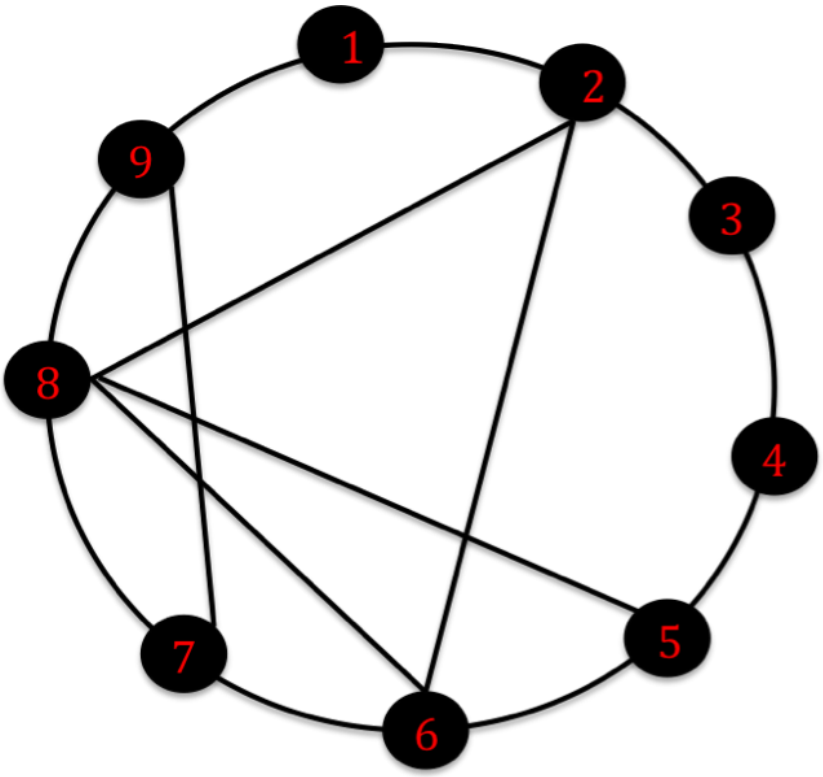
Homework 1: Networked Life (MKSE 112), Fall 2012, Prof. Michael Kearns

Issued Oct 2, 2011. Due as ***hardcopy*** in class Tuesday Oct 16. Remember that collaboration on homework is ***not*** permitted. Don't forget to ***staple and write your name.***

(40 points) Consider the undirected network shown below.



a) (5 points) What are the minimum and maximum degrees in the network? 2,5

(2.5 points each part)

b) (5 points) What is the *worst-case* diameter of this network (largest shortest-path distance between any pair of vertices)? Put a "*" next to two vertices whose distance equals the diameter. 3

(2 points for the diameter and 3 points for the pair of vertices.)

c) (15 points) What is the clustering coefficient of the network? What is the overall edge density of the network? Would you consider the clustering coefficient high or low, and why?

CC= (0 + 1/6 + 0 + 0 + 1/3 + 1/6 + 2/6 + 2/3 + 4/10 + 1/3)/9 (5 points)

Partial credit is given if you have failed to calculate the CC of some of the nodes. Interestingly most of you calculated the CC of node 6 wrong!

Edge density: 14/36 = 0.389 (5 points)

CC is low in comparison with the edge density. (5 points)

d) (15 points) A *coloring* of a network is an assignment of a color to each vertex such that all pairs of neighboring vertices have different colors. Coloring can be viewed as a problem of social differentiation --- for instance, nobody wants to wear the same outfit as a colleague on a work day, and it's better if everyone's cell phone ringtone is different than those of their close friends. On the network above, find a coloring that uses as few distinct colors as possible. Write each vertex's color next to it.

3 colors are required. If you used more than 3 colors or your coloring is not correct (i.e. there are neighboring vertices with the same color) you are not given any credit.

(30 points) In this problem, we will consider the directed network given by the web graph --- that is, the network in which the vertices correspond to web sites or pages, and there is a directed edge from page A to page B if there is a hyperlink on page A to page B (there may or may not be a reciprocal link from B to A). In this problem, we will be considering only directed paths --- that is, paths that follow the edges in the ``right direction''.

