A Multi-Layer Model for the Web Graph

L. Laura * S. Leonardi * G. Caldarelli † P. De Los Rios ‡

April 5, 2002

Abstract

This paper studies stochastic graph models of the WebGraph. We present a new model that describes the WebGraph as an ensemble of different regions generated by independent stochastic processes (in the spirit of a recent paper by Dill et al. [VLDB 2001]). Models such as the Copying Model [17] and Evolving Networks Model [3] are simulated and compared on several relevant measures such as degree and clique distribution.

1 Overview

The WWW can be considered as a Graph (WebGraph) where nodes are static html pages and (directed) edges are hyperlinks between these pages. This graph has been the subject of a variety of recent works aimed to understand the structure of the World Wide Web [3, 6, 9, 17, 18, 21]. Characterizing the graph structure of the web is motivated by several large scale web applications, primarily, mining information on the web. Data mining on the web has greatly benefited from the analysis of the link structure of the Web. One example is represented by the algorithms for ranking pages such as Page Rank [8] and HITS [15]. Link analysis is also at the basis of the sociology of content creation, and the detection of structures hidden in the web (such as bipartite cores of cyber communities and webrings[18]).

Developing a realistic and accurate stochastic model of the web graph is a challenging and relevant task for several other reasons:

- Testing web applications on synthetic benchmarks.
- Detecting peculiar regions of the WebGraph, i.e. local subsets that share different statistical properties with the whole structure.
- Predicting the evolution of new phenomena in the Web.
- Dealing more efficiently with large scale computation (i.e. by recognizing the possibility of compressing a graph generated accordingly to such model [1]).

*Dip. di Informatica e Sistemistica Università di Roma "La Sapienza" Via Salaria 113 00198 Roma Italy. E-mail: {laura,leon}@dis.uniroma1.it

†Sezione INFM and Dip. Fisica Università di Roma "La Sapienza", P.le A. Moro 2 00185 Roma, Italy. E-mail: gcalda@pil.phys.uniroma1.it

‡Institut de Physique Theorique Universitié de Lausanne, BSP 1015 Lausanne, Switzerland and INFM - Sezione di Torino Politecnico, Dip. di Fisica, Politecnico di Torino, C.so Duca degli Abruzzi 24, 10129 Torino, Italy. E-mail:PaoloDeLosRios@ipt.unil.ch

The study of the statistical properties of several observables in large samples of the WebGraph is at the basis of the validation of stochastic graph process models of the Web.

Kumar et. al. [18] and Barabasi and Albert [5] suggested that both the indegree and outdegree distribution of the nodes of the WebGraph follow a *power-law* distribution. Experiments on larger scale made by Broder et. al. [9] confirmed it as a basic web property. The probability that the indegree of a vertex is i is distributed by a power-law, Pr_u[in-degree(u)$= i] \propto 1/i^{\gamma}$, for $\gamma \approx 2.1$. The outdegree of a vertex is also distributed with a power law with exponent roughly equal to 2.7. The average number of edges per vertex is about 7.

In the same paper [9] Broder et. al. presented a fascinating pictures of the Web's macroscopic structure: a *bow-tie* shape, where the major part of the sites can be divided into three sets: a core made by the strongly connected components (SCC), i.e. sites that are mutually connected each other, and two sets (IN and OUT) made by the sites that can only reach (or be reached by) the sites in the SCC set. They also showed that for randomly chosen source and destination there is only 24% of probability that any path exists, and, in that case, the average length is about 16 (against the 19 of Barabasi [5]). The authors also find out that the WebGraph exhibits the *small world phenomenon* [24, 16] typical of dynamical social networks, only if the hyperlinks are considered undirected: within few edges almost all pages are reachable from every other page within a giant central connected component including about 90 % of the web documents.

A surprising number of specific topological structures such as bipartite cliques of relatively small size, from 3 to 10, has been recognized in the web [18]. The study of such structures is aimed to trace the emerging of a large number of still hidden *cyber-communities*: groups of individuals who share a common interest, together with web pages most popular among them. A bipartite clique is interpreted as a core of such community, defined by a set of fan, all pointing to a set of authorities, and the set of authorities, all pointed by the fans. Over 100.000 such communities have been recognized on a sample of 200M pages on a Web crawl from Alexa of 1997.

