CS224W - Social and Information Network Analysis
Fall 2010
Assignment 2

Due 11:59pm November 4, 2010

General Instructions

You are required to write the name of your collaborators for this assignment on your solution report. You are also required to submit the source code used to obtain your solutions along with your report. Your write-ups are to be original, and all external references must be duly cited.

The names beside each question point to the TA dealing with the corresponding question. It would help you to also plan your visit to our office hours accordingly.

Submission Instructions

We expect you to be able to access your Dropbox folder at http://coursework.stanford.edu .

Whichever files you want us to see are to be archived (or zipped) together into a single file and placed into this Dropbox folder. This file should be named of the format: **<SUNetId>_ <Last-Name>_<First-Name>_HW1** : where these are your last and first names respectively. Within this file, we require the following at least :

a. Your solution report. You can submit it as a .pdf or .doc. This report should be named as : **<SUNetId>_<Last-Name>_<First-Name>_HW1_Ans (.doc or .pdf)**
b. Any source code that you use towards obtaining your results stated in your solution report. Also include the mention of any tools whichever you use towards your solution(s).

Questions

Question 1. Power Laws and Preferential Attachment (30 points - Jen)

Part 1: Empirical Power Laws

Generate a dataset of 100,000 values following a power-law distribution: $h(x) \propto x^{-\alpha}$ with exponent α =2.5. Refer to the paper Power-law distributions in empirical data by Clauset, Shalizi and Newman for how to generate random numbers from a power-law distribution (note: reading this paper is very helpful in answering this question!). Since the probability density diverges as $x \to 0$, you will need to bound your distribution, so let $x_{min} = 1$.

(a) Plot on a log-log scale $P(X = x)$, the probability distribution function (pdf). Your plot can be a normalized histogram of the data you generated. To check if you generated your data correctly, you can optionally include the actual probability density function as a line on the same plot.

(b) Plot on a log-log scale $P(X < x)$, the cumulative distribution function (cdf). You can plot this as a normalized cumulative histogram of your generated data. As above, you can optionally include the actual cumulative density function as a line on the same plot.

(c) Make a plot of $P(X \geq x)$. This is called the complementary cumulative distribution function (ccdf). What do you notice about the ccdf in relation to the other two graphs? Show mathematically that if a distribution is a power law and the exponent of the pdf is α then the exponent of the ccdf is equal to α-1.

(d) One way to fit a power law distribution to data is to create a histogram of the frequencies of x. On a log-log plot this becomes $ln(h(x)) \propto -\alpha ln(x)$. You can extract the value of α by performing a least-squares

linear regression on the log-log histogram data.

(e) However, as discussed in Clauset et al., least-squares regression is usually not the best method. Instead, estimate a value for α using the maximum likelihood estimate. Report the value of α you found, as well as the standard error for your estimate.

Part 2: Preferential Attachment

Create the following evolving model for generating an undirected graph:
Begin with a complete graph of three nodes. At every time step, select an edge of the current network uniformly at random, and introduce a new node to the network. The new node should add an edge to one of the nodes on the selected edge; pick from the two possible nodes uniformly at random.

(f) Run your simulation for 10,000 time steps. Using the maximum likelihood estimate method from above, show that empirically, p_k , the fraction of nodes with degree k, follows a power law with exponent α between 3 and 4. Report the value of α you found.

(g) Provide an intuitive explanation as to why this model relates to the preferential attachment model, i.e., why it generates a graph with power-law degree distribution.

(h) In this question we explore the relationship between node age (how many time steps it was in the network) and degree. Run your simulation for 10,000 time steps, then generate a plot of node age vs. average degree of nodes of that age. When averaging, you can bin ages (e.g. average all nodes ages 1-100, 101-200, etc.). What do you notice?

(i) Select one of the original three nodes you started with and plot its age vs. degree across the 10000 time steps of the simulation (as opposed to above, where we only considered final age vs. degree). Also track the node that joined at time step 1000 and, on the same plot, plot its age vs degree for the 9000 time steps it was in the simulation. What do you notice about the rate of growth of degree for each node? Is this realistic/desirable?

