

Overview of Certain Properties of Scale-Free Graphs

Muhammad Mukarram Bin Tariq, Keshav Attrey
{mtariq, attrey}@cc.gatech.edu

November 22, 2005

Abstract

This lecture comprises two parts. In first part, we talk about certain connectivity properties of scale-free graphs that use the preferential attachment model. Specifically, we present the work by Mihail et.al. [1], where they show that these graphs have constant conductance, and they are frugal. The latter part of the lecture is concerned with graph models for Web graph, and is based on work by Kumar et.al. [3], where authors have presented graph models for the Web graph that can generate graphs with power-law degree distributions as well as bipartite cliques, which are both known to be hallmark properties of the Web graph. We expect the audience to take away the following from this lecture: For the first part, we would like the audience to appreciate that while having random graph models that can generate power-laws is quite remarkable, these models are quite useless unless we can say something useful about the graphs that are built using these graph models. The fact that [1] show constant conductance for these graphs, which in turn can be used to infer worst-case congestion as well as coverage time for the networks, is truly significant. In the second part of the lecture, our goal is show that there exist graph models that produce more properties found in real-world graphs than just the degree distribution. Specifically, we will talk about models that generate graphs with bipartite cliques, which are known to exist in the Web graph, but are absent in previously known graph models.

1 Introduction & Background

Certain large networks are known to have power-law degree distributions. The examples include mostly man-made systems, such as the AS-level topology of the Internet [5], the graph of hyperlinked web-pages on World-Wide-Web, the network of citations of scholarly publications, and even some natural systems, such as protein folding.

Since these networks are large and complex, they are difficult to analyze and simulate. Several researchers have proposed models that generate graphs that have power-law degree distribution. Broadly, these graphs can be categorized into two categories; 1) Structural and 2) Evolutionary. The structural models start with a given degree distribution and interpolate a graph to match that degree distribution. The evolutionary models identify certain growth primitives that give rise to skewed degree distributions. The evolutionary models can be sub-categorized into two types of models a) microscopic and b) macroscopic. The micro-scopic models, that identify microscopic level underlying processes that guide the growth of the graph. For example there are models that consider the growth process as one where the graph grows step by step, but at each step is guided by a multi-objective optimization problem; heuristically optimized tradeoffs [4] is one such model. The macroscopic evolutionary models, on the other hand, try to model the macroscopic process in the graph evolution. For example, the preferential attachment model, due to Bellabos et.al., uses

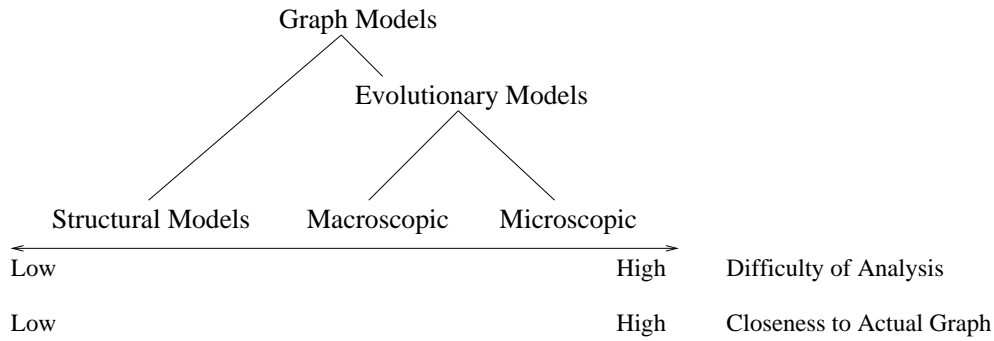


Figure 1: Different Types of Graph Models

the observation that the new vertices that join the graph prefer to join the vertices that are already well connected, i.e., they have a higher degree.

In general, the microscopic evolutionary models are closest to the graph that we try to model, but due to complex optimization problems solved at each step, they tend to have high and complex dependencies among edges and vertices. As a result, these models are hardest to analyze among all three categories. The macroscopic evolutionary models are still closer to the actual graph that we are trying to model, though not as accurate as the microscopic models. The macroscopic models still have some dependencies, but not as much as those in microscopic models; as a result these models lend themselves relatively easily for analysis. Structural models do not have such dependencies, so they are easiest to study, but at the same time they tend to be least accurate among all the three models, in terms of capturing the details of the original graph process.

Now we briefly introduce the two large and complex graphs that we discuss in this paper. The Internet is a large complex graph. As of year 2003 estimates, there are 580 million users on the Internet, and the world wide web hosts nearly 100 tera-bytes of data, and together with email, and instant messaging, Internet shuffled 532 peta-bytes of data! [6]. By some estimates, these numbers are growing exponentially. It is quite remarkable that the Internet is so complex, and yet works so effectively well, especially there is no single central authority that governs its growth. These facts make Internet as a fascinating topic of study. In this lecture, we will talk about certain properties of Internet topology, and some models for generating WWW type of graphs.

