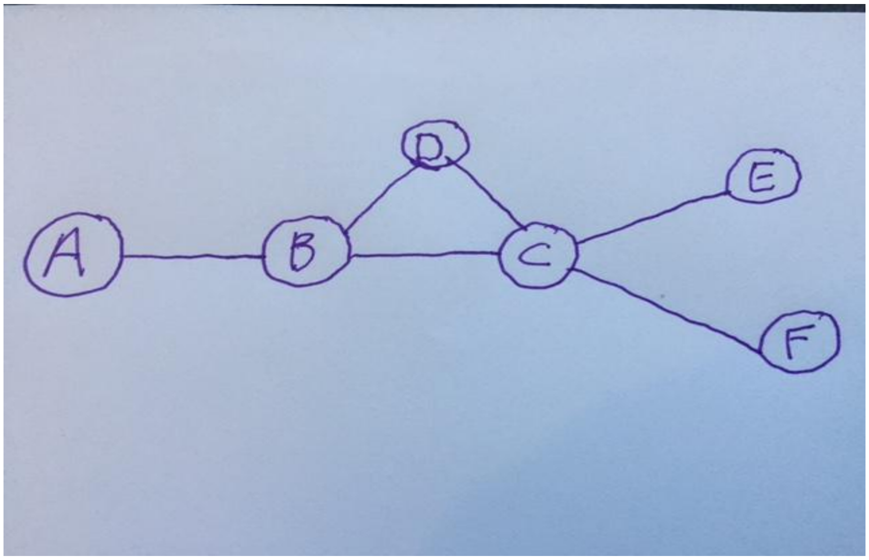
Homework 1
Networked Life (NETS 112)
Fall 2017
Prof Michael Kearns

Posted October 1, 2017. Due in hard-copy format at the start of lecture on Tuesday, October 17. Please don't forget to write your name and staple the pages together.

Collaboration of any kind is NOT permitted on the homework.

Your Name:

1. Consider the following network:



(a) Compute the average-case diameter of the network. Show your work.

There are 15 pairs of vertices. Compute the shortest path for each pair, sum up the path lengths, and divide by 15. The answer is 26/15 or about 1.733.

The lengths of the shortest paths between each pair of vertices are the following:

AB:1
AC:2
AD:2
AE:3
AF:3
BC:1
BD:1
BE:2
BF:2
CD:1
CE:1
CF:1
DE:2
DF:2
EF:2

The sum of these is 26, and there are 15 pairs of vertices, so we get 26/15

(b) Recall the “economic altruism” model in which each vertex starts with $1, and at each step divides its current wealth evenly amongst its neighbors. Compute the equilibrium wealth for each vertex. Show your work.

Recall that the formula for equilibrium wealth for a vertex v in a connected graph is $N*d(v)/D(G), where $N represents one dollar for each of the N vertices, d(v) is the degree of vertex v (the number of edges incident to v), and D(G) is the total degree of the graph (the sum of the degrees of all vertices in the graph, or twice the number of edges).

The graph has 6 vertices, so N=6
The total degree D(G) of this graph is 12

Vertices A, E, F have degree 1, so they each have equilibrium wealth $6*(1/12)=$.50
Vertex B has degree 3, so its equilibrium wealth is $6*(3/12)=$1.50
Vertex C has degree 4, so its equilibrium wealth is $6*(4/12)=$2.00
Vertex D has degree 2, so its equilibrium wealth is $6*(2/12)=$1.00