Now, we will change our simulation to see if we can increase the degree growth rate of late-joining nodes. We will again start with a fully-connected graph of three nodes, and at each time step a new node will enter the network. When each node enters the network, it will add one edge to another node with probability proportional to the degree of the other node. However, despite each new node only adding one edge when it joins, we will now allow it to enter the network with a virtual degree > 0. This "virtual degree" represents what is called the *fitness* of the node. The total degree of a node, in terms of the probability of it being chosen for attachment, is its fitness + its actual degree.

(j) Let half of the new nodes have fitness $= f > 0$, and the rest have fitness $= 0$. Repeat (i) above for $f = 5$, 20, and 50 – and make sure you're plotting actual degree of the node, not including fitness (so each late-joining node should start with only one actual edge, regardless of f). For the late-joining node you're observing, make sure it has fitness $= f$. You may want to run this a few times to get a better sense of what is happening since randomization will produce a different outcome each time (–although you only need to submit one set of plots). What do you observe?

Question - 2. Influence Maximization (20 points - Sudarshan)

Recall the greedy hill-climbing algorithm discussed in class, to determine the near optimal influence set for the network, starting from a bunch of nodes and their influence node sets. Also recall the tight online bound derived in class using the sub-modularity property, of identifying this optimal influence set.

NOTE - By failing, we mean that the greedy hill climbing method obtains a final solution (i.e., initial set of active nodes S) whose size (final number of active nodes) is below that of the optimal solution. Also, by an example, we mean to construct an influence set for each node:

A -> {A,B,C};

C -> {.....} etc...

where A is a node and when A is initially activated, then nodes A, B, C are finally active. A, B and C are the nodes it influences. This is its influence set. Same applies for C and the other nodes you'd define influence sets for.

The example would contain a set of nodes and their starting influence sets, with a step by step workout of how at subsequent iterations the final set of influence nodes (S) changes. You are also expected to define the optimal solution starter set (T) in the example, and eventually provide an estimate of the margin by which the size of the greedy method's solution set S falls short of that of the final solution (ie), if f(S) = x∗f(T), you have to show the calculation of x.

(a) For $k = 2$ (Recall that k stands for the size of the best starting set of influencer nodes that yields us the near optimal solution of the overall influence set.), is it possible to construct an example where the greedy hill climbing method (i.e., gives a suboptimal solution)? If yes, construct such an example?

(b) For $k = 3$, can you provide an example where the hill climbing method gives a final set whose size <= 0.85∗ size of the optimal solution set?

(c) For what general property of starting influence sets for nodes, does the greedy hill climbing method always result in the globally optimal solution ?

(d) Do part (a) using the following approach instead of using the greedy hill climbing approach In each iteration, to add extra nodes to the solution set, pick nodes in order of their decreasing size of influence sets. Again, is it possible to construct an example where this method gives a final influence set whose size <= 0.85∗ size of the optimal solution set? Does this perform better or worse than the greedy hill climbing method above ?

(e) Can you construct an example where the online bound (refer to lecture slides about its derivation) is arbitrarily loose?

Extra Credit: (10 points)

(f) Can you provide a general hypothesis on (ie, describe the algorithm for) generating starting influence sets for nodes such that the greedy hill climbing method always fails? We would admit that your hypothesis is valid if for an example you provide, you are able to show your hypothesis in action for at least 5 different values of k (k > 1).

Question - 3. Information cascades (18 points - Sonali)

Part A (10 points)

Consider a group of people (numbered 1, 2, 3, . . .) who will sequentially make decisions, that is, individual 1 will decide first, then individual 2 will decide, and so on. An individual can either accept(A) of reject(R) a decision based on some signals. The probability that accept is a good idea is 1/2 and the same is true for a reject decision. When individual i chooses he observes only his own signal and the action of previous individuals, 1 to i-1. (However, individual i does not see the actual private signals of any of these earlier people.) Specifically, there are two possible signals: a high signal (denoted H), suggesting that accepting is a good idea with a probability 4/7; and a low signal (denoted L), suggesting that rejecting is a good idea (with the same probability 4/7).

