Due October 29 2009 in class

CS322: Network Analysis

Problem Set 2 - Fall 2009

If you have any questions regarding the problems set, send an email to the course assistant: simlac@stanford.edu. Please write the name of your collaborators on your problem set. You can use existing software or code to compute the answers, you don't have to submit the source code.

The Problems

Problem 2.1

(From Easley and Kleinberg, Networks) In the basic "six degrees of separation" question, one asks whether most pairs of people in the world are connected by a path of at most six edges in the social network, where an edge joins any two people who know each other on a first-name basis.

Now let's consider a variation on this question. Suppose that we consider the full population of the world, and suppose that from each person in the world we create a directed edge only to their ten closest friends (but not to anyone else they know on a first-name basis). In the resulting "closest-friend" version of the social network, is it possible that for each pair of people in the world, there is a path of at most six edges connecting this pair of people? Explain.

Problem 2.2

You are developing a protocol to establish a peer-to-peer overlay network among n nodes. This protocol operates as follows.

Step 1: Each node flips a coin (n-1) times to decide whether it generates an edge to each of the other (n-1) nodes. The probability of doing so is p. Links are assumed undirected, regardless of which side establishes them. If two nodes flip their corresponding coins and both decide to connect to each other, only one edge is created.
Step 2: After this is done, every node not yet connected selects another node at random and establishes a link to this node.

If you let $p = \log n/(2n)$, does this protocol establish a connected network for large n? (Hint: determine what small components exist after Step 1, and in particular, the number of isolated vertices.)

What would your answer be if p was only $1/n$?

Problem 2.3 Generate a dataset of 1 million values following a power-law distribution with exponent 2.5. Then compute experimentally the exponent of the distribution, using the following 4 methods:

Refer to *Power-law distributions in empirical data* by Clauset, Shalizi and Newman for how to generate random numbers from a power-law distribution.

a) Fitting a line to the frequency distribution.
b) Fitting a line to the frequency distribution with logarithmic binning.
c) Using the complementary CDF.
d) Using the maximum likelihood estimate.

Problem 2.4 Consider the following evolving model for generating an undirected graph. Initially there are only three nodes connected into a triangle. At every time step, an edge of the current network is selected uniformly at random, and a new node is added to the network that links to both the endpoints of the edge. Prove that p_k, the fraction of nodes with degree k, follows a power law with exponent 3. Provide an intuitive explanation as to why this model is the same as the preferential attachment model.

Problem 2.5 In this exercise we will study the distribution of words in the English language. The data consists of a list of all the words in a dictionary and a text version of "A tale of Two Cities" by Charles Dickens (found at project Gutenberg). In the later, we have removed punctuation, apostrophes, etc... keeping only the 26 characters in the alphabet and the space.

(a) Write a program that reads the list of words provided and plot a graph showing the number of words that there exist of lengths between 3 and 8 (you can discard all other words). How fast does such number increase?

(b) Using the novel "A Tale of Two Cities" as a representative sample, we now plot how frequently each words is used in the English language. Sort the words in the novel along the x axis from the most frequent to the least, and plot their number of appearances (many words in the dictionary will not be in the novel. You should not take those into account). Does it follow a power law? If so, find an approximation for the exponent.

If you looked further into the previous plot, you would see that the most frequent words are usually shorter. We now develop models to explain why, if long words are more numerous in the dictionary, authors use short ones more often.

(c) Assume that a monkey typed one billion (10^9) random characters on a keyboard (26 letters + space bar), and call "word" any sequence of letters between two spaces. Find f(n), the expected number of times that a GIVEN sequence of length n would appear in the monkey's text (with spaces at both sides). Does f(n) follow a power law? If so, find an approximation for the exponent.

(d) In average, how many times would the 100-th most frequent word appear in the monkey's text? What about the 1000-th? (Hint: how long would those words be? Either simulate it or find an analytic expression) Is this a good model for the results in (b)?

(e) We will try to further improve the model by assigning different probabilities to different characters. Find the probability of each character (including space) in "A Tale of Two Cities" and generate ten thousand words according to that distribution. Repeat the plot in part (b) for this new text. Is the model better?