a network with N vertices can have and still be connected is N-1.

c. The paper "Graph Structure in the Web" cites the Strongly Connected Component of the Web as having approximately 91% of all pages.

d. In Kleinberg's model of navigation in networks, individuals may sometimes forward the message away from the target.

e. Standard notions of equilibrium do not require that individuals enjoy high satisfaction or payoffs.

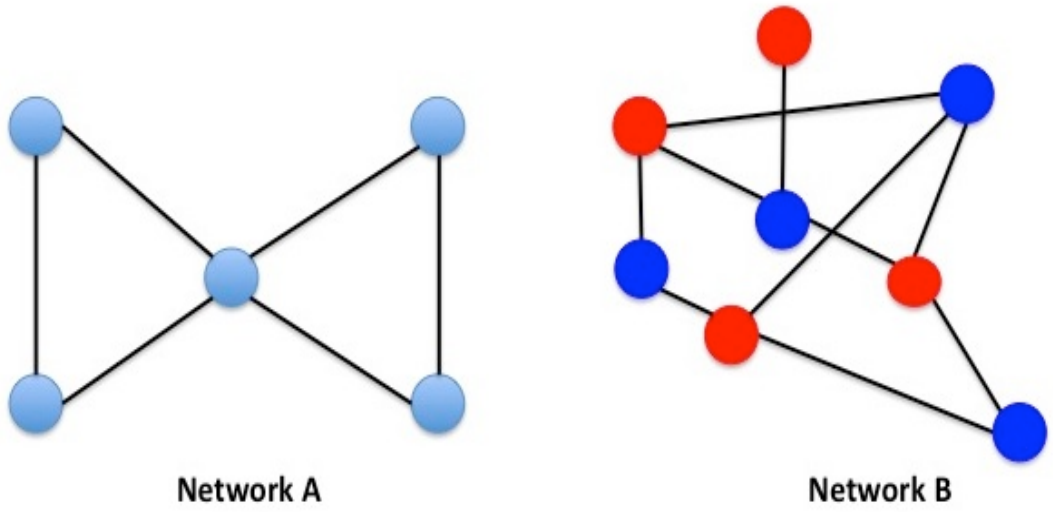
f. The PageRank of a vertex is determined entirely by its own degree.

g. If you have K friends or neighbors in a social network, the number of possible friendships among your friends grows roughly like K.

h. No matter how big N is, it is always possible to create networks of diameter N/2 in which every vertex has small degree.

i. The distribution of distances travelled by dollar bills in "Where's George?" is approximately normal (bell-shaped).

j. Zipf's Law states that the k-th most frequent English word is about 1/k as frequent as the most frequent English word.

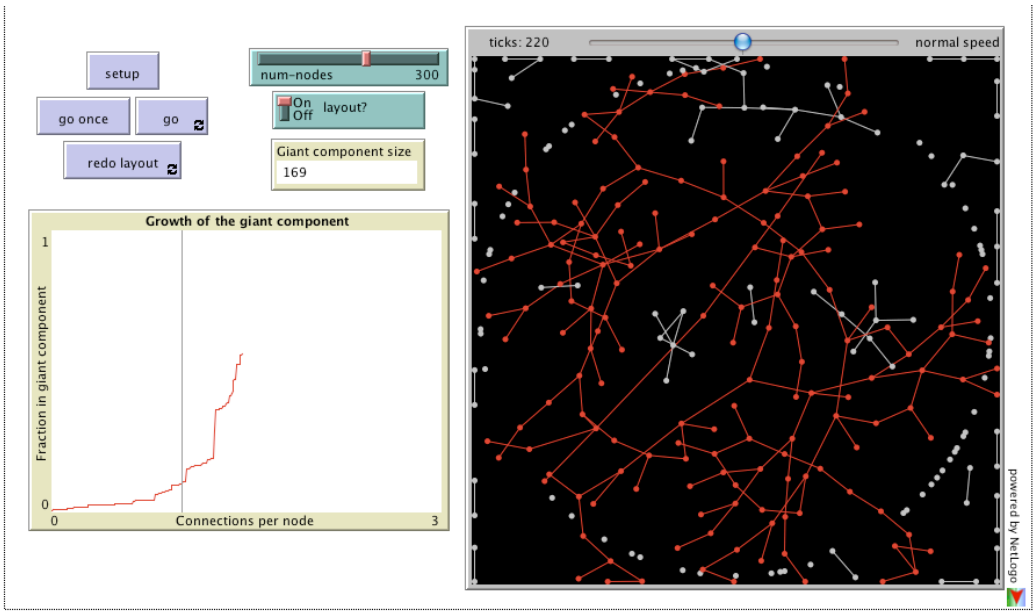


Problem 2 (10 points) For each of the two networks above, compute its clustering coefficient, and the equivalent edge density (value of p) in the Erdos-Renyi model. Say whether you think the network is highly clustered or not, and why.

(a) Network A

(b) Network B

Problem 3 (15 points) Describe a real-world, large-scale network (it can be social, technological, biological, economic, financial, etc.) that you believe does *not* have most or any of the main universal structural properties of typical social networks that we discussed in class and the readings. You should describe what the vertices and edges of your network are clearly, and then write a brief essay discussing why you think your network does not have the properties, and what might account for its difference with typical social networks.



Problem 4 (10 points) The image above is a screenshot from a simulator that we demonstrated in class. Briefly but clearly describe how this simulator works, including the underlying model it is illustrating. What is the main phenomenon about the underlying model that the simulator demonstrates? Be as detailed as possible. Then briefly explain in words why this phenomenon is common in large real-world social networks such as Facebook.

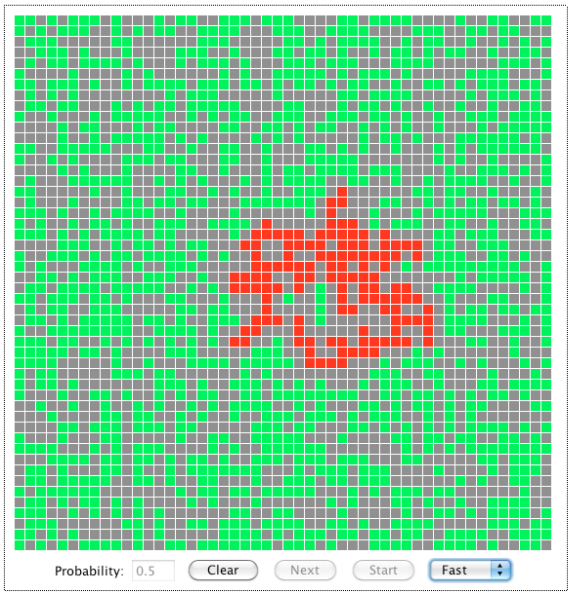
Problem 5 (15 points) Consider the following game of "competitive contagion" on networks. There are two competing players, Red and Blue. There is a network known to both players, and both players have some number of initial "seed" infections they can place in the network. Both players must choose where to place their seeds simultaneously. If both players choose to seed the same vertex, the vertex becomes infected with each color with probability ½. Once the seeds are placed, uninfected vertices are updated according to the following rule: any vertex adjacent to a vertex already infected with Red or Blue becomes permanently infected with that same color. If an uninfected vertex has both Red and Blue neighbors, it becomes infected with Red or Blue with equal probability. Updates of uninfected vertices occur as soon as they have an infected neighbor.

(a) Suppose the network is a "hub and spokes" network, where vertex 1 is connected to vertices 2,…,N and there are no other connections. Let each player have one seed infection. Describe the Nash equilibria of this game.

(b) Suppose the network is a simple cycle or ring of N vertices, and again let each player have one seed infection. Describe the Nash equilibria of this game.

(c) Suppose the network is again a cycle, and let the Red player have two seed infections and the Blue player one seed infection. Describe the Nash equilibria of this game.

