PageRank and Randomly Grown Graphs

Gokul G. Nair
Center for Applied Mathematics
Cornell University

Project Report: Top Ten Algorithms of the 20^{th} Century

Abstract

In the field of network science, there are several methods of randomly generating graphs with certain desirable properties. Google's aptly named PageRank algorithm is widely used to rank the vertices of a graph based on their connectivity to other vertices. In this report, we utilize PageRank to study the structures of graphs that are grown using different methods. In particular, we focus on graphs that are grown according to two models— with and without 'preferential attachment'. In the former model due to Barabasi and Albert, we notice that for small values of a parameter, the PageRank is smooth, but when the parameter is large, a discontinuity appears, which appears to be related to node-clustering in the graph. We estimate the size of this discontinuity using a simple approach. In the latter model (proposed by Callaway et al.), no such discontinuity arises in the PageRank.

1 Introduction: Random Graphs

1.1 Random Graphs

To say that networks are ubiquitous in nature would be an understatement. From social dynamics to the interactions between fundamental particles, the study of graph structures gives tremendous insight into the principles governing the world around us [1].

Random graphs have proven to be an interesting paradigm for researchers in a variety of disciplines. By the term 'random graph', we almost exclusively refer to the Erdös-Rényi graph, $G(n, p)$, which is generated by fixing n vertices and adding each possible edge with probability p [2]. It suffices to say that a great deal about the behaviour of such random graphs is known.

An important definition and result (without proof) for the purpose of this report are:

Definition 1 *The degree distribution, $P(k)$ of a graph G is the fraction of vertices that have degree k.*

Proposition 1 *The degree distribution of a $G(n, p)$ graph is Poisson with parameter np, for $n \to \infty$ [3].*

2 Randomly Grown Graphs

Although random graphs as defined in the previous section show many interesting phenomena, their properties often do not match those of networks in nature. In particular, observations suggest that many real-world networks have degree distributions that are power-laws rather than Poisson [4]. It was conjectured that the reason for this was that in reality, most networks grew from smaller ones by sequentially adding vertices. The sequence of adding vertices is often a random process and hence these graphs are referred to as *randomly grown graphs*. In this report we will study two such models and probe their structure using PageRank.

2.1 Barabási-Albert Model

The Barabási-Albert (BA) growth model is an archetypal example of *preferential attachment*— a phenomenon wherein vertices with a higher degree of connectivity have a higher tendency to form more links than vertices with lower degrees [1]. In colloquial speech this is referred to as the 'rich get richer' situation. Below is a more precise description of the growth process of the BA-model with parameters n and m:

```
Initialize graph with m disconnected vertices
while graph.size < n:
    Add new vertex
    for i from 0 to m:
        Pick an existing vertex-j with probability Pj
        Add edge from new vertex to vertex-j
```

Here,

$$p_j = \frac{d(j)}{\sum_{k \in V} d(k)}$$

where $d(k)$ is the degree of vertex k (and V is the vertex set). It is this form of p_j that manifests as preferential attachment.

The degree distribution of the BA-model has been shown to follow a power law asymptotically, with $P(k) \sim k^{-3}$ [1].

2.2 CHKNS Model

Callaway, Hopcroft, Kleinberg, Newman and Strogatz (CHKNS) proposed a graph growth model that does not show preferential attachment [5]. The model shows interesting critical phenomena [6] and is conjectured to be in the same universality family as the Kosterlitz-Thouless transition, showing an infinite-order phase transition in the size of its largest component [7].
In the original model, at every step a new vertex was added and a randomly chosen pair (not necessarily a new one) was connected by an edge with probability δ. We slightly modify this model to allow for m new edges to be added every step and also set $\delta = 1$.

```
Initialize graph with m disconnected vertices
while graph.size < n:
    Add new vertex
```

