Mathematical and Computational Models
Editors: R.Nadarajan, R.Anitha and C.Porkodi
PSG College of Technology, Coimbatore

RANDOM GRAPH MODELS FOR THE WEB GRAPH

Anthony Bonato
Department of Mathematics
Wilfrid Laurier University
Waterloo ON
CANADA
abonato@rogers.com

Abstract

One of the most extensively researched real-world networks is the web graph. The web graph has vertices representing web pages, with edges corresponding to the links between pages. We describe some of the key properties of the web graph and propose models to simulate its evolution. We survey more recent work on geometric models, where nodes occupy some geometric space and edges are created via a mix of preferential attachment and geometric distance.

Key Words: *web graph, random graph models, power laws, geometric models*

1. INTRODUCTION

In the last decade, there has been an enormous amount of research surrounding a particular real-world network. The *web graph*, which we will denote by W, has vertices representing web pages, and edges representing the links between pages. Unlike abstract graphs W is an experimental, real-world graph. The massive, evolving graph W has the following properties: it is a *sparse, self-organizing, small world, and power law network*. The network W is not the only real-world graph with these properties. Many technological, social, biological networks have properties similar to those present in the web. For example, power laws have been observed in protein-protein interaction networks, and networks formed by scientific collaborators. For more details on properties of W and other self-organizing networks, the reader is directed to the survey [5] or the forthcoming book [6].

In the next few subsections, we consider some the key properties observed in the web graph. We will discuss the order, power law degree distribution, and small world property of W. In Section 2

we survey a few random graph models for the web graph. Our focus is on recent geometric models, which incorporate both randomness and geometric position of nodes to define adjacency. Owing to space restrictions, proofs are omitted (although references are given).

1.1 The Order of the Web Graph

Even a casual surf on the web reveals that there are an enormous number of web pages and links. As W is growing rapidly, it seems difficult to obtain an exact, current estimate of the number of web pages. Nevertheless, in 2005, [13] reported that the web graph had about 11.5 billion pages. A recent study of Hirate et al. [14] found 53.7 billion web pages, with 34.7 billion pages indexed by the search engine *Google*.

1.2 Power Law Degree Distributions

Arguably the most important property observed in W is its power-law degree distribution. Given an undirected graph G of order t and a non-negative integer k, we define

$$N_{k,t} = |\{x \in V(G) : \deg_G(x) = k\}|.$$

The *degree distribution of G* is the sequence $(N_{k,t} : 0 \le k \le \Delta(G))$, where $\Delta(G)$ is the maximum degree of G. We say that the degree distribution of G follows a *power law* if for each degree k,

$$\frac{N_{k,t}}{t} \sim k^{-\beta}$$

for a fixed real constant $\beta > 1$. Since W is a massive graph, we are more interested in an approximate rather than an exact value. In both real-world networks and graphs generated by theoretical models, the power law may only fit for a certain range of degrees, with discrepancies for small or large degree vertices. The graph W may clearly be viewed as either a directed or undirected graph. Observe that if G is directed graph, then we may discuss power laws for the in- and out-degree distributions in the obvious way.

The empirical study of W began before the turn of the millennium. Based on their crawl of the domain of Notre Dame University, Albert, Barabási, and Jeong [2] claimed that the web graph exhibited a power law in-degree distribution, with $\beta = 2.1$. The exponent of $\beta = 2.1$ was further corroborated by a larger crawl of the entire web (including 200 million web pages) reported in Broder

et al. cite [7]. There was some evidence presented in both studies that the out-degree distribution follows a power law with [2] reporting $\beta = 2.45$ and [7] reporting $\beta = 2.72$.

1.3 The Small World Property

With technologies available such as the e-mail, and cell-phones, the world definitely feels like a smaller place. The term *small world graphs* was first introduced by social scientists Watts and Strogatz [16] in their study of various real-world networks, such as the network of Hollywood movie actors. The diameter of a graph is a well-known and easily defined global measure of distances in a graph. Small world graphs G of order t should satisfy

$$\text{diam}(G) = \Theta(\log t).$$

Despite this condition, data from [7] suggests that $\text{diam}(W) > 900$. In a real sense the diameter of W is infinite: simply create a web page p with no links, and ensure that no one knows about it. Then p is an isolated vertex in W!