Internet topology is a complex graph of routers, switches, hosts, and links. The overall Internet topology is a result of inter-networking of smaller autonomous systems or AS. Each AS has its own infra-structure, (internal topology), that this AS manages, and a set of customers, or end systems that these AS serve. In order to provide global connectivity, these AS connect with each other, and allow, either direct access to each others customers (peering relationship), or a transit to other AS, (transit relationship). The graph resulting from this interconnection at AS level is referred to as the AS-level topology.

In 1999, Faloutsos et.al. [5] first reported existence of power-law in distributions of degree for AS in AS-level Internet topology; see figure 2 and 3. The average degree at AS-level, as of October 26, 2005 is 4.82, and median degree is only 2. The power-laws imply that there are significant number of AS which are orders of magnitude larger (in degree), than the average AS. This raises the important questions of i) do these AS become central, and causes of congestion? How does the congestion scale with the size of the network? ii) how well connected is the Internet? iii) do these large ISPs become monopolies, and what would be effect on your connectivity if one of these ISP decides to disconnect you.

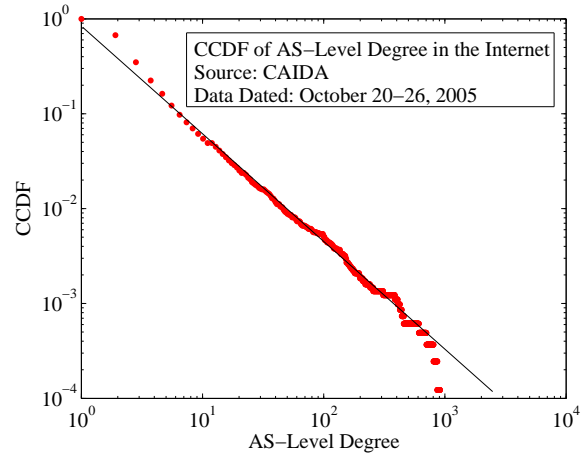


Figure 2: Power Laws in AS-level degree distribution

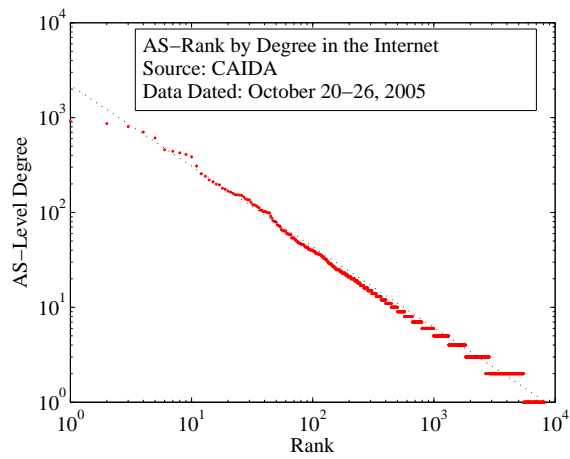


Figure 3: Power Laws in AS-level degree based rank (Zipf Law)

Note that in a constant degree tree, the congestion grows as n^2 with nodes, while in a constant degree expanders, this growth is close to the theoretical minimum of $n \log n$. Mihail et.al. [1] answer the first two questions by establishing constant conductance for an Internet-like graph model. This has an immediate implication that there would exist a multi-commodity routing with congestion $O(n \log n)$. Also constant expansion implies constant spectral gap between the first and second Eigen-values of the stochastic normalization of the adjacency matrix of the graph. Not only is this in line with measurement studies that were performed before [1], but can also be used to infer the graph coverage time.

As to the question of monopoly, or existence of competition in Internet connectivity business, Mihail et.al. [1] provide a partial answer. They show that for Erdos-Renyi $G_{n,p}$ model, when $np = \omega(\log n)$, the Vickery-Clarke-Grove over-payment (see later for details), over all origin-destination pairs, and all edges in shortest path for these pairs, is bounded from above by non-increasing function of the expected degree. Obtaining similar results of scale-free graphs remains an open problem.

In second half of the paper we talk about graph models for the Web-graph. Besides the existence of power-law degree distributions, a defining property of the Web graph, is the presence of a large number of bipartite cliques, which represent “communities” on the Web with a common interest. This property seems to be much less universal than the power-law distribution. In particular, a multitude of bipartite cliques have not been found on the AS-level topology of the internet. In the Erdős-Rényi model and similar random graph models, edges are chosen independently of each other, and these models produce neither a power law distribution nor a multitude of bipartite cliques comparable to that found on the web.

Previous models for web graphs are known not to generate bipartite graphs. So the question is how may we model these graphs. Might we model the Web graph effectively by attempting to mimic some of its microscopic processes? Might we create a model that produces the statistics observed on the web as an emergent feature rather than as a starting point of the model?

In 2000, Kumar et.al. [3] presented an influential model for the Web graph in which edges for new nodes were created by copying links from old nodes .

This model produces not only a power law degree distribution but a multitude of bipartite cliques significantly larger than those produced by previous models.

2 Conductance and Frugality in the Internet Graph

2.1 Graph Model

Mihail et.al. [1] use a macroscopic evolutionary graph model, that uses preferential attachment. Consider $G_{d,n}$, a graph that is grown using preferential attachment. The graph grows as following. $G_{1,n} = T_n$, where T_n is a tree grown in n time steps, one vertex at a time. The vertices are named after the time that they are created. At time $t = 0$, there is just one vertex with a self loop. At a later time, t , a new vertex arrives and attaches to one of the existing vertices. The probability that a new vertex attaches to a existing vertex is proportional to the degree of the existing vertex at the time of arrival of new vertex.