In a more recent paper Dill et al. [12] explain how the web shows a fractal structure in many different ways. Graph can be viewed as the outcome of a number of similar and independent stochastic processes. At various scales we have that there are "cohesive collections" of web pages (for example pages on a site, or pages about a topic) and these collections are structurally similar to the whole web (i.e. they exhibit the *bow-tie* structure and follow power-law for the indegree and outdegree). The central regions of such collections are called "Thematically Unified Clusters" (TUCs) and they provide a navigational backbone of the Web.

The correlation between the distribution of PageRank [8] (as computed in the Google search engine) and the in-degree in the web graph has also been considered [22]. They show that the PageRank is distributed with a power law of exponent -2.1, but very surprisingly there is very little correlation between the PageRank and the in-degree of pages, i.e. pages with high in-degree may have very low PageRank.

Most of the the statistical properties observed in the WebGraph cannot be found in traditional stochastic graph models. Firstly, traditional models such the random graph model of Erdös and Rényi are static models, while a stochastic model for the WebGraph evolves over time as new pages are published on the web or are removed from the Web.

Secondly, the random graph model of Erdös and Rényi fail to capture the self-similar nature of the web graph. Signature of the self similar behavior of the structure is the ubiquitous presence of power laws. Indeed, power law distributions have also been observed for describing the popularity (number of clicks) of web pages or the topological structure of the graph of the Internet [14, 10].

Aiello, Chung and Lu [2] proposed stochastic graphs appropriately customized to reproduce the power law distribution on the degree. They present a model for an undirected graph, meant to represent

the traffic of phone calls, in which the degree of the vertices is drawn from a power law distribution.

Albert and Barabasi and Jeong [3] started this study by presenting the first model of Evolving Network in which at every discrete time step a new vertex is introduced in the graph, and connects to existing vertices with a constant number of edges. A vertex is selected as the end-point of the an edge with probability proportional to its in-degree, with an appropriate normalization factor. This model shows a power law distribution over the in-degree of the vertices with exponent roughly -2 when the number od edges that connect every vertex to the graph is 7.

The Copying model has been later proposed by Kumar *et al.* [17] to explain other relevant properties of the WebGraph. First property is the amazing presence of a large number of dense subgraphs such as bipartite cliques. The Copying model is also an evolving model in which for every new vertex entering the graph one selects randomly a prototype vertex p amongst the ones inserted in previous timesteps. A constant number d of links connect the new vertex to previously inserted vertices. The model is parametric over a *copying factor* α. The end-point of the lth link, $l = 1, ..., d$, is either copied with probability α from the corresponding lth out-link of the vertex prototype p or it is selected at random with probability $1 - \alpha$.

The copying event is trying to model the formation of cyber communities in the web, web documents linking a common set of authorative pages for a topic of common interest. The model has been analytically studied and showed to hold power law distribution on both the in-degree and the number of disjoint bipartite cliques.

Very recently Pandurangan, Raghavan and Upfal [22] proposed a model based on the rank values computed by the PageRank algorithm used in search engines such as Google. They propose to complement the Albert,Barabasi and Jeong model in the following manner. There are two parameters $a, b \in [0, 1]$ such that $a + b \leq 1$. With probability a the end-point of the the lth edge is chosen with probability proportional to its in-degree, with probability b is chosen with probability proportional to its PageRank, with probability $1 - \alpha - \beta$ at random.

The authors show on computer simulation that with an appropriate fitting of the parameters the graphs generated capture distributional properties of both Page Rank and in-degree.

1.1 Our work

As a matter of fact the models presented so far correctly reproduce few observables as degree and Page Rank distribution.

We cite from the conclusions of the work of Dill et al. "There are many lacunae in our current understanding of the graph theoretic structure of the web. One of the principal holes deals with developing stochastic models for the evolution of the web graph (Extending [18]) that are rich enough to explain the fractal behavior of the web .....".

It is still missing a model that is rich enough to explain the self-similarity nature of the web and reproduce more relevant observables, for instance clique distribution, distance, connected components.

The recent study of Dill et al. [12] gives a picture of the web explaining its fractal structure as produced by the presence in the web of multiple regions generated by independent stochastic processes. The different regions being different in size and aggregation criteria, for instance topic, geography or Internet domain. All these regions are connected together by a "connectivity backbone" formed by pages that are part of multiple regions.

