

Math 262 Lectures Notes

The Vertex-Duplication Model

1 Motivations

When new users start up a website, they generally create their pages based at least in part on existing pages. The significance of this in terms of modeling internet graphs is that new nodes will tend not to be born with random edges, but with edges identical in part or fully with those of an already-existent node.

This turns out to also be a useful technique in computational biology. The *gene regulatory network* is a system in which each vertex represents a gene, and the edges represent regulatory relations; similarly, the *protein interaction network* has vertices representing proteins and edges representing interactions. In each case, a new vertex is likely to be a modification of an old one, rather than a completely new entity, so these systems are well modeled by vertex-duplication as well.

2 The model

The *partial-duplication* model builds a graph through the following steps, based on a duplication-probability p :

- At time t_0 , we start with an initial graph G_{t_0} on t_0 vertices.
- At each time t , we do the following:
 - Select a *sample vertex* v uniformly.
 - Add a new vertex u .
 - For each neighbor w of v , add the edge uw with probability p .
 - With probability p , add the edge uv .

3 Properties of the model

The most straightforward question to ask is: at time t , how many edges and vertices are there? We expect our model to be similar to other power-law graphs, and knowing the relationship between

the numbers of vertices and edges may help to determine which power-law graphs it resembles. The number of vertices at time t follows easily from our algorithm: at time t we have t vertices. The expected number of edges can be derived via a recurrence relation. Let τ_t be the number of edges at time t . Then, since at time $t + 1$, we duplicate some vertex v randomly, we may express the expected value of τ_{t+1} as such:

$$\begin{aligned}
E(\tau_{t+1}) &= E(\tau_t) + \sum_{v \in G_t} \frac{p}{t} E(d_{G_t}(v) + 1) \\
&= E(\tau_t) + \frac{p}{t} E\left(\sum_{v \in G_t} d_{G_t}(v)\right) + \sum_{v \in G_t} \frac{p}{t} \\
&= E(\tau_t) + \frac{p}{t} E(2\tau_t) + p \\
&= \left(1 + \frac{2p}{t}\right) E(\tau_t) + p
\end{aligned}$$

If we assume $p < \frac{1}{2}$, we can conjecture that

$$E(\tau_t) = \tau t + o(t),$$

that is, that $E(\tau_t)$ is approximately linear with respect to t . The validity of this conjecture may be ascertained by plugging it in to the recurrence relation above:

$$\begin{aligned}
\tau(t+1) + o(t) &= (\tau t + o(t)) \left(1 + \frac{2p}{t}\right) + p \\
\tau(t+1) &\approx \tau t \left(1 + \frac{2p}{t}\right) + p \\
\tau t + \tau &\approx \tau t + 2p\tau + p \\
(1 - 2p)\tau &\approx p \\
\tau &\approx \frac{p}{1 - 2p}
\end{aligned}$$

Since we now have a good idea of what τ should be, we must confirm that the error term is in fact $o(t)$. Let $f(t) = \tau t - \frac{p}{1-2p}t$. Then $f(t+1) = f(t) \left(1 + \frac{2p}{t}\right)$ by our previous recurrence, so expanding out this relationship we find that

$$\begin{aligned}
f(t) &= f(t_0) \prod_{j=t_0}^{t-1} \left(1 + \frac{2p}{j}\right) \\
&\approx f(t_0) \prod_{j=t_0}^{t-1} e^{\frac{2p}{j}} \\
&\approx f(t_0) e^{\sum_{j=t_0}^{t-1} \frac{2p}{j}} \approx f(t_0) e^{2p \ln t} \approx f(t_0) t^{2p} = O(t^{2p}) = o(t) \text{ for } p < \frac{1}{2}
\end{aligned}$$

We thus know that $E(\tau_t) = \frac{p}{1-2p}t + o(t)$, but not the distribution of τ_t about this mean. To determine the distribution, we must apply a variant of Azuma's inequality.

Definition 1. For a martingale $X = (X_1, X_2, \dots, X_n)$ with decision tree T , and a set of coefficients c_1, \dots, c_n , an edge between the k th and $(k+1)$ th level of T is bad if and only if $|X_{k+1} - X_k| > c_k$; that is, if it's an edge which prevents c_k from being a Lipschitz coefficient.

A leaf of the decision tree is bad iff the path to it from the root is incident to a bad edge.

Theorem 1 (The Generalized Martingale Inequality). For a martingale $X = (X_1, X_2, \dots, X_n)$ with decision tree T , and a set of coefficients c_1, \dots, c_n ,

$$\Pr(|X - E(X)| > a) \leq 2e^{-\frac{a^2}{\sum_{i=1}^n c_i^2}} + \Pr(B)$$

where B is the set of bad leaves in T .

Using this theorem, we consider the martingale generated by this process for the random variable τ_t . Letting our choice of coefficient c_i be such that $\Pr(B) = o(1)$, we get that

$$\Pr(|\tau_t - E(\tau_t)| > a) \leq 2e^{-\frac{a^2}{\sum_{i=1}^n c_i^2}} + o(1)$$

which we may constrain with, for instance, letting $a = \sqrt{\sum_{i=1}^n c_i^2}$, so that

$$\Pr\left(|\tau_t - E(\tau_t)| > \sqrt{\sum_{i=1}^n c_i^2}\right) \leq \frac{2}{e} + o(1)$$

which indicates a strong, if not necessarily narrowly distributed, central tendency.

References

- [1] Fan Chung, Linyuan Lu, Gregory Dewey, and David J. Galas. "Duplication models for biological networks", *Journal of Computational Biology*, **10** (2003), no. 5, pp. 677-688.