NWLife Forest Fire Demo

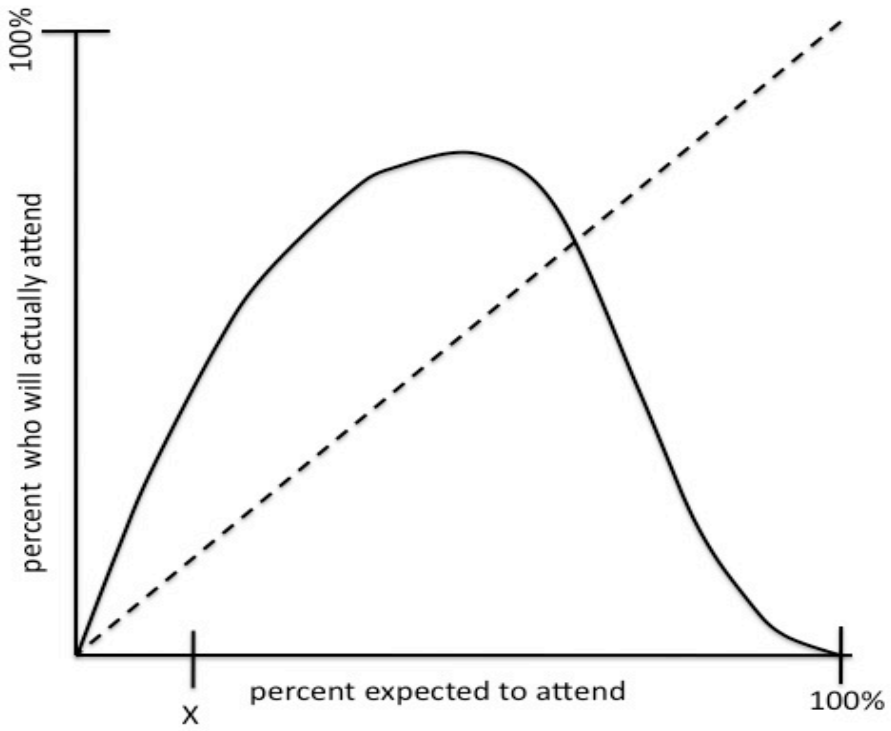


Problem 6 (15 points) The screenshot above shows the end result of running the forest fire simulator that we demonstrated in class.

(a) Briefly but clearly describe how this simulator works, including initialization and the process simulated.

(b) What value for the "Probability" (which has been blacked out) do you think was used for this particular run of the simulator? Why?

(c) Briefly but clearly describe how this simulator can be formalized as a model of random network formation, and the examination of a particular structural property (which you should name or describe) of the networks it generates.



Problem 7 (15 points) Consider the curve shown in the figure above, which is similar to diagrams we have analyzed in a recent lecture.

(a) Give an example of a real-world activity that might roughly correspond to the dynamics represented by this diagram.

(b) Consider a typical starting point to the left of the peak of the curve, such as the point labeled X. Describe in words how the attendance dynamics evolve from this starting point. You may want to annotate the diagram.

(c) Are there any equilibria in this system? If so, mark them on the diagram. Is it clear from the diagram whether or not the attendance will ever stabilize?

Problem 8 (10 points) For any value of N, describe a network of N vertices in which the diameter is as small as possible, the clustering coefficient is as high as possible, and there is at least one "connecter" vertex. The total number of edges in your network should grow proportionally (linearly) with N and not larger. Describe your network as precisely and succinctly as possible, and give the values of the diameter and clustering coefficient. You are free to provide an illustrating diagram.

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.

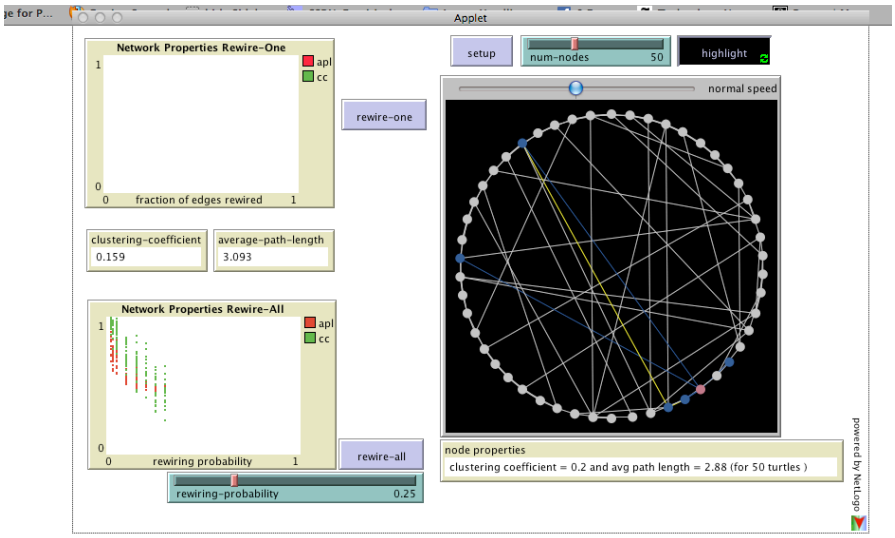
TRUE

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 :

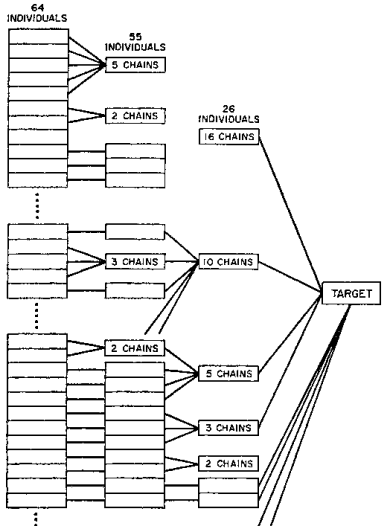
- *Specifying formation - 1 point*
- *Specifying starting configuration - 2 points*
- *Discussing Rewiring parameter - 2*

(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:

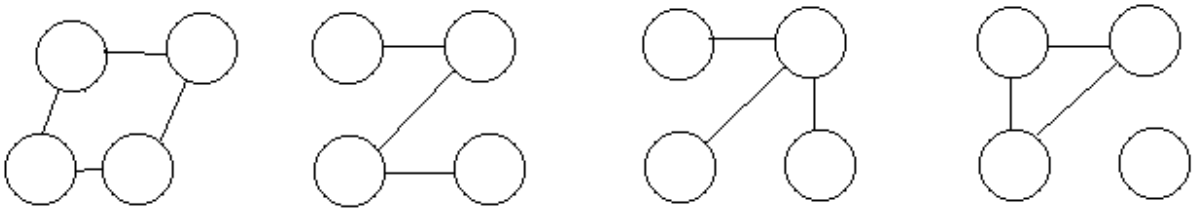
- *Mention that everyone who stood up had to remain standing till the target was reached (Points have been taken off if this wasn't clear)*
- *Bring up the question of scale in real networks in some form (Just saying it is not feasible or 'everyone can't be asked to stand up together' is not sufficient).*

.

Problem 5 (15 points)

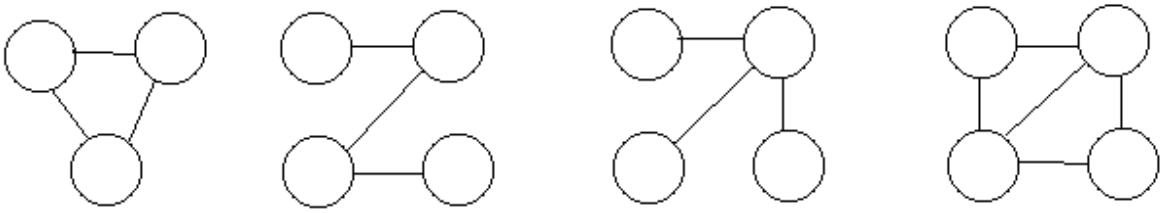
There is no need for any calculations here, and grades are based on whether you get the answers right. (i.e. there is no partial credit for this question.) Each part has 5 points, and the answer to part a,b and c is G4, G1, G4 respectively.