In fact all previous models present the Web as a flat organism, every page may potentially connect to every other page of the Web. This is indeed far from reality. We propose a "Multi-Layer" model in which every new page that enters the graph is assigned with a constant number of regions it belongs to and it is allowed to link only to vertices in the same region. When deciding the end-points of the edges

we adopt a combination of Copying and Evolving Network in the subgraph of the specific region. In particular, if an edge is not copied from the prototype vertex, its end-point is chosen with probability proportional to the in-degree in the existing graph. The final outcome of the stochastic process is the graph obtained by merging the edges inserted between vertices of all layers.

This model is explained in details in Section 2.

We then provide the results of an extensive experimental study of the distributional properties of relevant measures on graphs generated by several stochastic models, including the Evolving Network model [3], the Copying Model [17] and the Multi-Layer model introduced in this paper and make a comparative analysis with the experimental results presented in earlier papers and with the real data obtained from Notre Dame University domain on which the first analysis on the statistical properties of the web have been performed [5]. All models have been simulated up to 300,000 vertices and 2,100,000 edges.

1.2 Structure of the paper

This paper is organized as follows: in section 2 we introduce the Multi-Layer WebGraph model. In section 3 we describe the models we simulate and the collection of metrics that we have computed. In section 4 we show the experimental results.

2 The Multi-Layer model

In this section we present in details our model.

The model evolves in discrete time steps. The graph will be formed by the union of L regions, also denoted as layers. At each time step t a new page x enters the graph and it is assigned with a fixed number l of regions and d of edges connecting to previously existing pages.

Let $X_i(t)$ be the number of pages assigned to region i at time t. Let $L(x)$ be the set of regions assigned to page x.

We repeat l times the following random choice:

- $L(x) = L(x) \cup i$, where region i is chosen in $L/L(x)$ with probability proportional to $X_i(t)$ with a suitable normalization factor.

The stochastic process above clearly defines a Zipf's distribution over the size of the population of the regions, i.e. the value $X_i(t)$.

The d edges are evenly distributed (up to 1) between the l regions. Let $c = \lfloor d/x \rfloor$ and α be the copying factor. Consider each region i to which vertex x is assigned. Vertex x will be connected by c or $c+1$ edges to other vertices of region i. Denote by $\mathcal{X}$ the set of $X_i(t)$ vertices assigned to region i before time t. The layer i graph denoted by $G_i(t)$ is formed by the vertices of $\mathcal{X}$ and by the edges inserted before time t between edges of $\mathcal{X}$. We choose a prototype vertex p in $\mathcal{X}$. If we connect vertex x with c edges to region i, then for every $l = 1, ..., c$, with probability α, the lth edge is copied by the lth edge of vertex p in $G_i(t)$. Otherwise, the lth endpoint is chosen amongst those vertices in $\mathcal{X}$ not already linked by x with probability proportional to the in-degree in $G_i(t)$ (plus 1) with a suitable normalization factor. If $c+1$ edges need to be inserted and the prototype vertex is connected with only c edges, the $(c+1)$th edge is chosen with probability proportional to the in-degree.

The resulting graph has Edge set given by the union of the edges of all layers.

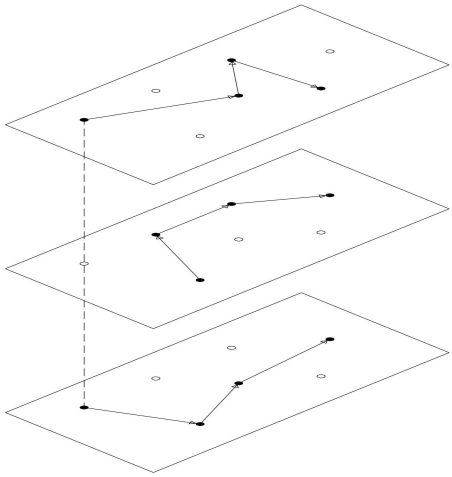

Figure 1: A Multi-Layer view of a graph.

3 Data Set and simulated models

We study the graph properties of several data sets of about 300k vertices and about 2,1 million edges, therefore yielding an average degree of 7. In particular, we consider the following data sets:

1. **ABJ** A graph generated according to Evolving Network model by Albert, Barabasi and Jeong model [3].

