3/18/10

CIS 112 - Networked Life

Midterm Solutions

Questions 2,3,4,7 were graded by Mickey Brautbar
Questions 1,5,6,8 were graded by Kareem Amin

Question 1)

a. False : Preferential Attachment does not, for example, explain high clustering coefficients.
b. False : There are six categories: SCC, IN, OUT, TENDRILS, TUBES, DISC.
c. True
d. False : Empirically, people tend to use geographic information more towards the beginning of the chain.
e. False : The number of possible friendships among your friends grows like K squared.
f. True
g. False : No, but it might have a high hub score.
h. False : Christakis is a sociologist (and a physician). Fowler is a political scientist.
i. False: In a social setting, clustering coefficients measure how often a friends share a mutual friend. Alternatively, it measures how many triangles are in the network. This does not necessitate that the vertices necessarily have anything "similar" about them.
j. False : In Gladwell's terminology, a "connector" has high degree.

Grading policy:

i. Each wrong answer [-1 point].

Question 2)

a. "www.wheresgeorge.com" is an on-line bill tracking website. A participant with a one dollar bill logs into the website service and records the bill's number and current location.
b. The authors are interested in tracking human movements across the USA. The locations of the dollar bills are considered as a proxy to the real movement patterns of their owners.
c. The main finding is that over a short time frame (up to two weeks) the distance traveled by a bill, between its current location and the next, follows a power-law distribution with an exponent of roughly 1.6.
d. Prof. Kleinberg describes a theoretical model where people are located on a two dimensional grid and a person gets an additional k random long distance acquaintances according to a power-law distribution with exponent alpha. Kleinberg's shows that efficient navigation based on local knowledge is only possible when alpha equals 2.

   We can draw a connection between the paper and Kleinberg's model by interpreting the dollar bill travel distance as a long-distance connection between acquaintances. The findings of the authors of "The Scaling Laws of Human Travel", that human migration patterns follow a power law with exponent 1.6, can then be interpreted as an instantiation of the Kleinberg's

model with parameter alpha of 1.6. Under this interpretation one should then discuss whether 1.6 is close enough to 2 to get an efficient navigation. Any argument is fine as long it is stated clearly.

Grading policy:

i. Ill or partial description of each question part [up to -5 points for each].

Question 3)

a) Recall that in the Erdos-Renyi model an edge appears with probability p and is absent with probability of 1-p.
We say that a property P has a tipping point at q if:
When p < q the probability that the network obeys P is close to zero.
When p > q the probability that the network obeys P is close to one.

b)
When p is about 1/N we start seeing a large component of size bigger than N/2.
When p is about log(N)/N we start seeing a connected network.
When p is about 2/square-root(N) we start seeing a diameter two network.

Grading policy:

ii. Ill description of tipping phenomena [up to -4 points].
iii. Wrong order of tipping properties [-2 points each].
iv. Wrong property/missing property description [up to -2 points each].

Question 4)

Two such structural properties that frequently occur simultaneously but appear to be in tension are small vertex degrees and low diameter. The Erdos-Renyi model can generate networks with these two properties for a range of its parameter p.

Grading policy:

i. Suggesting two properties that are not "in tension" with each other, or suggesting improper properties (such as non structural ones) [up to -5 points].
ii. Providing a model that cannot explain for the two properties proposed [up to -5 points].
iii. Partial / insufficient description of model [up to -5 points].

Question 5)

1. PageRank is the most famous "signal" used by the Google search engine. PageRank measures the "importance" of a webpage based on its location in the directed hyperlink network. One interpretation of the PageRank of a website is that it is the probability that after surfing the web long enough, a random surfer would have landed on that website. This is helpful for web search because it allows Google to determine the relevance of a website using something beyond just on-page textual clues (which are easily manipulated by the website's owner).
2. Another signal that Google uses is whether a page includes terms that are related to, or

synonyms for, the query that was searched. Here such synonyms are deduced from the data collected from Google's massive user base. People searching for "dogs" might subsequently search for "puppies". After examining enough such searches the search engine will learn that "puppies" is related to "dogs."

3. Another useful signal is what bi-grams (tri-grams, etc) exist in a query, or in the text of a webpage. For example, Google can interpret the query "new york times" as consisting of a single group of words that go together. However, "new york times square" should be interpreted as two groups of words "new york" and "times square," thus sparing the user articles about squares printed in *The New York Times*.

Grading policy:

Some people who did not read the article or who had trouble recalling its details began "core-dumping" anything they had ever read about the Google search engine. Because of this, only the **first three** signals listed were graded. The first two signals were worth three points each and the third was worth four points.

i. Writing about any signal that could be plausibly used in the Google search engine (regardless of whether it was mentioned in the article) [+1 point].
ii. Writing about a signal that was actually mentioned in the article, and suggesting why it might be helpful in web search [+ 3 (or +4) points].