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



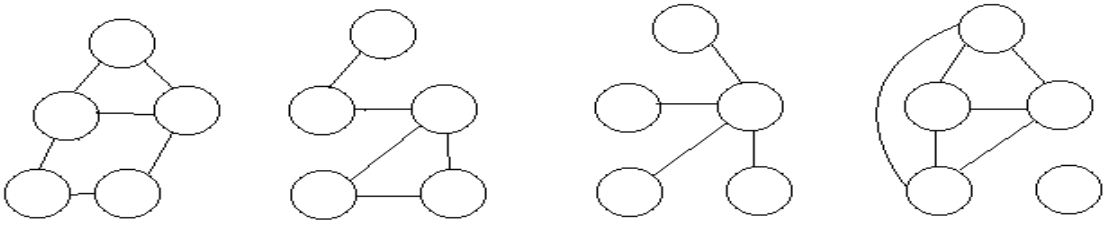
G1 G2 G3 G4

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



G1 G2 G3 G4

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



G1 G2 G3 G4

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.

SOLUTIONS for MIDTERM EXAMINATION
Networked Life (NETS 112)
October 21, 2014
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: _____ **/10**

Problem 2: _____ **/10**

Problem 3: _____ **/10**

Problem 4: _____ **/15**

Problem 5: _____ **/10**

Problem 6: _____ **/15**

Problem 7: _____ **/10**

Problem 8: _____ **/20**

Total: _____ **/100**

Problem 1 (10 points: Graded by Ryan) Indicate whether the following statements are *True* or *False*.

(a) Any planar graph can be colored using at most 3 colors.
F

(b) If a social network permits efficient navigation from only local, distributed information, it necessarily has small diameter.
T

(c) PageRank is the single most important element in determining relevance in Google's search engine.
F

(d) Prisoner's Dilemma has exactly one pure-strategy Nash equilibrium.
T

(e) As long as the housing preferences of each individual are mild, there is a global solution in which everyone is happy.
F

(f) Klienberg's navigation model might appeal more to mathematicians, and the Watts, Dodds, Newman navigation model might appeal more to sociologists.
T

(g) The Backstrom et al. Facebook diameter study showed strong evidence of increasing diameter over time.
F

(h) Having a heavy-tailed degree distribution is a monotone network property.
F

(i) If the fraction of people who want to participate next time increases with the fraction participating this time, the only equilibrium is 100% participation.
F

(j) Each time we experimented with tennis ball navigation in networks, the class found a near-shortest path.
F

Problem 2 (10 points: Graded by Shahin) Compute the clustering coefficient for the network given in Figure 1. Be sure to write the clustering coefficient for each node along with your work.

Solution. We have

$$cc(A) = 0 \quad cc(D) = 2/3 \quad cc(E) = 3/10$$
$$cc(B) = cc(C) = 1/3 \qquad cc(G) = cc(F) = 1$$

So the clustering coefficient of the whole network is $\frac{0+2/3+3/10+2/3+2}{7} = 109/210$.

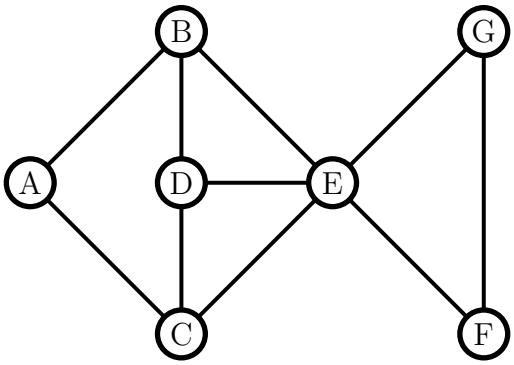


Figure 1: Compute the Clustering Coefficient for this Network

For grading: one point for the clustering coefficient of each node and 3 points for the clustering coefficient of the network. □

Problem 3 (10 points: Graded by Shahin) Consider the network in Figure 2. Determine the nodes with the smallest and largest rank after running the PageRank algorithm on the network. Briefly justify your answer.

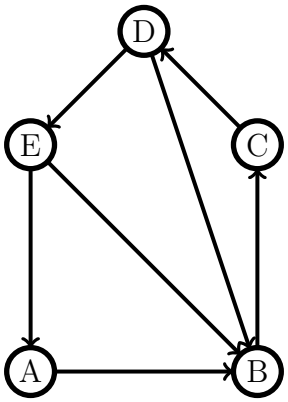


Figure 2: Page Rank

Solution. Node B has the largest page rank because it gets in-flows from nodes A, E and D. A has the smallest page rank because its in-flow comes from the stream connecting D to E and then A and half of this flow is already going to B in each step before getting delivered to A.
Common mistake 1: B, C and D all have the same page rank. (I deducted two points for this) Although their page ranks are close to B, but C and D does not have the same page rank as B. C and D do not have the same in-degree compared to B. Furthermore, although their getting a flow from B but as we are getting further away from B, the page rank decreases. So $B > C > D$ if we sort by page rank.
Common mistake 2: E and D have the smallest page rank because they have the largest out-degree. Page rank does not only depend on in and out degree. It also depends on what other nodes a node is connected to and how important those nodes are (where importance means having a high rank). I gave no partial credit for this answer. □

Problem 4 (15 points: Graded by Ryan) For this problem, consider the online graph coloring experiments you participated in.

(a) (5 points) Briefly but precisely describe the graph coloring problem.

(b) (5 points) In class there was a discussion of strategies employed by students in the class. Briefly but precisely describe at least two different strategies that were articulated.

(c) (5 points) Recall that for each network, there were additional points awarded for the three fastest completion times. Discuss the distribution of these additional points across the class, including a comparison to the distribution expected if all students were equally skilled at graph coloring. Based on this distribution, do you think graph-coloring ability would be better modeled as a normal or heavy-tailed distribution?

Solution.

(a) We are given a graph G with nodes V and edges E. We want to color each node of the graph such that no two nodes of the same color share an edge. We want to do this in as few colors as possible, or color each node from a predefined set of colors as the web app had.

(b) Color a high degree node one color, say Red, and try to color as many of its neighbors the same color that is different from Red. Find triangles and color each node of the triangle differently. Try to partition the nodes into different sets so that each set has no nodes that share an edge, then color each set differently.

(c) We consider the Ranked Points Total plot given in lecture. In this plot, the blue graph represents what would happen if everyone had the same coloring ability and completion times for everyone was normally distributed around the same mean for everyone. The red represents the real data. Note that the top ranked students were several times higher than what we would have expected from the blue plot. Recall that the top ranked student got 26 points and the points dropped pretty quick after that.

This problem involved looking at the points distribution and the ability distribution. You were asked to conclude how the ability distribution looked given how the points distribution looked. From the lecture slides, you are only given the ranked points distribution plot. From this, you can tell that the points distribution has a heavy tail because few people got way above zero medals.

In order to see a points distribution with a heavy tail, we may expect there to be just a few people that are slightly above average at graph coloring ability. Note that we do not need someone many many times above average in coloring ability for her to accrue a majority of the points, but merely slightly above average. We are not asking about the finish times distribution but the points distribution. This would be an arguement for why a *normal* ability distribution would make sense, because maybe a student is slightly above average at skill and so would gain a majority of the points every time.

You could also argue that the ability distribution could be heavy tailed. This would mean that there are very few people that are way above average in coloring ability. Hence, of course those people would get all the medals, so everytime there would be always the same awesome colorer getting gold, then the second getting silver and third getting bronze for every graph. In our data we did see people get a majority of the points, but more than 3 people got additional points. This may be a reason against ability being heavy tailed.

This problem was graded *very* leniently. If you were clear what distribution you were talking about (points or ability) then you typically got full credit.

□

Problem 5 (10 points: Graded by Ryan) Consider the curve in Figure 3 where the horizontal axis shows the percentage of the students who attend today's class and the vertical axis shows the expected percentage of student who will attend the next class.

