An Introduction to "Growing Protean Graphs"

Andy Wilson

February 7, 2012

This paper consists of a short introduction to the recent paper "Growing Protean Graphs" by Pawel Prałat and Nicholas Wormald ([4]). In this paper, Prałat and Wormald introduce a random graph model that gains vertices over time[1] and emulates the web graph, the real-world graph in which vertices are websites and edges are hyperlinks. The key features that the authors wish to emulate are the *power law* on vertex degrees and the unique *giant component*. The first property means that the number of vertices of degree d is approximately $d^{-\gamma}$ for some positive constant γ, and the second means that over half of the vertices of the graph are in the same connected component. Many classical graphs, such as $\mathcal{G}_{n,p}$ do not have these properties, but there are other recent models that do attain these properties, such as preferential attachment, which is analyzed by Chung and Lu in [2]. Protean graphs are unique in that "older" vertices are preferred when edges are added. We will not prove that protean graphs have these properties in this paper, but [4] and [3] contain proofs of these results.

Instead, in this paper we will present the algorithm on how to grow a protean graph and then use the differential equations method to analyze two key properties of the vertices of a protean graph. A *protean graph* is a sequence of undirected graphs G_t created by the following process:

- Choose $p \in (0.5, 1]$, $d \in \mathbb{N}$, $\eta \in (0, 1)$.
- Let G_0 be the graph with one vertex, v_1.
- Given G_t, randomly add a vertex (with probability p) or remove a vertex (with probability $1 - p$.)
  - If adding a vertex, label the vertex v_i, where i is the smallest integer not yet used to label a vertex. Then randomly add d edges from v_i to existing vertices, where v_j gets an edge with probability $j^{-\eta}/\left(\sum_{v_k} k^{-\eta}\right)$. (Technically, these edges are added randomly one at a time.)
  - If removing a vertex, choose an existing vertex uniformly at random. Assume we choose v_l. Delete v_l and all its edges. Then decrement the label on each vertex with label larger than l by one.

[1] In [3], the first author introduced a similar model that had a fixed number of vertices.

We will alter the above process slightly in order to do our analysis. Instead of starting with the graph on a single vertex, we will now begin with G_{t_0}, a graph on n vertices. Let v be the vertex with the largest label in G_{t_0}, i.e. the "newest" vertex in G_{t_0}. We essentially follow the same process as above, except we never remove v. We will consider the random variables

$$\begin{aligned} N_t &= \text{number of vertices of } G_t - 1 \\ J_t &= \text{number of vertices older than } v \text{ in } G_t \end{aligned}$$

Also let $t_f = t_0 + \lfloor cn^{2p}/\log^3 n \rfloor$.

Theorem 1. *With probability at least* $1 - exp(-\theta(log^2 n)$*, we have*

$$\begin{aligned} N_t &= n + (2p-1)(t-t_0) + O(n^p logn) \\ J_t &= n\left(\frac{N_t}{n}\right)^{\frac{p-1}{2p-1}}(1 + O(log^{-1/2} n)) \end{aligned}$$

for all $t \in [t_0, t_f]$.

The proof will employ the differential equations method, a method used to obtain concentrations for random variables of a discrete random process. For more information, see work of Wormald ([5]) and Bohman ([1]). Essentially, the idea is to notice that

$$\begin{aligned} E(N_{t+1} - N_t | G_t) &= 2p - 1 \\ E(J_{t+1} - J_t | G_t) &= -(1-p)\frac{J_t}{N_t} \end{aligned}$$

and then define deterministic functions that have the above as their derivatives, i.e.

$$\begin{aligned} z'(x) &= 2p - 1 \\ y'(x) &= -(1-p)\frac{y}{z} \end{aligned}$$

with initial conditions $z(t_0/n) = y(t_0/n) = 1$. (We have scaled the system by a factor of n here.) We can then obtain the following solution to the differential equation involving z and y.

$$\begin{aligned} (2p-1)\log y + (1-p)\log z &= C_1 \\ z - (2p-1)x &= C_2 \end{aligned}$$

We will then use the following lemma to show that N_t and J_t are concentrated around the "expected" solutions $z(t/n)n$ and $y(t/n)n$, respectively.

Lemma 1. *Let $G_0, \ldots, G_t$ be a random process and X_i a random variable determined by $G_0, \ldots, G_i$. If there are $b, c_i \in \mathbb{R}$ such that*

$$\begin{aligned} E(X_i - X_{i-1}|G_0, \ldots, G_{i-1}) &< b \\ |X_i - X_{i-1} - b| &\leq c_i \end{aligned}$$

then

$$P(\exists i \text{ such that } X_i - X_0 \geq ib + \alpha) \leq exp\left(-\frac{\alpha^2}{2\sum_{j=0}^{i} c_j^2}\right)$$

See [5] for a proof. Now we will prove (half of) the main theorem - in particular, we will show concentration for N_t.

Proof. Define $F(N_t) = N_t - (2p-1)t$. We calculate

$$\begin{aligned} F(N_{t+1}) - F(N_t) &= N_{t+1} - N_t - (2p-1) \\ E(F(N_{t+1}) - F(N_t)) &= E(N_{t+1} - N_t) - (2p-1) \\ &= 0 \\ |F(N_{t+1}) - F(N_t)| &\leq |N_{t+1} - N_t| + (2p-1) \\ &\leq 1 + 2p - 1 \\ &\leq 2p \\ &\leq 2 \end{aligned}$$

because N_t can change by at most 1 in each step. Now we apply the lemma with $b = 0$, $c_i = 2$, and $\alpha = n^p \log n$ to both $F(N_t)$ and $-F(N_t)$ to get

$$|N_t - N_{t_0} - (2p-1)(t - t_0)| \leq n^p \log n$$

with probability greater than

$$\begin{aligned} 1 - \exp\left(-\frac{n^{2p}\log^2 n}{4(t_f - t_0)}\right) &\geq 1 - \exp\left(-\theta(\log^5 n)\right) \\ &\geq 1 - \exp\left(-\theta(\log^2 n)\right) \end{aligned}$$

by the definition of t_f. Since $N_{t_0} = n - 1$, this gives us the concentration of N_t that we desired.

The analysis for the other half of the proof is slightly more difficult. It requires an additional parameter, a stopping time T. T is defined so that the desired concentration occurs for all t between t_0 and T. Then Prałat and Wormald show that, with high probability, $T = t_f$. For more details, see [4].

□

References

[1] Tom Bohman, *The triangle-free process*, Advances in Mathematics **221** (2009), 1653–1677.

[2] Fan Chung and Linyuan Lu, *Coupling online and offline analyses for random power law graphs*, Internet Mathematics **4** (2004), 409–461.

[3] Pawel Prałat, *Protean graphs*, Internet Mathematics **3** (2006), 21–40.

[4] Pawel Prałat and Nicholas Wormald, *Growing protean graphs*, Internet Mathematics **4** (2007), 1–16.

[5] Nicholas Wormald, *The differential equation method for random graph processes and greedy algorithms*, Lectures on Approximation and Randomized Algorithms (1999), 73–155.