Proceedings of the 2006 Winter Simulation Conference
A. Kogias and D. Anagnostopoulos

A METHODOLOGY FOR THE EVALUATION OF WEB GRAPH MODELS AND A TEST CASE

Antonios Kogias
Dimosthenis Anagnostopoulos

70 El. Venizelou Str.
Harokopio University of Athens
Athens, 17671 GREECE

ABSTRACT

Valid models of the WWW are important for creating WWW-like representations, upon which new algorithms and applications for searching, indexing, compression etc. can be tested, but also for predicting the evolution of the web and the emergence of important new phenomena. In this study we introduce a validation process for web-graph models and use it to analyze the behavior of the Exponential Growth Copying Model, which has been explicitly designed to model the WWW. We study the effect of individual parameters on its effectiveness, suggest appropriate parameter values for the creation of web-like graphs and indicate inherent deficiencies of the model.

1 INTRODUCTION

The World-Wide Web (WWW) has shown a tremendous growth in late years and estimates of its size are currently at the billion web-pages scale. No crawl or search engine can chart its entirety, a problem that is magnified by its ever increasing dynamic content. Therefore, it has become an extremely tedious task for researchers to obtain and manage real-world data (i.e. the web itself). The most promising solution for this problem is the use of models that create realistic representations of the WWW, where new algorithms and applications (for searching, compression, etc.) can be tested. These models can further enhance our understanding of the sociology of content creation on the web, our predictions of its evolution and the emergence of important new phenomena. The need for valid models of its structural evolution is much more pronounced today; existing technologies need to be thoroughly tested prior to deployment and future needs to be accurately predicted and taken into consideration.

In detail, the development of realistic and accurate stochastic models for the web-graph could enhance:

- Testing web applications with synthetic benchmarks (Laura et al. 2002).
- Detecting peculiar regions of the web-graph (local subsets that share different statistical properties with the whole structure).
- Analyzing the behavior of search algorithms that make use of link information, e.g. PageRank (Brin and Page, 1998), HITS (Kleinberg, 1998).
- Designing crawling strategies.
- Predicting the evolution and the emergence of important new phenomena in the web.
- Dealing more efficiently with large scale computations (e.g. by recognizing the possibility of compressing a graph generated by such a model).

In (Adler and Mitzenmacher, 2001) a characteristic case demonstrating the usefulness of web-graph models is presented: the problem of efficient algorithms for the compression of graphs with the link structure of the WWW. Motivated by the random graph models proposed in (Kumar et al. 1999b), the authors devised a compression algorithm based on finding similarities among the links of web-pages and tested it on graphs created by the model and compared its effectiveness with other widely used compression schemes, thus verifying its suitability for real web data.

In this study we introduce a validation process for web-graph models based on empirical data and use it to analyze the behavior of the Exponential Growth Copying Model (EGC), which has been explicitly designed to model the WWW (Kumar et al. 2000). We study the effect of individual parameters on its effectiveness, suggest appropriate parameter values for the creation of more web-like graphs and indicate inherent deficiencies of the model.

2 WORLD-WIDE WEB

The WWW is traditionally modeled as a graph, the so-called web-graph, where each static HTML page is a vertex and each hyperlink an edge of this graph (either directed or undirected). Directed edges are defined by vertex of origin (tail) and vertex of destination (head), whereas

undirected edges are defined by the two vertices they connect. For undirected graphs, the degree of a vertex is the number of distinct edges incident at the vertex. For directed graphs, the out (in) degree of a vertex is the number of edges having this specific vertex as tail (head).

The most widely accepted structural characteristic of the web-graph, reported by various researchers in web crawls (Broder et al. 2000; Kleinberg et al. 1999; Kumar et al. 1999a), is the existence of power-laws for vertex degrees. The power-law for in-degree states that "the probability that a vertex has in-degree i is proportional to i^{-x}, for some $x > 0$". The power-law for out-degree is similar, though for a different value of x. The values of x for in-degree power-law (x_{in}) and out-degree power-law (x_{out}) have been reported to be $x_{in} = 2.1$ and $x_{out} = 2.7$ (Broder et al. 2000). The existence of power-laws for vertex degrees and the corresponding exponent values are widely accepted in the literature as salient WWW characteristics.

