Power-law Degree Distributions

1 Complex Networks - A Brief Overview

Complex networks occur in many social, technological and scientific settings. Examples of complex networks include World Wide Web, Internet, movie actor collaboration network, science collaboration network, cellular networks, protein folding.

There are some measures and properties that we can focus on related to complex networks:

Small worlds: Even though the complex networks are often large, in most networks there is a relatively short path between any two nodes. One of the most popular manifestations of small worlds is the "six degrees of separation" which concludes that, in a social network, there is a path of acquaintances with a length of about six between most pairs of people . Especially the existence of nodes with very high degrees contributes to this fact (Ex: Degrees of Kevin Bacon in the actor collaboration network, Erdős in science collaboration network).

Clustering: A common property of social networks is that cliques form, representing circles of friends in which every member knows every other member. This property is quantified by the clustering coefficient which is defined as follows: For a node i that has k_i edges incident on it, if all its neighbor nodes were part of a clique, there would be $k_i(k_i-1)/2$ edges between them. The ratio between the actual number of edges E_i and the total number gives the clustering coefficient of node i:

$$C_i = \frac{2E_i}{k_i(k_i-1)} \tag{1}$$

The clustering coefficient of the whole graph is the average of individual C_i's.

Degree Distributions: This spread in the node degrees is characterized by a distribution function $P(k)$, which gives the probability that a randomly selected node has exactly k edges. Degree distributions are the main topic of this lecture.

1.1 Random Graphs

Random graphs were originally introduced by Paul Erdős and Alfréd Rényi in the 50s. In their classic first article, they define a random graph with n nodes connected by m edges, which are chosen randomly from the $n(n-1)/2$ possible edges. That means there are $\binom{n(n-1)/2}{m}$ graphs with n nodes and m edges. This forms a probability space where each graph is equiprobable.

An alternative definition of random graphs is the binomial model. In this case, we start with n nodes and every pair of nodes is connected with probability p. Then the expected total number of edges in the graph is $E(m) = p[n(n-1)/2]$. If, say, a graph G' has n nodes and m edges, the probability of obtaining this particular graph by this construction process is $P(G') = p^m(1-p)^{n(n-1)/2-m}$.

Mathematicians are interested in the properties of random graphs as $n \to \infty$. One of the important results is that the degree distribution of random graphs can be approximated by Poisson distribution:

$$P(k) = \binom{n-1}{k} p^k (1-p)^{n-1-k} \simeq e^{-pn}\frac{(pn)^k}{k!} = e^{-z}\frac{(z)^k}{k!} \tag{2}$$

where z is the average degree of a node and $p = z/(n-1)$. For n large, $z = pn$.

2 Power Law

It was initially thought that the Internet is an Erdős - Rényi random graph. However, it is later observed that the degrees have a power law distribution, that is, the probability that a degree is larger than D is about $cD^{-\beta}$ for some c and $\beta > 0$. For the Internet graph, it is observed that β is between 2.15 and 2.48 for various years. The same distribution is observed for some other Internet-related quantities such as the number of hops per message and largest eigenvalues of the Internet graph. This contradicts the random graphs model, which yields an exponential degree distribution.

Power laws are not observed only in Internet-related activities but also in income distributions, city populations, word frequencies and the world-wide web graph. One interesting situation where power laws result is as follows: consider the first digit of every number that person sees during the day. The distribution of these digits is a power law.

The power laws are explained by generative models that fall into one large category called *scale-free growth* or *preferential attachment* or *"the rich get richer"*: if the growth of individuals in a population follows a stochastic process that is independent of the individual's size (so that larger individuals attract more growth), then a power law will result. [1]

We will now go into the details of a model of Internet growth in the next section and then we will prove a similar result for a simple model of file creation.

3 A Model of Internet Growth

We will give a method for constructing a graph that will exhibit similar structural properties to the internet graph, namely the degree distribution will follow a power law. The entirety of this section is based on [2], and this paper will be referred to throughout the section.