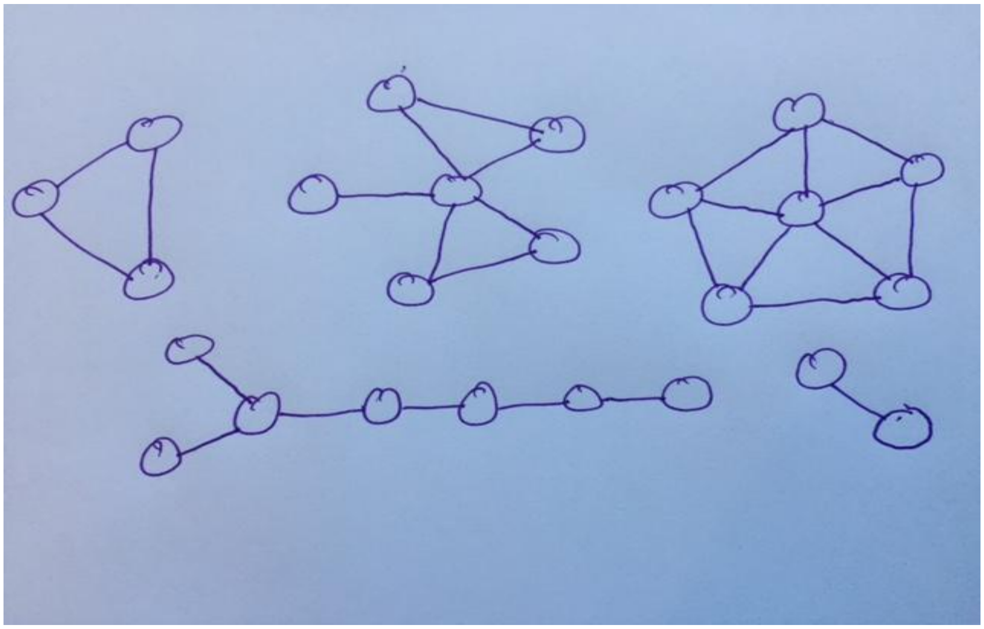
(c) Consider modifying the network by deleting some edges, and adding some new edges. Clearly indicate the modification with the *fewest* deletions and additions that causes all vertices to have the same equilibrium wealth, and indicate what that wealth is.
You can remove edge CD and add edges AD, AF, DE, and EF to make every vertex degree 3 using 5 changes. The equilibrium wealth in any setting where every vertex has the same degree will be $1 at each vertex. You can also remove BC and CF/CE, and add AF/AE and EF.

(d) Draw a connected network with the fewest vertices you can in which the number of distinct or different equilibrium values is exactly 4. Hint: it can be done with N = 6; I’m not sure if smaller N is possible or not.

There are lots of ways to do this with 6 vertices. One acceptable answer is the graph provided by the question. In general, such a graph must be connected and there must be at least four different values for the degrees of the vertices. The graph provided by the question has vertices with degree 1,2,3, and 4.

2. Consider the following network, which consists of multiple connected

components:



(a) Consider the contagion process considered in class, in which a vertex v is chosen at random to be infected, and the infection then spreads deterministically to kill all vertices in the connected component of v. What is the expected or average number of vertices killed for this network? Show your work.

The expected number of vertices infected is the sum over each component of the probability of that component being selected times the number of vertices in that component. The answer is 67/12, or about 5.583

To get this, we see that there are 24 total vertices, and we pick one to infect at random.

If we pick one from the first component, we infect 3 vertices, and the probability of this happening is 3/24, so this component contributes 3*(3/24)=9/24 vertices to the expectation.

If we pick one from the second component, we infect 6 vertices, and the probability of this happening is 5/24, so this component contributes 6*(6/24)=36/24 vertices to the expectation.

If we pick one from the third component, we infect 6 vertices, and the probability of this happening is 5/24, so this component contributes 6*(6/24)=36/24 vertices to the expectation.

If we pick one from the fourth component, we infect 7 vertices, and the probability of this happening is 7/24, so this component contributes 7*(7/24)=49/24 vertices to the expectation.

If we pick one from the fifth component, we infect 2 vertices, and the probability of this happening is 5/24, so this component contributes 2*(2/24)=4/24 vertices to the expectation.

Finally, we take the sum (9/24)+(36/24)+(36/24)+(49/24)+(2/24)=(134/24)=(67/12)

(b) Suppose you are allowed to "immunize" exactly one vertex that can no longer be infected. Which vertex would you choose to make the average number of vertices killed as small as possible? What would the new value for this average be? Show your work.

You should immunize the fourth vertex from the right in the lower-leftmost component. Now that this vertex is "immunized" and can no longer be selected, there are 23 vertices which can be chosen. The expected number infected is 103/23 or about 4.478

We repeat a similar calculation from before. Four of the components are unmodified, so we have a similar calculation for those, except now there are 23 vertices:

The first component contributes 3*(3/23)=(9/23) to the expectation
The second component contributes 6*(6/23)=(36/23) to the expectation
The third component contributes 6*(6/23)=(36/23) to the expectation
The final component contributes 2*(2/23)=(4/23) to the expectation

Now we've created two new components, each of size 3. Therefore, they each contribute 3*(3/23) to the expectation, for a total of (18/23).

Finally, we just sum up these numbers: (9/23)+(36/23)+(36/23)+(18/23)+(4/23)=(103/23)

(c) Suppose you are forced to add an edge between vertices in different connected components. Which edge would you choose to make the average number of vertices killed as small as possible? What would the new value for this average be? Show your work.