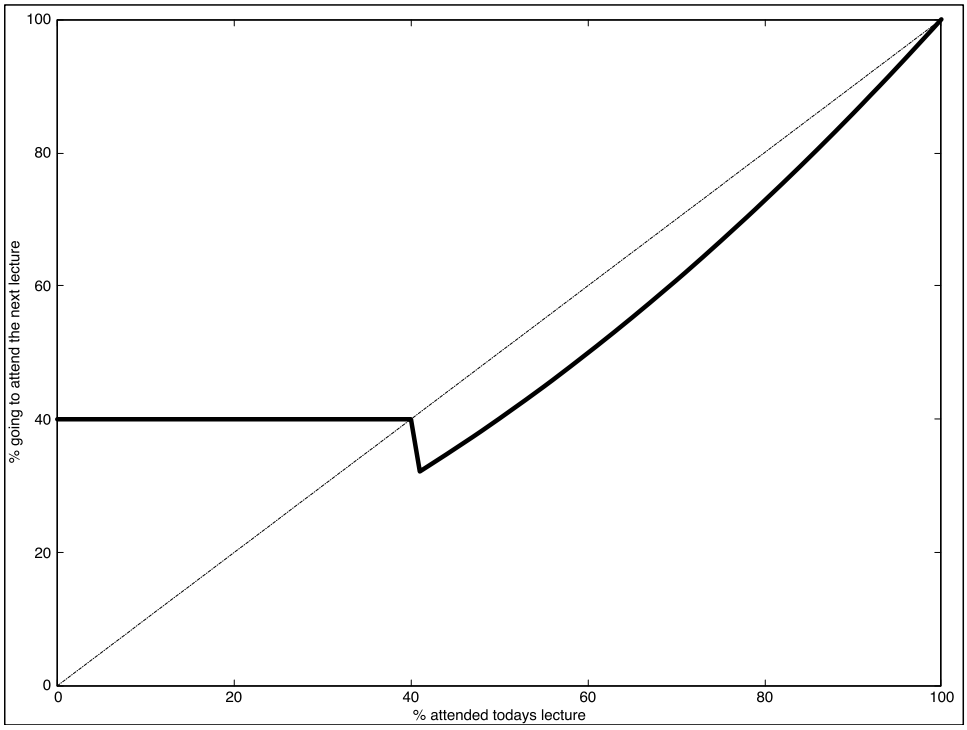


Figure 3: Equilibrium Analysis

(a) (5 points) Are there any equilibria in the system? If so, mark them on the diagram and mention whether the equilibria is stable or not?

(b) (5 points) Suppose 80% of the students attended the first lecture. Briefly describe the dynamic of participation for the future classes.

Solution.

(a) There are two equilibria, one at 40% which is stable and one at 100% which is not stable.

(b) The participation decreases until it reaches somewhere below 40% (corresponding to the bottom of the v-shape in the diagram), then it increases to 40% and stabilize at 40%.

□

Problem 6 (15 points: Graded by Ryan) Consider the following 2-player game matrix:

Row Player/Column Player	A	B
A	-1,-1	-1,+1
B	+1,-1	-10,-10

(a) (5 points) Does this game have any pure-strategy equilibria? If so, what are they?

(b) (5 points) Do you think there is a mixed-strategy equilibrium that is different from any pure-strategy equilibria? If so, broadly describe what the mixed-strategy equilibrium is. If not, why not?

(c) (5 points) Imagine that the game describes the contest known as Chicken: two players drive towards each other at high speed, and must choose either to keep going straight or swerve to their right at the last second. Indicate which of actions A and B correspond to straight and swerve, and explain why the payoffs in the table model this contest.

Solution.

(a) It has 2: Row Player plays A, Column Player plays B; Row Player plays B, Column Player plays A.

(b) There is a mixed strategy Nash equilibrium. Note that there are two pure strategy Nash equilibria, one where the Row Player is happy and the Column Player is unhappy, and one where the Row Player is unhappy and the Column Player is happy. However, we could imagine that the Column Player randomizes between his strategies to try to sometimes get the payoff where she is happy and in response to this the Row Player will also randomize between his strategies so that he will sometimes get the better payoff. In the mixed strategy Nash equilibrium, each player would need to ensure that the chance that they both end up playing B is small, because this is where both players are *very* unhappy.

(c) B would be going straight and A would be swerving. It makes sense because if both go straight, this corresponds to entry (B,B) and both end up with lots of damage to each other's cars (not to mention the medical bills). However, if one swerves and the other goes straight, then there is no collision and the one that goes straight gets "steet cred"

worth utility 1, while the other person that swerves will always be referred to as "chicken." In the last scenario when both swerve, there is no collision, but also both are called "chicken."

□

Problem 7 (10 points: Graded by Shahin) Describe, as precisely as possible, a network in which there is a special vertex v that you would argue is important, despite having very low degree. Give a schematic diagram of this network with v clearly identified. Describe, as precisely as possible, a general measure of vertex importance that would identify v as important in your network.

Solution. The network given in Figure 4 has a node v that has a low degree compared to all the other nodes but may be referred to as important because it connects two large connected components (in this case two cliques) that otherwise would not be connected. (5 points)

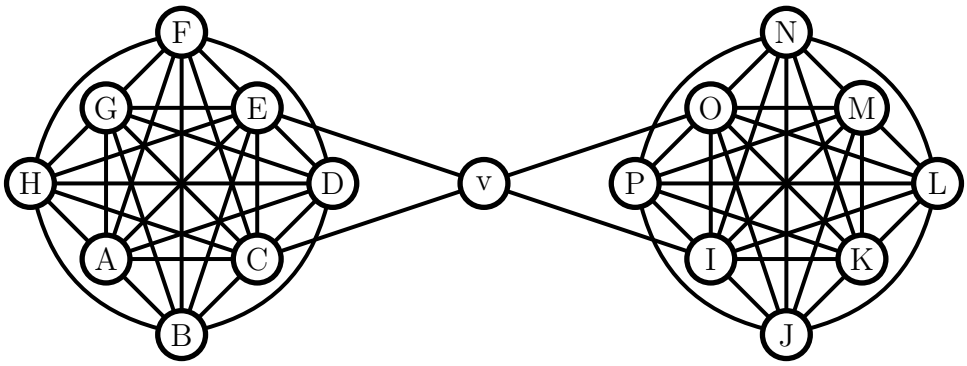


Figure 4: Solution for Question 7

Perhaps one definition for importance that would make v important compared to the other nodes would be to compute the size of the largest connected component in a graph G with and without the node v and compare these two numbers together. Hence, in Figure 4 v is important because the size of the largest connected component will drop from 17 to 8 after v being removed. Note that the deletion of no other node will decrease the size of the largest connected component. Another definition of importance would be as follows. Consider any two pairs of nodes in the graph and compute the shortest path between them. For each node v, computer the total number of shortest path among all the pairs that involves node v and refer to this number as $\text{Im}(v)$. A node is important if it has a low degree and the $\text{Im}(v)$ is high *e.g.* in Figure 4, $\text{Im}(v) = 48$ which is much higher than other nodes. Note that E, C, O and I also have high Im but they are not of low degree. (5 points)
Common mistake: define a node important if it has a low degree and removing the node makes the diameter graph disconnected (or increase the worst case diameter to infinity). Although v in Figure 4 indeed has such property, it is not hard to see why this is not a good notion of importance *e.g.*, node Q in Figure 5 will be defined as important with the latter definition but it is clearly not an important node in the network.