3.1 Construction

We will discuss a simple model of Internet growth and show that it results with power-law degree distribution under very general assumptions (the probability that a degree is larger that D is at least $D^{-\beta}$ for some $\beta > 0$). We will consider the unit square in R^2 with points p_i $i \in \{0, \ldots, n\}$ distributed randomly and uniformly within it and give a method to connect them into a tree. When the i-th node arrives, it attaches itself to one of the previous nodes forming a tree based on minimizing the following two objectives:

- To minimize Euclidian distance
- To minimize hop distance (connect to a node that is centrally located)

We construct the graph by choosing an initial point p_0 and then for $i = 1, \ldots, n$ we connect p_i to the point p_j that minimizes:

$$\min_{j<i} \ \alpha \cdot d_{ij} + h_j \tag{3}$$

[1] One should not confuse the concept Zipf's Law and scale-free networks: Originally, Zipf's law stated that, in a corpus of natural language utterances, the frequency of any word is roughly inversely proportional to its rank in the frequency table. So, the most frequent word will occur approximately twice as often as the second most frequent word, which occurs twice as often as the fourth most frequent word, etc. The term has come to be used to refer to any of a family of related probability distributions. A scale-free network is a specific kind of complex network in which some nodes act as "highly connected hubs" (high degree), although most nodes are of low degree.

where d_{ij} is the Euclidian distance between p_i and p_j and h_j is the measure of the "centrality" of p_j. Possible measures for h_j are:

- average number of hops from other nodes
- maximum number of hops from another node
- number of hops from a fixed center of the tree

α is a parameter that determines the relative importance of the two objectives. Note that this is not meant to be an accurate representation of how Internet grows but it is an interesting and simple model that leads to power law distribution.

The behavior of the model highly depends of the relative importance of the two objectives, α.

1. If α is too low, less than $1/\sqrt{2}$ in this case, Euclidian distances are not important and the resulting network is a star network.
2. If $\alpha = \Omega(\sqrt{n})$, then Euclidian distances become too important and we end up with a dynamic version of Euclidian minimum spanning tree.
3. If α is in between, then degrees have a power law distribution.

3.2 Properties of the Graph

The following theorems examine the structural properties of the tree T formed by the method outlined in the previous section. Throughout the paper let $N(i)$ denote the neighbors of a point p_i in T.

Theorem 1 *If $\alpha < \frac{1}{\sqrt{2}}$ then T is a star centered about p_0.*

Proof: By its definition $d_{ij} < \sqrt{2}\ \forall i, j$, so $\alpha d_{ij} < 1$ and $h_j \geq 1\ \forall j \neq 0$. Therefore each node added to T after p_0 will connect directly to p_0 forming a star. □

This theorem shows that when α is too small, the euclidian distances are ignored and all vertices are connected to the most central vertex in the tree. The next theorem shows that if α is too large, then the centrality of vertices is overlooked and the graph behaves as though the euclidian distance is the only thing considered, which results in an exponential degree distribution.

Theorem 2 *If $\alpha = \Omega(\sqrt{n})$ then the degree distribution of T is exponential.*

Proof: To show that the degree distribution is exponential we need to verify that the expected number of nodes that have degree at least D is at most $n^2 e^{-cD}$ for some constant c. Since exponential growth is asymptotically lower than power-law growth, the authors only show the upper bound here.

In order to obtain this bound for each p_i we will divide its neighbors into two classes depending on the length of the edges linking them. Define $S(i) = |\{j \in N(i) | d_{ij} \leq \frac{4}{\alpha}\}|$ and $L(i) = |\{j \in N(i) | d_{ij} > \frac{4}{\alpha}\}|$.

By the union bound $Pr[degree(p_i) \geq D] \leq Pr[S(i) \geq \frac{D}{2}] + Pr[L(i) \geq \frac{D}{2}]$ So if we can show that each of these classes of vertices have exponential behavior, then we will have shown that the total degree distribution has exponential behavior.

First we show that $S(i)$ follows the exponential bound. Let i be fixed and let $\alpha \geq c_0\sqrt{n}$. Then any points in $N(i)$ linked by short edges must fall in a circle of area $\pi r^2 = \pi(\frac{4}{\alpha})^2 \leq \frac{p_i}{c_0 n}$ and thus