2. MODELS FOR *W*

Pioneering work on random graphs was first done by Erdős and Rényi [8,9]. We begin by recalling the random graph $G(n,p)$. We are given n vertices and a fixed real number parameter p in $(0,1)$. For each of the $\binom{n}{2}$ many distinct pairs of vertices, add an edge between them independently with probability p. The probability space $G(n,p)$, in a certain sense, is static or *off-line*: the number of vertices is fixed. Although usually n is taken as very large, and the number of edges is viewed as being variable over time, the number of vertices in $G(n,p)$ is off-line. Few techniques or models were available before the late 1990's for on-line random graph models. The degrees of vertices are binomially distributed. Hence, $G(n,p)$ is not appropriate as a model of the web graph W (after all, the study of random graphs predates the inception of the internet by several decades). Nevertheless, random graphs supply the mathematical subtext for these new models, and many of the techniques used to analyze them are also useful for models of W.

Over the last decade, a large number of rigorous models for the web graph W have been proposed. Such models deepen our understanding of the generative mechanisms driving the evolution of W, and provide insight into superficially unrelated properties observed in the web.

2.1 Preferential Attachment Models

Arguably the most important web graph models are ones incorporating some form of preferential attachment. The first evolving graph model explicitly designed to model W was given by Barabási, Albert [3]. The idea behind their model is an intuitively pleasing one: new vertices are more likely to join to existing vertices with high degree. In a slogan, *the rich get richer*. This model is now referred to as an example of a *preferential attachment* (or *PA model*). Barabási and Albert gave a heuristic description and analysis of their PA model (using mean field theory from physics), and concluded that it generates graphs whose in-degree distribution follows a power law with exponent $\beta = 3$. Although their proof was not rigorous, their important work set the stage for most of the mathematics regarding the modelling W to come. A rigorous analysis of the PA model was first given in [4].

2.2 Geometric Models for the Web Graph

While the web graph does not live in physical space, W may be viewed as occupying what might be called a *topic-space*, where pages are closer if they have common topics. For example, two news websites would be closer in topic space than a page on golf and one on graph theory. One approach to modelling W is to exploit a geometric model. In *geometric random graph models*, vertices are identified with points in some geometric space $\mathbb{S}$, and edges are determined via a mixture of probabilistic rules and the position of vertices in $\mathbb{S}$. Such models have been extensively studied; see the book [15] for more about them. We consider two models for the web graph, recently proposed in [1] and [11].

An interesting geometric model for the web graph was designed by Flaxman, Frieze, and Vera (see [11] and also [10,12]). In their PA model, new vertices may only join to vertices within a certain distance apart, and edges are then chosen by preferential attachment. We give an informal description of the model. The graphs they generate occupy the surface of a 3-dimensional sphere $\mathbb{S}$ with surface

area 1. For a point u on $\mathbb{S}$ and $r > 0$, we consider the *spherical cap* $B_r(u)$ *of radius* r *centred at* u, defined as

$$\{x \varepsilon \ S: \|x\text{-}u\| \leq r\},$$

where $\| \cdot \|$ is the Euclidean norm in $\mathbf{R}^3$. Note that r is fixed, and we will let A_r be the area of $B_r(u)$. The parameters of the model consist of a positive integer m, a positive real number r, and $\alpha \geq 0$ which controls the number of loops. Define G_0 to be K_1. To define G_{t+1}, choose a vertex x_{t+1} uniformly at random (u.a.r.) from $\mathbb{S}$, and add it to G_{t+1}. As in any PA model neighbours of x_{t+1} are chosen via preferential attachment, but with the additional restriction that they are chosen only from within $B_r(x_{t+1})$.

This mix of preferential attachment and geometry leads to a power law, as described in the following theorem. The probability of an event A in a probability space is denoted by $\mathbf{Pr}(A)$, while the expected value of a random variable X is written $\mathbf{E}(X)$.

Theorem 2.1. [11] Assume that $0 < \beta < 1/2$ and $\alpha > 2$ are constants and $r \sim t^{\beta - 1/2}\log t$.