In (Broder et al. 2000) the authors present, to the best of our knowledge, the single most extensive and in-depth analysis of large-scale WWW characteristics, based upon two Altavista crawls (May and October 1999), consisting of over 200 million pages and 1.5 billion links each. They confirm that the power-law exponent for in-degree is ≈ 2.1 and for out-degree ≈ 2.7 (using the over 1 billion distinct links of each crawl).

That study also introduced the large-scale structure of the web-graph: the "bow-tie" shape. The (directed) web-graphs in it comprise of:

- SCC (Strongly Connected Component), consisting of all vertices reachable through directed paths from each other.
- IN component, consisting of all vertices that can reach the SCC via directed paths but are themselves unreachable from it.
- OUT component, consisting of all vertices that are reachable from the SCC via directed paths but cannot reach it themselves.
- TENDRIL components, consisting of vertices that are either reachable from the IN but cannot reach the SCC and the OUT components, or can reach the OUT but cannot be reached from the SCC and the IN components.
- DISCONNECTED components, consisting of vertices that do not belong in any of the above components.

It was noted that sizes of the SCC, IN, OUT and TENDRIL components are comparable (about 21% - 28% of the total number of vertices), while the DISCONNECTED component is much smaller (about 8% of the total number of vertices). This means that over 75% of the time there exists no directed path from a random start vertex to a random finish vertex; but if it does, its length was estimated to be 16 on average. Another issue explored was the diameter of the web-graph (defined as the maximum shortest path between any two vertices that such a path exists) by breadth-first searches a number of randomly chosen start points; it was estimated about 905.

In a preceding paper (Kumar et al. 1999a), a study of a 1997 web-crawl had been presented, emphasizing in the number of bipartite cores existing in the web-graph as signatures of emerging cyber-communities. A bipartite core $C_{i,j}$ consists of i vertices (fans) that all point to the same j vertices (centers). Among the about 200 million web-pages of the crawl, they found more than 130 thousand bipartite cores $C_{i,j}$ of fans $i \geq 3$ and centers $j \geq 3$. They also proved that the copying models they proposed (such as the EGC) are rich in such micro-structures.

The existence of power-laws in vertex degree distributions has been also recognized in various other network graphs. In (Faloutsos Faloutsos and Faloutsos, 1999) it was empirically shown that certain properties of the AS-level Internet topology are well described by power-laws. The power-law exponent was computed by linear regression of degree frequencies on the logarithmic scale. Because these distributions are heavy-tailed, calculations of best linear fit were restricted to the top roughly 75% of degree frequencies.

In (Bu and Towsley, 2002) it is stated that one should not attempt to fit a power-law to a degree frequency distribution unless sure that it is indeed a power-law distribution and not some other heavy-tailed one. The empirical complementary distribution (ecd) of vertex degree frequencies is proposed as the criterion of the existence of power-law: if it is a straight line, then it is a power-law. Analytically, let f(d) be the fraction of vertices with degree d; the ecd is $F(d) = \Sigma f(i)$ $\{i = d \text{ to } \infty\}$, i.e. the fraction of vertices with degree equal or greater than d.

3 EXPONENTIAL GROWTH COPYING MODEL

Previous to the extensive contemporary research of WWW structure, traditional random graphs had been considered adequate for its modeling. It became apparent though that these models do not give birth to power-law degree distributions, so other models were proposed, in varying degrees of complexity. For a brief review of the most typical random graph models for the web, please see (Kogias, Nikolaidou and Anagnostopoulos, 2005). Presently we describe the Exponential Growth Copying (EGC) Model (Kumar et al. 2000) which was used as the test case of our methodology.

Evolving Copying Models in general have been explicitly designed to model the WWW. It has been shown that they have a large number of complete bipartite subgraphs, as has been observed in the crawls, whereas several other models do not. Their development was based on the following very realistic intuitions about the WWW:

- Although some page creators may create content and links to other pages regardless of the already represented topics on the web, many will be drawn to existing topics of interest and link to pages within some of these existing topics (Kumar et al. 1999b).
- Due to the exponential growth of the WWW, a page creator will not "see" the most recent "epoch" of pages (i.e. will not be aware of the existence of pages created most recently) (Kumar et al. 2000).