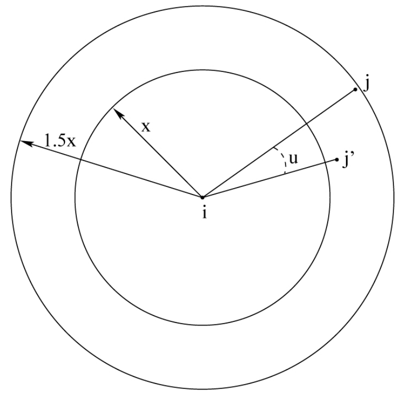

Figure 1: Proof of Thm. 2

$S(i)$ is bounded by a sum of bernoulli trials that represent the probability that each point falls within this circle. Therefore

$$E[S(i)] = n\pi(\frac{4}{\alpha})^2 \leq n\frac{pi}{c_0 n} = c$$

Where c is constant in c_0. Therefore by the Chernoff-Hoeffding bound, if we assume $D > 3c$ then:

$$Pr[S(i) > \frac{D}{2}] \leq e^{-\frac{(D-2c)^2}{D+4c}} \leq e^{-\frac{D}{21}}$$

Which establishes our exponential bound on $S(i)$.

Next we show that $L(i)$ also follows an exponential bound. Define $L_x(i) = |\{i \in N(i) | d_{ij} \in [x, 1.5x]\}|$. Consider Figure 1, and let x be any number larger than $\frac{4}{\alpha}$. Given p_j and $p_{j'}$ in this region between distance x and $1.5x$ from p_i, if the angle $\angle p_j p_i p_{j'}$ is small enough (namely when $\angle p_j p_i p_{j'} < c = cos^{-1}(\frac{43}{48})$) then p_j will connect to p'_j over p_i. This is because the bound on the angle would make $\alpha d_{ij'} > \alpha d_{jj'} + 1$ while $|h_i - h_j| \leq 1$.

We can verify that $c > \frac{2\pi}{14}$ and hence for any x we get $L_x(i) < 14$ meaning every such region has at most 14 points in it connected to p_i. Now if we consider the sum of the points in all of these regions which would contributed to our long edges we get at most an exponential number of them. Namely setting $\delta_i = \max\{\frac{4}{\alpha}, \min_j d_{ij}\}$ then the sum over all regions is:

$$L(i) = \sum_{k=1}^{-\log_{\frac{3}{2}} \delta_i} L_{\frac{3}{2}^{-k}}(i) \leq -14 \log_{\frac{3}{2}} \delta_i$$

.

Since the points are distributed randomly we get:

$$Pr[\delta_i \leq y] \leq 1 - (1 - \pi y^2)^{(n-1)} \leq \pi(n-1)y^2$$

The second term meaning one minus the chance that each of the remaining $n-1$ points falls in the region outside of the radius y from the point p_i. Finally combining this with our result about the total number of points depending on δ_i we have:

$$Pr[L(i) \geq \frac{D}{2}] \quad \leq \quad Pr[-14 \log_{\frac{3}{2}} \delta_i \geq \frac{D}{2}] \tag{4}$$

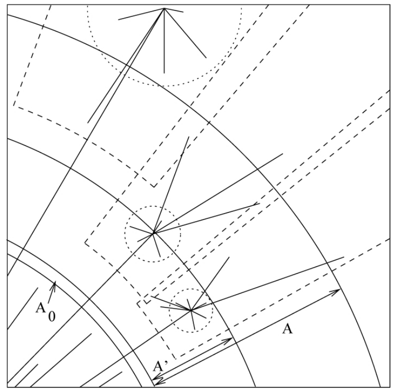

Figure 2: Proof of Theorem 3

$$= Pr[\frac{3}{2}^{-14}\delta_i \geq \frac{3}{2}^{\frac{D}{2}}] \quad (5)$$

$$= Pr[\delta_i \geq \frac{3}{2}^{-\frac{D}{28}}] \quad (6)$$

$$\leq \pi(n-1)(\frac{3}{2})^{-\frac{D}{14}} \quad (7)$$

$$(8)$$

Which establishes our exponential bound on the long edges. As mentioned earlier the union bound tells us that these two bounds can be combined to establish the exponential bound for the degree sequence. □

Next we can move on to the main result. The proof has more intricate details than the previous ones so we will only give a geometric sketch of the proof.

Theorem 3 *If $\alpha \geq 4$ and $\alpha = o(\sqrt{n})$ then the degree distribution of T is a power law.*

What we need to show is that the expected number of nodes with degree at least D is at least $c(\frac{D}{n})^{-\beta}$ for constants c and β. To show this bound we only examine neighbors of p_0 that fall within a certain distance of p_0 and show that enough of them have degree at least D to satisfy our bound.

