CS/EE 144 **Guru: John/Aditya**

HW 1: Warming up

Assigned: 01/06/16 Due: 01/15/16 at the start of lecture

We encourage you to discuss these problems with others, but you need to write up the actual solutions alone. At the top of your homework sheet, please list all the people with whom you discussed. Crediting help from other classmates will not take away any credit from you.

*You may turn in a hard copy of your assignment at the beginning of class or email your assignment to cms144.caltech@gmail.com. If you turn in a hard copy, **remember to keep a copy of your homework to use in your self-evaluation.**. If you turn in an electronic copy, **be sure to submit it as a single file.***

Start early and come to office hours with your questions! We also encourage you to post your questions on Piazza, as well as answer the questions asked by others on Piazza.

1 Warmup [20 points]

In this class, we will depend on you remembering your probability and graph theory basics. This question is meant to help you recall some of the basic results from these areas.

(a) (**10 points**) Let $X_1, X_2, \ldots$ be independent and identically distributed (i.i.d.) random variables with finite mean $E[X]$ and finite variance σ_X^2. Let $S_n = \sum_{i=1}^n X_i$. The goal of this problem is to prove the weak law of large numbers, i.e., to prove that

$$\lim_{n\to\infty} Pr\left(\left|\frac{S_n}{n} - E[X]\right| \geq \epsilon\right) = 0 \qquad \forall\, \epsilon > 0.$$

i) Prove Markov's inequality, which says that if X is a non-negative random variable with finite mean, then for all $t > 0$,

$$Pr(X \geq t) \leq \frac{E[X]}{t}.$$

ii) Prove Chebyshev's inequality, which says that if random variable X has finite mean $E[X]$ and finite variance σ_X^2, then for all $t > 0$,

$$Pr(|X - E[X]| \geq t) \leq \frac{\sigma_X^2}{t^2}.$$

iii) Use Chebyshev's inequality to prove the weak law of large numbers as stated above.

(b) (**10 points**) In a connected, undirected graph G, the *distance* between two vertices is length of the shortest path connecting them. The *diameter* of G is the greatest distance between two vertices. Distance and diameter are often encountered in various applications of graph theory, especially in communication and social networks.

Your task is to construct a graph where the diameter is more than 3 times the average distance between nodes. Can you generalize this to a case where the diameter is more than c times the average distance, where c is any positive constant?

2 Job Interview Questions [40 points]

This is a collection of four "job interview" type questions that will test to see how well you remember your probability.

(a) **Remove the bias (10 points)** Suppose you are out to dinner with your cheapskate friend and he suggests that you flip a coin to see who pays. Since your friend is really cheap you know he's trying to cheat you and will probably use a biased coin. So, you need to come up with a way to use a coin with an unknown bias p to get an unbiased random bit 0 / 1. How can you do this? What is the expected number of flips (in terms of p) to get one bit under your scheme? **Extra credit:** Is it possible to improve upon this?

(b) **A peer-to-peer system (10 points)** Suppose you are trying to download a file from a peer-to-peer system that works as follows. A file is separated into $n > 1$ chunks, which are then saved in n different servers. When downloading a file, you query the n servers in a sequence of rounds. During each round r you make a request for a chunk of the file and are allocated to one of the n servers uniformly at random. If you do not have the file chunk at the randomly selected server, you download it, and the total round takes time t_1. If you have already downloaded the file chunk at the selected server, you do not download it but have still spent time t_2 during the round. Your download completes when you have downloaded all n chunks successfully. What is the expected time to get all n chunks?

(c) **Techer's tactics (20 points)** Your dream car has just arrived at your favorite dealership and you decide to check it out. The dealer knows you are from Caltech, so his chances of tricking you are slim at best, but he decides to try anyway. His devious pricing offer is outlined below:

- *Dealer's Offer:* The dealer offers to sell the car for N dollars, where N is defined as follows. You get to describe any continuous random variable X (by specifying its density function). Then, the dealer generates a sequence $\{X_i\}_{i\in\mathbb{N}}$ of independent and identically distributed random variables according to X. A record is said to occur at time step $n > 1$, if $X_n > \max(X_1, X_2, \ldots, X_{n-1})$. That is, X_n is a record if it is larger than each of $X_1, \ldots, X_{n-1}$. N is defined as the time step at which a record occurs for the first time, that is,

$$N = \min\{n \ : \ n > 1 \text{ and a record occurs at time } n\}.$$

Your task: Determine whether this game is in your favor by calculating $E[N]$.

