Problem Set 1

Collaborated with: Steven Brotz, Abraham Hussain, David Kawashima

Problem 2

a.

i. This is very much so Simpson's paradox and in this case, the paradox can be attributed to the discrepancy in sample sizes. In the year 2015-2016, it may have been that a lot of people applied for Amazon (the higher accepting company) and a small amount of people applied for Facebook while in 2016-2017, less people applied to Amazon and more people applied to Facebook (the less accepting one). Because percentages do not take into account sample size, it may appear that the percentages for each company have improved over the year, but if you aggregate the two companies together each year, the company that had a lot more applications will contribute the majority to the overall percentage of the year so in the case above, Amazon's rate will contribute more to the yearly rate from 2015-2016 while Facebook will contribute more to the yearly rate from 2016-2017. The missing study is the number of applicants to each company.

ii. Let's say in year 2015-2016, Facebook was still kind of small (yet still very selective because it did not have the resources for a large intern class) while Amazon really took off. Thus, out of 500 applications, 400 applications went to Amazon and 100 applications went to Facebook. With 40% acceptance for Amazon and 10% acceptance for Facebook, 160 were accepted to Amazon and 10 were accepted into Facebook for an overall yearly acceptance of $170/500 \approx 0.34$. Now, in 2016-2017 there is a sudden reversal of roles. There was some scandal that swept Amazon so less people applied to Amazon (Amazon tried to retain all the talent they could get so they increased their acceptance rate). On the other hand, Facebook grew a lot over the year and because of increased popularity, received a lot more applications. With more money now, Facebook can now take more interns and so their acceptance rate also increases from last year. In this case, out of the 500 applications, we have 100 Amazon applications and 400 Facebook applications. With Amazon's acceptance rate of 75%, we get 75 acceptances and with Facebook's acceptance rate of 20%, we get 80 acceptances. Thus, our overall yearly acceptance rate is $155/500 = 0.31$. With this, we see that even though individual acceptance rates increased for both companies from the 2015-2016 to 2016-17 years, the overall acceptance rate calculated by aggregation of the two companies' applications resulted in a decrease from 2015-2016 to 2016-2017.

b.

a. Let's say we have N students. We let X represent the number of students who get back their own exam. We let x_i be an indicator variable that represents whether the i^{th} exam handed back belongs to student i. More specifically $x_i =$:

$$\begin{cases} 1 & \text{if student } i \text{ gets back own exam} \\ 0 & \text{if student } i \text{ does not get back own exam} \end{cases}$$

Realize that this problem is asking for the expected number of students who get back their own exam or $E[X]$. Realizing that expectation is a linear operator, we get:

$$E[X] = E[\sum_{i=1}^{N} x_i] = \sum_{i=1}^{N} E[x_i] \tag{1}$$

Now, to calculate $E[x_i]$: because x_i is an indicator variable, $E[x_i] = 1$ * probability that student i gets back own exam + 0 * probability that student i does not get back his own exam = probability that student i gets back own exam. Now, the probability that student i gets back own exam = probability that the exam was not previously taken by another student * probability that of the remaining exams to pick from, student i picks his own. The probability that the exam was not previously taken is the probability that his exam still remains of the remaining $(N-i-1)$ exams which is $\frac{N-i-1}{N}$ and the probability that he picks his own exam out of the remaining exams is $\frac{1}{N-i-1}$. Thus, $E[x_i] = \frac{N-i-1}{N}\frac{1}{N-i-1} = \frac{1}{N}$. Hence, plugging this into the above equation, we get:

$$E[X] = E[\sum_{i=1}^{N} x_i] = \sum_{i=1}^{N} E[x_i] = \sum_{i=1}^{N} \frac{1}{N} = 1 \tag{2}$$

Hence, we expect one student to get back their own exam.