1. If m is sufficiently large, then there exist constants $c, \gamma, \varepsilon > 0$ such that for all $k = k(t) \geq m$,

$$\mathbf{E}\left(\frac{N_{k,t}}{t}\right) = C_k k^{-(1+\alpha)} + O(t^{1-\gamma}),$$

where $C_k = C_k(m,\alpha)$ tends to a constant $C(m,\alpha)$ depending only on m and α as $k \to \infty$.

2. The random variable $N_{k,t}$ concentrates around its expected value via the following inequality:

$$\mathbf{Pr}\left(\left|\frac{N_{k,t}}{t} - \mathbf{E}\left(\frac{N_{k,t}}{t}\right)\right| \geq t^{1-\gamma}\right) \leq \exp(-t^{\varepsilon}).$$

In addition to the power law degree distribution exhibited by Theorem 2.1, the graphs G_t provably low diameter verifying one part of the small world property. An event holds *asymptotically almost surely* (*a.a.s.*) in a probability space if the probability it is satisfied tends to 1 as $t \to \infty$.

Theorem 2.2. [11] If $\alpha \geq 0$, $r \geq t^{-1/2}\log t$, and $m \geq K\log t$ where K is sufficiently large, then a.a.s. G_t has diameter $O(\log t/r)$.

A novel feature of the model is that it has *sparse cuts*, which is a property observed in *W*.

Theorem 2.3. [11] If $\alpha \geq 0$, and $r = o(1)$, then a.a.s. $V(G_t)$ can be partitioned into sets T_1 and T_2 so that $|T_i| \sim t/2$, and there are at most $4\sqrt{\pi rtm}$ edges between T_1 and T_2.

The *Spatial Preferred Attachment* (*SPA*) geometric web graph model [1] was proposed as an alternative to the geometric models of [10,11,12]. In the SPA model, each vertex is placed in space and surrounded by an *influence region*. The area of the influence region is determined by the in-degree of the vertex. Unlike the models of [10,11,12], in each time-step all regions decrease in area as a function of time. A new vertex v can only link to an existing vertex u if v falls within the influence region of u. If v falls within the influence region of u, then v will link to u with probability p. Hence, the SPA model utilizes the preferential attachment principle, but only implicitly: vertices with high in-degree have a large region of influence, and therefore are more likely to attract new links.

We emphasize two features that distinguish the SPA model from previous models. First, a new vertex can choose its links purely based on *local* information. The influence region of a vertex can be seen as the region where a web page is *visible*: only web pages that are close enough (in topic) to fall within the influence region will be aware of the give page, and thus have a possibility to link to it. Further, a new vertex links independently to each vertex visible to it. It follows that the new vertex needs no knowledge of the *invisible* part of the graph (such as in-degree of other vertices, or total number of vertices or links) to determine its neighbourhood. Second, since a new vertex links to each visible vertex independently, the out-degree is not constant.

We now give the formal definition of the SPA model. Let $\mathbb{S}$ be the surface of the sphere of area 1 in $\mathbf{R}^3$. The SPA model has parameters A_1, A_2, $p \geq 0$ such that p in (0,1], $A_1 \leq 1$, and $A_2 \geq 0$. Let $G_t = (V_t, E_t)$, and $V_t \subseteq \mathbb{S}$. Let $d^-(v,t)$ be the in-degree of vertex v in G_t. We define the *influence region* of vertex v at time $t \geq 1$, written $R(v,t)$, to be the cap around v with area

$$|\,R(v,t)\,| = \frac{A_1 d^-(v,t) + A_2}{t},$$

or $R(v,t) = \mathbb{S}$ if the right-hand-side is greater than 1.

The random process begins at $t = 0$, with G_0 equalling the empty graph, and we let G_1 be K_1. Time-step t, $t \geq 2$, is defined to be the transition between G_{t-1} and G_t. At the beginning of each time-step t, a new vertex v_t is chosen uniformly at random from $\mathbb{S}$, and added to V_{t-1} to create V_t. Next, independently, for each vertex u in V_{t-1} such that v_t in $R(u,t-1)$, a directed edge (v_t,u) is created with probability p. Thus, the probability that a link (v_t,u) is added in time-step t equals $p|R(u,t-1)|$.