□

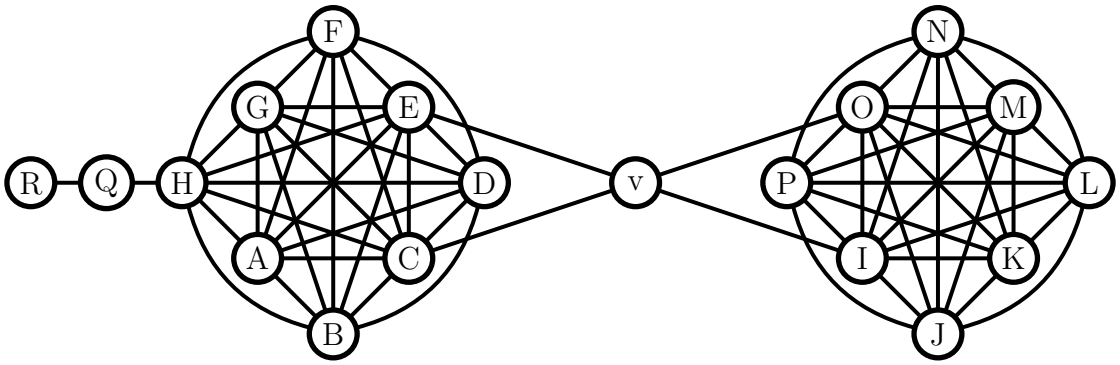


Figure 5: Is the node Q important?

Problem 8 (20 points: Graded by Shahin) In a recent lecture, we played a participatory game based on Schelling's segregation model.

(a) (5 points) Briefly but precisely describe the rules of the game.

(b) (5 points) Which of the following best describes the resulting outcome we observed: full segregation; partial segregation; full integration.

(c) (5 points) Briefly but precisely describe what was shown in the corresponding computer simulation in the same lecture.

(d) (5 points) Briefly but precisely describe the main points that this simulation is meant to illustrate.

Solution.

(a) We considered the network where students were nodes and a student's neighbors were all those that sat in the 8 adjacent seats around the student (1 point). We considered a student A to be *happy* if A had at least three neighboring seats that were occupied (loneliness) (1 point) and at least two of A's neighbors were of the same gender as A (homophily) (1 point). We then had the students move until all the students were *happy* (2 points).

(b) We saw partial segregation. (No partial credit)

(c) We again have a grid network where every node has 8 neighbors which are adjacent to it, and each node may be empty or be one of two colors. Of the nonempty nodes, 50% were one color, Red, and the other 50% Green. Each node desire a certain percentage p of its 8 neighbors to be of the same color and we can adjust this percentage. The nodes then can move to empty cells until all the nodes have at least p% of the same color.

The outcome of the simulation was an equilibrium: given what everyone else is doing, some nodes may want to move or not. We saw a tipping point when we set the p to be near 51%, just over a majority. This caused nearly everyone (95%) to be segregated from nodes with different colors. Where as when p was smaller, we observed partial

segregation. Also as we increased the p it took more time for nodes to reach an equilibrium.

For grading, I deducted points for not mentioning the details of the experiment as in part (a). Also, I deducted points if you have not mentioned the tipping point and the threshold between partial and full segregation. I did not deduct any point if you mentioned the tipping point in the answer to part (d). Please refer to my comments in your exams.

(d) The main point of the simulation was that we could not infer the individual preferences from the outcome of the simulation *i.e.,* we saw almost full segregation even when the preferences of the individuals were mild (p was close to 50%). Also, there might be other equilibria for small values of p with less segregation, but since the nodes were not cooperating with all the other nodes in the network, we did not achieve those equilibria. Again for grading, please refer to my comments in your exams.

□

MIDTERM EXAMINATION
Networked Life (NETS 112)
October 20, 2016
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:

Problem 1: _____/10

Problem 2: _____/20

Problem 3: _____/20

Problem 4: _____/20

Problem 5: _____/15

Problem 6: _____/15

TOTAL: _____/100

Problem 1 (10 points). Answer "true" or "false" to each of the following assertions.
[Graded by Adel]

(a) The web-based version of the Travers and Milgram experiment demonstrated that people forwarded messages based on professional ties early in the chain, and geographic proximity later in the chain.

False

(b) An Erdos Number of 5 or less is rare among published mathematicians.

False

(c) The squash network exhibited homophily of ratings.

True

(d) The value of the exponent r in Kleinberg's model that permits efficient navigation is 3.

False

(e) Being shown real-world friends who are poorly connected made people more likely to join Facebook.

True

(f) The main research are of Paul Erdos was statistical physics.

False

(g) Structural diversity refers to the many ways content can spread through social media.

True

(h) Cascades started by celebrity tweets tend to be broad but shallow.

True

(i) Gradually increasing the fraction of forest gradually increased the amount burned.

False

(j) A photo of Daniel Ellsberg appears on the course home page.

True

Problem 2 (20 points). This problem refers to the assigned reading "Can Cascades Be Predicted?" [Graded by Adel]

(a) What exactly were the authors trying to predict? Be as precise as you can.

Predicting ultimate size of viral cascade based on a number of variables that capture cascade properties; percentage of cascades that exceed f(k) if is k is median cascade size

(b) What were the five categories of features from which the authors tried to make their predictions? Give an example of a feature in each category.

Content: Type of picture
Root/author: who originally posted it
"Resharer": type of originator (page vs person)
Structural: network structure (depth vs breadth)
Temporal: how fast the shares spread

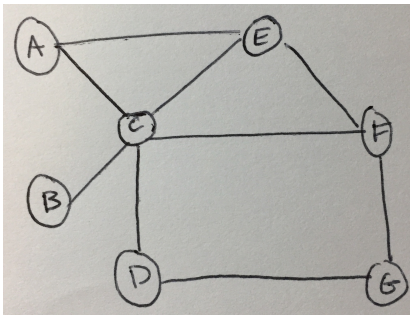
(c) What category of features was most predictive? What category was least predictive?

Most = temporal, least = content

(d) Can you draw any inferences from the answer to part (c) regarding how effective it might be to "engineer" viral content? Why or why not?

The paper doesn't provide negative evidence for engineering viral content because it was not a controlled experiment; in particular, there's no reason to think that most of the photos in the dataset were engineered for virality, so the fact that content features were not predictive doesn't mean they couldn't be.

Problem 3 (20 points). For the network shown below, numerically calculate each of the following quantities.
[Graded by MP]



(a) Diameter (average-case)

Sum of each possible path between all pairs of vertices divided by the total number of paths = 35/21 = ***5/3***

(b) Edge density

Number of edges present in network / total possible number of edges = 9/21 = ***3/7***
[Total possible number of edges = N(N-1)/2 = 7(6)/2 = 21]

(c) Clustering coefficient (assume degree 1 vertices have clustering coefficient 1)

cc(A) = 1
cc(B) = 1
cc(C) = 1/5
cc(D) = 0
cc(E) = 2/3
cc(F) = 1/3
cc(G) = 0
→ Global clustering = sum of individual clustering coefficients / number of nodes = ***16/35***

(d) Equilibrium wealth in the model discussed in class

Assume every vertex starts with $x. So total wealth = $7x
*For every vertex I, equilibrium wealth, W(i), equals Total Wealth * (degree(i) / total degree)*
W(A) = 7x(2/18) = 7x*(1/9)*
W(B) = 7x(1/18)*
W(C) = 7x(5/18)*
W(D) = 7x(2/18) = 7x*(1/9)*
W(E) = 7x(3/18) = 7x*(1/6)*
W(F) = 7x(3/18) = 7x*(1/6)*
W(G) = 7x(2/18) = 7x*(1/9)*

[We awarded to any answer that gave these ratios and a logical values for initial and total

wealth states.]

Problem 4 (20 points). In recent lectures, we have articulated five "universal" empirical properties of large-scale social networks.
[Graded by Brad]

(a) Briefly but clearly name and describe/define each of these properties. Be as precise as you can, using expressions involving the population size N where appropriate.

Small diameter: diameter << N (# of vertices), log(N), or some constant
Heavy-tailed degree distribution: largest degree >> average degree; polynomial decay of network degrees; modeled by power law
Giant component: existence of a large connected component that is much larger than the second largest component
Sparsity: average or typical # edges << possible # edges; or, average # of edges grows linearly in N edges added
Clustering: density of connectivity w/i a community >> connectivity b/w communities, local edge density much larger than global edge density

(b) One of the five properties is apparently "challenging", in the sense that it's not obvious how to satisfy it and all or some of the other properties simultaneously. State what the challenging property is, and justify your answer.