(15 points) Beginning at www.upenn.edu, find the *shortest* directed path you can that arrives at a web page hosted in the Internet domain of the *smallest* country (by population) that you can. By a directed path, we mean a sequence of complete URLs such that the page corresponding to each URL contains a link to the page corresponding to the next URL in the sequence. Your path may *not* contain URLs in the domain of any major social media site or search engine --- i.e. Facebook, Twitter, Google, etc. Write down the full URLs of the web pages along your path. If you like you may give a couple of different paths. You may use any methods you like, including search engines, to *find* your path, but the path itself must obey the rules specified above. Your path will be graded for its correctness, as well as how short it is and how small the destination country is. To find country-level Internet domains, use the following page:

http://en.wikipedia.org/wiki/List_of_Internet_top-level_domains#Country_code_top-level_domains

To find country populations, use the following:

http://en.wikipedia.org/wiki/List_of_countries_by_population

Main constraint is that you carefully follow the rules --- only links are allowed. If you include a non-link step like requiring entering text into a search box, you should receive 0 points since we explicitly forbade this.

(a) Most paths with lengths between 4-6 have been awarded points between 13-15 depending on how small your destination country was.

(b) Some of you have violated the links-only rule. Some of you seem to have (arbitrarily) chosen a small destination country without mentioning why (and stuck to it even when the path lengths become fairly large). Similarly getting stuck with trying to reach a website you thought would

be a good connector, even when it increased the path length, wasn't necessarily a good idea. Some of you ended up on sites with large paginated lists, ended up manually clicking through all pages to look for a particular entry (or hoping to find a suitable one) and counted all those links as collectively as one hop. A few of you also specified a relatively long path (8-10 hops) to an arbitrarily chosen small country and then claimed 'I don't think there can be a smaller path' without saying why or exploring alternative destinations or paths. Points have been taken off in all these instances.

Some of you tried creating a web crawler. This would be ok, although unnecessary, if you succeeded. If not, no credit has been given for trying to build one and eventually not coming up with a path.

(15 points) Discuss how you went about finding your short path(s) above --- what information you used about the pages visited, how you searched locally from each page, what other tools or methods you employed, etc.

(30 points) In this problem, you are asked to consider your own *Facebook neighborhood network*. This is the network of connections between your Facebook friends, but *excluding yourself and all your connections* (you are needed only to identify your friends).

(15 points) Estimate the edge density of your FB neighborhood network, and describe your methodology and calculations carefully. (Note that the edge density of your FB neighborhood network is the same as your own clustering coefficient.)

The methodology we were expecting here is that you somehow choose a "random" subset of your friends to compute the cc's and use the average as an estimate. The main thing you should NOT do is select a "biased" subset, like all their friends in your fraternity for example.

Choosing the right set and proper discussion of it has 10 points. Correct calculations has 5 points.

Some of you have utilized various software and apps to get the data you need. It is fine and you are given full credit if you worked on the data properly. Very few of you have used a software that gives the answer directly. For this one time, since we had not explicitly forbidden this you are given the grade. But do not repeat this for next time, the question is intended to be solved by you, not a computer program.

Some of you have chosen your sample set in an (unnecessarily) complicated way: e.g. splitting the friends into a few main groups, then choosing a random sample set from each group and combining all these sets as the final sample set. There are a few common mistakes that you have made with this approach: first note that you final sample is biased, unless the number of sample

you choose from each group is proportional to its size, or you take the weighted average of your samples as your estimation. If you did not do that 3-5 points are taken away from you.

The other mistake that many of you made was to estimate the numbers you need without providing any/sufficient evidence why it might be close to reality. Depending on where and how many times you did this, 4-10 points are takes from you.

If you included your own vertex in the calculations but the rest of your answer is correct, 3 points is taken away.

If edges are counted twice 1 point is taken away.

(15 points) Identify a group of at least 10 vertices for which you suspect the clustering of connectivity of that group is much higher than overall edge density computed in Part b). Verify this suspicion numerically, and describe your methodology and calculations carefully. Discuss how you identified this group, and why you think their connectivity is clustered.

Here of course you SHOULD choose a biased group where you suspect that there is greater interconnectivity among that subset of friends.

Choosing the right set and proper discussion of it each has half of the credit.

The calculations have 8 points. Many of you have computed the CC of the group with respect to the entire network. This is wrong and therefore you are not given the credit of discussion.