Lecture 8: Growing Network Models

36-720, Fall 2016
Scribe: Mo Li

26 September 2016

The statistical network models we've seen in the previous lectures are models with little "story" in the sense that vertices take one latent attribute at the beginning of time and edges are independent given attributes. Examples of the type include blocks in stochastic block models (SBM) and locations in constrained latent space models (CLSM). In today's lecture, we're going to see models where the network gets to its state because of explicit mechanisms for adding and removing nodes and edges. The inference is on the growth process rather than the observed configuration of the graph and the network characteristics depend on the growth mechanisms.

1 Multiplicative Growth and Power-Law Distribution

Many real world network models have degree distributions that are very right-skewed and heavy-tailed[1]. For example, consider the scientific citation network using databases provided by the Institute of Scientific Information (ISI). Totally around 5 million scientific papers have been recorded. Minimum citation number among all these papers is 0. The mean citation number is around 5. There are a few hundred papers with cititation numbers greater than 1000. The maximum citation number reaches around 20,000. All these observations demonstrate well the right-skewed and heavy-tailed behaviors of the real network models, which have actually been well documented as early as 1960's [2]. Usually, the familiar distributions (such as Poisson distribution) won't be such heavy-tailed and right-skewed. Thus it is natural for people to ask how such heavy-tailed distributions arise. Multiplicative growth, where continuous-size lumps grow in continuous time (Figure1), is one way.

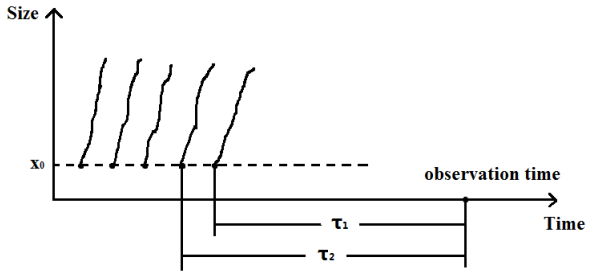

Figure 1: An illustration of multiplicative growth.

Consider a simple model where the growth rate is proportional to the size of the object. Denote $X_i(t) =$ size of object i at time t, so $\frac{dX_i(t)}{dt} = \mu X_i(t)$ for some $\mu > 0$. Initially, $X_i(0) = x_0$. Therefore, at age τ_i, we have

$$X_i = x_0 e^{\mu \tau_i} \tag{1}$$

We further assume that objects appear at uniform rate λ, which means that the appearance times are a Poisson process with intensity λ. Thus, looking backward from a given observation time, age distribution is exponential, $\tau_i \sim \text{Exp}(\lambda)$.

So far we have $X_i = x_0 e^{\mu \tau_i}$ at time of observation, and $\tau \sim \text{Exp}(\lambda)$ (its cdf is $1 - e^{-\lambda \tau}$). Then

$$\begin{aligned}
P(X \geq k) &= P(x_0 e^{\mu \tau} \geq k) && (2)\\
&= P(e^{\mu \tau} \geq \frac{k}{x_0}) && \text{(note that } x_0 > 0) \quad (3)\\
&= P(\mu \tau \geq log \frac{k}{x_0}) && (4)\\
&= P(\tau \geq \frac{1}{\mu} log \frac{k}{x_0}) && (\mu > 0) \quad (5)\\
&= 1 - P(\tau \leq \frac{1}{\mu} log \frac{k}{x_0}) && (6)\\
&= 1 - (1 - e^{-\frac{\lambda}{\mu} log \frac{k}{x_0}}) && \text{(plug in the cdf of } \tau) \quad (7)\\
&= e^{-\frac{\lambda}{\mu} log \frac{k}{x_0}} && (8)\\
&= (\frac{k}{x_0})^{-\frac{\lambda}{\mu}} && (9)
\end{aligned}$$

Therefore, X has a power law distribution or Pareto distribution. (Note that continuous power law distribution is also called Pareto distribution.) If we plot $P(X \geq k)$ against k, we get a distribution curve with long tail to the right. The plot of $\log P(X \geq k)$ against $\log k$ is a straight line with slope equals to $-\lambda/\mu$ (Figure2).