Sparsity or small diameter was accepted. Credit assigned depended on argument given.

Problem 5 (15 points). Consider the network formation model in which start with N vertices and no edges, and at each step, we pick a pair of vertices not already connected by an edge randomly among all such pairs, and add the edge between them. Suppose we run this process for E steps, at which point we will have added exactly E edges to the network.
[Graded by MP]

(a) For approximately what value of E would you expect the network to first have a giant component? Explain your reasoning by arguing why it is difficult for two large components to coexist at your chosen value for E.

Any answer on order N; the probability of adding an edge between two vertices and having one of them not in the connected component is very small

(b) Do you think there is a value of E at which the clustering coefficient of the network will be much higher than the overall edge density? Why or why not?

No, as discussed in lecture, there is simply no force in this model that would cause clustering — the expected clustering coefficient of a vertex is exactly the background edge density.

(c) Do you think there will be a value of E at which the degree distribution is heavy-tailed? Why or why not?

No, edges are just added at random so there is no "richer get richer process" in terms of individual degree.

Problem 6 (15 points). Consider the following model of network formation. We start with two vertices connected by an edge. At each step, we add a new vertex u with a single edge to the current network. The vertex v that we connect to u is determined as follows: with probability ½, we choose v to be the vertex that currently has the highest degree (breaking ties randomly); with probability ½, we choose v randomly among all vertices in the network so far.
[Graded by Brad]

(a) Do you think this model generates networks with small diameter? Why or why not? What do you think the diameter will be for large N?

Yes, there is a high probability that most vertices will be within 1 or 2 steps of the hub or highest degree vertex.

(b) Do you think this model generates networks with high clustering coefficient? Why or why not?

No, there will be one hub with none of its neighbors not connected to each other, implying low clustering.

(c) As precisely as you can, describe the degree distribution for large N.

Hub in middle will have ~ E/2 (or ~N/2) as degree and all other vertices will have degree values significantly less than ~E/2 or N/2 and their degrees will approximately have a Poisson distribution and not have a power-law distribution

Midterm Solutions and Grading Guidelines

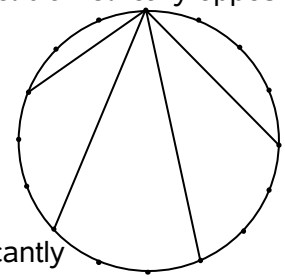
Problem 1 (10 points) For each lettered item on the left below, write the number of the item on the right that matches it best.

a) shortest path distance _18_	1) small changes have large effects
b) worst-case diameter _12_	2) heavy tail
c) a disconnected network _13_	3) exponential decay away from the mean
d) degree of vertex _16_	4) opposite of Solaria
e) scaling laws of human travel _14_	5) persuasive
f) hub _20_	6) PageRank
g) these tip _7__	7) monotone properties
h) amplification of the incremental _1__	8) an unfortunate equilibrium
i) connector _15_	9) rich get richer
j) maven _19_	10) fashion as epidemic
k) salesman _5__	11) the original “random” network model
l) power law distribution _2__	12) length of the largest shortest path
m) Normal distribution _3__	13) Infinite worst case diameter
n) Erdos-Renyi _11_	14) Where's George
o) preferential attachment model _9__	15) many friends and acquaintances
p) Caveman _4__	16) number of network neighbors
q) Paul Erdos _17_	17) Kevin Bacon
r) holiday greeting cards _8__	18) fewest hops between vertices
s) Hush Puppies _10_	19) information specialist
t) a random surfer model _6__	20) a vertex with very high degree

Problem 2 (10 points) Suppose we start with a network that is a simple cycle (ring) of N vertices. You are then allowed to add some fixed number of additional edges to the network. Feel free to illustrate your answers with diagrams.

a. Briefly describe the pattern of edges you would add if the goal is to make the diameter of the resulting network as small as possible.

One solution is to pick a vertex and make it a hub by adding edges that connect that vertex to other vertices. The path connecting the two most distant vertices A and B would go from A to the nearest vertex with an edge to the hub, then the hub vertex, then the vertex nearest to B that is connected to the hub and finally B. Other acceptable solutions are to randomly add edges, or to add edges that connect diametrically opposite vertices.



5 points for a correct solution

3 points for a solution that decreases the diameter but not significantly

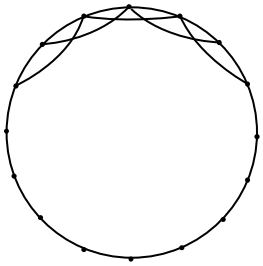
-2 points for inadequate or no explanation

-2 points for modifying the graph by adding vertices

-2 points for adding all possible edges to make the graph fully connected

b. Briefly describe the pattern of edges you would add if the goal is to make the clustering coefficient of the resulting network as large as possible.

Add edges that will connect the vertices with the neighbors of their neighbors. Initially this means connecting every second vertex. If more edges are available, connect every third vertex and so on. The clustering coefficient will increase because the new edges introduce connections between the neighbors of a vertex.



5 points for connecting neighbors of neighbors

3 points for solutions that increase the clustering coefficient but not significantly

No points for just giving the definition of the clustering coefficient

-2 points for inadequate or no explanation

-1 point for unclear explanation

Problem 3 (15 points) Consider the "People You May Know" functionality on Facebook, which suggests friends to add to your network based on common friendships you already have. Discuss what effects you think the introduction of this service has had on the following structural properties of the Facebook network. Briefly justify your answers.

a. Degree distribution

The degree of every vertex will increase. Therefore the average degree will increase and the degree distribution will be shifted towards higher degrees. The shape of the distribution will not be affected significantly because the people who already have more friends are likely to get even more because the service can make more suggestions, whereas people with fewer friends will not receive as many suggestions.

5 points for saying the average degree will increase but the distribution shape will not be affected much

3 points for just saying the vertex degrees will increase but not discussing the effect on the distribution

-3 points if no justification was given or it was wrong

-1 point for confusing the degrees with the degree distribution

b. Clustering coefficient

The service will be suggesting connections between friends of yours that are not already connected, so your clustering coefficient will increase. The same applies to everyone so the clustering coefficient for the overall network will also increse.

5 points for saying the clustering coefficient will increase

-3 points if no justification was given or it was wrong

c. Diameter

Adding edges can only decrease the diameter of the network. Since the new edges are between people who already had a common friend, they will not provide huge shortcuts and the diameter will not decrease significantly.

5 points for saying the diameter will decrease and discussing how much

4 points for just saying the diameter will decrease

-3 points if no justification was given or it was wrong

Problem 4 (10 points) The abilities of Wall Street workers in primarily quantitative roles (a.k.a. "quants") naturally vary considerably from individual to individual. Furthermore, if compensation on Wall Street degrades sufficiently (e.g. due to the current financial crisis), some quants may consider alternatives (going to grad school, joining a start-up, etc.). Apply the concept of the "Market for Lemons" discussed in Schelling to the hiring of quants on Wall Street. Be sure to discuss the asymmetry of information and the dynamics of the process, including the behavior and decisions of both quants and employers. Who leaves the market first, and who leaves last?

- While the employers only have limited information about their employees through the hiring process, the employees know how strong/weak their own quantitative skills are. An employee therefore knows if he or she is an above average quant, or if he or she corresponds to a "lemon" in the workplace.
- When the compensation decreases, the best quants realize they can get hired by other places like start-ups and get paid more for their skills. So, the best quants quit their Wall Street jobs and leave the market.
- It is now not worth it for new workers who have strong abilities to enter Wall Street, so only the relatively weaker potential quants apply for quant jobs.
- The employers realize this and know that the employees that are left are now not as good on average, and so lower the compensation even more.
- The better of the employees that are left now are undervalued and so leave the market. The cycle continues, with the proportion of bad quants getting larger and larger, and the compensation getting lower and lower, until only bad quants remain.
- When only bad quants are left, it is no longer worth it for Wall Street firms to employ quants at all, and so the market disappears.

