COMS W4995-1: Intro to Social Networks — Due on October 13th, 2011

Assignment #3 – Evolution

A. Chaintreau (instructor)

Why there is two parts in this assignment: Each part fulfills one of the two objectives of the class:

- **Manipulate concepts:** Getting Familiar with the technical concepts used in class, by reproducing similar arguments. Being proficient by manipulating the object to answer some small-size problem.
  You are expected to answer this question rigorously, the answer can be quite short.
- **Connect the concepts to real-life:** Interpret a problem you find in light of the notions you have learned. Develop some critical eye w.r.t. how the concepts introduced are useful in practice.

How to read this assignment : Exercise levels are indicated as follows

($\rightarrowtail$) "elementary": the answer is not strictly speaking obvious, but it fits in a single sentence, and it is an immediate application of results covered in the lectures.
Use them as a checkpoint: it is strongly advised to go back to your notes if the answer to one of these questions does not come to you in a few minutes.

($\curvearrowright$) "intermediary": The answer to this question is not a simple application of results covered in class, it can be deduced from them with a reasonable effort.
Use them as a practice: how far are you from the answer? Do you still feel uncomfortable with some of the notions? which part could you complete quickly?

($\twoheadrightarrow$) "tortuous": this question either requires an advanced notion, a proof that is long or inventive, or it is still open.
Use them as an inspiration: can you answer any of them? does it bring you to another problem that you can answer or study further? It is recommended to work on this question only when you are done with the rest.

Part A
Practicing the concepts

Exercise 1: A variant of the community guided attachment model

In class, we have analyzed a first community guided attachment model (CGA) which connect a set of nodes in a way that relates to a hierarchy. More precisely we have assumed that the set of nodes are *leaves* of a tree and that the tree allows to define a natural distance between them. As an example two leaves from the exact same parent are close whereas leaves of branch of the tree that met only at the root are far. More generally we can always define the distance between two nodes u and v, (*i.e.*, two leaves of the tree) as the number of edges in the path that connects them in the tree, which we denote $d(u, v)$.

We now consider a model where the set of nodes are not only leaves, but all nodes in the tree, including interior nodes. To simplify the model we assume, as made in class, that this tree representing the hierarchy is a complete b_ary tree. For a node u we denote its height by h_u (which is 0 if u is a leaf, 1 if u is the parent of a leaf, and so on). The total height of the tree is $h_{\max}$, so that there are $b^{h_{\max}}$ leaves in the tree, $b^{h_{\max}-1}$ nodes of height $h = 1$ and so on.

We will make the same assumptions as in the model seen in class:

- We are interested in large set of nodes, so we will assume that $h_{\max}$ is going to infinity. This implies that the number of nodes N goes to infinity as well, as we have:

$$N = \sum_{h=0}^{h_{\max}} b^{h_{\max}-h} = \frac{b^{h_{\max}+1}-1}{b-1}\,.$$

- The graph between the nodes is created recursively. It initially contains a single node, which is also a leaf. A new "generation of leaves" is then added to the graph by creating b children for each leaf. Each new leaf u creates an edge to another node v (a leaf or an interior node), with probability

$$\mathbb{P}\left[u \rightsquigarrow v \in E\right] = \frac{1}{c^{d(u,v)/2}}\,.$$

  These events are independent among all pairs (u, v). Note that this probability decreases with the distance $d(u, v)$ which is also natural since there are more and more nodes at a given distance d as d grows. Note also that the edge is directed and that by definition v has height at least the same as u.

The goal of this exercice is to understand under which condition this model creates edge densification. In other words, as $h_{\max}$ increases, we wish to determine for which choices of b and c the number of edges in the graph (or equivalently the average degree of a nodes) grows as a power function of N.

We remind that, when f and g are two functions of N, we denote by $f(N) = \Theta(g(N))$ the fact that two constants $M > 0, M' > 0$ exist such that $M \cdot g(N) \leq f(N) \leq M' g(N)$.

1. ($\curvearrowright$) Prove that for a leaf (*i.e.*, $h = 0$), the number of nodes at distance $d \leq 2h_{\max}$ is,

$$\texttt{NeighIn}(h_{\max}, 0, d) = \sum_{j=\lceil \frac{d}{2} \rceil}^{\min(d-1, h_{\max})} (b-1)b^{d-j-1} + \mathbb{I}_{\{d \leq h_{\max}-h\}}\,.$$

   Deduce that there exist $A > 0, A' > 0$ such that for any $d \leq 2h_{\max}$:

$$A \cdot b^{\frac{d}{2}} \leq \texttt{NeighIn}(h_{\max}, 0, d) \leq A' \cdot b^{\frac{d}{2}}\,.$$

2. ($\curvearrowright$) Deduce for $c < b$ that there exist $B > 0.B' > 0$ such that the average out-degree of a leaf, denoted $\texttt{DegOutLeaf}(h_{\max}, 0)$, satisfies:

$$B \cdot \left(\frac{b}{c}\right)^{h_{\max}} \leq \texttt{DegOutLeaf}(h_{\max}, 0) \leq B' \cdot \left(\frac{b}{c}\right)^{h_{\max}}\,.$$

   We assume that $h_{\max} = \frac{\ln(N)}{\ln(b)}$, which is a good approximation as N becomes large. Deduce that, as a function of the number of nodes N, the average out degree $\texttt{DegOutLeaf}(N)$ of a leaf satisfies $\texttt{DegOutLeaf}(N) = \Theta\left(N^{1-\frac{\ln c}{\ln b}}\right)$.