A more general case, $G_{d,n}$, is where at each time step, a vertex comes and joins with d existing vertices. Mihail et. al. do this in following way. First construct a tree T_{dn} , grown using the preferential attachment model in dn time steps, and then for all $1 \geq \tau \leq n$, contracting all the mini-vertices $(\tau d - i)$, for $0 \leq i \leq d - 1$. The contracted vertices may have self-loops and multiple edges to vertices in rest of the graph, but they are retained.

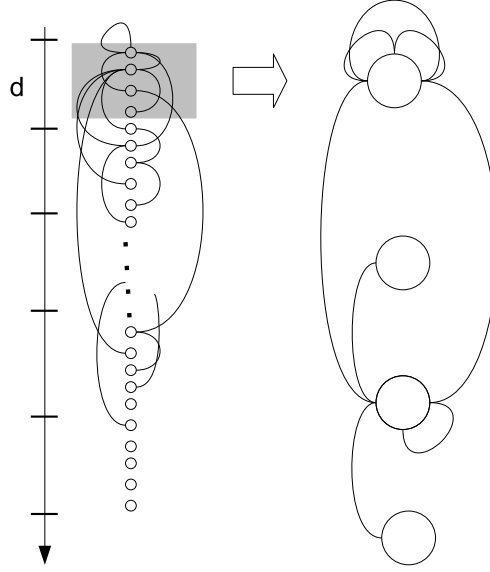


Figure 4: Growth of $G_{d,n}$: Contraction Step

2.2 Conductance of Scale-free Graphs

Let $G(V, E)$, be a undirected graph, with self-loops (as our $G_{d,n}$ has). Let $d_G(u)$ be the degree of a vertex $u \in V$. Then the for some $S \subset V$, volume of S is defined as $vol_G(S) = \sum_{u \in S} d_G(u)$. Also, for $S \subset V$, $C_G(S, \bar{S})$, is the multiset of edges with one endpoint in S and other in \bar{S} . The edge expansion ρ_G and conductance Φ_G of graph G are given as following:

$$\rho_G = \min_{S \subset V, |S| \leq \frac{|V|}{2}} \frac{C_G(S, \bar{S})}{|S|}$$

$$\Phi_G = \min_{S \subset V, vol_G(S) \leq \frac{vol_G(V)}{2}} \frac{|C_G(S, \bar{S})|}{vol_G(S)} \quad (1)$$

In plain words, edge expansion is a measure of connectedness of the graph, in the sense that higher the edge expansion, the more number of edges there exist between any two partitions of the graph.

Conductance, may be new to many audience, so lets describe it a bit here. If we have a graph with a some set of vertices. Every vertex with degree d_i can be thought of having d_i customers, and each customer wants to have a unit flow to all other customers. Then then there is $d_i d_j$ units of flow between vertices i and j . What we want to know is what is flow over the most congested link, and that is the measure of congestion of network.

Suppose we have $d_i d_j$ demand, between vertices i and j . Also consider we partition the network, in S , and \bar{S} , and obtain the cut $C(S, \bar{S})$. Then the question is that how much traffic wants to flow from one partition to another. The traffic is $\sum_{i \in S} d_i \sum_{j \in \bar{S}}$, and the total capacity between the partitions is $|C(S, \bar{S})|$. So there must be an edge that carries at least $(\sum_{i \in S} d_i \sum_{j \in \bar{S}} d_j) / |C(S, \bar{S})|$ traffic. So maximum congestion, C_{max} is

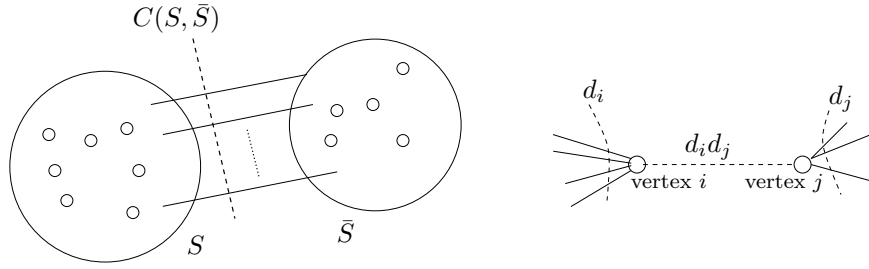


Figure 5: Illustration: Relation of Conductance and Congestion

$$\max_{S, \bar{S}} \frac{\sum_{i \in S} d_i \sum_{j \in \bar{S}} d_j}{|C(S, \bar{S})|} \leq C_{max} \quad (2)$$

greater than equal to this fraction. In fact, there is a result from theory of approximation, (according to [7]), that says that

$$C_{max} \leq \log n \frac{\sum_{i \in S} d_i \sum_{j \in \bar{S}} d_j}{|C(S, \bar{S})|}$$

So C_{max} is bounded from above and below.

$$\frac{\sum_{i \in S} d_i \sum_{j \in \bar{S}} d_j}{|C(S, \bar{S})|} \leq C_{max} \leq \log n \frac{\sum_{i \in S} d_i \sum_{j \in \bar{S}} d_j}{|C(S, \bar{S})|}$$

Relating 2 with the definition of conductance in eq. 1, we see that conductance is basically a bound on congestion.