You should add an edge between the two smallest components: the triple in the upper-left and the pair in the lower-right. The new expected number of vertices infected is 73/12, or about 6.083.

Again, the calculation process is very similar, except we again have 24 vertices. For the three unmodified components, the calculation is identical to that of part a: the two components of size 6 contribute (36/24) each to the expectation and the component of size 7 contributes (49/24) to the expectation.

We now also have a new component of size 5, which contributes 5*(5/24)=(25/24) to the expectation.

Summing these up, we get (36/24)+(36/24)+(49/24)+(25/24)=(146/24)=(73/12)

3. Carefully consider the assigned reading “Can Cascades be Predicted?”.

(a) In 5 brief bullet points, succinctly describe the main questions the authors are interested in.

-How accurately can the growth of cascades be predicted?
-What factors contribute to the growth of a cascade?
-What features are useful in predicting cascade growth?
-Does the future behavior of a cascade become more or less predictable as the cascade unfolds?
-Can the network structure of a cascade also be predicted?
-Do user-initiated cascades differ in the predictability and underlying structure from page structure cascades?

(b) In 5 brief bullet points, succinctly describe their problem formulation and methodology.

-Author’s prediction problem asks: Given a cascade of size k, will the cascade double in its size to reach at least 2k nodes
-Start experiment with 150,575 photos from Facebook that were shared 5 times
-Tried to predict whether each would hit 2 * 5 reshares
-Used a set of features and machine learner to predict future size
-Analyzed which features were actual strong determinants of whether a cascade would double in size

(c) In 5 brief bullet points, succinctly describe the main findings of the study.

-Actual content features i.e. what is a weak predictor of how widely disseminated a piece of content would be
-As the number k (reshares) grows, the features of the original poster becomes less important
-It is easier to predict the behavior of a cascade, the bigger it gets
-Initial structure heavily influences eventual size-successful cascades get many views in a short amount of time and achieve high conversion rates
-The deeper the cascade grows, the more likely it is to last longer and continue being reshared

(d) Briefly discuss some of the limitations or drawbacks of the methodology and study that were discussed in lecture.

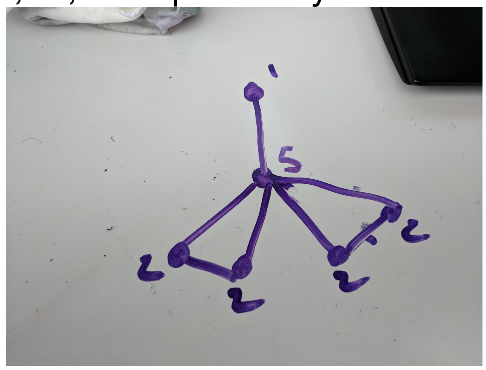
-Limited by choice of data
-Only one social media platform was used
-Not a clear distinction between a cascade being viral versus a cascade being popular
-The diffusion within Facebook was driven by mechanics (distinction between [pages and users and mechanisms by which they interact; liking and sharing are specific to Facebook)
-It only examines cascades independently; could examine the interaction
-For each category of features, we don't really know whether other features might have been more useful than the ones chose. In other words, the conclusion that the content is unimportant may just be a consequence of the authors' choosing bad/incorrect/uninformative features.

(e) Discuss what, if any, conclusions can be drawn from the study regarding the possibility of designing or selecting images that can go viral strictly on the basis of their content.

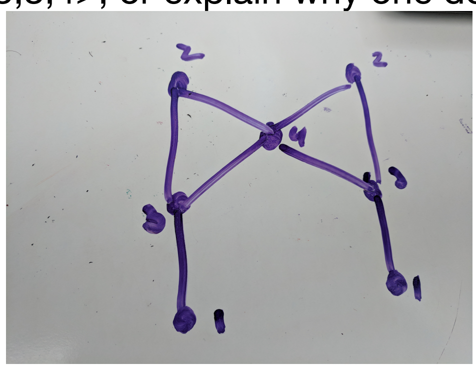
-Content is not a good predictor
-Temporal features are a good predictor
-80% accuracy of prediction
-The earlier the uploads of the same photo, the larger the cascade
-Independent resharings of the same photo can generate different cascade sizes
-The paper doesn't provide negative evidence for engineering 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

4. The *degree sequence* of a network is a list of the degrees of all the vertices. For example, the degree sequence of the network in Problem 1 is <1,3,4,2,1,1>. Note that the order of the degrees is unimportant; i.e. we could have given it as <1,1,1,2,3,4>.