b. To gain some intuition on how to solve this, we first look at the case where we just calculate the expected number of swaps for the first student who got their exam back to get his own exam. We let n be the total number of students/exams. To calculate the expected number of swaps for the first student, we note that there is a $\frac{1}{n}$ probability that it takes 0 swaps to get his own exam back (only if he gets his own exam back directly from teacher). To calculate the probability that it takes 1 swap for the first student to get his own exam back, we have to multiply the probability that the first student did not get his own exam back from the teacher $(\frac{n-1}{n})$ by the probability that he picks the right classmate to swap with $\frac{1}{n-1}$ which gets us $\frac{1}{n}$. Likewise, to calculate probability of two swaps, we multiply the probability that the first student did not get his own exam back from teacher $(\frac{n-1}{n})$ by the probability that he did not get his own exam after just one swap $(\frac{n-2}{n-1})$ multiplied by the probability that he gets his own exam on the second swap $(\frac{1}{n-2})$ which gives us $\frac{1}{n}$ probability that it takes two swaps. We can easily see from this, that the probability of any number of swaps from 0 to $(n-1)$ occurs with probability $\frac{1}{n}$, so our expected number of swaps for the first student to get back his own exam is:

$$\sum_{i=0}^{n-1} \frac{i}{n} = \frac{n-1}{2} \tag{3}$$

However, that's just the expected number of swaps for the first student. To extrapolate this to get the expected number of swaps for the second student, we realize that for the second student, because our first student is already fixed with his own exam, that there are only $(n-2)$ other students to swap with. Thus, the expected number of swaps for the second student is:

$$\sum_{i=0}^{n-2} \frac{i}{n-1} = \frac{n-2}{2} \tag{4}$$

Generalizing, we see that the expected number of swaps for the k^{th} student is $\frac{n-k}{2}$. Thus, to get the total number of swaps for all n students to get their own exams, we have to sum over the expected number of swaps for each of the n students:

$$\sum_{k=1}^{n} \frac{n-k}{2} = \frac{n(n-1)}{4} \tag{5}$$

Thus, we expect $\frac{n(n-1)}{4}$ swaps.

c.

i. We let X be the number of times Prof. Weirman sets a record in the coming year. We let x_i be an indicator variable that represents whether the i^{th} triathlon was a PR. More specifically $x_i =$:

$$\begin{cases} 1 & \text{if triathlon } i \text{ was a PR} \\ 0 & \text{if triathlon } i \text{ was not a PR} \end{cases}$$

Realize that for this problem we want $E[X]$. Realizing that expectation is a linear operator, we get:

$$E[X] = E[\sum_{i=1}^{N} x_i] = \sum_{i=1}^{N} E[x_i] \tag{6}$$

To calculate $E[x_i]$, because x_i is an indicator variable, $E[x_i]$ = probability that the i^{th} triathlon was a PR. Looking at a few cases: $E[x_1] = 1$ because the first run is guaranteed to be a PR, $E[x_2] = \frac{1}{2}$ because after the second run, there have been only two runs and for the second run to be a PR happens with probability 1/2. Likewise $E[x_3] = \frac{1}{3}$ and $E[x_4] = \frac{1}{4}$. Thus, $E[x_i] = \frac{1}{i}$. Thus, $E[x] = \sum_{i=1}^{N} E[x_i] = \sum_{i=1}^{N} \frac{1}{i}$.

ii. No, it does NOT matter that the finish times follow a normal distribution. If we look at how we came up with the answer for part *i.*, we did not make use of the information about the underlying distribution.

iii. From the previous part, we got that the expected number of PRs by Prof. Wierman is $\sum_{i=1}^{N} \frac{1}{i} \approx \ln N$ for large N. To get how many races (N) I should encourage Prof. Wierman to run to get more than M expected PRs, we set $\ln N = M$ and solve for N.

$$\ln N = MN = e^M \tag{7}$$

Thus, we should encourage him to run $\lceil e^M \rceil$ triathlons.