The proof of this theorem is largely geometric. Some basic properties of the construction give us the following two results. If we consider the point p_i connected directly to p_0 then consider Figure 2.

Lemma 1 *Points connected to the graph in the circular dashed region around any point p_i that arrive after p_i will connect directly to it.*

Proof: The radius in this region will be small enough that a point falling in this region must not be in the corresponding region of another point connected directly to p_0 and since any other point will have hop distance at least one larger than p_i the new point will always have its cost function minimized when it connects to p_i. □

Lemma 2 *Points arriving after a point p_i outside of the larger dashed region will never connect to it.*

Proof: Whenever a point p_j is either too close to p_0, or the angle $\angle p_j p_0 p_i$ is too large, larger than $\sqrt{2.5\alpha(d_{i0} - 1/\alpha)}$, p_j will always prefer connecting to p_0 or another point over p_i. □

With these two Lemmas established we can say a few more things about the behavior of the model.

Sketch of Proof for Theorem 3: Consider Figure 2 depicting three regions about p_0 labeled A_0, A' and A. The radii of these regions are defined by $[1/\alpha, 1/\alpha + \rho]$, $(1/\alpha + \rho, 1/\alpha + .5\rho^{\frac{2}{3}}]$ and $(1/\alpha + \rho, 1/\alpha + .5\rho^{\frac{2}{3}}]$ respectively where $\rho = 4\sqrt{D/n}$. By our choice of ρ and Lemma 1 it is evident that for any point p_i linking directly to p_0 contained in A', it will have region of influence at least $\frac{\pi D}{n}$, which is a probability that any given point will land in that region. It is therefore expected to have at least $\frac{\pi D}{2}$ points that link to it in the final tree, since at least this many points are expected to arrive in this region after it, and other points could possibly connect to it as well, by Lemma 1 and 2.

- Any point in A' can only link to either p_0 or a neighbor of p_0 contained in $A_0 \bigcup A$. This follows from Lemmas 1 and 2 which characterize what points can possibly link to.
- Any point arriving in A' can only connect to another point in $A_0 \bigcup A$ if the angle between them and p_0 is small. More specifically this angle can be no larger than $\sqrt{10\alpha}\rho^{1/3}$ by Lemma 2.
- Points arriving in $A_0 \bigcup A$ can be thought of claiming sectors of A'. When a point arrives in $A_0 \bigcup A$ if no other point already present forms a small angle with it and p_0 then we know it must connect directly to p_0, and points arriving within that small angle of it and p_0 may connect to it. With high probability the number of points that land in A' is proportional to the regions area since the points are uniformly distributed in the unit square.
- If we partition A' into regions with rays separated by angles of degree equal to $\sqrt{10\alpha}\rho^{1/3}$ coming out of p_0. By our earlier statement about what points landing in $A_0 \bigcup A$ can connect to, we know that any point landing in a partition must connect directly to p_0 whenever the partition it lands in and the two partitions next to it empty. We see that there are $N = 1/(8\sqrt{\alpha}\rho^{1/3})$ partitions of angle $16\pi\sqrt{\alpha}\rho^{1/3} > \sqrt{10\alpha}\rho^{1/3}$.
- It can then be shown by the Chernoff bound that by the end of the construction of the graph at least half of these partitions should have a point in them and therefore at least one sixth of these partitions, $N/6$, contain a point in $A_0 \bigcup A$ linking directly to p_0.
- With high probability at most $N/12$ of these points are contained in A_0, since A_0 only occupies a small area compared to A.
- Therefore there are at least $N/12$ points in A that are neighbors of p_0 and by our previous argument their expected degree is at least $\pi D/2$ in the final tree.
- By the Chernoff bound the probability that any one of these $N/12$ points with expected degree $\pi D/2$ has degree less than D is exponentially small.
- With high probability we expect that at least $N/24n$ of these points has degree at least D and by choice of N and ρ, if we consider n and α as fixed, we get $N/24n = CD^{-1/6}$ where C is constant in n and α.

And this concludes the proof. □

3.3 Experimental Results

The authors implemented the stated model along with a few simple modifications to it and generated graphs for large values of n. The results all exhibited a power law for values of α in the range for which we proved they should.