Here I'd also like to mention that the assumption that flow between vertices i and j is $d_i d_j$ may seem artificial to some, but it is reasonable under assumption that all edges (links) are similar. Some well reputed work on traffic matrix estimation in the networking area also makes such assumptions. In particular, the *tomogravity* model [11], shows that it is reasonable to assume for real networks that the demand from i to j is proportional to product $f_i f_j$, where f_i is net flow through i , f_j is net flow through j .

2.2.1 Interesting Implications of Constant Conductance

Mihail et.al. [1] establish constant conductance; this has a couple of very interesting implications.

1) This can be used to show that there is a polynomial time routing algorithm that can route on G , with maximum link congestion of $O(n \log n)$, if traffic between any two vertices is proportional to product of their degree! This result is based on Leighton and Rao's work on designing approximation algorithms using the Multicommodity max-flow min-cut theorems [8].

2) We can show that the second eigen-value of the matrix corresponding to the random walk on G is $1 - c$, almost surely, for some positive constant c . This means that the cover time of the graph using a random walk is $O(n \log n)$. From the guest lecture that Sam gave on Expanders, recall that the constant degree expanders are desirable. This corollary seems to imply that scale-free graphs with constant average degree might behave like constant degree expanders!

Please refer to Corollaries, 2.3 and 2.4 in [1] for further details.

2.2.2 Establishing Constant Conductance

Theorem 1 *There is a positive constant α , such that, for any $d \geq 2$, the random graph $G_{d,n}$ has edge expansion $\alpha = \alpha(c, d)$, and conductance $\frac{\alpha}{d+\alpha}$, almost surely.*

Specifically, for $\alpha \leq \min\{\frac{d-1}{2} - \frac{c+1}{4}, \frac{1}{5}, \frac{(d-1)\ln(2) - \frac{2}{5}\ln 5}{2(\ln d + \ln 2 + 1)}\}$

$$Pr\{\rho_{G_{d,n}} < \alpha\} \leq o(n^{-c})$$

and

$$Pr\{\Phi_{G_{d,n}} < \frac{\alpha}{d+\alpha}\} \leq o(n^{-c})$$

Proof: From now on, lets just use G to refer to $G_{d,n}$. Let S be a subset of n vertices in G , and $vol_G(S) \leq dn/2$. Every vertex in S contributes d to the volume of S , then $d|S| \leq vol(S) \leq d|S| + C_G(S, \bar{S})$. It immediately implies that $|S| \leq n/2$. Also we can use it to bound conductance (Φ_G) in terms of edge expansion (ρ_G).

$$\begin{aligned} \Phi_G &= \min_{S \subset V, vol_G(S) \leq dn/2} \frac{C_G(S, \bar{S})}{vol_G(S)} \\ &\geq \min_{S \subset V, vol_G(S) \leq dn/2} \frac{C_G(S, \bar{S})}{d|S| + C_G(S, \bar{S})} \end{aligned}$$

The second step follows from $vol(S) \leq d|S| + C_G(S, \bar{S})$. Multiplying and dividing by $|S|$, we obtain:

$$\Phi_G \geq \frac{\rho_G}{d + \rho_G} \quad (3)$$

Now lets try to bound ρ_G . Mihail et. al. do this using a counting argument. Lets call a set $S \subset n$ BAD, if $|C_G(S, \bar{S})| < \alpha|S|$. Also lets define a mini-vertex t , from T_{dn} be BAD if t is associated with a vertex in S , but father of t is in \bar{S} , or vice versa. If a vertex is not BAD, then it is GOOD. Now for a set S , $|S| = k \leq n/2$, let A be set of good vertices for S , so that $|A| \leq \alpha k$. Lets denote a a vertex t to be bad as \tilde{t}

$$Pr\{t \text{ is BAD} : t \in \{n\}, t \notin A\} \leq \frac{\binom{dk}{\alpha k}}{\binom{dn-\alpha k}{dn-\alpha k}} \quad (4)$$

The above relation comes from lemma 1, that we describe a little later.

There are $\binom{n}{k}$ for S for each $k \leq n/2$. Then for each choice of S , we have $\alpha k \binom{dn}{\alpha k}$ choices for A . For a set to be bad, there has to exist the condition of 4, above.