(a) Draw a connected network with N=6 vertices in which the degree sequence is <1,2,2,2,2,5>, or explain why one does not exist.

One possible answer: 

(b) Draw a connected network with N=7 vertices in which the degree sequence is <1,1,2,2,3,3,4>, or explain why one does not exist.

One possible answer: 

(c) Draw a connected network with N=7 vertices in which the degree sequence is <1,1,2,2,3,3,5>, or explain why one does not exist.

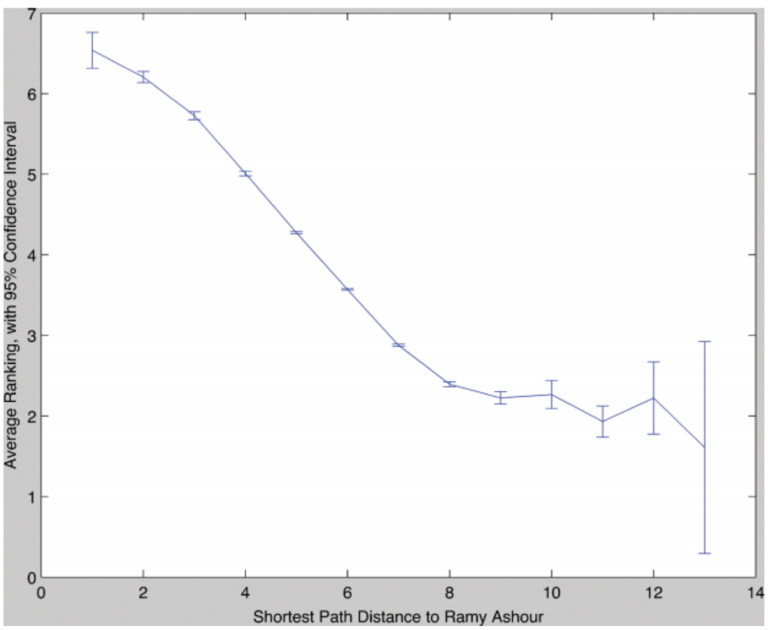
This is impossible. The total degree (the sum of the degree sequence) must be even. The sum of this degree sequence is odd, therefore no graph with this degree sequence exists.

(d) Think of N as being large, and consider connected networks in which the all of the degrees are 4, i.e. the degree sequence is <4,4,4,…,4>. Clearly describe the networks meeting this condition that have the lowest and highest diameters you can design.

There are a few ways to approach this problem. The first is to start with a graph we discussed where every vertex has the same degree: a cycle. To make the diameter of this graph smaller, we add long-distance edges. Each vertex already has degree 2, so if we add to each vertex an edge which goes ⅓ and ⅔ of the way around the circle, no two vertices are more than n/6 hops apart. To keep the diameter of this graph high, we add a lot of short-distance edges. If we add an edge from each edge to the neighbors of its neighbors (i.e. if A-B-C-D-E are a portion of the cycle, we add edges A-C and C-E), we don't make existing paths much shorter, and no two vertices are more than n/4 hops apart.

The other way to approach it is to start with a grid graph, which is a graph where most vertices already have degree 4. For a low diameter graph, we start with a square grid graph and we connect vertices along the border to vertices on the opposite border. In this graph, no two vertices are more than about sqrt(n) hops apart. For a high diameter graph, we start with a very long and skinny grid graph, like 2xN and then connect vertices on the border with nearby vertices. In this graph, no two vertices are more than about n/4 apart.

5. Consider the figure below, which is from the assigned article "The Small-World Network of Squash".



(a) Carefully and clearly describe exactly how this plot was generated, and what point it is making. Use precise terminology and definitions from class.
-Graph represents squash players' distance to Ramy Ashour
-Axes are players' distance to Ramy Ashour and confidence interval
-Players' distance to Ramy Ashour is measured by shortest distance paths
-The graph shows that players' Ashour number is good measure of their ranking

(b) Why are the vertical bars wider at the left and the right?
-The vertical bars are wider at the left and right because their is more variance the closer and farther you are form Ramy Ashour

