Modelling and simulation of the web graph: evaluating an exponential growth copying model

Antonios Kogias* and Mara Nikolaidou

University of Athens, Panepistimiopolis, 15771, Athens, Greece
E-mail: coyas@hua.gr E-mail: mara@di.uoa.gr
*Corresponding author

Dimosthenis Anagnostopoulos

Harokopio University of Athens,
70 El. Venizelou Str., 17671, Athens, Greece
E-mail: dimosthe@hua.gr

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, and mostly for predicting the evolution of the web and the emergence of important new phenomena. Researchers have viewed the WWW as a graph, the so-called web graph. We present a brief review of the most typical random graph models for the web and introduce a validation process for web graph models. We evaluate the behaviour of the Exponential Growth Copying (EGC) model, which has been explicitly designed to model the WWW, and analyse the effect of individual parameters on its effectiveness through simulation modelling. Specifically, we derive the in and out degree distributions of the resulting graphs for various parameter values and measure them against the empirical analytical results from the real web (i.e. power laws for in and out degrees). Finally, we suggest appropriate values to improve EGC effectiveness and deliver a realistic model of the web graph.

Keywords: copying models; modelling and simulation; web modelling.

Reference to this paper should be made as follows: Kogias, A., Nikolaidou, M. and Anagnostopoulos, D (2005) 'Modelling and simulation of the web graph: evaluating an exponential growth copying model', *Int. J. Web Engineering and Technology*, Vol. 2, No. 1, pp.29–49.

Biographical notes: 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' specialty 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.

Mara Nikolaidou received a Degree and a Doctorate Degree, both in computer science, from the University of Athens in 1990 and 1996, respectively. She is currently in charge of the Library Automation Centre of the University of Athens. Her research interests include distributed systems, digital libraries, modelling and simulation, and workflow systems.

Dimosthenis Anagnostopoulos is an 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 modelling and simulation, business process modelling, object-oriented systems and distributed systems and networks, as well as modelling and performance evaluation of transportation systems.

1 Introduction

The WWW has shown a tremendous growth in late years and current estimates of its size are at the billion web pages scale. This increase in size however, has been unregulated – the more information content the web contains, the more difficult it is to locate it. As it is extremely difficult for researchers to obtain and manage real-world data, it is important that models of the web be used for creating WWW-like representations, upon which new algorithms and applications for searching, indexing, compression etc. can be tested. Furthermore, models are needed for understanding the sociology of content creation on the web, predicting its evolution and the emergence of important new phenomena.

Researchers view the WWW 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 their vertex of origin (tail) and their 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 as tail (head) this specific vertex. Crawls of the web (Broder et al., 2000; Kleinberg et al., 1999; Kumar et al., 1999a) report that 'power laws' exist for in and out degrees. A power law for in-degree is 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 x.) The values of x for the in-degree (x_{in}) and out-degree (x_{out}) have been reported to be $x_{in}=2.1$ and $x_{out}=2.7$ (Broder et al., 2000). These power laws are widely accepted in the literature as salient WWW characteristics.

The development of realistic and accurate stochastic models of the web graph is a challenging task for the following reasons:

- 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)
- analysing the behaviour of search algorithms that make use of link information (e.g. page rank (Brin and Page, 1998), HITS (Kleinberg, 1998) etc)
- designing crawl strategies
- predicting the evolution and the emergence of important new phenomena in the web
- dealing more efficiently with large scale computations (i.e. by recognising the possibility of compressing a graph generated by such a model).

Adler and Mitzenmacher (2001) present a characteristic case demonstrating the usefulness of web graph models. They considered the problem of compressing graphs of the link structure of the WWW and providing efficient algorithms, motivated by the random graph models proposed in (Kumar et al., 1999b). They designed a compression algorithm based on finding similarity among the links of pages and tested it on graphs created by the model. They compared it with other widely used compression techniques and verified its suitability for real web data using the TREC-8 WT2g dataset (Hawking et al., 2000).