$$Pr\{\exists \text{ a BAD Set } S\} \leq \sum_{k=2}^{n/2} \binom{n}{k} \alpha k \binom{dn}{dk} \frac{\binom{dk}{\alpha k}}{\binom{dn-\alpha k}{dn-\alpha k}}$$

using $\binom{n}{k} \binom{(d-1)n-\alpha k}{(d-1)k-\alpha k} \leq \binom{dn-\alpha k}{dk-\alpha k}$, we get,

$$Pr\{\exists \text{ a BAD Set } S\} \leq \sum_{k=2}^{n/2} \alpha k \binom{dn}{\alpha k} \binom{dk}{\alpha k} \binom{dn - \alpha k}{dn - \alpha k}^{-1}$$

using the bounds $\binom{n}{k}^k \leq \binom{n}{k} \leq \frac{en^k}{k}$, we get,

$$\begin{aligned} Pr\{\exists \text{ a BAD Set } S\} &\leq \sum_{k=2}^{n/2} \alpha k \left(\frac{n}{k}\right)^{\alpha k} \left(\frac{ed}{\alpha}\right)^{2\alpha k} \left(\frac{(d-1)k - \alpha k}{(d-1)n - \alpha k}\right)^{(d-1)k - \alpha k} \\ &\leq \sum_{k=2}^{n/2} \alpha k \left(\frac{n}{k}\right)^{\alpha k} \left(\frac{ed}{\alpha}\right)^{2\alpha k} \left(\frac{k}{n}\right)^{(d-1)k - 2\alpha k} \\ &\leq \sum_{k=2}^{n/2} \alpha k \left(\frac{ed}{\alpha}\right)^{2\alpha k} \left(\frac{k}{n}\right)^{(d-1)k - 2\alpha k} \\ &= \sum_{k=2}^{n/2} \alpha k \left(\frac{ed}{\alpha}\right)^{2\alpha k} \left(\frac{k}{n}\right)^{(d-1-2\alpha)k} \end{aligned}$$

There are $O(n)$ terms in the above summation. We can bound this summation to $o(n^{-c})$, if we can bound the largest term by $o(n^{-(c+1)})$. It can be seen that for small enough α , all the terms in the summation are smaller than the term for $k = 2$, if $(2d - 2 - 4\alpha) > 0$, which is true for small enough α , {at least that is what Mihail's paper says, but this condition does not seem sufficient to me}. Hence, we need to bound $n^{-(2d-2-4\alpha+1)}$. This can be bounded by $n^{-(c+1)}$, for $c < 2(d-1) - 4\alpha - 1$.

Note: there are some problems with these bounds as they are presented in the conference version of the paper [?] that we used to prepare this lecture. We talked with Prof. Mihail, and she pointed us to corrected proof that is in their journal version. That paper can be found in [?], page 4.

Lemma 1 For a fixed subset $S \subset [n]$, $|S| = k$, and for a fixed subset $A \subset [dn]$, $|A| \leq \alpha k$, the probability that all the mini-vertices associated with $[dn] \setminus A$ are BAD in $G_{d,n}$ is $\binom{dk}{\alpha k} / \binom{dn - \alpha k}{dk - \alpha k}$.

Proof: The detailed proof is fairly involved can be found in [1, ?]. There is no point in repeating it verbatim here. Instead, we give a high level overview of the the approach. The bound is obtained by finding the probability that $x_i \in S \setminus A$ is BAD, given A is set of all good mini-vertices upto x_i . Similarly for $\bar{x}_i \in \bar{S} \setminus A$. These probabilities turn out to be $< \frac{i+|A|}{z_i+1|A|} = \frac{\text{volume of } S \text{ when } x_i \text{ arrives}}{\text{volume of Graph at that time}}$

Now probability for all A to be good and \bar{A} to be bad is $\Pi\left(\text{Prob}\{x_i \in S \setminus A \text{ is BAD}\}\right) \Pi\left(\text{Prob}\{\bar{x}_i \in \bar{S} \setminus A \text{ is BAD}\}\right)$ and this after some clever algebraic manipulation comes out to be $\frac{(dk)!(dn-dk)!}{(dn-|A|)!(|A|)!}$ which is essentially what the lemma says (replacing $|A|$ with αk).

2.3 Frugality of Scale-free Graphs

Consider $v(e, u, v)$, the Vickery-Clark-Groves(VCG) over-payment of edge an $e \in E$ with respect to vertices $u, v \in V$, if e is on the shortest path P from u and v , and the cost of shortest path between

u and v increases by $v(e, u, v)$ if e is removed from the graph $G(V, E)$. Mihail et.al. deal rather interestingly with edges that are bridges. The VCG over-payment for these edges is taken as 0, instead of infinity, as the standard definition for VCG over-payment suggests. This treatment allows analysis of general random graphs where small components and bridges occur with small probability. Also it allows us to more appropriately use VCG over-payment as a measure of “monopoly” in the network, without being concerned with bridges that are result of conscious decision by the client ISP’s to be singly homed, and are not in anyway a result of “monopoly” or lack of competition. Mihail et.al. show that the average $v(e, u, v)$ for $G \in G_{n,p}$ (the Erdos Renyi random graph), is $O(1)$ and $\Omega(1/np)$.