2. **ACL** A graph generated according to the Aiello,Chung and Lu model [2].This model is for undirected graphs. To generate a directed graph we modified the model in the following way: we build 2 sets, OS and IS. In OS we put 7 copies of each vertex (to simulate the average out-degree of 7), while in IS we put $deg(v)$ copies of each vertex. $deg(v)$ is chosen accordingly to a power law distribution with exponent -2.1. Then we randomly match the two sets (i.e. every element in the set OS is matched with one element of IS).

3. **CL0N** A set of graphs generated according to the Kumar *et al.* model [17], with copying factor $\alpha = 0.N$: e.g. CL03 has $\alpha = 0.3$.

4. **ER** A graph generated according to the Random Graph model of Erdös and Renyi model.

5. **ML0N-**$\#L$ A set of graph generated according to the Multi-Layer model presented in this paper, with $\alpha = 0.N$ (*copy factor*). $\#L$ indicates the the number of layers, and, unless otherwise specified, we have 3 layers per page, with out-degree respectively 3,2,2 (that sum up to 7, the average out-degree of the Web). ML03-25 indicates a graph generated with $\alpha = 0.3$ and 25 Layers.

6. **SWP** A graph generated according to the small-world model [26]: we considered a bi-dimensional lattice and added three "distant" random node to every node.

7. **NotreDame** The nd.edu subnet of the Web [3].

We compare the synthetic data obtained by the models listed above and the **NotreDame** data set on the following measures:

a The degree distribution $P(k)$ giving the frequency of a certain degree k in the graph.

c The number of disjoint bipartite cliques in the graph. A bipartite clique (k, c) is formed by k vertices all connected by directed links to each of the c vertices. We implemented the method described in [18] to estimate the number of disjoint bipartite cliques in the graph.

b The number of vertices $N(d)$ within distance d from a certain vertex v_0 in a graph obtained by removing the orientation of the edges.

d The clustering coefficient [20] measures at which extent the neighbors of a vertex are connected each other. It is defined in the following way: consider a vertex v of the graph $G = (V, E)$. Now consider the k neighbors of the vertex v i.e. the vertices $v_1..v_k$ such that (v, v_i) is an edge of the graph G. The clustering coefficient is the average over all vertices of the graph of the ratio between the number of edges (v_i, v_j), with $i, j = 1..k$, that belongs to E, and its maximum possible value $k(k-1)/2$, that is the number of edges in the complete graph made with vertexes $v_1..v_k$. This measures of how much the neighbors of a vertex are connected each others. We measure both the directed Clustering Coefficient (C_d) than the the undirected Clustering Coefficient (C_u) of the graph obtained by removing the orientation of the edges.

4 Analysis of results

The simulation results are summarized in Table 1. In the first and second column we report the number of disjoint bipartite cliques $(3, 3)$ and $(4, 3)$. The number of cliques observed in the NotreDame data set is considerably smaller than that observed in CL07 where the power law distribution with exponent -2.1 is obtained. In [18] Kumar et al. analyze a 200M web pages sample from Alexa, and they measure 29769 (4,3) cores, that are four times those present in the CL07 set and ten times the number in the NotreDame set (but both CL07 and NotreDame have only 300k nodes). We also notice that the number of cliques in the Copying Model grows by orders of magnitude vs α. The Multi-Layer model shows a number of cliques that grows vs α. All other models do not show any consistent number of cliques.

Next in Table 1 we report the values of Clustering Coefficients. It is interesting to notice that, exactly as the number of cliques, the Clustering Coefficients for the Copying Model also grow by orders of magnitude vs α. In the Multi-Layer Model the Clustering Coefficients grow vs α and decreases vs L.

The last column of Table 1 reports the measure of the fitting of the distribution of the in-degree in the log log scale. Figures illustrating the frequency of the indegree in the graph generated by the various models and in the NotreDame subdomain are reported in appendix.

The NotreDame data set indegree follows a power law with exponent -2.1 as reported by several observation of the web graph. The evolving network exhibits a power law with in-degree -2.0. The ACL model is explicitly constructed with a power law distribution of exponent -2.1.