Problem 3

a. So we want the quantity $E[d_i]$. We first see that this is precisely equivalent to the quantity(n is number of vertices, $|E|$ is number of edges.):

$$\sum_{i=1}^{n} \frac{d_i}{2|E|} d_i \tag{8}$$

To see why, we note that there is a $1/|E|$ probability of picking each of vertex i's edges and i to be selected as the endpoint of that edge happens with probability $1/2$. Now, given that the edge relationships are symmetric (i.e. if v_i connects to v_j then v_j also connects to v_i) we can rewrite the above equation as

$$\frac{\sum_{i=1}^{n} d_i^2}{\sum_{i=1}^{n} d_i} \tag{9}$$

Now, seeing that we want to have a μ in our final expression, we add and subtract μ:

$$\frac{\sum_{i=1}^{n} d_i^2}{\sum_{i=1}^{n} d_i} + \mu - \frac{\sum_{i=1}^{n} d_i}{n} \tag{10}$$

Now, recognizing that the final result has a term with μ in the denominator and seeing that $\sum_{i=1}^{n} d_i = n\mu$, we get:

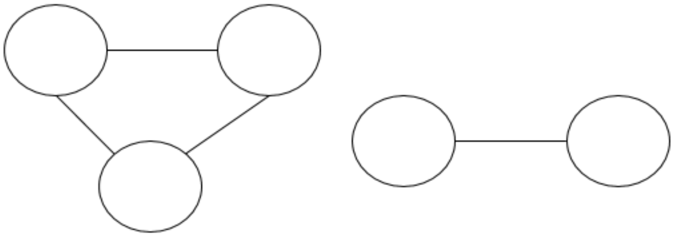
$$\mu + \frac{\frac{\sum_{i=1}^{n} {d_i}^2}{n} - \mu^2}{\mu} \tag{11}$$

Lastly, recalling that $\sigma^2 = E|d_i^2| - (E|d_i|^2)$, we arrive at:

$$E|d_i| = \mu + \frac{\sigma^2}{\mu} \tag{12}$$

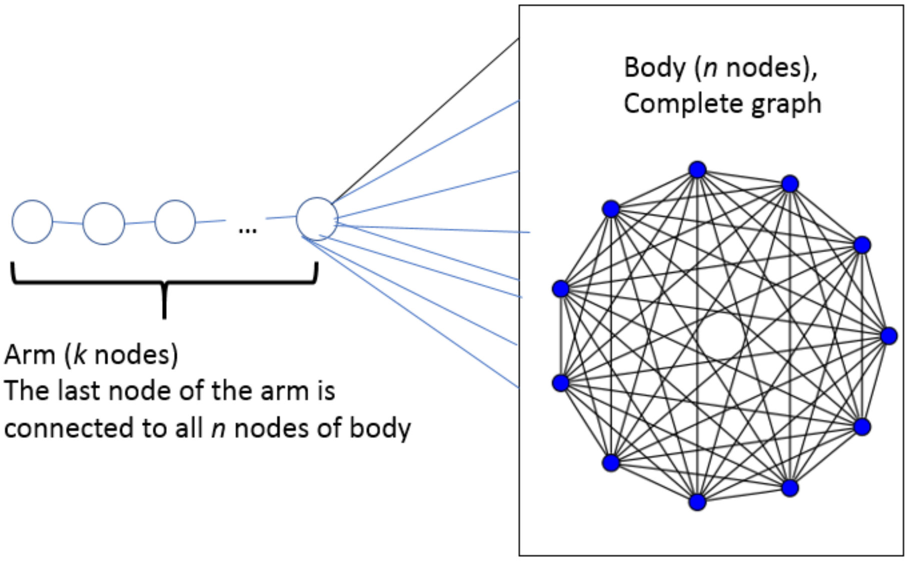
as desired.

b. We see such a graph below. We have 3 nodes with degree 2 and two nodes with degree 1 and each node has the same degree as its neighbor/s.



Problem 4

We will construct a general graph that we show will work for parts a, b, and c. Below is a diagram.


Body (n nodes),
Complete graph
...
Arm (k nodes)
The last node of the arm is
connected to all n nodes of body

We can divide up the graph into two portions: the arm and the body. First, let's start off with the arm. The arm will consist of k nodes in a line, with a single edge connecting adjacent nodes. The body is a complete graph consisting of n nodes. The last node of the arm has a total of $(n+1)$ edges, one edge from the previous node of the arm and n edges connecting itself to each of the n nodes in the body.