The contribution of this paper involves the following:

- Presenting a brief review of the most typical random graph models for the web and introducing a validation process for web graph models.
- Studying the behaviour of the Exponential Growth Copying Model (EGC) for the web graph (Kumar et al., 2000), which has been explicitly designed to model the WWW. We analyse the effect of individual parameters on its overall effectiveness, indicate weak points and suggest appropriate parameter values. The EGC model can then be used for creating web-like graphs.

Specifically, we derive the in and out degree distributions of the resulting graphs for various parameter values and measure them against the empirical analytical expressions of these distributions on the real web (i.e. power laws for in and out degrees). We then draw conclusions about the model behaviour through measuring the error distribution between experimental and empirical results. Finally, we suggest appropriate values for EGC parameters to enable the creation of more 'web-like' graphs.

2 Web graph models

We present a brief review of the most typical random graph models for the web.

2.1 Evolving network with preferential attachment (Barabasi and Albert, 1999)

Starting with a small initial graph, at each discrete time step a new vertex with a fixed number of outgoing edges is added, their heads chosen between the already existing vertices, with probability proportional to their in-degree. This was the first web model to use the evolving network and preferential attachment techniques to provide random graphs. It was shown that these two elements are enough to create power-law in-degree distributions. However, the in-degree exponent is ≈2.9 (not 2.1 of the real web) and the out-degree of vertices is constant (it does not follow any power law).

2.2 Copying models (Kumar et al., 1999b)

The General Copying Model starts with a small initial graph and, at each time step, a new vertex and k outgoing edges are added. As a result of a stochastic selection, either k vertices are randomly selected as heads, or one vertex is randomly selected and k outgoing edges are 'copied' as heads of the new edges. The α model is a simplified version of the general model where k=1 and the stochastic selection is between creating a self-loop edge for the new vertex or copying a random existing edge to the new vertex as outgoing. The α–β model has k=1, but discrete (although similar) stochastic selection of the head and

the tail of a new edge: the new head (tail) is either the new vertex or the head (tail) of randomly selected existing edge. These models were based on the realistic intuition that, although some page creators on today's web may create content and links to other sites without regard to the topics that are already represented on the web, many page creators will be drawn to the existing topics of interest and thus create links to pages within some of these existing topics. It is proven analytically that by using α=0.52 and β=0.58, the α–β model generates graphs that follow the power-laws for in and out degree with exponents 2.1 and 2.38, respectively. This model was used in (Adler and Mitzenmacher, 2001) to test the proposed web graph compression algorithm and compare its effectiveness with other widely-used compression schemes.

2.3 Evolving copying models (Kumar et al., 2000)

The Linear Growth Copying Model starts with a small directed initial graph and, at each time step, a new vertex with *d* outgoing edges is added. A prototype vertex is then chosen randomly among the existing vertices and the copying process begins, based on parameter α of the model: the destination of the new vertex's i^{th} edge is either (with probability α) a randomly selected existing vertex or the destination of the i^{th} edge of the prototype vertex. The Exponential Growth Copying Model (EGC) is described in detail in the next section. These models generate more bipartite cliques than any other model and are thus more compliant with the real web observations.

2.4 Randomly grown graphs (Callaway et al., 2001)

A small initial graph is continually expanded, at discrete time steps, with a new vertex and a new undirected edge between two vertices chosen uniformly at random. Its purpose is to provide a minimal framework of network growth for comparison with the traditional random graph, trying to explain the effect of growth on the graph structure (recall that the traditional random graph comprises of a static number of vertices and all possible edges are equiprobable (Erdös and Rényi, 1959). The model creators state that they do not claim that this model is an accurate reflection of any particular real-world system, but it provides useful insights – growth itself seems to enforce a measure of preferential attachment.

2.5 ACL models (Aiello et al., 2001)

In these four (*A–B–C–D*) models the graph is expanded at discrete time steps with at least one vertex and at least one edge. The *A–B–C* models are directed, whereas the *D* model is an undirected variation of the *C* model. We describe the *C* version (which is the most complex and considered as the most 'general' case): the algorithm begins with an initial graph of some vertices and edges. A new vertex is added and four numbers ($m^{e,e}$, $m^{n,e}$, $m^{e,n}$, $m^{n,n}$) are sampled from a probability distribution at each time step – each controlling a different action to be taken at this time-step, as follows: $m^{e,e}$ edges are added randomly (with preferential attachment on in and out degree of existing vertices), $m^{e,n}$ edges are added to the new vertex (with preferential attachment on out degree for the selection of their tails), $m^{n,e}$ edges are added from the new vertex (with preferential attachment on in degree for the selection of their heads) and $m^{n,n}$ self-loop edges are added to the new vertex. Let $\mu^{e,e}$, $\mu^{n,e}$, $\mu^{e,n}$ and $\mu^{n,n}$ be the expectations for the respective random variables.