1) Calculate the probability that the decision is Accept given that the first player sees a High signal using Bayes rule.

2) What is the probability of an accept decision, given that the third person gets a Low signal and the previous two people have made accept decision?

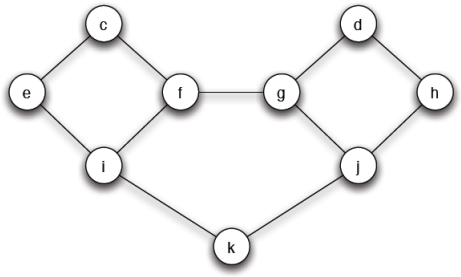
3) Find the smallest number x such that when the first x people guess that the decision is accept, an information cascade starts where every person guesses that it is a accept decision(A cascade).

4) What is the probability that the first x people guess decision is accept(A) when the decision is actually reject (R) (where x is the same as part 3)?

5) Suppose that you are the tenth person to make a choice and you have observed that everyone before you chose R. That is, we are in an R-cascade. Now lets suppose that before you (person 10) receive your signal, you decide to ask person 9 about the signal that they observed. Lets suppose that person 9 observed a High signal, that person 9 tells you that his signal was High, and that you know that person 9 is telling the truth. After this, you receive your own signal. What decision should you make, A or R, and how does it depend on which signal you receive?

Part B (8 points)

For this problem, consider the network shown in the next page. Two technologies A and B are competing in this network. Suppose that everyone in the network is initially using B as a default behavior. Technology A wants to take over the network. Payoff for using technology A is 3 and payoff for using technology B is 2.



1) Consider Points e and f are the early adopters of technology A. Illustrate the adoption pattern by the rest of the network. (Hint: You will see that that the cascade runs for a while but stops while there are still nodes using B. Why is that? What does it tell you about the structure of the network ?)

2) One strategy the firm producing A wanted to push its adoption past the point is to raise the quality of its product so that there is a complete cascade, in which every node in the network switches to A. What should the minimum threshold be for that to happen?

3) When its not possible to raise the quality of A in other words, when the marketer of A can not change the threshold a different strategy for increasing the spread of A would be to convince a small number of key people in the part of the network using B to switch to A. What minimum nodes of would you choose (give their labels)?

4) Suppose you were allowed to add a single edge to the given network, connecting one of nodes c or d to any one node that it is not currently connected to. Could you do this in such a way that now behavior A, starting from the early adopters, would reach all nodes? Give a brief explanation for your answer.

Question 4. Infection Diffusion (30 points - Nadine)

We consider the spread of an infection through a society, modeled by a network of nodes. The disease can be transmitted when two people, one of which is already infected, are in contact. The spread of the disease is governed by its infection rate, the immunity of individuls and the structure of the network. It is important to study the partition of the network into its separate connected components, as they provide a natural limit to the diffusion of the disease.

In this problem, you will study a diffusion procedure. The model is a network in which certain individuals can be immune to the infection. You will determine a threshold value of the fraction of immune people below which the disease still spreads to a significant part of the network nodes.

Part 1

Let G be a graph on n nodes. We add an edge between every pair of nodes with probability p.

a) What is the expected number of edges in the graph?

b) What is the probability $p_D(d)$ that a certain node has exactly d edges?