Now, from this construction, we know that our maximal diameter is k. Now, to calculate average distance, we must first calculate the total distance between nodes. We first calculate the distances between every pair of nodes comprising the arm. Given that we have k nodes in our arm, we will have 1 path of length $(k-1)$, 2 paths of length $(k-2)$ length, 3 paths of length $(k-3)$, ..., $(k-1)$ paths of 1 length. Thus, our total distance amongst nodes of the arm is:

$$\sum_{i=1}^{k-1} i(k-i) = \frac{1}{6}k(k-1)(k+1) \tag{13}$$

For the body, because it is complete, amongst nodes in the body, the total distance amongst nodes of the body is:

$$\binom{n}{2} = \frac{1}{2}(n-1)n \tag{14}$$

Now for distances between nodes of the body and nodes of the arm, we notice that there are n paths of length 1 (all from the node of the arm that is directly connected to each node of the body), n paths of length 2 (from the second to last node of the arm), n nodes of length 3, ... n nodes of length k. This is equivalent to:

$$\sum_{i=1}^{k} in = \frac{1}{2}k(k+1)n \tag{15}$$

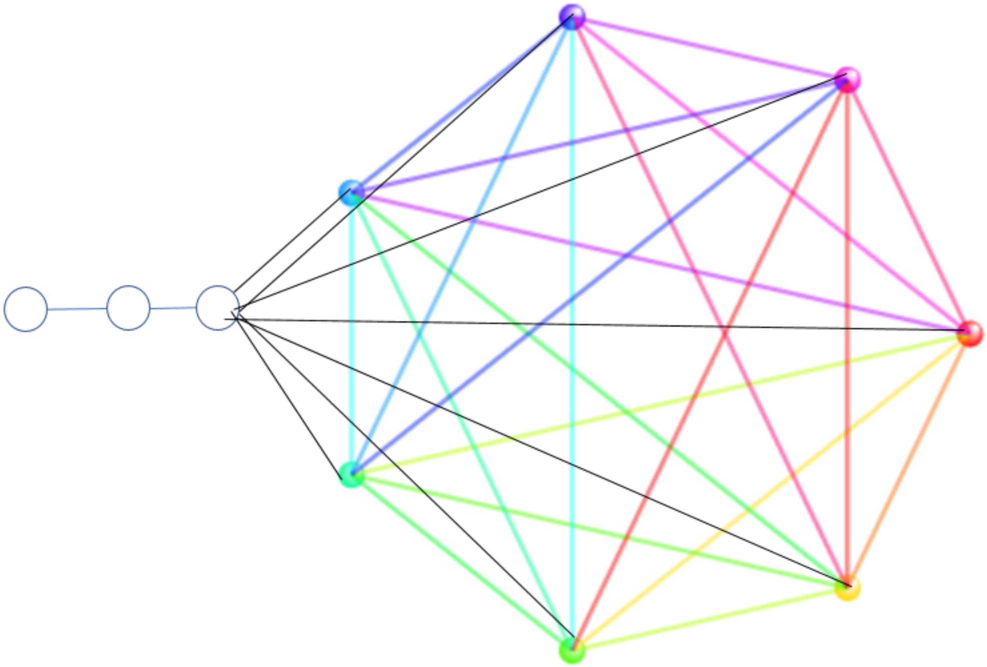
Thus, our total sum of distance is:

$$\frac{1}{6}k(k-1)(k+1)+\frac{1}{2}(n-1)n+\frac{1}{2}k(k+1)n \tag{16}$$

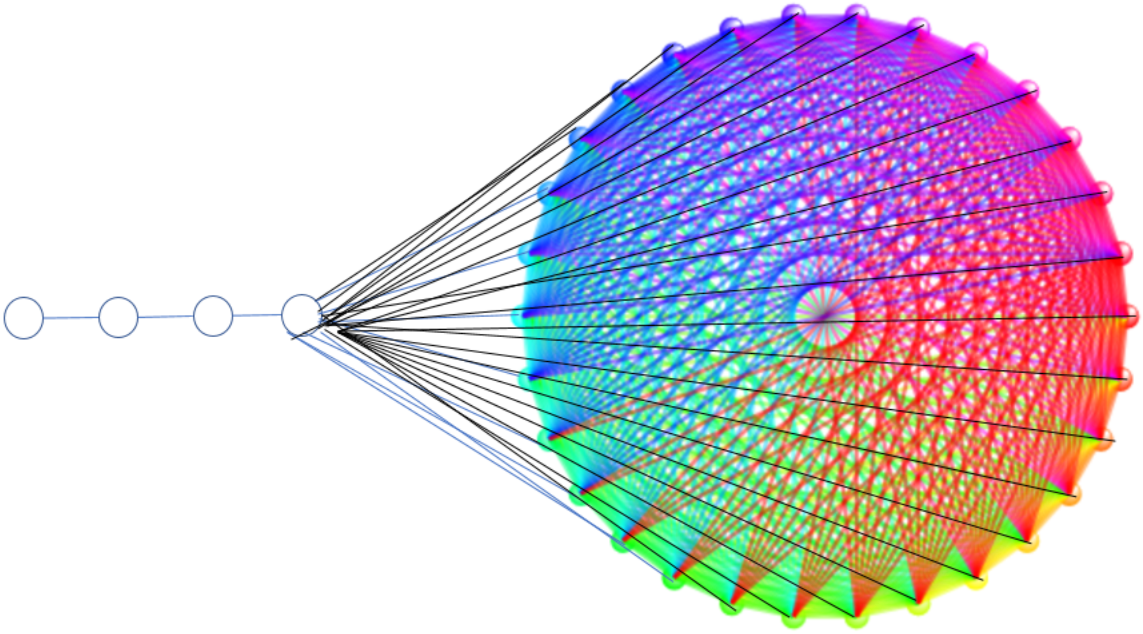
Our graph is connected so our total number of paths is $\binom{n+k}{2} = \frac{1}{2}(k+n-1)(k+n)$ and hence our average distance is:

$$\text{Average distance} = \frac{\frac{1}{3}k(k-1)(k+1)+(n-1)n+k(k+1)n}{(k+n-1)(k+n)} \tag{17}$$

a. Now, we construct an unweighted graph where the maximal diameter is more than 2 times the average distance between nodes. We set our arm to contain 3 nodes ($k = 3$) and our body to consist of 7 nodes ($n = 7$). Thus, our maximal diameter is 3 and plugging in $k = 3$ and $n = 7$ into the average distance equation, we get our average distance to be 67/45 ≈ 1.49 in which case we have that our maximal diameter is more than 2 times the average distance.



b. For a maximal diameter that is more than 3 times the average distance, we set $k = 4$ and $n = 30$. Our maximal diameter is 4, and our average distance is $\frac{745}{561} \approx 1.327$.



c. The general description is outlined above (the beginning of this answer's writeup). A few things to notice: if we fix k, as n approaches ∞, the average distance will approach 1. This can easily be seen by taking the limit as $n \to \infty$ of our average distance function while treating k as a constant. Thus, to get a graph where the maximal diameter is more than c times the average distance between nodes, we let $k = c + 1$, so we have $(c + 1)$ nodes in the arm and make n an extremely large number because we know that increasing n will make the average distance approach 1. Once the average distance is close enough to 1, we know that our maximal distance which is $(k = c + 1)$ is more than c times the average distance ≈ 1. In other words, if we take the limit as n goes to ∞ of the ratio of maximal diameter to average distance, we get:

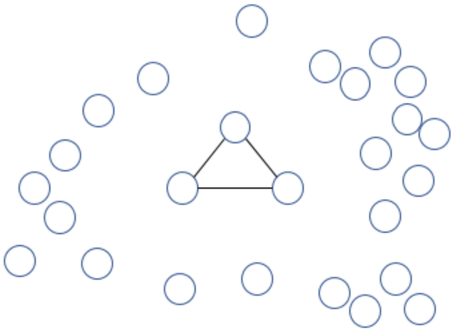
$$\lim n \to \infty \frac{k}{\frac{\frac{1}{3}k(k-1)(k+1)+(n-1)n+k(k+1)n}{(k+n-1)(k+n)}} = k \tag{18}$$

Thus, if we let $k = (c + 1)$, we can guarantee ourselves that we will be able to come up with a graph that has a maximal diameter at least c times the average distance.