It is also proven in (Aiello et al., 2001) that almost surely the in and out degree sequences follow power law distributions with exponents that are simple functions of these expectations.

2.6 Growth and redirection model (Tadic, 2001)

A small initial directed graph is grown at discrete time steps with one new vertex and a number of edges. A fraction of this is the number of new edges that will be created as outgoing to the new vertex. The rest are existing edges of the graph that will be updated, which are selected preferentially on the out degree of their tail. The heads of all these edges are chosen with preferential attachment on in degree. Parameter β is used to capture the ratio of new edges to updated edges and it is noted that when $\beta=3$, the in degrees follow a power law distribution with exponent ≈ 2.16 and that the out degrees follow a power law distribution with exponent ≈ 2.62. These values agree with real-web data. Therefore it is surmised that, in the current state of the web, a three updated to one added link ratio is observed at each evolution step in average. Presently, there are no empirical observations about this ratio from the real web.

2.7 Multi-layer model (Laura et al., 2002)

Based on a study by Dill et al. (2001), this model tries to capture the fractal structure of the web produced by the presence of multiple regions generated by independent stochastic processes, connected together by a connectivity backbone formed by pages that participate in multiple regions. At each discrete time step, a new vertex is added with a number of regions it belongs to and a number of outgoing edges to existing vertices. The edges are evenly distributed among regions. A prototype vertex is chosen for each region, and (based on a model parameter) either edges are copied from this vertex or destinations are chosen with preferential attachment on the in degree among the same region. If the prototype vertex has not enough outgoing edges, the rest are chosen among the vertices of the whole graph with preferential attachment on in degree. A comparison with other web models and a crawl on in degree distribution power law exponent and other structural aspects is provided in (Laura et al., 2002). However, the out degree of all vertices is constant, therefore not following any power law.

2.8 Steady-state model (Eppstein and Wang, 2002)

This model starts with a sparse undirected graph (randomly formed during initialisation). At each time step, a vertex with incident edges is selected as well as an existing edge of the graph. Next, another vertex is selected with preferential attachment on degree. If these vertices are different and an edge between these vertices does not exist, then the selected edge is redirected to connect the selected vertices. Its creators believe that the existing growth-based models are adequate to explain the web's current graph structure (Eppstein and Wang, 2002), but further state that it would be interesting to know if a different model is needed as the growth rate slows down while its link structure continues to evolve. This model is the only one that results in power laws without incremental growth, but it is an undirected model.

3 The EGC model

Evolving Copying Models 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 to them 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 in – say – the last week or two)(Kumar et al., 2000).

The EGC was proposed in 2000 (Kumar et al., 2000) and incorporates both intuitions; a detailed description of its algorithm follows.

At time step t a new epoch of vertices arrives, the size of which is a constant fraction of the current graph. Each of these vertices may be linked only with vertices of previous epochs. The EGC model is formally described parametrically, using a 'growth' factor $p>0$, the 'self loop' factor $\gamma>1$, the 'tail copy' factor $\gamma' \in (0,1)$ and the 'natural link' factor $d>0$. Let $G_t=(V_t, E_t)$ denote the directed graph created by the model at time step t with vertex set V_t and edge set E_t. The number of new vertices at time step t is determined by sampling from the standard binomial distribution $B(V_t, p)$, which means that for large t, V_t is well concentrated around its mean $(1+p)^t$. Since p is used in the standard binomial distribution, $p \leq 1$. For analysis simplification, it is assumed that $V_1=1$ and $V_t=(1+p)^t$ (Kumar et al., 2000). The expected number of edges generated at time step $t+1$ is $(d+\gamma)pV_t$.

