CS 5854: Networks, Crowds, and Markets
Homework 4

Instructor: Rafael Pass TAs: Cody Freitag, Drishti Wali

Assigned: November 26, 2019 Due: December 12, 2019, 11:59 pm

Late Deadline: At most one slip day can be used!
Submissions cannot be accepted after December 13, 11:59 pm

Homework Policies and Guidelines:

Submission: Your homework solutions must be **typed** and submitted as a single .pdf file in CMS. You must additionally submit a single .zip file containing all relevant files specified in the assignment for all coding problems. Template .tex and .py files will be provided containing an outline for your submission.

Late Days: Each student may use four "late days" in total as desired throughout the semester, each of which grants a 24-hour extension to an assignment's due date. Late work beyond this limit will still be accepted and graded until grades have been released for that assignment, but (unless discussed in advance with the TA and/or instructor) will have a negative impact on final grades.

Collaborations: You may work in groups of up to **4** students. Every member of the group must list all other collaborators at the top of their assignment. *(Note: the maximum size of a connected component of groups must be 4. If A,B,C, and D work together, it must not be the case that A and E work together.)*

Your submitted answers, explanations, and discussion for all written questions in the pdf must be your own, individual, unique solutions. You *may* share and even submit the same code and graphs with your collaborators. Additionally, you may make use of published material (papers, github, wikipedia, etc.), provided that you acknowledge and specifically cite all sources used. It is considered a violation of academic integrity to submit a problem solution that you are unable to explain orally to a member of the course staff.

To make this further clear, you are encouraged to discuss problems and examples and work on code in a group, but you must then separately write up the solution yourself. You may not copy any collaborator or external resource even if you cite them. It is up to the TAs and graders to determine whether or not two submissions (or a submission and an external source) are too similar to be considered "your own words." Furthermore, you may receive a 0 for any problem that the TAs and graders determine is not your own work even if cited.

How to receive credit: You must **justify all answers** to receive credit unless specified otherwise. We will do our best to make clear the level of justification we expect for each problem. For coding questions, please turn in complete, executable code for each part of the question that asks for an implementation, and include a .txt file containing any required outputs if not already included in the main .pdf file. *Using Python is strongly preferred.* If you want to use another language, you must first consult with the TAs. We will not grade code based on style, but we may mark down code if we are unable to understand what it is doing. You may use standard libraries to implement data structures such as graphs, but, unless otherwise specified, you may not use pre-existing implementations of any algorithms without express permission from the TAs. (If a problem asks you to implement X and you use a package that implements X for you, you will not get credit.)

Part 1: Voting

1. In the American presidential elections, while the popular vote is used up to the state level, the electoral college decides the winner at the national level. Assuming there are only two candidates, is this system strategy-proof? Does it elect a Condorcet winner? Justify your answers. (Note: You can assume for simplicity that each state gets a single "vote" in a national election, and the state then runs an election by popular vote with however many people are in that state to determine which vote to cast at the national level.)

2. Suppose we want to run a popular vote election between two candidates A and B. There are 1,000,000 eligible voters and suppose 52% of them prefer A to B. Instead of running a full election where every person casts a vote, we poll m randomly selected people (with replacement for simplicity) and ask them to report their preferred candidate. The result of the poll is the majority of the responses received.

   (a) Compute a value of m so that the result of the poll is incorrect with probability at most 1%? (Use the Chernoff bound in the book, show your work.)

   (b) Let n be the number of people in the population, ϵ be defined such that $(1/2 + \epsilon) \cdot n$ prefer A to B, and let δ be the desired accuracy (so the probability the result is incorrect is at most δ). Write m as a function of n, ϵ, and δ.

   If the number of people in the population increased by a factor of 10, how would that affect m? If ϵ decrease by a factor of 2, how would that affect m? If we want to increase our confidence by a factor of 10, how would that change m? If $\epsilon = 1/n$ (so 1 person would be the deciding vote), what would this imply about m given your bound from above?

   (c) In practice, what might be wrong with the above assumptions (i.e. why might we not we use polls to run our elections)?

Part 2: Stable Matchings

3. For the following setting, find **(a)** the male-optimal and **(b)** the female-optimal stable matching. For each part, simulate the Gale-Shapley algorithm (i.e. in words, clearly indicate what happens at each step) to show that it arrives at the matching you find.