All Copying Models incorporate the first intuition, but the EGC is the only one that incorporates both. It has been proven analytically (Kumar et al. 2000) that the graphs created by the EGC model follow some power law for in-degree with a bounded exponent and that they also contain a large number of bipartite cliques. Both conclusions agree with real WWW observations (Kumar et al. 1999b). For a detailed description of the graph creation algorithm please see (Kumar et al. 2000). The EGC model is formally described by four parameters:

- "growth" factor $p \in (0 , 1]$, used in the typical binomial distribution
- "self-loop" factor $\gamma > 1$, defines the initial "attractiveness" of each vertex and is used to control the amount of attractiveness gained by a new edge to or from a vertex
- "tail-copy" factor $\gamma' \in (0, 1)$, provides a way of tuning the out-degree distribution
- "natural link" factor $d > 0$, is the amount of natural (non self-loop) edges that are added to the graph for each new vertex

The EGC model is not hierarchical; it does not try to use structural properties of the WWW (e.g. web-sites that contain web-pages that have hyperlinks within and without the same web-site) to produce a web-like graph. Instead, it provides an evolutionary framework, based on realistic intuitions about the web, to capture its macroscopic structural characteristics. In a study of network models (Tangmunarunkit et al. 2002) it is concluded that degree based models (although using minimal information about the system they are trying to picture) behave substantially better than hierarchical models (that need a lot of information to start with and have more complicated algorithms) in picturing networks with loose (not strict) hierarchies (e.g. the web); however it is suggested that they be used only when the number of vertices is substantially large.

Compared with other models, EGC provides a very convenient way of adaptation to real-time evolution studies of the WWW. All other models grow by one vertex at each time-step, a fact that doesn't help when one must define how much real time passes at each time-step. Even if that was possible, since the WWW's growth is approximately exponential, the real-time duration of each time-step should be adjusted to reflect the onslaught of new vertices arriving in incrementally smaller real-time intervals. The EGC model suffers from no such drawbacks; one has only to decide the real time equivalent of one model's "epoch" and adjust the growth-factor p accordingly.

4 MODEL EVALUATION FRAMEWORK

In principle, there are two approaches that may be employed for model validation: analytical solution or simulation. Being more exact, the former approach is preferable, but it cannot always be employed because of the complexity of models. For instance, it is analytically proven that the EGC model follows some power law, with a bounded exponent for in degree distribution; the same is not proven for out-degree, due to the model's complexity of edge creation. Simulation can overcome these difficulties, but requires special attention to the selection of experimentation parameters and output analysis. As a "what-if" type of investigation, simulation may be used to narrow the search space of the problem under study and accordingly focus on specific parameters. In web graph models' validation, simulation may determine whether model results conform with empirical observations and, at a second stage, refine (i.e. appropriately parameterize) proposed models to exhibit greater efficiency. Finally, a valid model can be further used for studying additional features of the real web, such as diameter, number of small structures, clustering, components etc.

The established way of obtaining WWW data is by using web crawlers (a.k.a. spiders or robots) that run a continuous loop of downloading web pages, extracting URLs and in turn attempting to download these also. When a specified amount of time has passed or a specified amount of pages has been downloaded, the crawler terminates and exports the data set, which is subsequently processed to extract various statistics.

However, web crawlers have shortcomings too. They do not provide a WWW snapshot per se, as they cannot download each and every URL simultaneously; thus the temporal granularity they provide is usually very coarse. Furthermore, they cannot map the whole web: they cannot reach pages if there are no page references, web-site administrators can forbid them from entering their web-sites, they cannot capture page changes that occur in time intervals smaller than the crawling duration and are constrained by available secondary storage. Therefore, crawls provide a data set that is not the real web, but a sample -hopefully- big enough to draw valid conclusions from.

On the other hand, a valid model of the system in question may resolve many issues through simulation. Even if various models for the web-graph have been pro-

posed, we are still far from a widely accepted valid model. The optimal solution would be to compare model results with crawls of the whole WWW; however, as previously explained, this is infeasible. The usually adopted approach is to compare model results with various crawls; this is also cumbersome because the model needs to produce results of similar size to the crawl – and numerous runs must also be completed before any comparison is made, for the results to obtain statistical validity.

In this study we propose the framework of comparing the model's results against the valid statistical features of the real web, which are consistent across various crawls (Barabasi and Albert, 1999; Broder et al. 2000; Kleinberg et al. 1999; Kumar et al. 1999a; Kumar et al. 1999b; Laura et al. 2002), i.e. in and out degree power law distributions. Furthermore, we explore the compliance with other, less widely recorded, statistical properties of the WWW; namely large-scale structure, bipartite cores and diameter.

