Statistical Network Models - 9/26/16 Class Notes

Matthew Babcock

October 12, 2016

Network Growth Models

1 Introduction

To date we have focused on static graphs where the edges are independent given the attributes of the nodes they connect (what block/community/location those nodes are labeled). We would like to further explore how networks are formed over time. We are interested in models in which networks arrive at a specific state due to explicit mechanisms for adding (or removing) nodes and edges. One way of going about finding such models is to infer something about the growth mechanism based on the observed configuration of the network at a given time.

An important observation about real networks from earlier in the class is that the degree distributions of real networks are typically very right-skewed and heavy tailed. Take for example the directed network of scientific citations from the ISI: there are approximately 5 million recorded nodes (papers), but the modal number of citations is 0, the mean is around 5, the maximum citations is around 20,000, and there are a few hundred papers with over 1,000 citations. This is similar to other real networks that have various constraints such as friendship or biophysical networks. You can't get these types of distributions through combinations of Poisson's, so the question is what type of underlying model might explain how these networks came about.

2 Proposed Mechanisms

2.1 Simple Multiplicative Growth Model

In this type of model, we imagine continuous-sized objects growing in continous time.

In the simplest form:

$X_i(t)$ **= size of object i at time t,**

$\frac{\delta X_i(t)}{\delta t} = \mu X_i(t)$ **= growth** $\propto$ **size, large objects grow faster than small**

$X_i = X_0$**initially, and at age** τ_i**,** $X_i = X_0 e^{\mu \tau_i}$

Objects appear at uniform rate λ **so looking backward from a given time, the age distribution is exponential:**$\tau_i \sim Exp(\lambda)$**.**

Altogether:

$Pr(X \geq k) = Pr(X_0 e^{\mu\tau} \geq k) = Pr(e^{\mu\tau} \geq \frac{k}{X_0}) = Pr(\mu\tau \geq log\frac{k}{X_0}) = Pr(\tau \geq \frac{1}{\mu} log\frac{k}{X_0})$ **->** $e^{\frac{-\tau}{\mu} log\frac{k}{X_0}} = (\frac{k}{X_0})^{-\frac{\tau}{\mu}}$

or $Pr(X \geq k) \propto k^{-\alpha}$**(pareto/power law distribution)**

This helps to match the skewness of real networks, but something more is needed to explain heavy tails of real distributions.

2.2 Yule-Simon Model

Alternatively in this model, we imagine discrete-sized groups being formed in discrete time. An example would be the number of times a particular word is used in a given document as it is being written:

1. With probability ρ, pick a completely new word from the dictionary at random.
2. With probability (1-ρ), pick a word you have already used at random and copy it.

This results in the probabilty of picking a particular word type being proportional to the number of times the word is used (tokens). These word frequencies are some of the best robust power law distributions. For large k:

$Pk = O(k^{(-\alpha+1)})$ **which is a power law distribution**

2.3 Cumulative Advantage Models

The main concept behind this group of models is that the more connected a node is, the more likely it is to form connections to new nodes (example: when a paper is cited more times it is more likely to be cited in the future as well). It is a more specific form of the above model in which c steps are taken at once.

The preferential attachement version from Barabsi and Albert:

1. At each time step, add a node.
2. The new node has c edges (initially).
3. Assign edges to existing nodes at random, but with the probability of attachment to a degree k node $\propto$k (for undirected graphs).

This results in degree distributions $\sim k^{-\alpha}$for some α. But, also results in little correlation in neighborhoods.

There are some addtions to these models: non-linear preferential attachment, fitness to nodes (adds to the probability of attachment).

2.4 Vertex Copying Model

An alternative mechanism that results in networks with the same degree distribution as those formed by cumulative advantage is based on copying nodes (real world examples: genomes, errors in citations). How it works:

1. Pick node v uniformly at random.
2. Make a copy of v, and ensure that the copy has all of the same out-going edges.
3. For each edge attached to the copy of v, leave the edge in the network with probability δ, otherwise re-wire to totally random node.

This model results in networks with the same degree distribution as cumulative advantage but with the additional of correlated neighborhoods.