3. ($\curvearrowright$) Conclude that the same result holds for the average out-degree of any node. (Hint: use the fact that (1) leaves have maximum average out-degree among the nodes, and (2) there is a constant fraction of leaves in the graph for any $h_{\max}$.)

4. ($\curvearrowright$) What can you say about the average degree of a node as a function of N when $b < c$?

The rest of this exercice is only used for extra credit. Using similar arguments, it proves that this model generates a power-law distribution of in-degree of nodes.

5. ($\looparrowright$) For a node u with height h, show that the number of nodes v at a distance d and with height $h_v \leq h$, when we assume that $d \leq h$, is equal to:

$$\texttt{NeighIn}(h_{\max}, h, d) = b^d + \sum_{j=\max\left(1, \left\lceil \frac{d-h}{2} \right\rceil\right)}^{\min\left(\left\lfloor \frac{d}{2} \right\rfloor, h_{\max}-h\right)} (b-1) b^{d-1-j} \,.$$

   Remember that nodes at distance d may be in the subtree rooted in u as well as in other parts of the tree.

   Deduce that there exists $C > 0, C' > 0$ such that for all $h_{\max}, h, d$, such that $d \leq h \leq h_{\max}$ we have

$$C \cdot b^d \leq \texttt{NeighIn}(h_{\max}, h, d) \leq C' \cdot b^d \,.$$

6. ($\curvearrowright$) If we now assume $d > h$, how is the above expression modified? Deduce that there exists $D > 0, D' > 0$ such that for all $h_{\max}, h, d$, such that $h < d \leq 2h_{\max} - h$ we have

$$D \cdot b^{\frac{d+h}{2}} \leq \texttt{NeighIn}(h_{\max}, h, d) \leq D' \cdot b^{\frac{d+h}{2}} \,.$$

7. ($\curvearrowright$) Deduce that there exists $E > 0, E' > 0$ such that:

$$E \cdot \left(\frac{b}{c}\right)^{h_{\max}} c^{h/2} \leq \texttt{DegIn}(h_{\max}, h) \leq E' \cdot \left(\frac{b}{c}\right)^{h_{\max}} c^{h/2}$$

8. ($\curvearrowright$) Show that, when $h = h_{\max}$ we obtain the node with largest average degree, and by the previous inequality we have $\texttt{DegIn}(N, h = h_{\max}) = \Theta\left(N^{1-\frac{1}{2}\frac{\ln c}{\ln b}}\right)$.

   N.B.: More generally, we have $\texttt{DegIn}(N, h) = \Theta\left(N^{1-\frac{1}{2}\frac{\ln c}{\ln b}} c^{-\frac{h_{\max}-h}{2}}\right)$. This implies that the node with "depth" $h_{\max} - h$ (and, hence, which has rank i, where $i = b^{h_{\max}-h}$, in the decreasing sequence of average degree) has a degree $(\sqrt{c})^{h_{\max}-h}$ smaller than the node with larger average degree. In other words, the node with rank i has degree smaller by factor $(\sqrt{c})^{\frac{\ln i}{\ln b}} = i^{\frac{1}{2}\frac{\ln c}{\ln b}}$. Hence, the sequence of average degrees follows a power-law with coefficient $\frac{1}{2}\frac{\ln c}{\ln b}$.

Part B

Concepts at large

Exercise 1: How to avoid death by success? You are working in a social media start-up whose main product is a web content recommendation platform that outperform current method. The service was launch a year ago and you are now, after an initial adoption phase, in a steady increase of the number of users.

The main challenge you are facing is to keep up with the demand from an evergrowing amount of users. The current recommendation technique that you use is based on a set of shortest paths from each node to a set of landmark (which are representative nodes). The main cost per node comes from computing coefficients during each hop of these path, so that your cost of running the recommendation is a function of the number of hops from this node to a set of others.

A start-up claims to make a recommendation as accurate as yours, with a brand new technique which does not use path but only direct neighbors. The computation cost of this method per node is then a function of the total number of neighbors. You run a test and conclude that *today* with the current graph of friendship, the two methods have roughly the same computation cost, and similar accuracy.

1. As you expect the graph to expand, and hence its topology to evolve in the coming months, how can you predict the evolution of the computation cost for each method? Should you buy the recommendation engine developed by this other start-up?

2. This other start-up mentions that they are developing an extension that allows to reuse computed results whenever a new node is belonging to a triangle between two already computed node. Do you think that this method can result in important savings?