5 EGC MODEL SIMULATION

The EGC model needs to be appropriately adjusted to represent the WWW with as much effectiveness as possible. In this simulation-based study, we provide answers to the following issues:

- As the model always creates power-law distributions for vertex in-degrees with a bounded exponent, for which parameter values this exponent is realistic (-2.1)?
- As it is unknown whether the model creates power-law distributions for vertex out-degrees, is it possible to obtain them? If yes, what are the parameter values that the power-law exponent is realistic (-2.7)?
- Are the above parameter constraints sensitive to the initial graph size?
- Are the desired power-laws resilient over different magnitudes of final graph sizes?
- Is it possible, for any value of the growth factor p, to find proper values for the other parameters in order to obtain the desired power-laws?
- Based on results of experiments for the above parameter values, what predictions does the model offer for the structural properties (diameter, components, bipartite cores) and do these predictions agree with real WWW observations?

This paper extends and completes a preliminary study (Kogias, Nikolaidou and Anagnostopoulos, 2005) concerning the effect of EGC model parameters. The conclusion was that the EGC can easily produce power-laws for vertex in-degrees but not so easily for out-degrees. Parameter p was found to almost surely have no effect on the existence and appropriateness of power-laws for node degrees. Parameter d was found that degree distributions approached the desired power-laws when its value increased. Parameter γ was found that degree distributions approached the desired power-laws when its value decreased. Parameter γ' was found that out-degree distributions were very sensitive to it, only approaching power-law for its greater values. The present in-depth study focuses on the following:

- Experiments consist of 30 runs, thus strengthening statistical properties, with incremental final graph sizes up to 1,000,000 vertices.
- Parameter p is studied for the same values, so as to certify the model's appropriateness for real-time simulation (as previously explained).
- Parameter d in the WWW was found in 1999 (Broder et al. 2000; Kleinberg et al. 1999; Kumar et al. 2000) to be about 7. Thus we attempt to achieve the best power-law approximation using this value in all experiments, for different values of the other parameters that cannot be empirically measured.
- Parameter γ is a positive integer and defined $\gamma > 1$. Using the preliminary study's findings that low values increase the model's efficiency, we expect best results for values 2 and 3. Value of 1 degrades the model to a simple preferential attachment approach, but since some kind of interesting behavior might emerge in this degenerate case, we choose to use it in our experiments. For the sake of completeness, we also use the values 4 and 5.
- Parameter γ' defines the existence of power-law for vertex out-degree distributions and fine-tunes the asymmetry between the in-degree and out-degree power-law exponents. Using the preliminary study's findings that high values increase the model's efficiency, we try to pinpoint the exact area in the 0.5 to 0.95 range that this efficiency reaches its peak. Thus we use the value set {0.05 0.25 0.5 0.55 0.6 0.65 0.7 0.75 0.8 0.85 0.9 0.95}.

The EGC model was implemented in ANSI C for Windows 32bit and all the experiments were run in a Intel Xeon 3 GHz CPU with 1 GB RAM system. We used the Mercenne Twister (Matsumoto and Nishimura, 1998) random number generator whose C implementation is freely distributed for non-profit use. Each experiment consisted of 30 runs of the EGC algorithm for final graph size of 1,000,000 vertices starting from an initial graph of 1 vertex with γ self-loops (unless otherwise noted). Self-loops and duplicate edges are deleted prior to the computation of each run's results. This is a necessary step because we rather consider parameter γ as an evolutionary conditioner of the model than an intrinsic characteristic of the real WWW; therefore its value must be carefully tuned to pro-

vide vertices with the appropriate initial attractiveness. We average in-degree and out-degree frequency and ecd distributions for all runs of each experiment. Next we use either MATLAB or StatGraph for computing best-fit linear regressions in log-log plots. In degree ecd plots we also compute means and variations of the error of the fit, which will define the existence of power-laws in the corresponding degree distributions (Bu and Towsley, 2002). In degree frequency plots linear regression is done for the about 75% bottom distinct degree values, to diminish the "heavy tail" effect (Faloutsos Faloutsos and Faloutsos, 1999). For parameter values that both in-degree and out-degree ecd indicate the existence of power-laws (linear fit mean error ≈ 0 and error variation ≈ 0) and both power-law exponents are close to the desired ones (2.1 and 2.7 for in- and out- degree respectively), we do a further examination as follows. We explore the influence of both initial graph size (experiments with initial size of 1,000 vertices) and final graph size (experiments with final size of 1,000 50,000 100,000 and 500,000 vertices) and measure various structural graph characteristics to achieve accurate prediction of their values in graphs larger than 1,000,000 vertices; specifically, predict the values of these characteristics for final graph size of 200,000,000 vertices, which is the case where these same characteristics were measured on the real WWW (Broder et al. 2000).

