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

Ananya Shivaditya aks298

Collaborators: I worked with Alex Popeil amp453, Nathan Cinnamond nc532, and Shobhna Jayaraman sj747.
Outside resources: I used the following outside resources: none.
Late days: I have used zero late days on this assignment.

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.)

   Solution:

   1. Yes, the system is strategy-proof. There is no incentive for a rational player to strategically lie in this vote. We can show this by taking one player i who votes in their particular state. We know that the player i can only affect the result of their own state, determined by popular vote. Now i only has two options: to vote for A or B, and it clear if i prefers A, the incentive is for i to rank A first. If not, i will harm A's chance of winning.

      A Condorcet winner (Def 1.2) is a candidate who wins at least half the total number of votes in the state. If there are only two candidates, the popular choice vote in each state elects a Condorcet winner. But now, since each state casts a single vote each in EC, this is effectively popular voting hence the final elected candidate is a Condorcet winner. However, if different states have different number of EC votes, we do not necessarily elect a Condorcet winner. Taking the example of the 2016 US Presidential elections, we notice that the winner did not attain more than half the total number of votes, but only at the EC level did he win more number of states, and so went on to take the position.

   □

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)?

Solution:

(a) True fraction of population that prefers A to B $p_a = 0.52$. Take $\chi_i = Bernoulli(p)$, i.e. X_a, X_B are independent. Our poll result is the candidate with the majority of the preference, i.e. we can be inaccurate as far as 2% before our poll becomes incorrect.

So we want to be within 2% accuracy of the population with confidence 99%.

$$Pr[|\bar{X} - p| \leq \theta] \geq 1 - \delta \tag{1}$$

Use the two-sided Chernoff bound (ref Princeton)

$$Pr[|\bar{X} - p| \geq \epsilon p] \leq 2exp(-\frac{\epsilon^2}{2+\epsilon} \times pn) \tag{2}$$

We take $\theta = \epsilon p$, so take $\epsilon = \frac{\theta}{p}$.

$$Pr[|\bar{X} - p| \geq \epsilon p] \leq 2exp(-\frac{\frac{\theta^2}{p^2}}{2+\frac{\theta}{p}} \times pn) = 2exp(-\frac{\theta^2}{2p+\theta} \times n) \tag{3}$$

Because we are looking at bounds, and the biggest p can be is 1, we write

$$Pr[|\bar{X} - p| \geq \theta] \leq 2exp(-\frac{\theta^2}{2+\theta} \times n) \tag{4}$$

So if we want confidence $1 - \delta$ in estimate. $\delta = 0.01$.

$$\delta \geq 2exp(-\frac{\theta^2}{2+\theta} \times n) \tag{5}$$

$$n \geq \frac{2+\theta}{\theta^2} ln(\frac{2}{\delta}) \tag{6}$$

$$n \geq 26756.5 \tag{7}$$

(b)

$$Pr[Majority(x_1, \ldots, x_n) = W] \geq 1 - 2e^{-2\epsilon^2 n} \tag{8}$$

by Theorem 16.2. Similarly,

$$Pr[Majority(x_1, \ldots, x_m) = W] \geq 1 - 2e^{-2\epsilon^2 m} \tag{9}$$

The poll is accurate to δ given

$$Pr[Majority(x_1, \ldots, x_n) = W] - Pr[Majority(x_1, \ldots, x_m) = W] = \delta \tag{10}$$

$$1 - 2e^{-2\epsilon^2 n} - (1 - 2e^{-2\epsilon^2 m}) = \delta \tag{11}$$

$$2(e^{-2\epsilon^2 n} - e^{-2\epsilon^2 m}) = \delta \tag{12}$$

$$e^{-2\epsilon^2 m} = \frac{\delta}{2} + e^{-2\epsilon^2 n} \tag{13}$$

$$-2\epsilon^2 m = ln(\frac{\delta}{2} + e^{-2\epsilon^2 n} \tag{14}$$

$$m = -\frac{1}{2\epsilon^2}[ln(\frac{\delta}{2} + e^{-2\epsilon^2 n})] \tag{15}$$

$$m = \frac{1}{2\epsilon^2} ln[\frac{2}{\delta + 2e^{-2\epsilon^2 n}}] \tag{16}$$

If $n- > 10n$, m increases.
If $\epsilon- > \frac{\epsilon}{2}$, m also increases.
If $\delta- > 0.9\delta$, mm increases.
If $\epsilon = \frac{1}{n}$,

$$m = \frac{n^2}{2} ln[\frac{2}{\delta + 2e^{\frac{2}{n}}}] \tag{17}$$

We have that $e^{-\frac{2}{n}} >> \delta$ for big n. Hence,

$$m = \frac{n^2}{2} ln[2e^{\frac{2}{n}}] = n \tag{18}$$

Which we exactly what we anticipated.