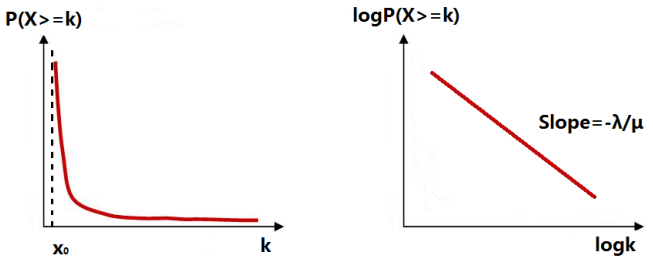

Figure 2: Plot of $P(X \geq k)$ against k on normal scale and log-log scale.

2 Yule-Simon Distribution

Yule-Simon distribution is a discrete distribution named after Udny Yule, a British statistician and Herbert A. Simon. It arose originally from the study of distribution of biological taxa and subtaxa[3]. Simon, an American political scientist, economist, sociologist, psychologist and computer scientist,

conducted research over a broad range of fields and was a Nobel Prize laureate because of his work in economics in 1978.

In above section we've considered a simplified "toy" case where continuous size lumps grow in continuous time, leading to power law distribution. In the case of Yule-Simon distribution, the variables have discrete sizes and grow in discrete time. Consider the problem of computing the number of uses of particular words in a given document as it is written.

Assume that with probability ρ, pick a completely new word from the dictionary at random. And also assume that with probability of $1-\rho$, pick a word you have already used at random from the text and copy it. Thus, in this case, the probability of picking a particular word type $\propto$ number of tokens of that words, i.e. the number of times that word has been used. For example, in the sentence "the cat sat on the mat", the word type "the" has two tokens.

Denote $N_k(t)$ = number of words used exactly k times in the first t words. Going from t to $t+1$, we either add a totally new word with probability ρ or copy an existing word with probability $1-\rho$, and this will make $N_k(t)$ change by either 1 or -1 or remain the same (in the case of adding a totally new word from the dictionary). Thus we can calculate that

$$P(N_k(t+1) = N_k(t)+1) = \frac{(1-\rho)N_{k-1}(t)(k-1)}{t} \qquad \text{for } k \geq 2 \tag{10}$$

$$P(N_k(t+1) = N_k(t)-1) = \frac{(1-\rho)N_k(t)k}{t} \qquad \text{for } k \geq 2 \tag{11}$$

Therefore,

$$E(\Delta N_k(t)|N_k(t), N_{k-1}(t)) = E(N_k(t+1) - N_k(t)|N_k(t), N_{k-1}(t)) \tag{12}$$

$$= \frac{(1-\rho)N_{k-1}(t)(k-1)}{t} - \frac{(1-\rho)N_k(t)k}{t} \tag{13}$$

$$= \frac{1-\rho}{t}(N_{k-1}(t)(k-1) - N_k(t)k) \tag{14}$$

Now claim that $N_k(t) \to tp_k$ for some time-invariant p_k as $t \to \infty$. Then if we substitute in above equality, we get

$$(t+1)p_k - tp_k = \frac{1-\rho}{t}(tp_{k-1}(k-1) - tp_k k) \tag{15}$$

$$\Leftrightarrow p_k = (1-\rho)(p_{k-1}(k-1) - p_k k) \tag{16}$$

$$\Leftrightarrow p_k(1+k(1-\rho)) = (1-\rho)(k-1)p_{k-1} \tag{17}$$

Therefore, we have

$$p_k = \frac{(k-1)(1-\rho)}{1+(1-\rho)k}p_{k-1} \tag{18}$$

Define $\alpha = \frac{1}{1-\rho}$, then we get

$$p_k = \frac{(k-1)/\alpha}{1+k/\alpha} p_{k-1} \tag{19}$$

$$= \frac{k-1}{k+\alpha} p_{k-1} \qquad \text{notice the nice recursive form in } p_k \tag{20}$$

$$= \frac{k-1}{k+\alpha} \cdot \frac{k-2}{k+\alpha-1} \cdots p_1 \tag{21}$$

$$= \frac{\Gamma(k)\Gamma(\alpha+1)}{\Gamma(k+\alpha+1)} p_1 \qquad \text{for } k \geq 2 \tag{22}$$

p_1 can be fixed either by normalization or by similar argument to above. As $k \to \infty$, according to Stirling's approximation for $log(\Gamma(x))$, we have

$$log(\Gamma(x)) = log(\sqrt{2\pi}) - x + (x - \frac{1}{2}) logx \tag{23}$$

Therefore, for large k

$$p_k = O(k^{-(\alpha+1)}) \qquad \text{where } \alpha = \frac{1}{1-\rho} \tag{24}$$