6 SIMULATION EXPERIMENTS RESULTS

6.1 In and Out Degree Distributions

Using an initial graph of ***1*** vertex with γ self-loops, final graph of ***1,000,000*** vertices and parameter $d = 7$, we made ***180*** experiments for all possible combinations of the values of the other parameters $\{p, \gamma, \gamma'\}$. Each experiment consists of ***30*** runs and each run uses a different value for initialization of the random number generator. We average and compute frequency and ecd distributions for in- and out-degrees over all runs of an experiment. We use *MATLAB* to perform linear regression and find the best linear fit in log-log plots, then compute error means and variances of these fits. If the linear fit of the ecd is good (error mean less than ***1*** and error variance less than ***0.1***), we surmise that the respective frequency distribution is a power-law and the exponent computed via linear regression is valid. We present the parameter values that we found to produce power-laws as well as the corresponding power-law exponents computed by the linear fit procedure in Table 1 (called 'optimum' henceforth) and an instructive out-degree log-log linear fit plot in Figure 1.

Table 1: Best Parameter Values and Corresponding Power-law Exponents of Degree Frequency Distributions for d = 7 and Final Graph of 1 million Vertices

p	γ		γ'		in-degree power-law exponent		out-degree power-law exponent	
0.05	2	3	0.70	0.80	2.08	2.24	2.73	2.63
0.50	2	3	0.65	0.70	2.04	2.16	2.73	2.73
0.95	2	3	0.60	0.70	2.00	2.12	2.78	2.66

Conclusions:

- Values of $\gamma > 3$ or $\gamma < 2$ create graphs that do not follow the desired power-laws for in- and out- degree frequency distributions simultaneously.
- Values of $\gamma' < 0.6$ create graphs that do not follow power-laws for out-degree frequency distributions at all (in-degree frequency distributions are not influenced by this parameter). Best results are for values of γ' in the range ***0.6*** to ***0.8*** and seem to be slightly influenced by the value of γ: if $\gamma = 2$ then γ' should be around ***0.65*** but if $\gamma = 3$ then γ' should be around ***0.75***.
- For all values used for ***p,*** there are values for γ and γ' that create graphs with the desired power-laws for in- and out- degree frequency. These values do not differ much between varying values of ***p***.

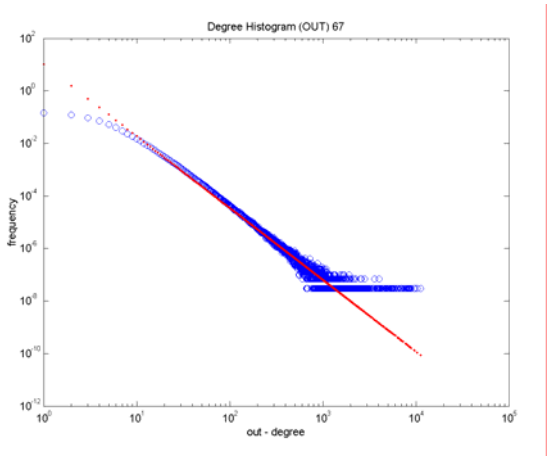

Figure 1: Linear Fit of out-degree Frequency log-log Plot for p = 0.50 γ' = 0.70 and γ = 3

6.2 Sensitivity to Initial Graph Size

For the optimum parameter values we made **6** additional experiments with the distinction that the initial graph has **1,000** vertices with **γ** self-loops each. Each experiment consists of **30** runs for a final graph of **1,000,000** vertices (results analysis and linear regressions as in subsection 6.1). The conclusion was that, not only out-degree fre-

quency distributions, but also in-degree frequency distributions do not follow power-laws when the initial graph is large (instructive exhibit Figure 2).