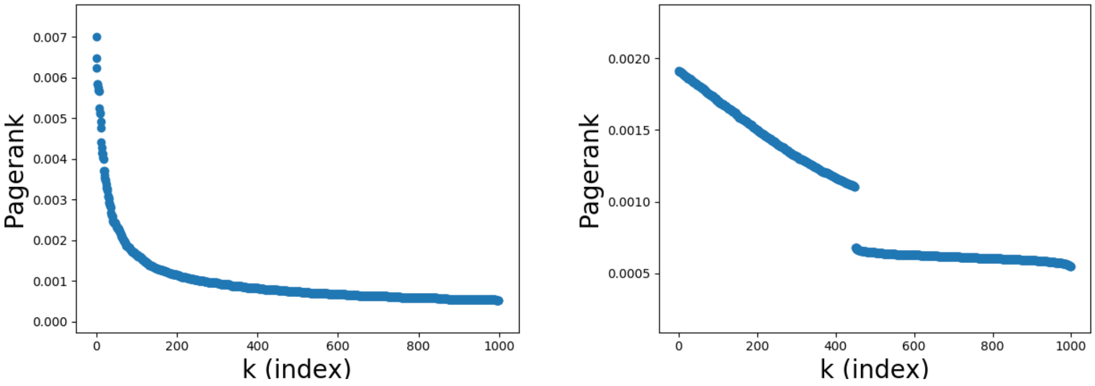
```
for i from 0 to m:
    Pick two vertices uniformly at random
    Add edge between chosen vertices
```

Note that this model doesn't show preferential attachment, since at every step, each existing vertex has the same chance of being linked. However, nodes that started out earlier in the graph have a higher chance of forming more links than new ones, which is in contrast to Erdös-Rényi graphs (which do not have a temporal aspect).

3 PageRank on BA and CHKNS graphs

We calculated the PageRank of graphs grown using the models presented in the previous section. For the purpose of this report, the only two parameters are n, the final size of the graph and m, the number of edges added at each step (also equal to the initial number of vertices). We first present the results of some numerical experiments and a simple theoretical argument to explain them.

When $m << n$, the PageRank of a BA graph is smooth and obeys a power-law decay (Figure 1a), however, when m is comparable to n, it has a discontinuity, with m vertices in the lower branch and $n - m$ in the upper.



(a) $n = 1000$, $m = 10$. The distribution is a power-law with exponent~ 0.5).

(b) $n = 1000$, $m = 550$. When m is comparable to n, the PageRank is discontinuous.

This discontinuity can be made sense of by drawing a force directed diagram of the graph (Figure 2b)— a useful way of drawing graphs wherein each edge has a spring-force associated with it. Approximately $n - m$ vertices are well connected to each other and to the rest of the graph and form a central cluster. The other m vertices behave as 'satellites' linking to the central cluster but not interacting among themselves. In order to estimate the gap in the PageRank, we digress to give an overview of random walks on graphs.

3.1 Random walks on Graphs

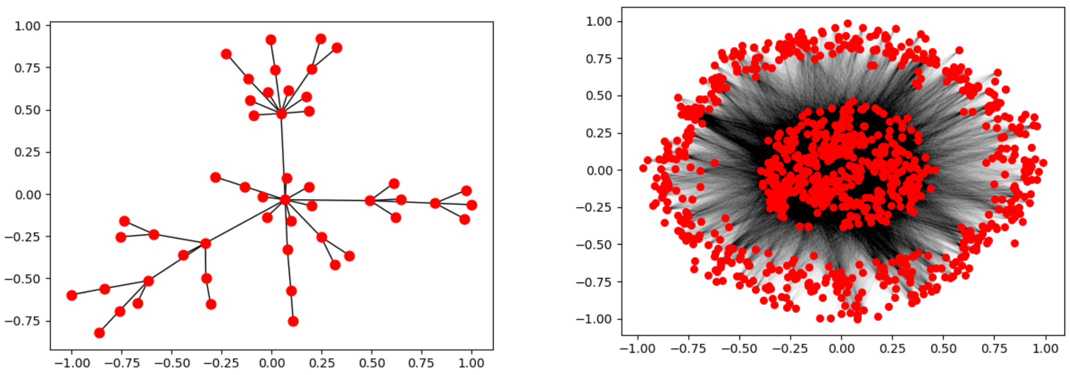
Consider a spider crawling on a graph such that when the spider gets to a particular node he/she picks an edge uniformly at random and follows it. The resulting Markov chain (provided it satisfies certain properties) possesses a stationary distribution π. If we denote the transition probability from node i to node j as p_{ij}, then the detailed balance equation is given by:

$$\pi(i)p_{ij} = \pi(j)p_{ji}.$$

Using the detailed balance equation and the normalization condition for probability distributions, one can derive an expression for π of a connected undirected graph [8]:

$$\pi(i) = \frac{d(i)}{2|E|}. \tag{1}$$

Here $d(i)$ is the degree of node i and $|E|$ is the cardinality of the edge set. In sufficiently dense graphs, the stationary distribution is close to the PageRank, so it is instructive to study π in order to understand the behaviour of the PageRank.



(a) A small BA graph, $n = 50$, $m = 1$.

(b) BA graph (Force directed diagram), $n = 1000$, $m = 550$. There are approximately m 'satellite' nodes.

3.2 Estimating the BA gap

In order to estimate the size of the gap in the PageRank of a BA graph, we explicitly construct a graph with two clusters. This graph approximately behaves like the BA graph for large m, but is more tractable analytically. The approximate BA model is constructed in the following way:

1. Start with m vertices (refer to these as initial vertices).
2. Add a new vertex.
3. Begin adding edges from the new vertex to **non**-initial vertices, in order of degree.
4. If all the **non**-initial vertices are exhausted, choose initial vertices uniformly at random and link to them until a total of m edges have been added.

5. repeat steps 2, 3 and 4 until the graph has n nodes in total.

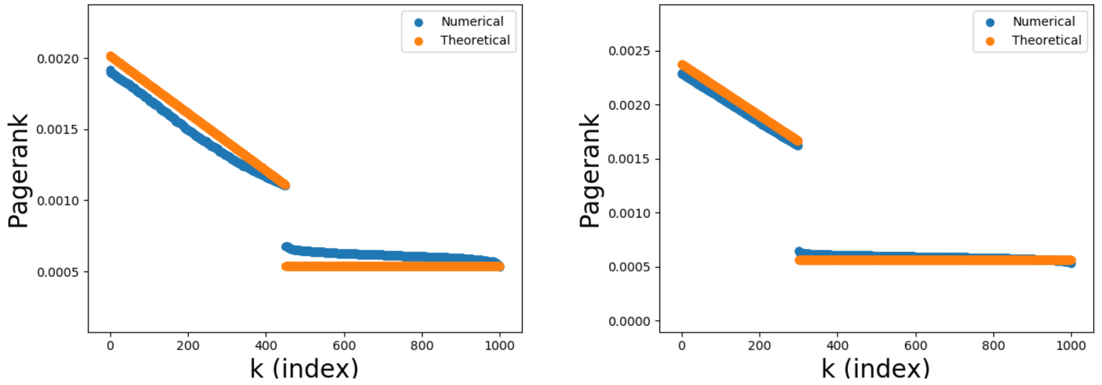
It is relatively direct to compute the expected degree of each vertex in this model and using Eq (1), the stationary distribution is given by

$$\pi(i) = \begin{cases} \frac{n-i}{2m(n-m)} & i \geq n-m \\ \frac{(n-m)(3m-n+1)}{4m} & i < n-m, \end{cases} \tag{2}$$

and the approximate gap is given by

$$\Delta = \frac{1}{4}\left(\frac{2}{n-m} + \frac{(n-1)n}{m} + 3m - 4n + 1\right). \tag{3}$$

Figure 3 shows that the approximation is very good for large values of m.