3 Cumulative Advantage Networks and Preferential Attachment

Cumulative advantage processes was first studies intensively by Derek John de Solla Price, who was a British physicist, historian of science and information scientist, credited as the father of scientometrics. He conducted quantitative studies of the networks of citations between scientific papers[2]. In the model he proposed, each paper cites on average c other papers. For each citation, with probability ρ it goes to a totally random paper, and with probability $1-\rho$ it goes to a specific paper with probability $\propto$ its in-degree. He found out that the in-degree and also out-degree of a citation network end up being proportional to $\frac{\Gamma(k)}{\Gamma(k+\alpha+1}$ for large k, which have power-law distribution. Later on in 1976, Price proposed a general mathematical theory of cumulative advantage processes in directed graph[4].

In 1999, Barabási and Albert, who claimed that they did not know Price, introduced a model similar to Price's model which is called preferential attachment nowadays[5]. The graphs they studied are undirected graphs. In the model, at each time step, a node is added to the network. The new node has exactly, at least initially c edges. The edges are assigned to existing nodes, and the probability of attaching to a node with degree k is proportional to k. In other words, the networks expand continuously by adding new nodes, and the new nodes attach preferentially to sites that are already well connected. They found out that $p_k \propto k^{-3}$ for large k. Therefore, it can actually be turned into a special case of Price model (with $\alpha = 2$).

To summary, cumulative advantage process/preferential attachment has degree distribution $\approx k^{-\alpha}$ for large k where α is some positive value. A key feature is that highly linked nodes attract more nodes, otherwise there's little correlation in neighborhoods. Of course, there are a variety of flavors of cumulative advantage process/preferential attachment, such as non-linear preferential attachment[6], fitness model where each node is assigned a fitness parameter capturing an intrinstic ability of compete for edges at the expense of other nodes[7], etc.

4 Duplication or Copying Model

Vertex copying mechanisms are originally motivated by the fact that some web page authors will note an interesting but novel commonality between certain pages and will link pages exhibiting this commonality and the also fact that most authors will be interested in certain already represented topics and will collect pages together to link to pages about these topics.

Consider the simple case where the nodes are never deleted.

- Pick a vertex v uniformly at random.
- Make a copy of v and all its out-going edges.
- For each edge, with probability δ to leave it alone or with probability $1 - \delta$ to re-wine it to a totally random node.

With such assumptions, the degree distribution turns out to follow the power-law distribution, which is the same as the cumulative advantage process. Examples of such type of networks include both directed and undirected networks such as the protein interaction networks, biological networks, citation networks, etc. As a side note, copying model was actually first proposed for a biological networks of genes.

References

[1] William J. Reed and Barry D. Hughes (2002), "From Gene Families and Genera to Incomes and Internet File Sizes: Why Power Laws are so Common in Nature", *Physical Review E* **66**: 067103.

[2] Derek J. de Solla Price (1965), "Networks of Scientific Papers", *Science* **149**: 510–515.

[3] G. Udny Yule (1925), "A Mathematical Theory of Evolution, based on the Conclusions of Dr. J. C. Willis, F.R.S", *Philosophical Transactions of the Royal Society B* **213**:21-87.

[4] D.J. de Solla Price (1976), "A General Theory of Bibliometric and Other Cumulative Advantage Processes",*Journal of the American Society for Information Science* **27**:292-306.

[5] A. Barabási and R. Albert (1999), "Emergence of Scaling in Random Networks", *Science* **286**:509-512.

[6] A. Barabási "Network Science" Ch. 5

[7] R. Albert and A. Barabási (2002), "Statistical Mechanics of Networks", *Reviews of Modern Physics* **74**:47-97.