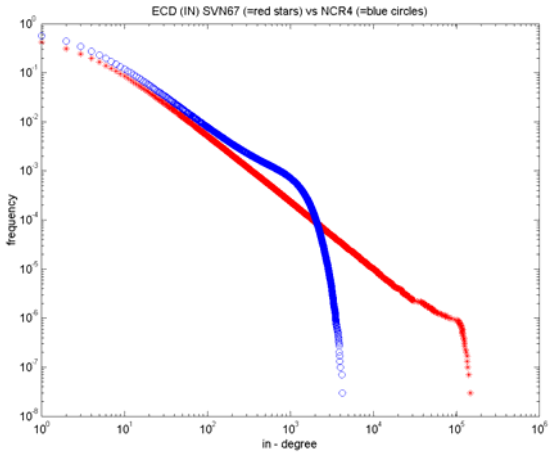

Figure 2: Plot of in-degree Frequency ecd for p = 0.50 γ' = 0.70 γ = 3 and Initial Graph of 1 (red stars) and 1,000 (blue circles) Vertices

6.3 Resiliency over Different Sizes of the Final Graph

To study the resilience of the various statistical properties over different magnitudes of the final graph size (and to also provide information for the later prediction activity) we made ***174*** more experiments of ***30*** runs each, for the optimum parameter values and initial graph of ***1*** vertex and ***γ*** self-loops. We varied the size (vertices) of the final graph in the set {***10,000 50.000 100,000 150,000 200,000 250,000 300,000 350,000 400,000 450,000 500,000 550,000 600,000 650,000 700,000 750,000 800,000 850,000 900,000 950,000 1,000,000***} and we analysed the results with *StatGraph* for the statistical properties of interest.

6.3.1 In and Out Degree Distributions

Power-law exponent values show stability and vary very close to the values presented in Table 1. We perform linear regression to discover the relationship of these exponents with final graph size, which outlines a logarithmic linear model of interdependence. Using this model, we predict that in- and out- degree frequency power-law exponents for final graph size of 200 million vertices are very close to those of 1 million vertices. Therefore we surmise that the EGC model is very stable on account of in- and out- degree frequency distributions, when its parameters take the values indicated in Table 1 or close to them. The general conclusion is that, for any value of parameter p, there exist values for the other parameters, so as the desired power-laws for vertices degree frequency distributions to exist in the final graph, regardless of its size. In Table 2 we present the predictions for power-law exponents.

Table 2. Prediction of Power-law Exponents for in- and out- Degree Frequency Distributions for the Optimum Parameter Values and Final Graph of 200 million Vertices (95% Confidence Interval)

p	**γ**		**γ'**		**in-degree power-law exponent**		**out-degree power-law exponent**	
0.05	2	3	0.70	0.80	2.13	2.29	2.77	2.73
0.50	2	3	0.65	0.70	2.08	2.22	2.78	2.81
0.95	2	3	0.60	0.70	2.00	2.17	2.81	2.72

6.3.2 Exhaustive Count of Bipartite Cores

The exhaustive count of bipartite cores $C_{i,j}$ ($i \geq 3$, $j \geq 3$) in the final graphs of all these experiments allows us to perform linear regression and establish a prediction model for a final graph of 200 million vertices. The best prediction model found was a multiplicative linear fit model and its predictions agree to the observed count of such bipartite cores in the real WWW (about 200,000 bipartite cores). Therefore we surmise that the EGC not only delivers in producing graphs much wealthier in micro-structures than all other models, but also in adequate quantity. We present predictions of bipartite cores count on a final graph of 200 million vertices in Table 3.

Table 3. Prediction of Bipartite Cores Count for the Optimum Parameter Values and Final Graph of 200 million Vertices (95% Confidence Interval)

p	**γ**		**γ'**		**$C_{i,j}$ (i≥3, j≥3) count (thousands)**	
0.05	2	3	0.70	0.80	204	346
0.50	2	3	0.65	0.70	237	212
0.95	2	3	0.60	0.70	172	375

6.3.3 Size of Large-scale Structural Components

Size of the structural components of the final graph is a major point that the EGC fails to even approach the desired values. In all experiments the SCC component's size is in giant proportion to the other components, resulting in the vast majority of vertices belonging to the SCC (and consequently the graph has a very small diameter). Linear regression of size of the SCC, IN, OUT and REST (all vertices that do not belong to the other components) vs. size of the final graph produces a multiplicative model of prediction as the best linear fit, which we use to predict component sizes in a final graph of 200 million vertices. The

EGC effectiveness towards these structural characteristics, however undesirable, appears stable. We present in Table 4 the predictions of SCC, IN, OUT and REST components' size and an instructive prediction plot in Figure 3.