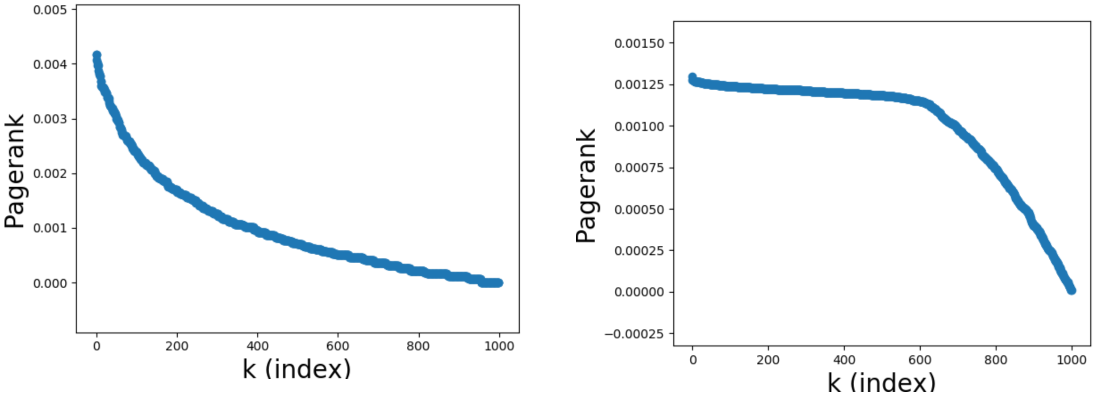
(a) $n = 1000$, $m = 550$

(b) $n = 1000$, $m = 700$

Figure 3: A comparison between the numerical and theoretical results for a BA graph.

3.3 Approximating the CHKNS PageRank

The PageRank of a CHKNS graph when $m << n$ is approximately exponential. However, when m is increased, the distribution changes drastically, having a shoulder-like structure— with a relatively flat portion of high PageRank and a smoothly decreasing tail (Figure 4). The force-directed diagram (Figure 5b) shows that there is a highly clustered subset of nodes, which corresponds to the flat portion of the shoulder. However, unlike BA, there is no sharp boundary between the cluster and satellite nodes, hence the shoulder is continuous.



(a) $n = 1000$, $m = 10$.The distri- -bution is exponential.

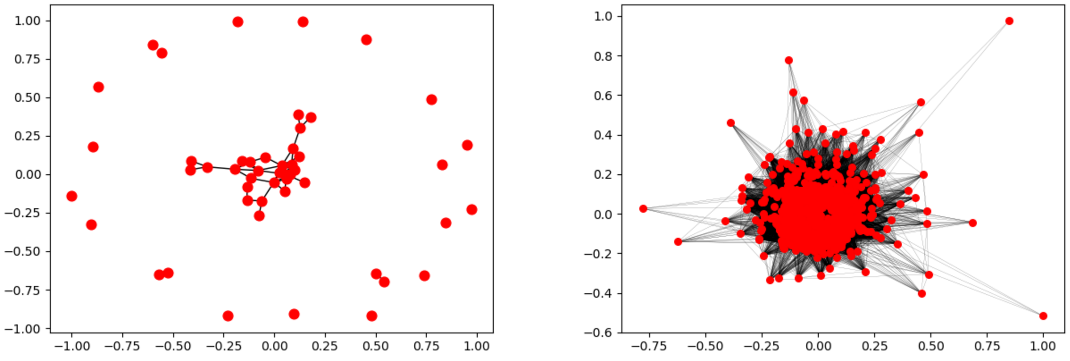
(b) $n = 1000$ and $m = 550$. (CHicKeNS' shoulder)

Figure 4: PageRanks for CHKNS graphs for different values of m.

Callaway et al. estimated the degree distribution for their model in the case where a single edge is added every step. They derived a rate equation for $d_k(t)$, the number of nodes of degree k at time t. In the more general case where m edges are added every step, one needs to account for the multitude of events that could result in a change in $d_k(t)$. However, in the case where m is small relative to t, one can write the following equations:

$$d_0(t+1) = d_0(t) + 1 - 2m\frac{d_0(t)}{t} + o\left(\frac{1}{t^2}\right) \tag{4}$$

$$d_k(t+1) = d_k(t) + 2m\frac{d_{k-1}(t)}{t} - 2m\frac{d_k(t)}{t} + o\left(\frac{1}{t^2}\right) \tag{5}$$



(a) A small CHKNS graph, $n = 50$, $m = 1$.

(b) CHKNS graph for $n = 1000$ and $m = 550$.