It is interesting to notice that the Copying model follows a power law distribution only when the copying factor *alpha* approaches 0.5, but only with a fairly large $\alpha \simeq 0.7 \div 0.8$ we have a slope close to -2.1. The Multi-Layer model in-degree follows a power law with exponent of -2.1 within a fairly large variation of the copying factor $\alpha \in [0.3, 0.7]$ and of the number of layers $L \in [25, 100]$.

In Figures 2 and 3 we see that all the simulated models and the NotreDame data set show a small world phenomena. All vertices of the graph are reachable within a distance of a few edges when the orientation of the edges is removed. This has already been observed for instance in [9].

	$(3,3)$	$(4,3)$	CC_d	CC_u	γ
ABJ	–	–	0.098559	0.076166	≈ 2
ACL	–	–	0.000269	0.000147	≈ 2.1
CL03	594	31	0.000489	0.000084	–
CL05	7322	1121	0.001782	0.000502	–
CL07	34420	7799	0.017311	0.008977	≈ 2.1
ER	–	–	0.260056	0.186878	–
ML03-25	6	1	0.037614	0.026046	≈ 2.1
ML03-50	21	3	0.034382	0.023676	≈ 2.1
ML03-100	25	4	0.033624	0.023507	≈ 2.1
ML05-25	66	3	0.066144	0.050971	≈ 2.1
ML05-50	82	7	0.059395	0.045663	≈ 2.1
ML05-100	117	14	0.053126	0.041604	≈ 2.1
NotreDame	6022	3154	0.324038	0.259829	≈ 2.1
SWP	–	–	0.000018	0.000017	–

Table 1: Data Sets

5 Conclusion and further work

In this work we present a Multi-Layer model of the Web graph and an attempt to compare several models of the web graph and real data set. We plan to further develop the idea of a Multi-Layer Model for the WebGraph and compare the simulated models with larger and different samples of the Web. We also plan to compare data set on other relevant metrics such as size of strongly connected components.

Our effort is aimed to design a model that resembles the complex nature of the Web Graph.

References

[1] M. Adler and M. Mitzenmacher. Towards compressing Web graphs. U. of Mass.CMPSCI Technical Report 00-39.

[2] W.Aiello, F. Chung, L.Lu. A random graph model for massive graphs. *Proc. ACM Symp. on Theory of computing*, pp.171-180, 2000.

[3] R. Albert, H. Jeong and A.L. Barabasi *Nature* **401**, 130 (1999).

[4] J.R. Banavar, A. Maritan and A. Rinaldo, *Nature* **399**, 130 (1999).

[5] A.L. Barabasi and R. Albert, Emergence of scaling in random networks *Science* **286** 509, (1999).

[6] B. Bollobas. Random Graphs, *Academic Press, London*, (1985).

[7] B. Bollobas, F.Chung. The diameter of a cycle plus a random matching. *SIAM Journal of Discrete Maths 1:328-333*, (1988).

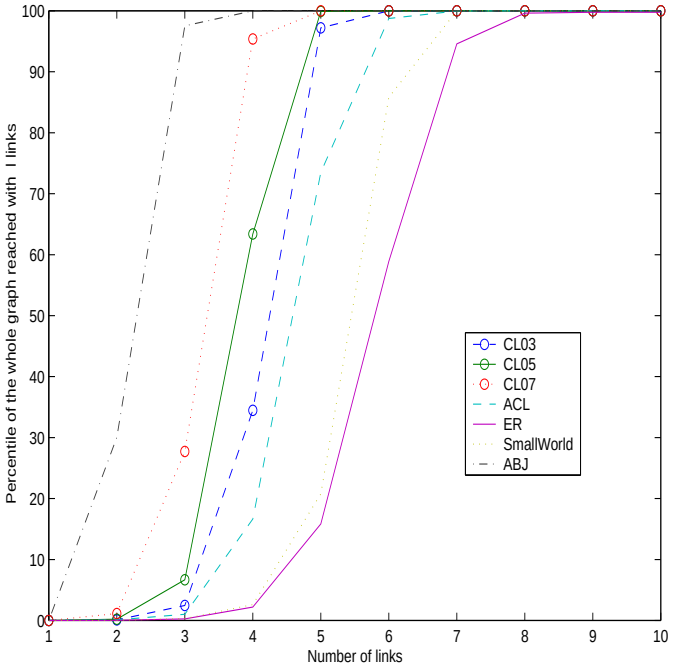

Figure 2: Distance of portions of the graph.