(c) In real life, these assumptions may prove to provide inaccurate results, since we have made assumptions about the chi squared distributions freedoms, and further sampled down the votes of the general population. This not only would raise a hug outcry, it would not be justifiable or honorary to the number of voters who were not chosen to cast their vote in the random selection.

□

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.

Solution:

3. Following the Gale Shapley Algorithm for stable matching:

   i. Either of two sets are chosen to make proposals, doesn't matter which one - both will produce stable matching if it exists.

   ii. Whoever is unmatched proposes to his preferred option who hasn't already rejected him.

   iii. The proposed will accept if it is his/her first proposal or if he prefers the proposer over his current match.

   iv. Iterate till match is stable.

   (a) Chosing males to make proposals:

   1- W proposes to D. D accepts (first proposal).
   2- X proposes to D. D rejects as D is currently matched with W which is more preferred.
   3- Y proposes to C. C accepts (first proposal).
   4- Z proposes to D. D rejects as D is currently matched with W which is more preferred.
   5- X proposes to B. B accepts (first proposal).
   6- Z proposes to B. B rejects as B is currently matched with X which is more preferred.
   7- Z proposes to C. C accepts over its current match Y.
   8- Y proposes to B. B rejects as B is currently matched with X which is more preferred.
   9- Y proposes to D. D accepts over its current match W.
   10- W proposes to B. B rejects as B is currently matched with X which is more preferred.
   11- W proposes to C. C accepts over its current match Z.
   12- Z proposes to A. A accepts (first proposal).

   This results in the following stable matching: $X - B, \quad Y - D, \quad W - C, \quad Z - A$

(b) Chosing females to make proposals:

1- A proposes to X. X accepts (first proposal).

2- B proposes to X. X accepts over its current match A.

3- C proposes to X. X rejects as X is currently matched with B which is more preferred.

4- D proposes to Y. Y accepts (first proposal).

5- A proposes to W. W accepts (first proposal).

6- C proposes to W. W accepts over its current match A.

7- A proposes to Y. Y rejects as Y is currently matched with D which is more preferred.

8- A proposes to Z. Z accepts (first proposal).

This results in the following stable matching: $B - X, \quad D - Y, \quad C - W, \quad A - Z$

We observe that the same stable matching outcome is reached irrespective of which set is initially chosen to propose.

4. Here is an example of a non-bipartite matching scenario for which no stable matching exists.

Players	Preferences
A	$B > C$
B	$C > A$
C	$A > B$

In the above scenario, we find no stable matching. It is a cyclic choice mapping, no one is ever happy with whosoever proposes to them.

A matches with B.
B matched with C.
C matched with A. But A wants B, and so on and so forth.

A non-bipartite graph must contain at least one cycle of odd length. Let $v_1, v_2, ..., v_k$ be such a cycle. We construct the preference table T such that v_i ranks its predecessor v_{i1} first and its successor v_{i+1} second. According to their mutual highest rankings, the members of this cycle prefer to stay among themselves. Therefore, we call such a cycle exclusive. In an exclusive cycle with an odd number of members at least one member must find a partner outside the cycle. Let us assume that, under a matching M, this poor chap is v_i. Then v_i prefers v_{i+1} to its current situation (whether v_i is covered by M). Since v_i is v_{i+1}'s first choice, v_{i+1} prefers v_i to its current situation and (v_i, v_{i+1}) form a blocking pair. Hence matching M is not stable.

$\square$

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 $1/2 + \epsilon$)?

Solution:

5. Let test result be T and actual condition be A.
   Given,
   $P[T = +|A = False] = \frac{15}{100}$
   $P[T = -|A = True] = \frac{25}{100}$
   $P[A = True] = \frac{1}{100}$

   With Bayes' rule:
   $P[A = True|T = +]$
   $= P[T = +|A = True] * \frac{P[A=True]}{P[T=+]}$
   $=$ (1 - $P[T = -|A = True]$) $* \frac{P[A=True]}{(P[T=+|A=True]*P[A=True])+(P[T=+|A=False]*P[A=False])}$
   $= (1\text{-}\frac{25}{100})*\frac{\frac{1}{100}}{(1-\frac{25}{100})(\frac{1}{100})+(\frac{15}{100})(\frac{99}{100})}$
   $= \frac{75}{75+(15*99)}$

   $= 0.0480769 = 4.808\%$

6. In case the people weigh their own evidence c times as heavily as others' prior guesses, it is equivalent to saying that there are c-1 people who came before the person, and were given the same evidence as the person.
   Therefore we can say, for incorrect cascading, we need everyone to choose the incorrect value, despite if receiving a correct signal. This can be summed up as
   $P[X_i = 0|W = 1] + P[X_i = 1|W = 0]$ for all i, which will only happen when:
   $(\frac{1}{2} - \epsilon)^k(\frac{1}{2} + \epsilon)^c < (\frac{1}{2} + \epsilon)^k(\frac{1}{2} - \epsilon)^c$
   where k = i-1. This will hold true if k > c. i.e. the person chooses incorrectly, based on all evidence as weighted above.
   Probability for this to occur $= (\frac{1}{2} - \epsilon)^{c+1}$, since the first c+1 people will need to receive the incorrect signal in order for everyone else to follow and cascade to occur.