The above equations are derived by considering the various events that can occur in a given step. Eq (5) says that the number of nodes with degree k in step $t+1$ is equal to the number of nodes with degree k in the previous step plus some additional terms accounting for loss/gain. The expected number of $k-1$ degree nodes that received a new edge and consequently became degree k nodes is given by the second term while the term that is subtracted accounts for degree k nodes that are lost by forming new edges. We do not include terms that are higher order in t^{-1}.
Eq (4) is similar to the Eq (5) except that there are no nodes of lower degree than 0, so there is no process that increases d_0 by adding edges. However, at every step we add a new node, therefore Eq (4) has the $+1$ term.

The above equations can be solved with the ansatz $d_k(t) = p_k \cdot t$ to obtain the following formula:

$$d_k(t) = \frac{(2m)^k t}{(1+2m)^{k+1}}, \tag{6}$$

which is an excellent approximation for small m (Figure 6). Notice that $d_k(t)$ is exponential in k (for fixed t).

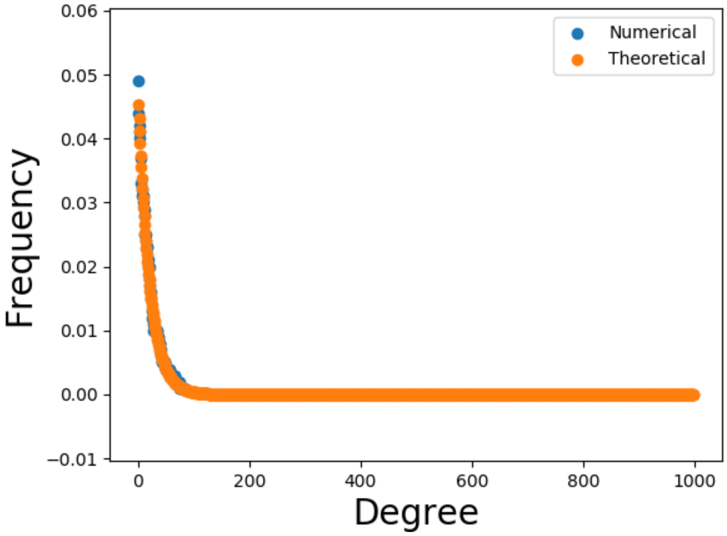


Figure 6: Degree distribution for CHKNS graph.

3.4 Limitations and Further directions

It goes without saying that the analyses shown in this report has several limitations. Firstly, although we have estimated the size of the gap in the BA PageRank, we have not addressed the issue of when the gap occurs. The model we proposed only predicts the existence of the gap for m comparable to n, but does not predict its non-existence for smaller m.

Another glaring issue is that in our analysis of the CHKNS model, we assume m is small compared to n. In the case where m is comparable to n, the rate equations would need to include several high order terms. As an example, the event where j nodes of degree k are lost by the addition of m edges would look like $j\binom{d_k(t)}{j}/\binom{t}{j}$. An interesting possible direction is to include a substantial number of such terms and either numerically or theoretically calculate the resulting distribution. This may explain the appearance and structure of the shoulder distribution.

References

[1] Mark EJ Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

[2] John Hopcroft. Computer science 485 lecture notes, cornell university, January 2006.

[3] Albert-László Barabási et al. *Network science*. Cambridge university press, 2016.

[4] Erzsébet Ravasz, Anna Lisa Somera, Dale A Mongru, Zoltán N Oltvai, and A-L Barabási. Hierarchical organization of modularity in metabolic networks. *science*, 297(5586):1551–1555, 2002.

[5] Duncan S Callaway, John E Hopcroft, Jon M Kleinberg, Mark EJ Newman, and Steven H Strogatz. Are randomly grown graphs really random? *Physical Review E*, 64(4):041902, 2001.

[6] Rick Durrett. Rigorous result for the chkns random graph model. In *DRW*, pages 95–104, 2003.

[7] Archishman Raju, Colin B Clement, Lorien X Hayden, Jaron P Kent-Dobias, Danilo B Liarte, D Zeb Rocklin, and James P Sethna. Normal form for renormalization groups. *Physical Review X*, 9(2):021014, 2019.

[8] Josep Fábrega. Random walks on graphs, upc barcelona, June 2011.