10 points total

1 point for acknowledging that employees know their skills, while employers don't

3 points for recognizing the best quants leave first

2 points for realizing that employers will then lower compensation further

3 points for realizing that this process will spiral, like the market for lemons

1 point for realizing the bad quants leave the market last

Problem 5 (15 points) Consider networks in which each vertex has at most K neighbors, and every pair of vertices is at most D hops apart.

a. Suppose K = 150 and D = 6. Is it possible to have such a network whose size is equal to that of the U.S. population? Simply answer "Yes" or "No".

Yes
3 points for writing "Yes"

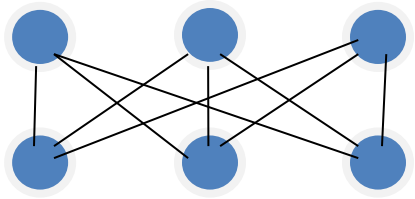
b. Why are K=150 and D=6 particularly interesting choices in light of course material and readings?

- K=150 is a plausible bound on the number of neighbors people have in real life social networks. Gladwell referred to it as "the magic number 150", and it is believed that humans have evolved to be able to keep track of social relationships in group sizes of up to 150. In particular, experiments have found a relationship between the neocortex size in primates and their average group size.
- 6 is a plausible worst-case diameter for social networks. Travers and Milgram found that the letters which reached the target generally arrived after about six hops, leading to the belief in pop culture that everyone is reachable within "six degrees of separation".

4 points for satisfactory explanation of 150, mentioning at least one of "magic number 150", brain size in primates, or Goretex's success in keeping group sizes below 150.
4 points for satisfactory explanation of 6, mentioning at least one of Travers and Milgram's experiment, the book Six Degrees, "six degrees of separation", or Kevin Bacon

c. If N is the network size, describe how to construct networks of arbitrarily large N satisfying K=N/2 and D=2. Feel free to illustrate with a digram.

The most natural answer is to construct a bipartite graph. Divide the vertices into two groups, so there are N/2 vertices in each group. Each vertex has an edge to every other vertex in the opposite group, so each vertex has N/2 neighbors. The distance between any two vertices in different groups is 1, and between any two vertices in the same group is 2, so it satisfies D=2. An example is drawn below:



2 points for describing (through words and/or through drawing) a network where each vertex has at most N/2 neighbors
2 points for describing (through words and/or through drawing) a network where the worst case diameter is 2.

Problem 6 (15 points) This problem asks you to discuss the various models for network formation discussed in class and the readings.

a. What assumptions does the Erdos-Renyi model make that are inappropriate for many networks? Discuss at least one specific naturally occurring network (social, biological, economic, organizational, etc.) and how the Erdos-Renyi model fails to represent it well.

The Erdos-Renyi model assumes that edges are formed independently and at random. One specific network that the Erdos-Renyi model fails to model well is the network of friendships on Facebook. In Erdos-Renyi, the expected value of the clustering coefficient of a vertex is the same as the baseline probability of edges. In the Facebook network, however, two people who have mutual friends are much more likely to be friends with each other than two people who do not have mutual friends. One often meets new friends by being introduced to them by an existing friend. In addition, people tend to become friends because of shared classes or activities, and thus all members of a group are more likely to be friends with each other than with an arbitrary person outside of the group.

2 points for mentioning at least one assumption of the Erdos-Renyi model (such as that edges are chosen at random, that edges are independent of each other, or that degrees can be modeled by a Poisson distribution)
1 point for mentioning a plausible naturally occurring network
1 point for describing a property of the chosen network
1 point for explaining that Erdos-Renyi doesn't have that property

b. What was Watts' motivation for introducing the alpha model? Discuss how the alpha model is a better/worse/equally good representation for the network you mentioned above.

Watts introduced the alpha model to better model the high clustering he found in several naturally occurring networks. In the alpha model, the probability of including an edge between two vertices is not a constant, but instead is dependent on the number of neighbors they share. The alpha model is a better representation than Erdos-Renyi for the Facebook friendship network I discussed above, because it models the high clustering found in the friendship network.

2 points for saying that Watts introduced the alpha model to model high clustering
1 point for choosing a plausible choice of better/worse/equally good for the network chosen above (most likely, the choice was better)
2 points for discussing a property of your network that the alpha models that Erdos-Renyi doesn't (or vice-versa, if you're arguing for worse)

c. What was the motivation for introducing the preferential attachment model? Discuss how the preferential attachment model is a better/worse/equally good representation for the network you mentioned above.

The preferential attachment model was introduced to model the heavy tail of degree distributions found in several naturally occurring networks. In many networks, there are a few vertices with degree much, much higher than the mean, which isn't reflected in either Erdos-Renyi or the alpha model. The preferential attachment model is therefore a better representation for the Facebook friends network, since it models the existence of Connectors who have several times more friends than the average person.

2 points for saying preferential attachment was introduced to model the heavy tail of degree distributions.

1 point for choosing a plausible choice of better/worse/equally good for the network chosen above (if you interpreted the comparison as being between Erdos-Renyi and preferential attachment, you probably chose better; if you interpreted the comparison as being between the alpha model and preferential attachment, you might have chosen better or worse or equally good).

2 points for discussing a property of your network that the preferential attachment model reflects but Erdos-Renyi (or the alpha model) does not.

Problem 7 (10 points) In the directed network below, the vertices represent web pages and the edges represent hyperlinks. In each of the following questions, you do not have to justify your answer --- simply answer each directly.

a. Which vertex has the highest PageRank? Which has the lowest?

F has the highest because it has only incoming edges. A has the lowest because it has only outgoing edges.

Now add a new vertex labeled G which has 2 edges. One edge goes from G to A, and the other from G to B.

b. Which of the vertices G and F has greater PageRank (or do they have the same)?

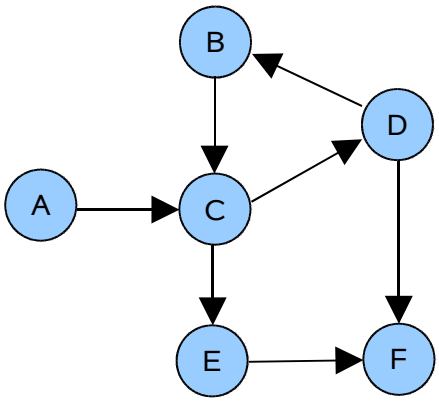
F has greater PageRank than G, because F has only incoming edges and G only outgoing.

When G and its edges are added, will the PageRank of the following vertices become higher, lower or remain the same?

c. Vertex B

d. Vertex C

e. Vertex D



B,C and D will have higher PageRank. The new vertex G sends more weight "downstream", so the PageRank of A and B will increase and subsequently so will the PageRank of C and D.

2 points for every question. No explanation required.

Problem 8 (15 points) Imagine a distributed population of people playing games over a network in which each player controls their own vertex. In each game, players must choose a color for their vertex from a limited set of colors. At all times players can see their own current color and those of their network neighbors, but nothing else. Players can change their color at any time. Feel free to illustrate your answers with diagrams.

a. Consider first a social coordination game in which there is a collective goal that every vertex in the network eventually chooses the same color simultaneously (unanimous consensus). Restricting your answer to connected networks, describe one network that you think should be relatively easy for this problem (in terms of the time it takes the population to reach consensus), and another network that you think should be relatively hard. Briefly justify your answers.

If players cannot see the degrees of their neighbors, a long chain structured network should make consensus hard because each node can only see the colors of up to two other nodes and therefore has a very limited view of the network and because chains have maximum average diameter which makes coordination take a long time to propagate across the network.

Conversely, networks which look approximately like a large clique will make coordination easy since players can see what global trends are forming and synchronize to them.

