CS/ENGRI 172, Fall 2002
11/1/02: Homework Five

Seek not for the favor of the multitude; it is seldom got by honest and lawful means. But seek the testimony of the few; and number not the voices, but weigh them.
— *Immanuel Kant*

When someone points an accusing finger, look where the second finger points.
—*proverb cited in Gilb and Weinberg,*
Humanized Input: Techniques for
Reliable Keyed Input *(1977)*

Due at the *beginning* of class on Wednesday, November 13. Hand in the parts separately, with your name on each part. Answers will be graded on both correctness and clarity. If a question asks for explanation, then no credit will be assigned for answers without explanation. In general, you should always include an explanation of your answers when appropriate (among other things, this helps in assigning partial credit).

Some potentially useful facts We use a, b, and c to denote constants, x to denote variables.

$$\begin{aligned}
\ln(x) &= \text{the logarithm of } x, \text{base } e \approx 2.718\ldots \\
a^{b+c} &= a^b a^c \\
e^{\ln(ax+b)} &= ax + b \\
\frac{d(ae^{bx})}{dx} &= abe^{bx} \\
\frac{d((ax)^b)}{dx} &= ab(ax)^{b-1} \\
\frac{d(\ln(ax))}{dx} &= a/ax = 1/x
\end{aligned}$$

Part A

1. In this question, we consider a specific example of how different the uniform and preferential attachment models are in terms of predicting in-degree as a function of time.

Suppose we start with one initial document, d_{-1}, and that for each new document $d_1, d_2, \ldots$ we give the new document 20 links to pre-existing documents, allowing repeated links to the same document. Assume timesteps correspond to seconds.

(a) How long would it take according to the uniform attachment model for at least one of the non-initial documents to achieve an expected in-degree of 2000 (which is approximately the in-degree of www.cs.cornell.edu, according to Google)? Show and explain your work.

(b) How long would it take according to the preferential attachment model for at least one of the non-initial documents to achieve an expected in-degree of 2000? Show and explain your work.

(c) Can you use exactly the same equations as you used above to compute how long it would take the *initial* document d_{-1} to achieve an estimated in-degree of 2000? Explain your answer.

2. Consider the set L consisting of all and only sentences consisting of a mixture of a's and b's, where a's can't be right next to each other, and b's can't be right next to each other. For example, *aba* and b are in L, but *abaa* and *ababba* aren't. Give a CFG that generates all and only the sentences in L, explaining how it works (e.g. writing down the role of each rewrite rule). Also, draw a parse tree according to your CFG for the sentence *aba*.

Part B

3. In class, we computed the estimated in-degree $I_j(t)$ of documents under the uniform attachment and preferential attachment models, but didn't actually compute the resulting in-degree distributions — that is, for a given in-degree, what fraction of documents have that in-degree. Here, we finish off the calculations.

The general idea of how to arrive at the in-degree distribution is as follows. Define the function $\text{Atmost}_t(D)$ to be the fraction of non-initial documents at time t that have in-degree *less than or equal to* D. Then, to compute for any given in-degree value D how many documents have *exactly* that in-degree — that is, to compute the in-degree distribution — we can take the derivative of $\text{Atmost}_t(D)$ with respect to D, because the derivative gives us the difference between $\text{Atmost}_t(D)$ and $\text{Atmost}_t(D + \Delta)$ for Δ arbitrarily close to 0.

For simplicity, assume the evolution process started with one initial document, and that for each new document that was created, only one link from it to a pre-existing document was chosen. We only look at the in-degrees for a *fixed* time $t > 1$, so we treat t as a constant, and have the t non-initial documents $d_1, d_2, \ldots, d_t$ to consider.

(a) Recall that $I_j(t)$ is the estimated in-degree of d_j. Under the uniform attachment model, for any given in-degree value D, for which documents d_j, $1 \leq j \leq t$, is it the case that $I_j(t) \leq D$? Explain your answer, showing your work.

Note: our answer is in terms of j; for example, an (incorrect) answer might be, "$I_j(t) \leq D$ when j is at least D".

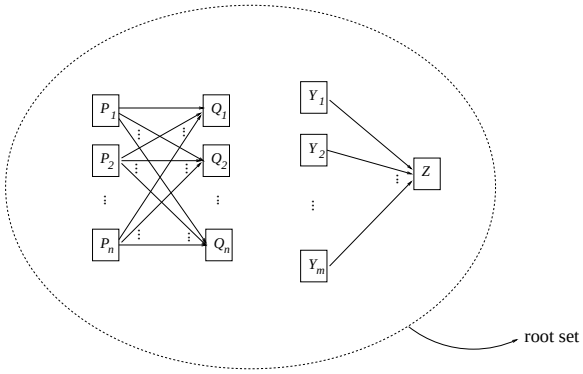
(b) Use the above subproblem to give $\text{Atmost}_t(D)$ as a function of D under the uniform attachment model, showing and explaining your work. For example, continuing with our incorrect answer above, we would observe that of the t non-initial documents, our computations imply that $t - D + 1$ of them would have in-degree at most D, which means that the fraction of non-initial documents with in-degree at most D is $\text{Atmost}_t(D) = (t - D + 1)/t$.

(c) Finally, arrive at the in-degree distribution function under the uniform attachment model by taking the derivative of $\text{Atmost}_t(D)$ with respect to D; show your steps and provide explanations as necessary. (Self-check: you should already know what kind of function to expect from lecture.)

(d) Repeat the above three subproblems for the preferential attachment model.

Part C

4. In this question, we consider tradeoffs between high in-degree and community membership with respect to the hubs and authorities algorithm (henceforth HA). Let m and n be two whole numbers bigger than 1. Consider the following set of $2n + m + 1$ documents, all deemed to be on the same topic by some content-based information retrieval system:



All of the n P-documents point to all the n Q-documents, and all of the m Y-documents point to document Z.

(a) Let $m = 2n$, where n is a fixed, unknown constant satisfying the conditions above. Suppose HA is run until the step where the hub scores are pseudo-normalized has occurred twice. Which documents have the highest hub scores at this point, and which have the highest authority scores? Explain your answers, making sure to show your computations of the hub and authority scores at each step. You should use the tabular format from the Lecture Twenty-Four handout, but make clear how you are computing the updates and the normalization factors at each step.

(b) Repeat the above subproblem, but where $m = n^2$.

(c) Suppose evil corporation X wishes to cause their webpage to undeservedly have the highest ranking according to an HA-based search engine. Their strategy is to set up enough shell companies with webpages that link to the X homepage that it becomes the webpage with the highest in-degree among all webpages mentioning X's product. Discuss in a few sentences whether X's strategy will be effective.