Note: the colored complete graphs were generated using an online tool.

Problem 5

a. One simple example is below. We have a single triangle with many disjoint vertices around that single triangle. Now, as we increase the number of disjoint vertices (let's call n), Cl^{avg} goes to 0. This is because for each disjoint vertex i that we add, we get that $Cl_i(G) = 0$ and thus as $n \to \infty$, these start to dominate the average and so we get that $Cl^{avg} \to 0$. On the other hand, $Cl(G)$ will always remain at 1 because no matter how large n gets, there is only one triangle and only three connected triples.



Problem 6

a.
First path (7 nodes): 3D printing (click on **Prototyping**) → Prototype (click on **Boeing 787 Dreamliner** → Boeing 787 Dreamliner (click on **Federal Aviation Administration**) → Federal Aviation Administration (click on **National Aeronautics and Space Administration**) → NASA (click on **SpaceX**)→ SpaceX (click on **Gwynne Shotwell**) → Gwynne Shotwell.

Shortest path (6 nodes): 3d printing (click on **Airbus A350 XWB**) → Airbus A350 XWB (click on **Federal Aviation Administration**) → Federal Aviation Administration (click on **National Aeronautics and Space Administration**) → NASA (click on **SpaceX**)→ SpaceX (click on **Gwynne Shotwell**) → Gwynne Shotwell.

First path/ shortest path (4 nodes): Steven H. Low (click on **Cornell University**) → Cornell University (click on **Superman**) → Superman (1978 film) (click on **Star Wars**) → Star Wars (film)

c.
Path (20 nodes, 01/06/2018 12:30 PM):
How Clean is the Cloud? - A. Wierman - 4/19/2017 → From the Big Bang to Black Holes and Gravitational Waves - K. Thorne - 3/11/2016 → The Absurdity of Detecting Gravitational Waves → Your Mass is NOT From the Higgs Boson → Can We Really Touch Anything? → Spinning Tube Trick → 5 Pictures To Test Your Intelligence → 14 TRICKS THAT WILL EMBARRASS YOUR TEACHER → 14 Weird Ways To Sneak Food Into Class / Back To School Pranks → 16 Edible School Supplies! Prank Wars! → Gummy Food vs Real Food Challenge! → ULTIMATE SQUISHY FOOD VS. REAL FOOD CHALLENGE!!! → Making REAL Play Doh Ice Cream AND Dessert Pie!!! → How to Make Play Doh Ice Cream with Molds Fun and Creative for Kids → Ice Cream Coloring Page — Learn Colors — How to Draw Ice Cream → Play Doh Rainbow Ice Cream Cone, Ice Cream Scoop, & Ice Cream Popsicle — Fun & Easy Pay-Doh! → How to Make Play Doh easy playdo by unboxingsurpriseegg! → Play Doh Shrek 2 Rotten Root Canal Playset Dentist Dr Drill N Fill Play

Dough Comparison toys Review → Shrek - All Stars → Smash Mouth- All Star.

Path (11 nodes, 1/06/2018 1:15 PM): Cassie Goes for a Walk → MIT cheetah robot lands the running jump → Real Dog Meets Boston Dynamics Robot Dog for First Time → Cat Vs. Dinosaur - Cat Spooked, Then Befriends a Robot Dinosaur - Maya The Cat → Cats and dogs react to RC toys - Funny animal compilation → CATS in Ridiculously Adorable COSTUMES [Funny Pets] → CATS vs BALLOONS (HD) [Funny Pets] → Top Cats Vs. Cucumbers Funny Cat Videos Compilation - Gatos Vs. Pepinos Vdeo Recopilacin → Top 200 Highlights of Animals - VERY FUNNY ANIMALS → What Does the Cat Say? - Ylvis - The Fox (What Does the Fox Say?) [Official music video HD] → Ylvis - The Fox (What Does The Fox Say?) [Official music video HD]