Theorem 2 *For $G \in G_{n,p}$, the Erdos-Renyi random graphs with n vertices, and edge probability p , and $np = \omega(n \log n)$, with probability $o(n^{-c})$ for some constant $c > 0$, the average $v(e, u, v)$, over all vertices, u, v and edges e on shortest path between u, v , is $O(1)$ and $\Omega(1/np)$*

Proof: Let P be the set of edges in the shortest path, and P_2 , be the set of edges on any other path connecting vertices u, v , so that $P \neq P_2$. For all edges in $P \setminus P_2$, $v(e, u, v) \leq |P_2| - |P|$, (because in worst case, there could be a single edge that is different in P_2 and P , and then its cost should be the difference of cost of two paths.) Also, for an edge $e \in P_2 \cap P$, $v(e, u, v) \leq \text{diameter}(G \setminus e)$ (reason: after removal of e , there should still be path between u and v , that is shorter than the diameter of $G \setminus e$, by definition of diameter. If e is a bridge, then again by definition of special treatment of bridges, the VCG cost would be $0 \leq \text{diameter}(G \setminus e)$). From these two observations, we get,

$$\frac{\sum_{e \in P} v(e, u, v)}{|P|} \leq |P_2| - |P| + \text{diameter}(G \setminus e) \frac{|P \cap P_2|}{|P|} \quad (5)$$

The term on right is for all $e \in (P_2 \cap P)$, and the two terms on left are for $e \in P \setminus P_2$.

Now since both $|P|$ and $\text{diameter}(G \setminus e)$ are $\Theta(\log n / \log np)$ with probability $1 - o(1/n)$, in order to get the bounds theorem, we must find a P_2 , for which $|P \cap P_2|$ is bounded from above by a constant in expectation.

Lets consider the frontier of breadth first search from u , as $\Gamma_i(s) = s \in V(G) : \text{distance}(u, s) = i$, and the set of all vertices within a certain distance from u as, $N_i(u) = \cup_{j=0}^i \Gamma_j(u)$. Using lemma 10.7 from Bellabos’s Random Graph text, for every u and any $\delta > 0$, with probability $1 - O(n^{-8})$, we have,

$$||\Gamma_i(u)| - (pn)^i| \leq \delta (pn)^i \quad (6)$$

Similarly, lets define the frontiers and set of discovered vertices doing BFS from v , as $\Gamma_i(v)$ and $N_i(v)$. The frontiers for u and v would meet for the first time for $i = \lfloor (|P| + 1)/2 \rfloor$ (as $|P|$ is the length of the shortest path). Lets continue to grow these frontiers, until they meet again, lets call this step i_2 . Let $P_2(u)$ be the shortest path between one of the vertices in $\Gamma_{i_2}(u) \cap \Gamma_{i_2}(v)$ and u , (the intersection has at least one vertex, since frontiers are meeting), and let $P_2(v)$ be the shortest path from the same vertex to v . Let $P_2 = P_2(u) \cup P_2(v)$.

$|P \cup P_2(u)| \geq i$ implies that P and P_2 go through the same vertex of $\Gamma_j(u)$ for some $j \geq i$. However, since vertices of $\Gamma_j(u)$ through which either P or P_2 pass, can be viewed as independent uniform tosses, the probability that they are the same in $\Gamma_j(u)$ is precisely $1/|\Gamma_j(u)|$. Therefore

the expected size of intersection $E(|P_2(u) \cap P|) \leq E(\sum_{i>1} 1/\Gamma_i(u))$. Same is the case for the intersection of path alternate path $P_2(v)$ and P . Together, we get with high probability that

$$E(|P \cap P_2|) \leq E\left(\sum_{i \geq 1} \frac{2}{\Gamma_i(v)}\right) = O\left(\frac{1}{pn}\right) \quad (7)$$

To bound $|P_2| - |P|$, let k be smallest i , such that, $|\Gamma_i(u)| \geq \sqrt{n}$. It suffices to consider the case in which $|P_2| \geq 2k$ and $|P| \leq 2k$, and for then to add the expectations of the deviations from these bounds. For P_2 , consider the sets of $(k+i)^{th}$ horizons, for all $0 \leq i \leq n$, $\Gamma_{k+i}(u)$, and $\Gamma_{k+i}(v)$, and calculate the probability that $|\Gamma_{k+i}(u) \cap \Gamma_{k+i}(v)| \leq 2$. For all but a fraction $O(n^{-8})$ of all graphs, these sets will have at least $g = \frac{3}{4}\sqrt{n}(pn)^i$ elements. Therefore, in these graphs the probability that the sets intersect in fewer than 2 points is at most $g(1 - \frac{g}{n})^{g-1} \leq e^{-(pn)^{2i/3}}$. Thus the expectation of $|P_2| - 2k$ is at most $2 \sum_{i>0} e^{-(pn)^{2i/3}} = O(1)$.

Now, in order to bound the expectation of $2k - |P|$, consider the $(k-i)^{th}$ horizons, for $1 \geq i \leq k$, of cardinality at most $h = \frac{5\sqrt{n}}{4(np)^{i-1}}$. The probability that these sets intersect is at most $1 - (1 - \frac{h}{n})^h \leq 1 - e^{-1/2(pn)^{2i/3}} \leq \max\{1, (np)^{2-2i}\}$. Therefore, the expectation of $k - |P|$ is at most $2 \sum_{i>0} \max\{1, (np)^{2-2i}\} = O(1)$. This completes the proof of the upper bound.

For lower bound, note that expectation of $k - |P|$ is $\Omega(1/np)$, and that indeed $|P| < k$ and $|P_2| > k$ with some non-vanishing probability.