□

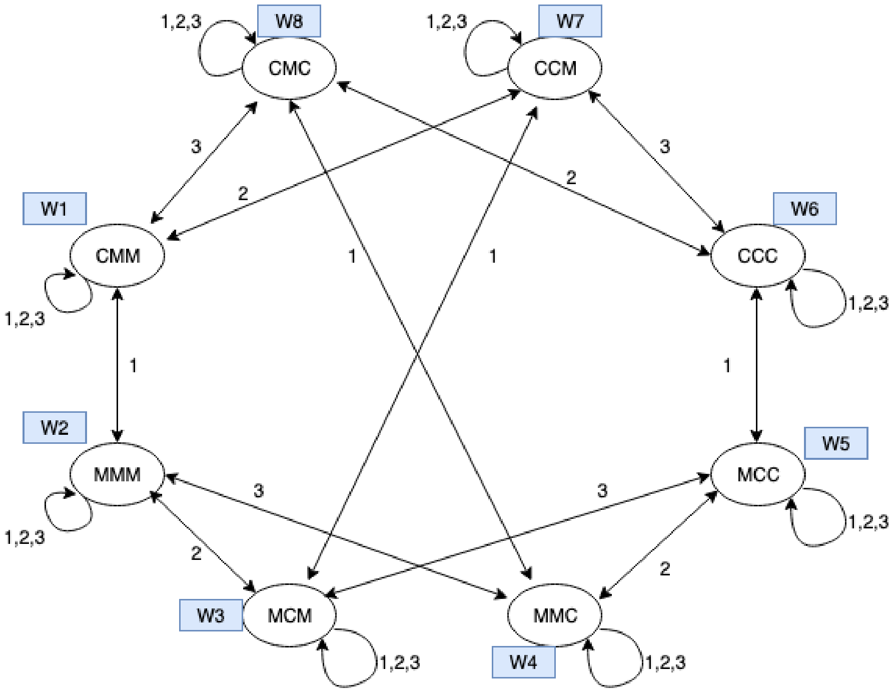
7. Recall the "muddy children" example covered in class and in the notes. Section 17.2 of the notes gives a formal argument that Claim 17.1 (that, if there are m muddy children, they will answer "yes" on and not before round m) holds for the case where there are two total children.

   (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).

Solution:

7. (a) Knowledge network for the case when there are 3 total children:



   (b) Before the father announces that atleast one child has mud on his/her forehead, all worlds are possible. After he announces, we know that W6 is not the case. After he asks for the first time if any child knows for sure if they have mud on their forehead, the possible worlds of W5, W7 and W8 are eliminated if no child yells yes. This is because the child with mud on their forehead would see that none of the other children have mud on theirs, so it must be himself.

   Now every child knows that atleast two children have mud on their foreheads. The second time father asks and no one yells yes, the worlds W1, W3 and W4 are gone because if any child had seen that one of the other two doesn't have mud, he would learn that he does have mud and say yes.

This must mean that all three children have mud, and would all say yes third time father asks, since only W2 remains.

This can be generalized to any number of children. After n questions, all states with n or less muddy children are eliminated. So this process will terminate after n questions, where n = the number of children with mud on their foreheads. This n could be as large as total, so this process will take at most the number of questions as there are total number of 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.

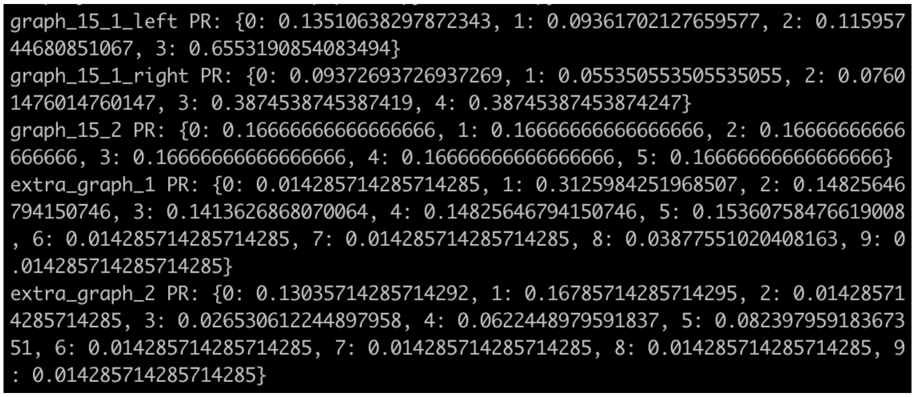
   (b) Now, run the PageRank algorithms from the last problem on this new graph. You shouldn't need n to be much higher than 10-20 for the algorithm to converge to a fixed point.