Table 4. Prediction of SCC, IN, OUT and REST Components Sizes for the Optimum Parameter Values and Final Graph of 200 million Vertices (95% Confidence Interval)

p	**γ**		**γ'**		**Component Size (millions)**							
					scc		**in**		**out**		**rest**	
0.05	2	3	0.70	0.8	203	189	5	3	5	7	7	11
0.50	2	3	0.65	0.7	195	179	21	20	2	3	7	9
0.95	2	3	0.60	0.7	192	178	31	26	1	3	6	10

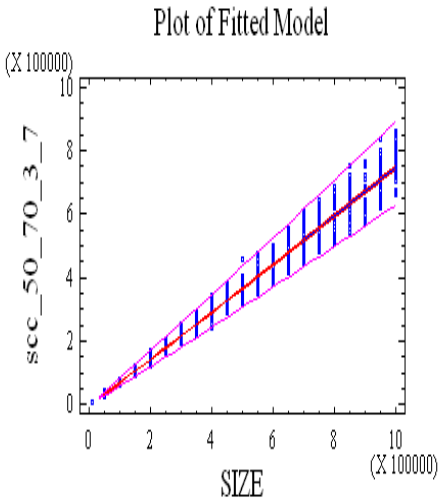

Figure 3: Plot of Linear Regression for Best Multiplicative Model Fit of SCC Component Size vs. Final Graph Size for p = 0.5 γ' = 0.7 and γ = 3

6.3.4 Diameter

The existence of a giant SCC component indicates that the diameter of the final graphs is too small because the graphs are too well connected. We certify this fact by running 30 experiments for the optimum parameter values but with final graph sizes of 10,000 50,000 100,000 500,000 1,000,000 vertices. Because diameter finding experiments are highly time-consuming, involving a BFS (Breadth-First Search) from every vertex in the graph, we only ran the BFS on one run of each experiment for indicative purposes. Results show diameter sizes, as expected, quite low. With linear regression vs. graph size and prediction for final graph of 200 million vertices, diameter estimates are still very low compared to the ones reported from the real WWW. Thus we surmise that the EGC model cannot provide graphs with the desired diameter (mainly due to the giant SCC component it creates). Experiment results and predictions are presented in Table 5.

Table 5. Prediction of Graph Diameter for the Optimum Parameter Values and Final Graph of 200 million Vertices (95% Confidence Interval)

p	**γ**		**γ'**		**200M Diameter**	
0.05	2	3	0.70	0.8	14	26
0.50	2	3	0.65	0.7	13	59
0.95	2	3	0.60	0.7	72	72

7 CONCLUSIONS

We presented a framework for evaluating web graph models and the test case of the EGC model, where we used the framework to assess its eligibility for simulating the WWW. The conclusions drawn from the experimental results are summarized below.

Concerning in- and out- degree power-law distribution, we concluded that power-law vertex degree distributions do exist; they have realistic exponents (for **d = 7**) when **γ'** lies between **0.6** and **0.8** and seem to be only slightly affected by the value of **γ**: if **γ = 2** then **γ'** should be about **0.65** while if **γ = 3** then **γ'** should be about **0.75.** Prediction of these exponents for a final graph of **200 million** vertices agrees with WWW observations. These parameter constraints are sensitive to the initial graph size. Not only out-degree frequency distributions, but also in-degree frequency distributions do not follow power-laws when the initial graph is large. The desired power-law degree distributions are resilient over different magnitudes of final graph sizes.

EGC demonstrated a very stable behavior in all properties we studied; based on its stability we made predictions for a final graph size of **200 million** vertices. Because values for all parameters were found for low, middle and high values of **p** such as to produce the desired observable graph characteristics, we conclude that for any value of **p**, appropriate values for the other parameters can be found. This property strengthens its usefuleness in temporal studies of the web.

Concerning structural properties (diameter, components, bipartite cores), EGC predictions agree with the real WWW only on the bipartite cores' count, where predictions for a final graph of **200 million** vertices are very close to reality. However, graphs produced display a great expansion of the SCC component in the expense of the other structural components, which is not a result of the increase of graph size but an intrinsic characteristic of the model. This property naturally affects the diameter of the graphs produced, which in all experiments and predictions for final graphs of **200 million** vertices is too small.

Overall, EGC succeeds in producing graphs with very realistic vertex degree frequency distributions and very satisfying quantity of bipartite cores. It is also a very good candidate for graph evolution versus time or size simula-

tion studies. However, we note that it fails to structure the produced graphs according to real WWW observations.