Question 6)

1. One example is the distribution of city population sizes in a country. A natural "Rich Get Richer" process might be that people are drawn to larger cities because of the greater availability of job opportunities or public resources. The large population, in turn, causes the city to generate more opportunities.
2. Another example is one we saw in class: the number of reviews of an iPhone application. The most reviewed iPhone applications are more likely to be downloaded, and subsequently reviewed, since a reliable base of people has already evaluated the product.
3. Many people wrote about Zipf's law in natural language, but had trouble giving a sensible "Rich Get Richer" process to explain it. This is not surprising, since there is no agreed-upon explanation for why Zipf's law holds. A reasonable explanation might be that when writing text, the writer is less inclined to use more archaic words (as her reader will have no idea what she's talking about). On the other hand, not only is she more inclined to use common words, but the absence of common words will make the text similarly incomprehensible (imagine a text with no articles or conjunctions). Thus, already relatively common words have a tendency to be penned more often.

Grading policy:

i. An overly abstract description of network formation, amounting to Preferential Attachment, was worth at most two points, since the question asked for two **other** examples.
ii. Providing a quantity that is plausibly heavy-tailed [+2 points].
iii. Describing a natural "Rich Get Richer" Process to explain the quantity provided (+3 points).
iv. Conflating exponential growth/decay with heavy-tailed distributions, even if the example provided was valid [-1 point].

Question 7)

a)

Draw a star graph where one vertex is directly connected with each of the other vertices. The worst-case diameter is then 2 (see figure below). The average clustering coefficient is zero since there are no triangles in the network.

b)

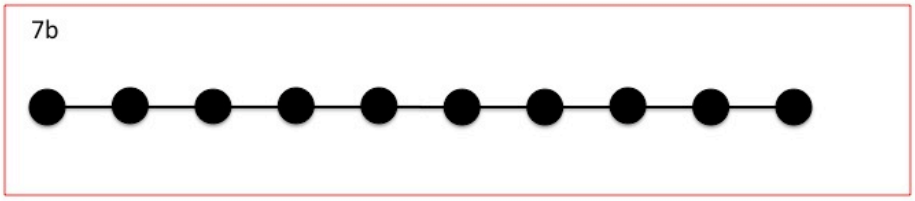
Take a line graph on the 10 vertices. The worst-case diameter is the worst possible and equals 9 (see figure below. The average clustering coefficient is zero since there are no triangles in the network.

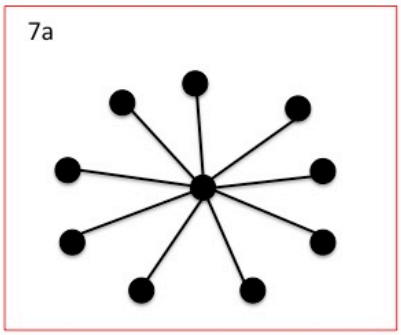
c)

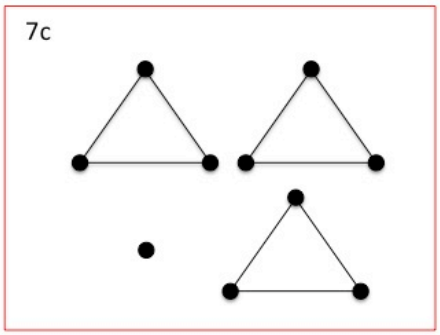
Divide 9 vertices into three groups of three vertices each. Then connect the vertices in each of the groups to form a triangle. Finally add one isolated vertex (see figure below). The clustering coefficient is (8*1 +1*0)/9 = 0.9.

Grading policy:

i. Not providing a network with worst-case diameter of 2 for part a [up to -3 points].
ii. Not providing a network with worst-case diameter of 9 for part b [up to -3 points].
iii. Not providing for part c a network with average clustering coefficient of 0.9 [up to -3 points].
iv. Not computing, or giving invalid computation of the clustering coefficient for the network in part a or b [-1 points each].







Question 8)

a) The Framingham Heart Study was the basis of much of the authors' research. For starters, the data included physiological information that would not be accessible on Facebook. The data is multi-generational and spans a longer time period than the lifespan of all online social networks. The data was also obviously geographically restricted. Furthermore, friendships were asymmetric in the Study (one person could cite another person as a

friend and not be cited back), while this is not true on Facebook. People also tend to have higher degrees (i.e. be "hyper-connected" on Facebook).

b) Three examples are: Smoking, Obesity, and Happiness.

c) The "Three Degrees of Influence" phenomenon is the tendency for members of a network to be influenced by people within three degrees of them. After three degrees, the influence is negligible.

Grading policy:

i. Mentioning the Framingham Heart Study by name, or providing a detailed description of the study [+4 points].

ii. Providing a vague, but reasonable description of the Framingham Heart Study (e.g. "The Authors used survey data") [+2 points].

iii. An overly vague description that could be used to describe any experiment (e.g. "The authors collected empirical data.") [+0 points].

iv. Each example exhibiting contagion behavior (mass-hysteria could only be used once) [+2 points].

v. Providing the "obvious" part of the answer to (c), that one can influence people three degrees from herself in a network [+2 points].

vi. Providing the slightly more nuanced part of the answer to (c), that the influence disappears at four degrees [+4 points].