Females	Preferences	Males	Preferences
A	$X > W > Y > Z$	W	$D > B > C > A$
B	$X > W > Y > Z$	X	$D > B > A > C$
C	$X > W > Z > Y$	Y	$C > B > D > A$
D	$Y > W > Z > X$	Z	$D > B > C > A$

4. Prove that there exists a non-bipartite matching setting (where every individual has preferences over all other individuals) for which no stable matching exists.

Part 3: Beliefs

5. There is a test for a certain disease that has a 15% false positive rate and a 25% false negative rate. (So, if someone has the disease, there is a 75% chance the test will return positive; if they do not, there is an 85% chance it will return negative.) If 1% of the population has this disease, what is the probability that someone who tests positive for the disease actually has it? (Use Bayes' Rule; show your work.)

6. In the "foolishness of crowds" example from section 16.2 of the notes, let's assume that people decide that their own evidence should instead be weighed c times as heavily as others' prior guesses, where c is an integer greater than 1. Now what is the probability that a cascade occurs and *everyone* is incorrect (in terms of ϵ, where the probability that a player receives correct evidence is $1/2 + \epsilon$)?

7. Recall the "muddy children" example covered in class and in the notes. Section 17.2 of the notes gives a formal argument that Claim 17.1 (that, if there are m muddy children, they will answer "yes" on and not before round m) holds for the case where there are two total children.

   (a) Now draw the knowledge network for the case where there are three total children.

   (b) Using a similar argument to the case of two children, use your network and the formal "possible worlds" model of knowledge to formally show that Claim 17.1 holds for the case when there are three total children (and any non-zero number of muddy children).

Part 4: PageRank and Social Networks

The objective of this question is to use the PageRank algorithm as a way to determine how "influential" a node is in a social network based on its in-links from influential nodes. For this question, we provided a template in python called 'hw4.py'. You must use the template to submit your code, as we will grade your code in a (partially) automated way. *You should submit the hw4.py file, not a .pynb or other python file.*

8. Design an algorithm that runs the iterative ϵ-scaled PageRank algorithm for a specified number n of rounds on a given directed graph, with $\epsilon = 1/7$. Run it (with $n = 10$) on the examples in figures 15.1 (both left and right) and 15.2 (the two disjoint triangle graph), as well as at least two other simple test cases with at least 10 nodes.

9. Now we'll run PageRank on the Facebook data.

   [http://snap.stanford.edu/data/egonets-Facebook.html].

   The file is called "facebook_combined.txt.gz"; remember that it has 4,039 nodes.

   (a) Once again, remember that this is an undirected graph! Before running your algorithms from the previous problem, implement a transformation into a directed graph, i.e. each undirected edge corresponds to two different directed edges.

   (b) Now, run the PageRank algorithms from the last problem on this new graph. You shouldn't need n to be much higher than 10-20 for the algorithm to converge to a fixed point.

   (c) Where did most of the score tend to end up in your experiments? Look at the nodes that have the highest or lowest scores; is there a consistent pattern among your trials?

(d) Intuitively explain your results in terms of a measure of influence in a social network. Do you think that this is an accurate measurement? How could we try to improve it (for instance, by incorporating link strengths or other measures of popularity)?

Part 5: Essay Question

(This problem should be completed individually and not in a group. However, it will be graded based on completion.)

Write 1/2 to 1 page discussing or analyzing one of the following prompts using any of the concepts taught in this class:

- Read the following New York Times article adapting a recent work from Nobel Prize winners Esther Duflo and Abhijit Banerjee:
  https://www.nytimes.com/2019/10/26/opinion/sunday/duflo-banerjee-economic-incentives.html
  The article makes the claim that people aren't driven by financial incentives as much as you would assume. In particular, they cite a study that claims that people believe that "Everyone else responds to incentives, but I dont." Why might this undermine certain assumptions made in this class? How can we model this situation using ideas from this class?

- Watch the following speech from Sacha Baron Cohen:
  https://www.youtube.com/watch?v=ymaWq5yZIYM
  The speech discusses the relevance of the spread of (mis)information in social network. For example, he makes the claim that "fake news outperforms real news because lies spread faster than truth." How can use tools from this class to explain this phenomena? How could we use tools from this class to identify the spread of misinformation?

- Pick one or more recent news article (within the last few months) related to networks, markets, or beliefs to analyze and discuss. (In particular, you may look at one of the above examples and discuss a different aspect if you wish.)