This is a binomial distriution. The limit of the previous distribution for large n and small p is a poisson distribution with parameter $(n-1) \times (p)$ which is of the form:$p_D(d) = \frac{e^{(n-1)p}((n-1)p)^d}{d!}$. Suppose now that we have already connected $n-1$ nodes of the graph and assume that we have a large connected component. Let β be the fraction of nodes in that component.

c) Assume the n'th node to be added has degree d. What is the probability that it does not get included in the connected component, given that the fraction of nodes in the largest component is β? Using the same argument on all nodes, express the overall fraction of nodes outside the largest component in terms of p_D. (Hint: sum over all possible degrees d the probability that a node has degree d and is not included in the largest component)

d) Rewrite the fraction of nodes outside the connected component in terms of the poisson distribution. What happens as $(n-1) \times p$ tends to infinity?

e) Run simulations for graph constructions by varying the parameter $\lambda = (n-1)p$ (fix p=0.1 and increase n). For each graph, determine the size of the largest component. Plot the fraction of nodes in your largest component with respect to λ. What do you observe to be a critical value for λ above which the fraction of nodes in the largest connected component increases suddenly?

Conclusion: We conclude from part 1 that for this network model, the fraction of the nodes belonging to the connected component tends to 1 as the number of network nodes increases.

Part 2

Let the nodes of graph G represent n individuals of a society. Denote the node degree distribution by p_D. When an edge exists between an infected node and a non-infected non-immune one, the disease is immediately transmitted (so infection is certain once there is an edge). You will study the random immunization of nodes.

Let α be the probability that an individual is immune to an infection and assume we start by infecting a single individual. If the first infected person belongs to the largest connected component, the infection will end up spreading to a significant fraction of the population. Thus a disease outbreak is tightly related to the presence of a large connected component in the graph. Bearing in mind that some nodes in our network are allowed to be immune to the infection, we consider a sub-graph of G, G', which consists of non-immune nodes.

We construct G' by removing each of the nodes from the network with probability α. We try to determine a threshold value for α below which we still have a large connected component which allows the infection to spread.

a) Let p_α be the degree distribution of the nodes in the graph G'. Write $p_\alpha(d')$ in terms of p_D.
(Hint: a node with a degree d' in G' corresponds to a node of degree d in G that lost $d-d'$ edges when immune nodes and their corresponding edges were removed).

b) Fix an edge in the graph G' and choose one of the nodes it connects to at random. Let $p^*_\alpha(d)$ denote the probability that the chosen node has degree d. Prove that:

$$p^*_\alpha(d) = \frac{d \times p_\alpha(d)}{E_{p_\alpha}[d]}$$

where $E_{p_\alpha}[d]$ is the expected value of the degree of a node with respect to the distribution p_α in the graph

G'. This value will be useful in subsequent questions when for a given edge, we try to evaluate the expected degree of a node it connects to in order to see whether the component is growing.

We will now give a heuristic argument to find the threshold value for the appearance of a large connected component. We infect one node v in the network. Let's look at the nodes it is connected to. If we choose an edge that connects it to w, we need the degree of w to be at least equal to 2 for this component to keep growing. Thus, we require the expected degree of the node w to be greater than or equal to 2 to have a connected component: $E^*_{p_\alpha}[d] \geq 2$.

c) Write $E^*_{p_\alpha}[d]$ in terms of $E_{p_\alpha}[d]$ and $E_{p_\alpha}[d^2]$. Conclude that the threshold value for α satisfies $E_{p_\alpha}[d^2] = 2 \times E_{p_\alpha}[d]$.

d) We now need to write $E_{p_\alpha}[d]$ and $E_{p_\alpha}[d^2]$ in terms of α. Prove that: (1) $E_{p_\alpha}[d] = (1-\alpha)E_{p_D}[d]$ and (2) $E_{p_\alpha}[d^2] = E_{p_D}[d^2] \times (1-\alpha)^2 + E_{p_D}[d]\alpha(1-\alpha)$.

e) Find an expression for the threshold value α by plugging in the values from part d into the threshold equation from c. For a fixed number of nodes $n = 100$ and $p = 0.1$, plot the fraction of nodes in the largest connected component versus different values of α.

Conclusion: We have found a threshold value for the fraction of immune individuals, below which we have a large connected component. The disease spreads to a significant fraction of the population even after the removal of immune nodes from the network.