To see whether this game is in your favor, it may be useful to consider the following:

(i) What random variable should you choose? Does it matter?

(ii) Let p_n denote the probability that a record occurs at time step $n > 1$. Let $X_n^{\max} = \max_i\{X_1, \ldots, X_n\}$ Can you write p_n as a simple function of X_n and X_{n-1}^{max}?

After considering this game, you propose a counter offer.

- *Techer's Offer:* The procedure is similar with some changes - the dealer gets to pick X, and then you get to choose a positive integer M. The dealer generates the sequence as above, and you pay a dollar amount equal to the expected number of records in M time steps.

The dealer reluctantly accepts this offer.

Your task: Ideally, you would like to pay x dollars for the car. What M should you choose?

Note: You may use the approximation $\sum_{i=1}^{n} \frac{1}{i} \approx \ln n$ for large n.

3 Understanding Clustering [20 points]

We will see in class that many networks exhibit "homophily", i.e., if a vertex A is connected to vertices B and C, then it is very likely that B and C are connected as well. An example of this is that if B and C are both friends of A, then it is likely that B and C are friends too.

The notion of "clustering coefficient" attempts to provide a measure of this tendency for the formation of triangles in the network. In this exercise, we introduce two different definitions of clustering coefficient and see that, despite the fact that they appear very similar, they can yield very different results.

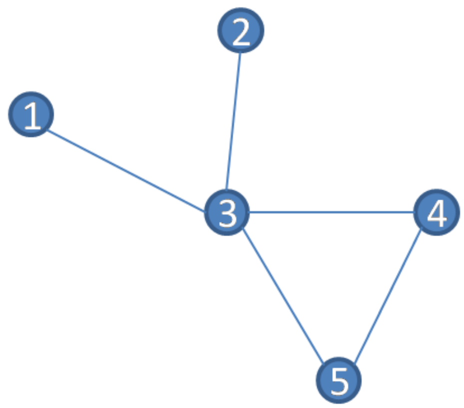

Figure 1: *Consider the graph G depicted above. Since vertex 3 has 4 neighbors, the number of triples centered at vertex 3 is 6, implying $Cl_3(G) = 1/6$. $Cl_1(G) = Cl_2(G) = 0$, $Cl_4(G) = Cl_5(G) = 1$, implying $Cl^{avg}(G) = 13/30$. The number of triangles in G is 1, node 3 contributes 6 connected tuples, nodes 4 and 5 contribute 1 each, implying $Cl(G) = 3/8$.*

Consider an undirected graph G with vertices labeled $1, 2, \cdots, n$. We define the *average clustering coefficient* of G as follows.

$$Cl^{avg}(G) := \frac{\sum_{i=1}^{n} Cl_i(G)}{n},$$

where

$$Cl_i(G) := \frac{\text{number of triangles centered on vertex } i}{\text{number of triples centered on vertex } i},$$

where a triple centered at vertex i is an unordered pair of vertices that are connected to i. Intuitively, $Cl_i(G)$ is the probability that two 'friends' of i are 'friends' with each other. For vertices i with degree 0 or 1, define $Cl_i(G) = 0$.

We define the *overall clustering coefficient* of G as follows.

$$Cl(G) = \frac{3 \times \text{ number of triangles in the graph}}{\text{number of connected triples of vertices}},$$

where a connected triple refers to a vertex connected to an unordered pair of vertices. Note that $Cl(G)$ is the fraction of connected triples that have the third edge filled in to complete the triangle. See the example in Figure 1 for an illustration of the two definitions.

(a) **Contrasting the two definitions [20 points]** Note that for any graph, $Cl^{avg}(G), Cl(G) \in [0,1]$, with a greater value indicating more 'clustering.' However, despite the fact that these two measures seem very similar, they are not. Your task in this exercise is to illustrate that these two definitions of the clustering coefficient can be very different.

Construct an example (a family of graphs) showing that the average and overall clustering coefficients can be as different as possible. Specifically, construct an example where as the number of nodes in the graph becomes large, one of clustering definitions approaches 1 while the other approaches 0.

(b) **Visualizing your ego network [Optional]** Beyond simply computing the clustering coefficients for a given graph, inspecting the "clusters" in a graph can often reveal interesting information, i.e., looking for sets of nodes that have lots of internal connections but very few connections to nodes outside of the set. (Note that clusters will play a crucial role when we study the spread of epidemics later in the course.)