3 Models for the Web Graph

3.1 Introduction

The World Wide Web may be viewed as a directed graph, each of whose vertices is a static HTML web page, and each of whose edges corresponds to hyperlink from one web page to another. The out-degree of a node corresponds to the number of hyperlinks (or ‘‘outlinks’’) on a web page, and the in-degree to the number of other web pages that have hyperlinks to that web page.

The Web is huge, with around 2 to 3 billion nodes. The Web is also a sparse graph, meaning that the average degree has remained constant (around 7 links per page) as the number of nodes on the Web has grown from a few hundred thousand to its current size. (Equivalently, a sparse graph is one in which the sum of the degrees of the nodes, or the ‘‘volume,’’ is $O(n)$.)

Power laws have been found for many of the Web’s statistics: its in-degree, its out-degree, its strongly connected components, the number of web pages in a site, the number of visitors to a site in a day, and the number of links clicked by web surfers. [9]

A defining property of the Web graph, however, is the presence of a large number of bipartite cliques, which represent ‘‘communities’’ on the Web with a common interest. This property seems to be much less universal than the power-law distribution. In particular, a multitude of bipartite cliques have not been found on the AS-level topology of the internet. In the Erdős-Rényi model and similar random graph models, edges are chosen independently of each other, and these models produce neither a power law distribution nor a multitude of bipartite cliques comparable to that found on the web. The Aiello, Chung, and Lu graph model [10], which first generates a power-law degree sequence and then generates a graph matching the degree sequence does not produce many bipartite cliques [3]. And the preferential attachment model used by Mihail et al. to model the AS-level topology has not been shown to produce many bipartite cliques either.

We need something different. Might we model the Web graph effectively by attempting to mimic some of its microscopic processes? Might we create a model that produces the statistics observed on the web as an emergent feature rather than as a starting point of the model?

In 2000, Ravi Kumar along with other researchers from IBM Almaden Research Center, Verity Inc., and Brown University, presented an influential model for the Web graph in which edges for new nodes were created by copying links from old nodes [3].

This model produces not only a power law degree distribution but a multitude of bipartite cliques significantly larger than those produced by previous models.

3.2 Motivations for Modeling the Web Graph

The web is enormous and growing rapidly, so assessing the scalability and performance of algorithms targeted toward the Web is critical.

Only companies of Google’s magnitude have the capacity to, for example, test the performance of their new Web crawling algorithms by running them directly on the entire Web. But with an appropriate model of the web, even a small company or research team could run its algorithms on small graphs produced by the model to get a sense of how well their algorithms perform and scale with an increasing number of nodes.

Even companies like Google want to know whether their current algorithms scale well. Just because their algorithms perform well and are cost-effective today does not mean that they will be in five years.

A good graph model can also help us develop algorithms better targeted toward the Web. Sub-graph enumeration problems, for example, seek to discover all instances of some type of topological structure on the web (such as bipartite cliques), and are highly valuable in classifying the nodes on the web semantically, but are in general are computationally difficult to solve. By utilizing a good graph model of the Web, perhaps we can come up with efficient algorithms that leverage the assumptions of these models.

A good model can suggest new properties of the Web, and it can provide a framework that we can use to prove properties of the Web.

Interestingly, the copying model described below also has direct applications to the field of Genomics. This model can be used to explain how existing genes mutate during duplication to form new genes with new behaviors.

3.3 Main concepts

- The power law degree distribution exists on the web.
- Many bipartite cliques, which represent web “communities”, exist on the web.
- In evolving graph models, a graph grows in stages over time.
- In copying graph models, edges are added to new nodes by copying links from old nodes.

3.4 Bipartite Cliques and Copying Models

A large number of bipartite cliques appear on the web. A bipartite clique $K_{i,j}$ is a set of vertices V partitioned into $V = V_1 \cup V_2$ where all possible $i * j$ edges starting in V_1 and ending in V_2 are present.

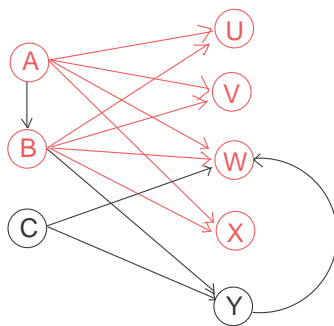


Figure 6: A 2,2 bipartite clique $A, B \cup U, V, W, X$ within a graph

Each bipartite clique represents a web “community”, in that it is made up web pages A_1, A_2, \dots, A_n , all of which point to the same set of web pages, B_1, B_2, \dots, B_n . In other words, web pages A_1, A_2, \dots, A_n all are “interested” in the same “topic,” because they all point to web pages B_1, B_2, \dots, B_n .

“Evolving” or “growth” graph models introduce (and remove) vertices and edges in discrete stages, as time evolves. They start with 1 vertex and grow the graph by adding vertices in each stages.

Kumar et al. have used the process of “copying” links to provide a cause for the multitude of bipartite cliques on the web, by mimicking what we might presume to happen when a new web page is added to the web: the author of a new web page finds an existing web page that deals with his topic of interest, copies some of the links from that web page, and then adds some additional links that he found through other means [3].

There is some debate over whether this is actually the most common process on the web to produce bipartite cliques. For example, some models take into account the use of search engines such as Google, which automatically bias a search engine user toward looking at web pages that are already popular.