The model easily generalizes to other metric spaces. See Figure 2.1 for a drawing and degree distribution (as a log-log plot, and so revealing the characteristic straight line fit of a power law) of a graph generated by a simulation of the SPA model. In the simulation, vertices were drawn on the unit square with 5,000 vertices, $p = A_1 = 1$, and $A_2 = 0$.

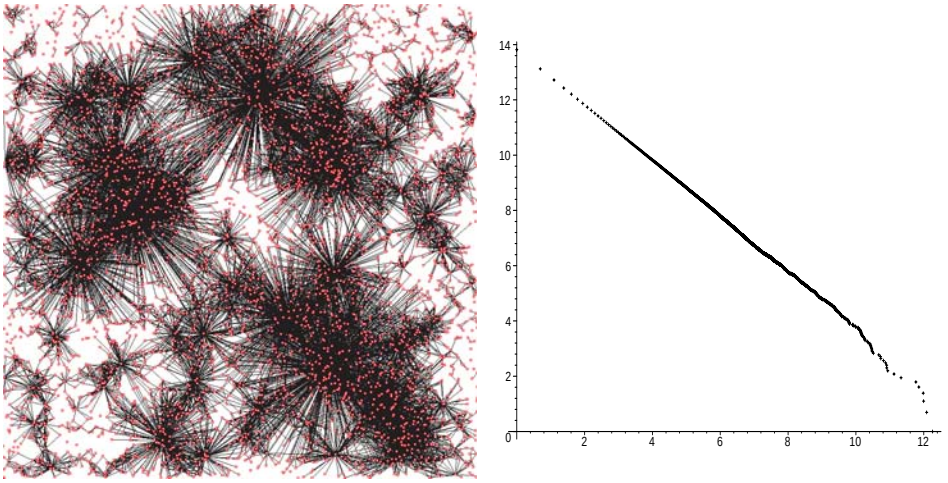

Fig 2.1 A simulation of the SPA model.

For an integer $n \geq 0$, define

$$k_f = k_f(t) = \left(\frac{t}{\log^4 t} \right)^{pA_1/(6pA_1+2)}.$$

The main result of [1] is that with high probability a graph G_t generated by the SPA model has an in-degree distribution that follows a power law in-degree distribution with exponent $1+1/(pA_1)$, with concentration up to t^{k_f}. If $pA_1 = 10/11$, then the power law in-degree exponent is 2.1, the same as observed in the web graph (as described in Section 1.2 above).

Theorem 2.4. [1] Fix p in (0,1]. Then for any $k \geq 0$,

$$\mathbf{E}(N_{k,t}) = c_k t(1+o(1)),$$

where

$$c_0 = 1/(1+pA_2),$$

and for $1 \leq k \leq t$,

$$c_k = \frac{p^k}{1 + pA_2 + kpA_1} \prod_{j=0}^{k-1} \frac{jA_1 + A_2}{1 + pA_2 + jpA_1}.$$

For $k = 0, \ldots, k_f$, a.a.s.

$$\mathbf{E}(N_{k,t}) = c_k t(1+o(1)).$$

Since it can be shown that

$$c_k = ck^{-(1+1/(pA_1))}(1 + o(1))$$

for some constant c, Theorem 2.4 shows that for large k, the proportion $\frac{N_{k,t}}{t}$ follows a power law with exponent $1+1/(pA_1)$, with concentration for all values of k up to k_f.

The proof of Theorem 2.4 in [1] follows by using a relaxation of the well-known Azuma-Hoeffding martingale techniques. The random variables $N_{k,t}$ do not a priori satisfy the c-Lipschitz condition where c is positive integer: it is possible that a new node may fall into many overlapping regions of influence. It is shown in [1] that the deviation from the c-Lipschitz condition occurs with exponentially small probability, which is enough to guarantee concentration around the expectation of $N_{k,t}$ (for values of k up to k_f).

With positive probability a new node will land in an area of $\mathbb{S}$ not covered by any influence regions, and thus have out-degree zero. Therefore, the underlying undirected graph of G_t is not connected. In fact, we expect that for the majority of distinct pairs u,v, there will not be a directed path from u to v. Since this is a property also observed in the web graph, it does not detract from the SPA model, but rather suggests that we should consider another parameter other than diameter to indicate a small world property. Thus, we focus on the (geometric) distance, in $\mathbb{S}$, spanned by the links.