To illustrate this, one natural place to start is with your own social network. If you have a Facebook account, Wolfram Alpha provides an interesting visualization of your "ego network", i.e., the network created by connections among you and your friends. You can look at a report of your ego network by going to `http://www.wolframalpha.com/facebook/`. (Note that you have to sign up for a Wolfram Alpha account, but it is free.) The most interesting part of the report comes near the bottom, where the ego network is visualized and clusters are identified.

It is often quite surprising how much can be learned simply by studying the clusters of nodes in social networks. For instance, in Adam's ego network, one cluster corresponds to high school friends, one cluster to college friends, one to friends in Pasadena, one to family, and one to in-laws.

If you do look at your report, please let us know in your solution set if you see something interesting (or even show us the picture and discuss it if you don't mind).

4 Six degrees of separation [20 points]

In most real-life network graphs, we find that we can reach from almost any node to any other node within a very small number of hops (often six). This is known as six degrees of separation and was first illustrated experimentally in 1967, when Stanley Milgram performed a famous experiment where he asked a randomly selected set of 296 people from the midwest to try to forward a letter to a *target*, a stockbroker in Boston. The participants were given some personal information about the target (including his name, occupation and address). They were asked to forward the letter to someone they knew on a first-name basis and to pass along the same instructions, so that the letter reached the target as quickly as possible. Eventually, 64 letters made it to the target, with the median number of intermediaries for these letters being around 6 and hence six degrees of separation.

The Milgram experiment shows six degrees of separation in the social network, but it has also been shown to occur in many other networks, e.g., six degrees of Kevin Bacon for actors/actresses. In this problem, you explore six degrees of separation in a few other networks.

Collaboration policy: In this problem, you are not allowed to collaborate with anyone. You should also attempt to do these searches without consulting more information about the topic. Please cite any tools you end up using.

Grade policy: In this problem, you are only are required to complete 2 of the 3 problems.

(a) **Wiki your way through**

The existence of short paths between nodes is a characteristic feature of a number of networks around us. In this exercise, we will look for them in the Wikipedia web graph.

We have listed below ordered pairs of 'entities' (objects or people), each having a dedicated web-page on Wikipedia. For each pair, starting from the web-page for the first entity, you have to find a sequence of links (only to other Wikipedia pages) that will take you to the Wikipedia web-page corresponding to the second. Write down (i) the first path that you found, and (ii) the shortest path that you found. In your submission, specify your paths precisely, i.e., write the sequence of links you used.

- Network congestion → Capacitor
- Paul Erdős → K. Mani Chandy

You get 1 bonus points each if your shortest path is the shortest among all the homework submissions. Note that you are *not* allowed to edit any Wikipedia pages to 'create' your path!

(b) **Know your professors**

In this problem, let us consider the graph formed by co-authorship in published papers. We say person a is connected to person b with an edge if and only if there is a published paper with both a and b as authors. This graph is commonly referred to as the *co-authorship network* and has been the subject of a lot of analysis for the research community.

Your job in this task is to find paths between professors of CS at Caltech and other well-known names. The shorter the path you find, the better it is! As before, write down both the first path that you found, as well as the shortest path that you found between:

- Adam Wierman → Erwin Schrödinger
- Paul Erdős → K. Mani Chandy

You get 1 bonus point each if your shortest path is the shortest among all the homework submissions. Include the names of the papers with the list of authors that form the path between the ordered pairs given above. Hint: Google “Erdos number”.

(c) **A marketing perspective**

In this task, find a path between the following products on *www.amazon.com*. An ordered pair of products $a \to b$ shares a directed edge if and only if b appears in the "Customers Who Bought/Viewed This Item Also Bought/Viewed" section of the Amazon page for product a. Your task is to find a path from:

- ‘Star Wars: The Complete Saga (Episodes I-VI) [Blu-ray]’ $\to$
  ‘Colgate Extra Clean Toothbrush, Full Head, Soft, 6 Count’

along edges as described above. These products have dedicated pages in Amazon. The path should only go through products on Amazon and not ads from external websites. Record the date and time at which you discovered this path since the links in the graph change over time.

Note that if you are signed in to your Amazon account, chances are the recommendations are customized to you, so you should sign out of your account while searching for a path. It is also recommended that that you clear your cache/cookies, or instead, use the ‘Incognito mode’ in Chrome, ‘Private browsing’ in Firefox/Safari, or ‘InPrivate browsing’ in IE. You are *not* allowed to place a spurious order of the second product from the page of the first product to artificially ‘create’ a path!

You get 1 bonus point if your path is the shortest among all the homework submissions. Clearly enlist the exact names of the intermediate products in your submission.