COMS W4995-1: Intro to Social Networks — Due on October 4th, 2011

Assignment #2 – Scale

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?

($\nrightarrow$) "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: Analysis the copying model Through the analysis of the Yule process, we have seen in class the consequence of reinforcement. Reinforcement here denotes the fact that a difference between two entities (e.g. the size of two genus, the number of links received by two webpages) is itself biasing the dynamics so that the difference continues to increase. As a consequence, even starting from a small initial set of equivalent entities, minor difference created by randomness could further lead to major differences. In the case of the Yule process, it provided a simple model explaining the imbalance of species among genus which is characterized by a power law.

In this exercice, we conduct a very similar analysis to model edges created in a graph. The main result is to show that a very simple copying strategy leads to big imbalance, characterized by a power law degree distribution of nodes' in-degree.

The copying model We analyze the following dynamics. We start from a directed graph containing N_0 nodes such that each of these nodes has exactly one outgoing edge. We introduce at each time step, denoted by $t = 1, 2, \ldots$, a new node $v(t)$ with a single outgoing edge $e(t)$ that is initially connected to another node chosen uniformly at random (that we denote by $u(t)$). We then assume the following evolution:

- with probability p, the process stops there, and the new edge connects $v(t)$ to $u(t)$.
- otherwise (hence, with probability $(1-p)$), $v(t)$ examines the edge that is starting at $u(t)$ and decides to *copy* this edge. This means that the edge from $v(t)$ to $u(t)$ is changed to one that goes from $v(t)$ to the destination of the edge starting in $u(t)$.

Evolution of node's degree Since the graph is directed all nodes both have an out-degree and an in-degree. The out-degree of all nodes in the graph remains constant equal to 1. The interesting problem is to analyze the evolution of the in-degree of nodes in the graph as t becomes large.

Let us denote for any $i \geq 0$ by $X_i(t)$ the number of nodes in the graph with an in-degree equal to i.

1 ($\rightarrowtail$) How many nodes (denoted by $N(t)$) and edges (denoted by $E(t)$) are there in the graph as a function of t?

2 ($\curvearrowright$) Assuming that $X_0(t)$ (*i.e.*, the number of nodes with no incoming edge) is known, what are the possible values of $X_0(t+1)$ and what are the probability that these values occur?

3 ($\curvearrowright$) Derive from the previous question the evolution equation giving the expectation $\mathbb{E}[X_0(t+1)]$ as a function of $\mathbb{E}[X_0(t)]$.

We now assume that $p < 1$.

4 ($\curvearrowright$) Let us introduce, for a given constant c_0, the sequence $\Delta_0(t) = \mathbb{E}[X_0(t)] - c_0 t$. Show that there exists a value of the constant c_0 such that:

$$\exists A > 0 \text{ such that for any } \quad t > 0 \, |\Delta_0(t)| \leq A \ln(t) \, .$$

What is the value of c_0? (*i*) $c_0 = \frac{1}{1-p}$ (*ii*) $c_0 = \frac{1}{2-p}$ (*iii*) $c_0 = \frac{1}{1+p}$

(N.B.: Note that the following fact is useful: If $(x_n)_{n\geq 0}$ is a sequence of real number such that $(\forall n \geq 0 \, , \, x_{n+1} = x_n r_n + s_n \, , \quad \text{and} \quad |r_n| \leq 1)$, then we have $|x_n| \leq |x_0| + \sum_{i=1}^{n} |s_i|$.)

5 ($\rightarrowtail$) Deduce that the following hypothesis is true for $i = 0$:

$$\forall \varepsilon > 0, \exists A > 0 \text{ such that } \quad |\Delta_i(t)| \leq A t^{\varepsilon} \, .$$

6 ($\curvearrowright$) For a sequence of constant $c_0, c_1, \ldots$, let us define $\Delta_i(t) = \mathbb{E}[X_i(t)] - c_i t$. Show that for any $i > 0$, if the sequence satisfies $c_i = c_{i-1}\left(1 - \frac{2-p}{(1+p)+i(1-p)}\right)$ then we have:

$$\Delta_i(t+1) = \Delta_i(t)\left(1 - \frac{p + i(1-p)}{N(t)}\right) + \Delta_{i-1}(t)\frac{p + (i-1)(1-p)}{N(t)} + \frac{A}{N(t)} \tag{1}$$

where A is a constant.

7 ($\curvearrowright$) Deduce by recurrence that the hypothesis of question 5 is true for all $i \geq 0$.

8 ($\nrightarrow$) We admit the following concentration result on the random sequence $(X_i(t))_{t>0}$ for any $i \geq 0$

Lemma 1. *For any $M > 0$, we have* $\mathbb{P}\left[|X_i(t) - \mathbb{E}[X_i(t)]| > M\right] \leq 2\exp\left(-\frac{M^2}{8t}\right)$

Prove that almost surely there exists a value T such that for $t > T$ we have

$$|X_i(t) - \mathbb{E}[X_i(t)]| \leq 4\sqrt{t \ln(t)}\,.$$

Conclude that for any $i \geq 0$, almost surely, the fraction of nodes with in-degree equal to i converges to c_i as t grows large.

9 ($\rightarrowtail$) Assuming that $p < 1$, show that for $i > 0$ we have:

$$c_i = c_{i-1}(1 - \frac{\beta}{i} + \varepsilon(i)) \text{ where } |\varepsilon(i)| \leq \frac{A}{i^2} \text{ and } \beta = \frac{2-p}{1-p}\,.$$

As a consequence, as shown in the lecture, if we neglect the error term $\varepsilon(i)$ we have that c_i is approximately following a power-law with coefficient β.

For which values of p does the power law becomes the most imbalanced? Does this correspond to your intuition about the dynamics of copying.

10 ($\curvearrowright$) Assuming now $p = 1$, how could you characterize the decrease of c_i as a function of i? Relate this behavior to the dynamics of the copying model.

Part B

Concepts at large

Exercise 1: Effect of Recommendation Engine You are working on a start-up that deals with a large number of user-generated videos. An important part of your website is that users are able to post recommendation for others, in which they "endorse" videos.

Based on your study of some users feedback, you have noticed that users on the website typically prefer videos that receive some endorsement as opposed to none, essentially because it avoids spam. You have also noticed that not so many users are interested by the most endorsed videos, that happen to be hugely popular, because they tend to be somewhat too broad to really be in their interest. You would like to keep this popularity somewhat balanced so that users can enjoy the breadth of your catalog.

One person suggests to introduce a random catalog browsing, in which every user will be shown an endorsement that would be chosen uniformly at random among all endorsements that exist at this time, together with a quick link to endorse this videos as well.

1 Using elements from the lecture, can you predict how the popularity will change as you introduce this feature?