However, within the genomics community it is agreed that this process of “copying links” actually provides a good model for what occurs during gene replication and mutation. Gene and protein interactions can be modeled as a graph where each node is a gene or protein, and each edge between two nodes is present if there is an interaction between that gene and protein. Mutations that occur during duplication of genes can be modeled as a new node being generated with most of the same linkage as an existing node and a few different links.

In the evolving, copying model described below, a power law distribution appears among the in- and out-degrees of the nodes, and bipartite cliques appear significantly more often than they do in prior models.

3.5 Algorithms

The linear growth model starts with 1 vertex and adds one vertex in each time step. At time t , the number of vertices is t . The exponential model starts with 1 vertex and adds $O(|V|)$ in each stage, which causes exponential growth with respect to t .

3.5.1 Linear Growth Model

Start with a single vertex.

For $t = 1$ to n :

1. Add 1 new vertex v to the graph $G = \langle V, E \rangle$
2. Select an existing (old) vertex p uniformly at random. (Below, we will copy outlinks from p to v .)
3. Add d outlinks starting at the new vertex v and ending at old vertices. (i.e., d is the average degree of G). Select the ending vertices as follows.
For $i = 1$ to d :
 - (a) With probability c (the 'copy factor').
Copy the i th outlink from p to the new vertex v . This produces a new edge from v to a vertex that p points to.
 - (b) With probability $1 - c$
Create a outlink from v to a random existing vertex.

3.5.2 Exponential Growth Model

Start with a single vertex.

For $t = 1$ to $\log n$:

1. Add $p * |V|$ new vertices to the graph $G = \langle V, E \rangle$ where constant $p > 0$ is a growth factor.
2. Let constant $d > 0$ be the out-degree factor; i.e., on average, we want $d * p * |V|$ new edges to be created in each stage. Assume at time 1, $|V| = 1$, and at time t , $|V| = (1 + p)^t$ and $|E| = d * |V|$.
Then $(1 + p)^t + p * |V| = (1 + p)^t + p(1 + p)^t = (1 + p)^{t+1}$
3. Iterate through all old edges.
For each old edge xy ,
 - (a) With probability p :
Create a new edge and set the ending vertex to y . (This will create $d * p * |V|$ new edges in each stage, on average.)
Then select the starting vertex of this edge as follows:
 - i. With probability k : Choose one of the new vertices uniformly at random
 - ii. With probability $1 - k$: Choose one of the old vertices with probability in proportion to the old vertices' current out-degree.

3.6 Degree Distributions and Number of Bipartite Cliques

Let $E[N_{t,k}]$ be the expected number of nodes of degree k at time t .

In the linear model, the degree distribution becomes

$$E[N_{t,k}] = O(tk^{-(c+1)/c})$$

where c is the copy factor (the fraction of outlinks on a new node that result from copying).

In the exponential model, the degree distribution becomes

$$E[N_{t,k}] = O(tk^{\log(1+p)})$$

where p is the growth factor.

Let $K_{t,i,j}$ be the number of bipartite cliques at time t with ixj vertices (i.e., partitioned into one set of i vertices and one set of j vertices). Let d be the constant out-degree.

Then in the linear model

$$K_{t,i,d} = \Omega(te^{-i})$$

for $i \leq \log t$ In other words, there are an exponentially increasing number of bipartite cliques as the size of the clique decreases.

References

- [1] M. Mihail, C. Papadimitriou, A. Saberi. On Certain Connectivity Properties of the Internet Topology. Proc. 44th IEEE Symposium on FOCS. October 2003. pp 28-35
- [2] M. Mihail, C. Papadimitriou, A. Saberi. On Certain Connectivity Properties of the Internet Topology. Journal Version Available Online: <http://www.cc.gatech.edu/~mihail/jcss.pdf>
- [3] R. Kumar, P. Raghavan, S. Rajagopalan, S. Sivakumar, A. Tomkins, E. Upfal. Stochastic Models for the Web Graph. Proc. 41s IEEE Symposium on FOCS. 2000.
- [4] A. Fabrikant, E. Koutsoupias, C. Papadimitriou. Heuristically Optimized Tradeoffs: a new paradigm for Power-laws in the Internet. Proc. 29th International Colloquium on Automata, Languages and Programming. 2002.
- [5] On Power-Law Relationships of Internet Topology. Proc. ACM Sigcomm'99.
- [6] Telchordia Netsizer. www.netsizer.com
- [7] Personal Communication with Prof. Milena Mihail.
- [8] Multicommodity Max-flow Min-cut theorems and their applications in designing approximation algorithms. JACM Vol. 46. 1999. pp 787-832.
- [9] J. Vera, Variations of the Preferential Attachment Model, Presentation at Georgia Tech, Nov. 7, 2005.
- [10] W. Aiello, F. Chung, and L. Lu. A random graph model for massive graphs. Proc. ACM Symp. on Theory of Computing, pp. 171-180, 2000.
- [11] Y. Zhang, M. Roughan, N. Duffield, and A. Greenberg. Fast Accurate Computation of Large-Scale IP Traffic Matrices from Link Loads. In ACM SIGMETRICS, 2003.