[8] S.Brin and L.Page. The anatomy of a large-scale hypertextual Web search engines. In *Proceedings of the 7th WWW conference, 1998*

[9] A.Broder, R.Kumar, F. Maghoul, P.Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener, *Graph structure in the web*

[10] G. Caldarelli, R. Marchetti, L. Pietronero, *Europhys. Lett.***52**, 386 (2000) .

[11] C. Cooper, A. Frieze. A general model of web graphs. (2001)

[12] S. Dill, R. Kumar, K. McCurley, Sridhar Rajagopalan, D. Sivakumar, A Tomkins. Self-similarity in the web. (2001)

[13] P. Erdös and R. Renyi, *Publ. Math. Inst. Hung. Acad. Sci* **5**, 17 (1960).

[14] M. Faloutsos, P. Faloutsos and C. Faloutsos. On Power-Law Relationships of the Internet Topology. *ACM SIGCOMM* (1999).

[15] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. Journal of the ACM, vol. 46 n. 5 pag 604-632 (1997).

[16] J.Kleimberg. The Small World Phenomenon: an algorithmic perspective.

[17] R.Kumar, P.Raghavan, S.Rajagopalan, D. Sivakumar, A. Tomkins, E. Upfal. Stochastic models for the web graph.

[18] S.R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the Web for Emerging Cyber Communities. *Proc. of the 8th WWW Conference*, pp. 403-416, (1999).

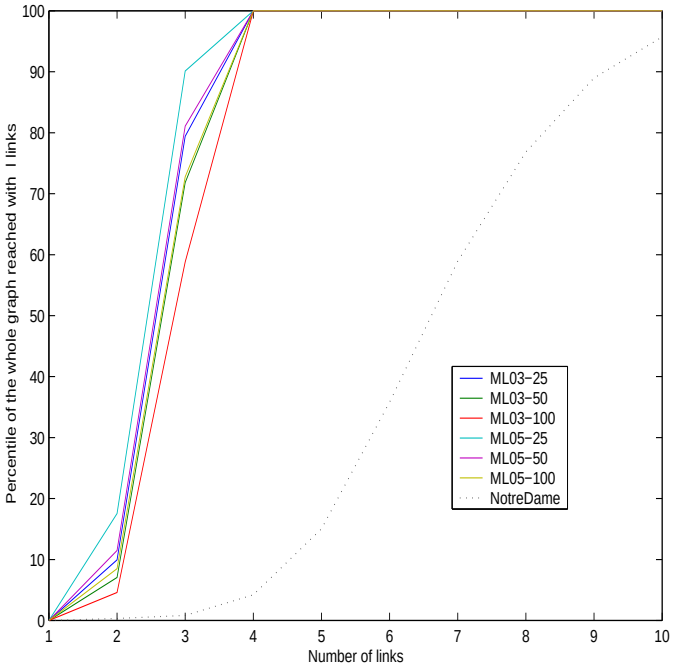

Figure 3: Distance of portions of the graph.

[19] T.A. McMahon and J.T. Bonner, On Size and Life Freeman, New York (1983).

[20] M.E.J.Newman, Models of the small world. *J. Stat. Phys. 101, 819-841 (2000).*

[21] C.H. Papadimitriou, Algorithms, Games and the Internet. STOC'01

[22] G. Pandurangan,P. Raghavan,E. Upfal. Using PageRank to Characterize Web Structure.

[23] I. Rodriguez-Iturbe and A. Rinaldo, Fractal River Basins, Chance and Self-Organization, Cambridge University Press, Cambridge (1997).

[24] D.Watts, S. Strogatz, Collective Dynamics of small-world networks, *Nature***393** 440, (1998).

[25] G.B. West, J.H. Brown and B.J. Enquist, *Science* **276**, 122 (1997).

[26] D. J. Watts and S. H. Strogatz, *Nature* **393**, 440 (1998).

[27] A description of the project, together with maps done by Cheswick B. Burch H.,is available at http://www.cs.bell-labs.com/who/ches/map/index.html.

Log-log plot of the indegree distribution.