The authors observed that the range of α that produces the power law behavior was larger than what it was proven for, and they believe their proofs could be tightened but would become much more complicated. This behavior also is apparent in higher dimensions and by using other metrics for d_{ij}.

4 A Model of File Creation

The power law phenomenon can be seen in other generative models. We present a simple model of file creation so that we can see connections between seemingly different places where power laws arise.

Assume that we have n data items and we want to partition them into files. We are given the popularity p_i for each data item i (this can be the expected number of times the data item will be retrieved for Internet transmission each day). Our objectives are to minimize:

- Total transmission cost
- Total number of files

That is, we would like to find a partition Π that minimizes

$$\left[\sum_{S\in\Pi}\left(|S|\cdot\sum_{i\in S}p_i\right)\right]+\alpha|\Pi| \tag{9}$$

This captures the trade-off between the transmission cost and the file creation overhead. Again, α designates the relative importance of the two objectives.

Proposition 1 *The optimum solution partitions the p_i's sorted in decreasing order and can be found in $O(n^2)$ time by dynamic programming.*

Now, assume that p_i's are i.i.d. from a distribution f. Suppose that the optimal partitions $S_1,\dots,S_k$ have sizes $s_i=|S_i|$ and the average item in S_i is a_i.

Lemma 3 $s_i+s_{i+1}\geq\sqrt{\alpha/a_i}$ *and* $s_i\leq\sqrt{2\alpha/a_i}$

The proof uses that, by optimality, it is not advantageous to merge two sets or to split one in the middle.

Now consider the cumulative distribution Φ of f and its inverse Ψ ($\Psi(x)$ is the least y for which $\Pr[z\leq y]\geq x$). It may be useful to view $\Psi(y/n)$ as the expected number of elements with popularity smaller than y. Let $g=\Psi(\log n/n)$.

Lemma 4 *In the optimum solution, that are at least $y/2\sqrt{2\alpha/g}$ sets of size at least $\frac{1}{2}\sqrt{\alpha/\Psi(2y/n)}$ almost certainly.*

Sketch of proof: With high probability, the popularity of the smallest element is no bigger than g. For large enough $y \leq n$, there are at least y elements with popularities smaller than $\Psi(2y/n)$. By the previous lemma, the sets that contain these elements have sizes that satisfy $s_i + s_{i+1} \leq \sqrt{\alpha/\Psi(2y/n)}$ and $s_i \leq \sqrt{2\alpha/g}$. Thus, these elements are divided into at least $y/\sqrt{(2\alpha/g)}$ sets (by the second equality), half of them of size at least $\frac{1}{2}\sqrt{\alpha/\Psi(2y/n)}$ (by the first inequality). $\square$

From this lemma, we get

$$\Pr\left[\text{size of a file} \geq \frac{1}{2}\sqrt{\alpha/\Psi(2y/n)}\right] \geq y/2n\sqrt{2\alpha/g} \tag{10}$$

Set $x = \frac{1}{2}\sqrt{\alpha/\Psi(2y/n)} \Rightarrow \Psi(2y/n) = \alpha/4x^2 \Rightarrow 2y/n = \Phi(\alpha/4x^2) \Rightarrow y = n\Phi(\alpha/4x^2)/2$. Therefore,

Theorem 4 *In the distribution of file sizes induced by the optimal solution,*

$$\Pr\left[\text{size of a file} \geq x\right] \geq \Phi(\alpha/4x^2)\sqrt{g/32\alpha}. \tag{11}$$

Therefore, if $lim_{z\to 0}\Phi(z)/z^c > 0$ for some $c > 0$ then the file sizes have a power law distribution. Any continuous distribution f that has $f(0) > 0$ (e.g. exponential, normal, uniform etc.) gives a power law.

References

[1] Reka Albert and Albert-Laszlo Barabasi. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47–97, 2002.

[2] Alex Fabrikant, Elias Koutsoupias, and Christos H. Papadimitriou. Heuristically optimized trade-offs: A new paradigm for power laws in the internet. *Proc. 29th International Colloquium on Automata, Languages and Programming*, 2002.

[3] M. Newman, D. Watts, and S. Strogatz. Random graph models of social networks. *Proceedings of the National Academy of Sciences*, 99, 2002.