From a simulation-based standpoint, we have contributed in presenting an evaluation framework, more detailed than simple in- and out- degree distribution analysis. We have also contributed by evaluating the EGC model's behavior in so far uninvestigated areas. Although we provided no new insights about the real www, this paper contributes in the accreditation of such studies, by setting up a validation framework for web models used to draw inferences about the www. However, updating this framework with contemporary measurements proves to be problematic since no such measurements currently exist; more data collection and analysis research is needed in that direction Extending the framework to allow comparison of models is another open research area that may provide valuable insights in establishing a widely-accepted world-wide web graph model.

ACKNOWLEDGMENTS

This research was supported in part by Pythagoras program (MIS 89198) co-funded by the Greek Government (25%) and the European Union (75%).

REFERENCES

Adler M. and Mitzenmacher M. 2001. Towards compressing web graphs. *Proceedings IEEE Data compression Conference*, 203-212.

Barabasi A. and Albert R 1999. Emergence of scaling in random networks. *Science 286,* 509-512.

Brin S. and Page L. 1998. The anatomy of a large-scale hypertextual web search engine. *Proceedings 7th WWW Conference*.

Broder A., Kumar R., Maghoul F., Raghavan P., Rajagopalan S., Stata R., Tomkins A. and Wiener J. 2000. Graph structure in the web: Experiments and models. *Proceedings 9th WWW Conference*, 309-320.

Bu T. and Towsley D. 2002. On distinguishing between Internet power-law topology generators. *Proceedings INFOCOM 2002*.

Faloutsos M., Faloutsos P. and Faloutsos C. 1999. On power-law relationships of the Intenet topology. *Proceedings ACM SIGCOMM '99*.

Kleinberg J. 1998. Authoritative sources in a hyperlinked environment. *Proceedings 9th Annual ACM-SIAM Symposium on Discrete Algorithms*, 668-677.

Kleinberg J., Kumar R., Raghavan P., Rajagopalan S. and Tomkins A. 1999. The web as a graph: Measurements, models and methods. *Proceedings International Conference on Combinatorics and Computing*, 1-18.

Kogias A., Nikolaidou M. and Anagnostopoulos D. 2005. Modelling and simulation of the web graph: evaluating an exponential growth copying model. *International Journal of Web Engineering and Technology Vol. 2 No 1*, 29-49.

Kumar R., Raghavan P., Rajagopalan S. and Tomkins A. 1999a. *Trawling the web for emerging cyber-communities. Proceedings 8th WWW Conference*, 403-416.

Kumar R., Raghavan P., Rajagopalan S. and Tomkins A. 1999. Extracting large-scale knowledge bases from the web. *Proceedings 25th VLDB Conference*, 639-650.

Kumar R., Raghavan P., Rajagopalan S., Sivakumar D., Tomkins A. and Upfal E. 2000. Stochastic models for the web graph. *Proceedings 41st Annual Symposium on Foundations of Computer Science*.

Laura L., Leonardi S., Caldarelli G. and De Los Rios P. 2002. A multi-layer model for the web graph. *Proceedings 2nd International Workshop on Web Dynamics*..

Matsumoto M. and Nishimura T. 1998. "*Mercenne Twister: a 623-dimensionally equidistributed uniform pseudo-random number generator*" ACM Transanctions on Modeling and Computer Simulation Vol. 8, No 1, January 1998, pp. 3-30.

Tangmunarunkit H., Govindan R., Jamin S., Shenker S. and Willinger W. 2002. Network topology generators: degree-based vs structural. *Proceedings SIGCOMM '02*.

AUTHOR BIOGRAPHIES

ANTONIOS KOGIAS graduated from the Hellenic Army Academy in 1993 and became an officer of the Engineer Corps in 1994. He was awarded the 'Computer Programmer and Analyst' speciality in 2000. He received an MSc in the area of 'New Technologies in Informatics and Telecommunications' from the University of Athens in 2004. Following that, he was transferred to the Research and Informatics Corps where he is currently employed. He is a PhD candidate in Harokopio University of Athens. His research interests lie in the area of distributed systems and computer simulation. <coyas@hua.gr>

DIMOSTHENIS ANAGNOSTOPOULOS is Associate Professor at Harokopio University of Athens. He received a Degree and a Doctorate Degree, both in computer science, from the University of Athens in 1991 and 1996, respectively. He has published more than 50 papers in refereed journals and conference proceedings. His research interests include modeling and simulation, business process modeling, object-oriented systems and distributed systems and networks, as well as modeling and performance evaluation of transportation systems. <dimosthe@hua.gr>