In the case that players can see the degrees of their neighbors, then preferential attachment style networks will also make coordination easy since players with low degrees can play "follow the leader" and copy the color of their highest degree neighbor who is likely to have a more complete picture of the network than they do. Additionally, preferential attachment style networks have small diameter which allows coordination to propagate quickly.

+2.5 points for giving a network structure that makes coordination easy and for providing a reasonable justification of why it makes it easy.

+2.5 points for giving a network structure that makes coordination hard and for providing a reasonable justification of why it makes it hard.

b. Now consider a social differentiation game in which each player's goal is to be a different color than all of its neighbors. Again restricting your answer to connected networks, describe one network that you think should be relatively easy for this problem (in terms of the time it takes the population to reach a point where everyone is a different color than all their neighbors), and another network that you think should be relatively hard. Briefly justify your answers.

In the differentiation game, a chain structured network makes differentiation easiest since each player has at most two nodes to differentiate itself from, colorwise. Conversely, highly entangled networks (e.g., a single large clique) will make players very constrained in their color choices and hence make differentiation hard.

+2.5 points for giving a network structure that makes differentiation easy and for providing a reasonable justification of why it makes it easy.

+2.5 points for giving a network structure that makes differentiation hard and for providing a reasonable justification of why it makes it hard.

c. Consider both of the games above on networks generated according to the Erdos-Renyi and Preferential Attachment network formation models. For each game, do you think that game should be easy or hard on each formation model? Why? You may refer to your answers to parts a. and b. if desired

Erdos-Renyi Networks:

Erdos-Renyi networks with large "p" (i.e. dense networks) could make coordination easy since players will be able to see the colors of many nodes in the network and can therefore follow whatever trends are developing. Additionally, such networks have small diameter making coordination propagate quickly. Erdos-Renyi networks with large "p" could make differentiation hard since such networks will be very tangled and thus creating many constraints in terms of what colors are available to players.

Erdos-Renyi with small "p" should make coordination hard since all players will have a small number of neighbors and will therefore have a very limited view of the network. Additionally, such networks have large diameters, making coordination take a long time to propagate. Erdos-Renyi with small "p" should make differentiation easy since players will have a small number of numbers to differentiate themselves from.

+1.25 points for stating that Erdos-Renyi networks can make coordination easy if "p" was large due to density of edges or for stating that Erdos-Renyi networks can make coordination hard if "p" was large due to "over-cluttering" of edges.

+1.25 points for stating that Erdos-Renyi networks can make differentiation hard if "p" was large (due to density of edges) or for stating that Erdos-Renyi networks can make differentiation easy if "p" was small due to sparsity of edges.

Preferential Attachment Networks:

In the case that players cannot see the degrees of their neighbors, preferential attachment networks could make consensus hard since most players have a small number of neighbors and thus have a limited picture of the network. At the same time, preferential attachment networks do have small diameter which could facilitate coordination. For the differentiation game, preferential attachment networks could make differentiation easier since most players in a PA network will have small degree and will therefore have few neighbors to differentiate themselves from. Additionally, such networks are tree structured and therefore have zero clustering. Consequently, they do not create tangled cycles of color constraints. The downside to PA networks for

differentiation is that they invariably possess a small number of nodes with very high degree who are extremely constrained in their color choices. Such overly-constrained vertices can become stuck in terms of what colors they can choose.

In the case that players can see the degrees of their neighbors, both coordination and differentiation might be easier on preferential attachment networks since players can defer to their high degree neighbors when making color choices.

+1.25 points for stating that Preferential attachment networks can make coordination easy due to hubs and small diameter.

+1.25 points for stating that Preferential attachment networks can make differentiation easy due to absence of clustering, or for stating that Preferential attachment networks can make differentiation hard due to overly constrained hubs with many neighbors.

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.

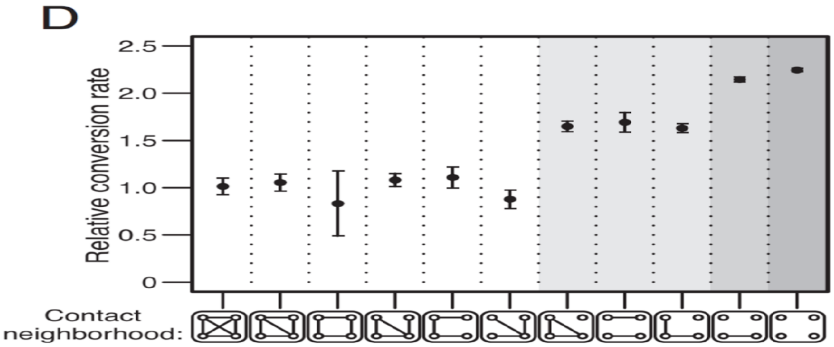
i. The only properties known to have a tipping point or threshold behavior in the Erdos-Renyi model are giant component and small diameter.

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)

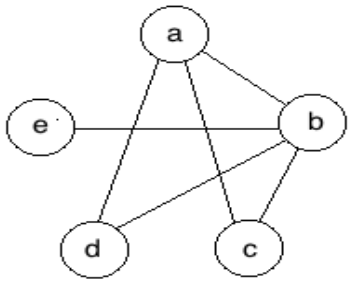


This diagram is taken from the article "Structural Diversity in Social Contagion." The goal of the experiment is to analyze the growth of Facebook. Facebook recruits new users by emailing them and showing that some of their real world friends are already using Facebook. The authors find that recruitment 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.

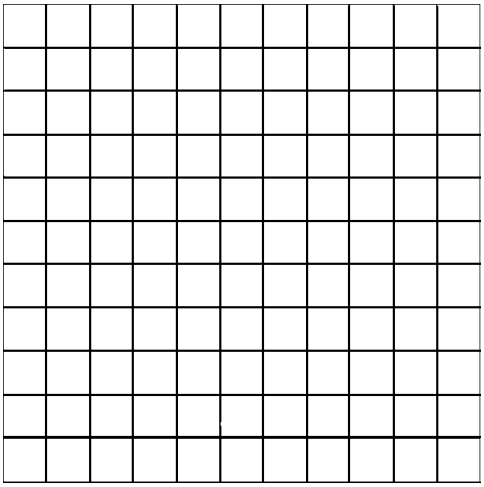
a. Is the property of containing H as an induced subgraph a monotone property? Why or why not?

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)

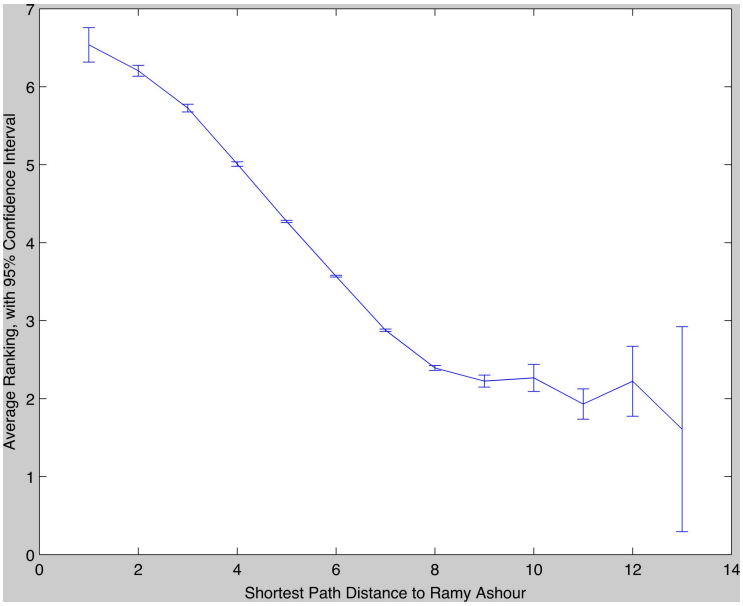
In class and the readings, we examined three different network formation models that will all yield a clustering coefficient higher than the overall edge density. Briefly but as precisely as you can, describe each model, and for each, say whether you think the model generates networks with clear "community" 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.