For a pair of points u,v in $\mathbb{S}$, let $L(u,v)$ be the length of the shortest curve embedded in the surface of $\mathbb{S}$ that connects u and v. Define

$$L_t = \sum_{(v_t, v_i) \in E_t} L(v_t, v_i).$$

That is, L_t is the sum of the lengths of new edges added at time t in the SPA model. Note that L_t is a continuous random variable.

Theorem 2.5. [1] Suppose that $pA_1 > 2/3$. Then

$$\mathbf{E}(L_t) = \Theta\left(t^{-\left(\frac{1-pA_1}{pA_1}\right)} \right).$$

Theorem 2.5 contrasts with the analogous result for graphs generated with a similar process to the SPA model, but where all influence regions have area d/t for $d > 0$ a constant. We call this a *threshold model*. In the threshold model, $E(L_t)$ decreases much faster than for the SPA model with p large, such as when $p > 2/3$ and $A_1 = 1$. For example, if $pA_1 = 1$, then $E(L_t) = O(1)$. We refer to this property of the SPA model as the *geometric small world property*.

Theorem 2.6. [1] In the threshold model with areas of influence d/t, where d is a constant, there is a constant c so that

$$\mathbf{E}(L_t) \sim ct^{-1/2}.$$

It is likely that the maximum length of a directed path at time t in the SPA model is $O(\log t)$. This conjecture, along with an analysis of sparse cuts, is left for future work.

3. REFERENCES

[1] W. Aiello, A. Bonato, C. Cooper, J. Janssen, P. Prałat, A spatial web graph model with local influence regions, In: *Proceedings of The 5th Workshop On Algorithms And Models For The Web-Graph,* 2007.

[2] R. Albert, H. Jeong, A. Barabási, Diameter of the world-wide web, *Nature* **401** (1999) 130.

[3] A. Barabási, R. Albert, Emergence of scaling in random networks, *Science* **286** (1999) 509-512.

[4] B. Bollobás, O. Riordan, J. Spencer, G. Tusnády, The degree sequence of a scale-free random graph process, *Random Structures and Algorithms* **18** (2001) 279-290.

[5] A. Bonato, A survey of models of the web graph, In: *Proceedings of Combinatorial and Algorithmic Aspects of Networking*, 2004.

[6] A. Bonato, *A Course on the Web Graph*, AMS Graduate Series in Mathematics and AARMS Monograph series, in press.

[7] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener, Graph structure in the web, *Computer Networks* **33** (2000) 309-320.

[8] P. Erdős, A. Rényi, On random graphs I, *Publicationes Mathematicae Debrecen* **6** (1959) 290-297.

[9] P. Erdős, A. Rényi, On the evolution of random graphs, *Publ. Math. Inst. Hungar. Acad. Sci.* **5** (1960) 17-61.

[10] A. Flaxman, A. Frieze, J. Vera, A geometric preferential attachment model of networks, *Lecture Notes in Computer Science* **3243** (2004) 44-55.

[11] A. Flaxman, A. Frieze, J. Vera, A geometric preferential attachment model of networks, *Internet Mathematics* **3** (2007) 187-205.

[12] A. Flaxman, A. Frieze, J. Vera, A geometric preferential attachment model of networks II, In: *Proceedings of The 5th Workshop On Algorithms And Models For The Web-Graph*, 2007.

[13] A. Gulli, A. Signorini, The indexable Web is more than 11.5 billion pages, In: *Proceedings of the 14th International Conference on World Wide Web*, 2005.

[14] Y. Hirate, S. Kato, H. Yamana, Web structure in 2005, In: *Proceedings of the 4th Workshop on Algorithms and Models for the Web-Graph*, 2006.

[15] M. Penrose, *Random Geometric Graphs*, Oxford University Press, Oxford, 2003.

[16] D.J. Watts, S.H. Strogatz, Collective dynamics of 'small-world' networks, *Nature* **393** (1998) 440-442.