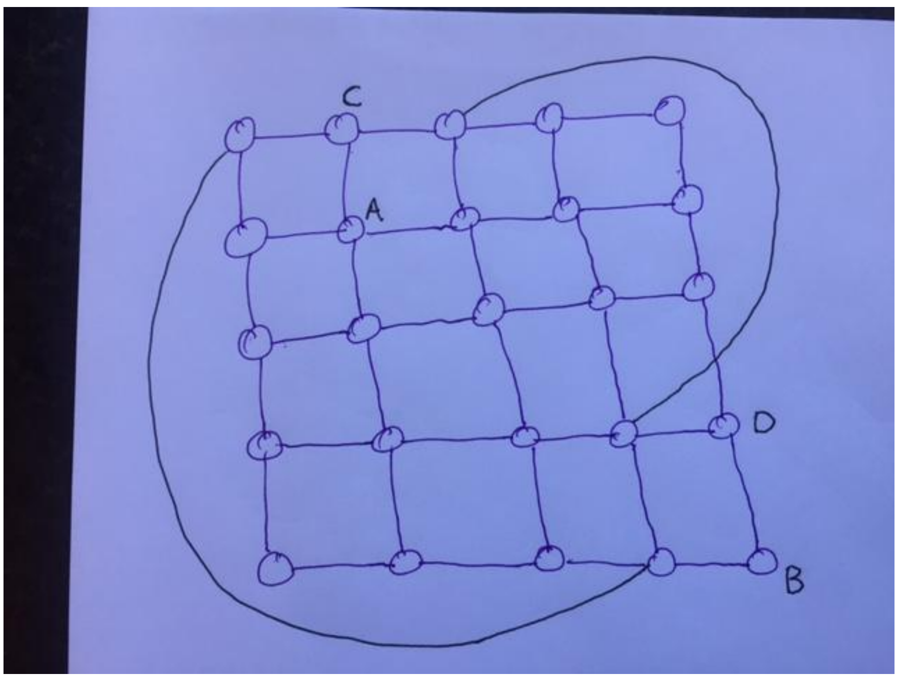
-More variance leads to a wider range of confidence intervals
-Location and skill level can also lead to variation

(c) Suppose that we used the same methodology to generate the figure, but used a mediocre player as the focal vertex instead of Ramy Ashour. Discuss how you think the figure would now look, and justify or explain your answer. Both “mathematical” (e.g. in which you make assumptions about the distribution of ratings) or “sociological” (e.g. in which you discuss the possible forces governing the arrangement of matches) answers are acceptable; just be clear about your assumptions.

What follows is an outline of just one acceptable answer; others are also fine as long as they are well-argued using course concepts

We would expect the leftmost point on the graph to be lower, as a mediocre player certainly is not as highly ranked as a professional like Ashour.
- We expect there to be forces which would cause a trendline to increase, as a mediocre player probably plays other mediocre players who are slightly higher ranked, and those players play against other individuals who are slightly higher ranked than themselves, and so on, up to the level of professional players.
-We also expect there to be forces which would cause the trendline to decrease, as a mediocre player probably plays players slightly worse than himself, and those player play against individuals slightly worse than themselves, and so on.
- Putting these together, the slope of the trendline is inconclusive, but we should not be surprised if it is flat or only slightly increasing or decreasing due to these competing forces.
- We expect the error bars to be narrower on the left than the right. This is because a mediocre player may play against only other mediocre players, so there is only a small variance in skill. However, he may play people slightly better and slightly worse, so there is a wider range of talent at a distance of one hop. Those players at distance 1 will play against a still wider range of talent, so the error bars should be wider at distance 2 than at distance 1. The same reasoning applies for the error bar being wider at distance 3 than 2, and so on. additionally, using any player as a starting point, we might expect there to be some low ranked players at a high distance, such as a community of recreational players with very weak connections to the tighter cluster of professional and high-ranked play. This would also cause higher variance at the right side of the chart

6. Consider the assigned article “Navigation in a Small World” in the context of the network below, which is an underlying 5 by 5 grid augmented by two “long distance” edges. Using the algorithm for navigation examined in the article, for each source-destination pair below, write the length of the *shortest* and *longest* possible paths found by the algorithm.



In each of these, the longest paths found by the algorithm are those which uses only the grid lines. In the cases where the shortest path found by the algorithm is shorter than one which uses grid lines, the algorithm gets lucky and finds a useful long-distance edge.

(a) From A to B Shortest: 6 Longest: 6

(b) From B to A Shortest: 4 Longest: 6

(c)	From C to B	Shortest: 4	Longest: 7
(d)	From B to C	Shortest: 3	Longest: 7
(e)	From A to D	Shortest: 5	Longest: 5
(f)	From D to A	Shortest: 4	Longest: 5
(g)	From C to D	Shortest: 3	Longest: 6
(h)	From D to C	Shortest: 3	Longest: 6