   (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)?

Solution:

8. The results for each graph specified in the documents are the the image below. The exact structures of the extra graphs can be seen in the python file.



```
graph_15_1_left PR: {0: 0.13510638297872343, 1: 0.09361702127659577, 2: 0.115957
44680851067, 3: 0.6553190854083494}
graph_15_1_right PR: {0: 0.09372693726937269, 1: 0.055350553505535055, 2: 0.0760
1476014760147, 3: 0.3874538745387419, 4: 0.38745387453874247}
graph_15_2 PR: {0: 0.16666666666666666, 1: 0.16666666666666666, 2: 0.16666666666
666666, 3: 0.16666666666666666, 4: 0.16666666666666666, 5: 0.16666666666666666}
extra_graph_1 PR: {0: 0.014285714285714285, 1: 0.3125984251968507, 2: 0.14825646
794150746, 3: 0.1413626868070064, 4: 0.14825646794150746, 5: 0.15360758476619008
, 6: 0.014285714285714285, 7: 0.014285714285714285, 8: 0.03877551020408163, 9: 0
.014285714285714285}
extra_graph_2 PR: {0: 0.13035714285714292, 1: 0.16785714285714295, 2: 0.01428571
4285714285, 3: 0.026530612244897958, 4: 0.0622448979591837, 5: 0.082397959183673
51, 6: 0.014285714285714285, 7: 0.014285714285714285, 8: 0.014285714285714285, 9
: 0.014285714285714285}
```

Figure 1: Scaled Page Rank Results

9. (a) To create this graph, we parsed the facebook file and used the same DirectedGraph interface as we did for problem 8. For each pair of nodes we created two edges to preserve the network structure in a directed format.

   (b) Results in q9.txt file.

   (c) Convergence is reached after n = 20. These scores were not consistent or regular. No node exceeded a score of 0.007. It is apparent that the nodes with the highest scores typically had the largest number of connections in the network. However, amongst this group, having the largest amount of connections did not mean having the highest score. For instance the node with the largest amount of friends was node 1045, however 1045 had only an approximate score of 0.002, far below the maximum. Through further digging and experimentation of independent social networks, we found that scores were maximized when nodes had a large amount of connections, and those connections had little to no connections of their own besides the maximal node.

   (d) To explain the above in terms of a social network influence. The most influential nodes(the ones with the highest page rank score) are those that have large amounts of friends, and those friends have a small amount of friends of their own. Meaning the most influential node is one that has a monopoly of influence on its followers. This generally is an adequate measure of influence in social networks, and really captures the idea of monopoly influencers. However, this algorithm is ambiguous when nodes are equally influenced by other influential nodes. We could add a metric that takes into account the friends shared between the influential nodes and the shared node and gives the influencer with more shared connections with the common node.

□

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 don't." 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.)

Solution: Choosing prompt 1 - the 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
"Economic Incentives Don't Always Do What We Want Them To"

Exploring the above article in terms of the study of markets and networks, we realise a few key insights.

1. Incentives are not easily captured. By intrinsic virtue, people are incentivised by such varied things that it is not really possible to understand exactly which incentive or the combination of which incentives would be the most important for a person, let alone an entire network of people. The ways in which we can capture the trend itself could be via a neural network, wherein no factor is identified initially, just the input and output conditions. The model would learn and mimic the outcomes of the real world, and we would be better able to model the real world.

2. It is clear also, that the assumption that there is an important factor of influence in people's decisions from incentives of financial or other nature may be misplaced. There are more factors at play here that outweigh those of social/financial incentives. Previously in this study, we attempted to model a person's decision based on the information gained by decisions of others, as well as the signal of the correct decision (i.e. intuition) received by the individual. While this simplified model may be extendable, it is important to note the claim of the article that each person considers him or herself very little or even negligibly prone to influence by financial incentives. In other worlds, markets are incapable of bringing about the most socially optimal solution solely by themselves. There is intervention required by governance nad poilcy makers, as well as awareness among the general population.

3. Using the above, we attempt to better model real world situations. It is seen among all economic classes of society - financial incetivization is short-lived, and insufficient to promote work and welfare. The factors that seem to matter more are status, dignity, and social connections. Policy and governance is required to intervene when people lose the their jobs and are unwilling to allow change of location or occupation to make a living. These additional factors, weighted according to current trends would heavily enhance our study to model the real world networks and markets.

□