Each new vertex is generated with γ self-loop edges. The heads and tails of the remaining edges are chosen according to the following process:

For each edge directed to $u \in V_t$ at time t, a new edge directed to u is generated with probability $dp/(d+\gamma)$. Assuming that the expected number of edges at time t is $(d+\gamma)V_t$, the expected number of edges generated in this process is dpV_t. The tails of the new edges generated in this step are determined as follows:

- randomly, among the pV_t new vertices of this step, with probability $1-\gamma'$
- randomly, among the vertices created in the previous steps, with probability γ'.

In both cases, vertices are chosen with probability proportional to their current out-degree. Considering the new self-loop edges, the expected number of edges at time $t+1$ is $(d+\gamma)V_{t+1}$.

It has been proven analytically that the graphs created by the EGC model follow some power law for in-degree (Kumar et al., 2000) with a bounded exponent and also that they contain a large number of bipartite cliques. Both conclusions agree with the real WWW observations (Kumar et al., 1999b). However, it is not clear whether the out-degrees satisfy the power law distribution (Aiello et al., 2001), as this has not been examined yet, as well as whether the bounded exponent for in degree power law distributions is close to the value observed in the real web.

4 Web graph model validation

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; it cannot always be employed though, due to 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 but not for out-degree (due to the model complexity of link creation). Simulation can overcome these difficulties, but it 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 the validation of web graph models, simulation may determine whether model results conform with empirical observations and, at a second stage, to refine (i.e. appropriately parameterise) the proposed models to exhibit greater efficiency. At the end, 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.

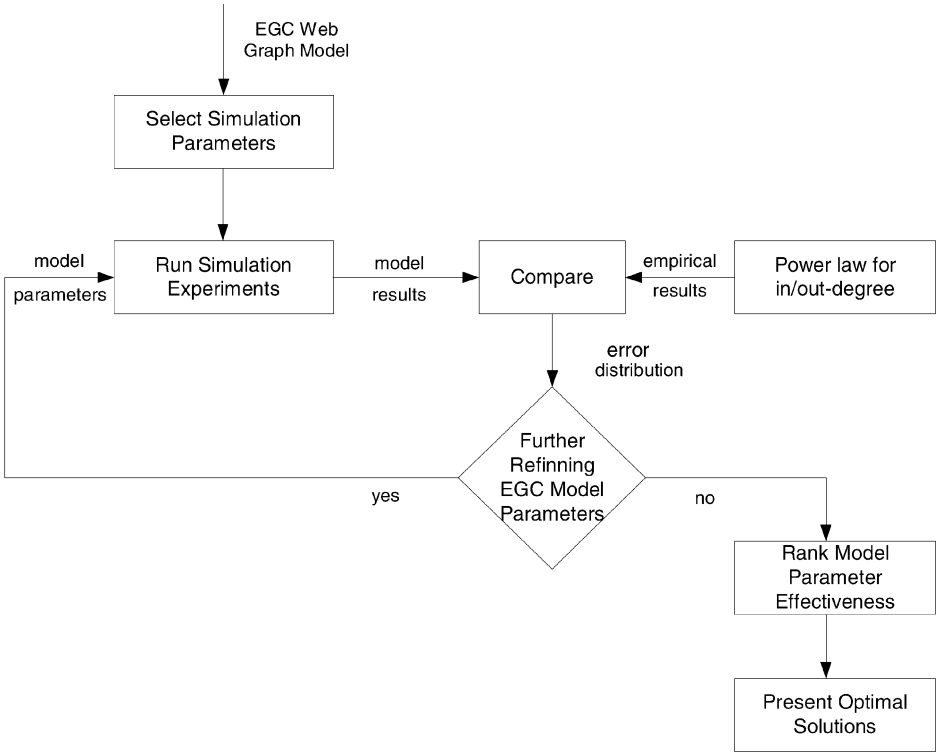
The long established way of obtaining WWW data is by using web crawlers (a.k.a. spiders or robots) that run a continuous loop of downloading web pages, extracting URLs and in turn attempting to download them also. When a specified amount of time has passed or a specified amount of pages has been downloaded, the crawler terminates and exports the data set. These data sets are then 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 when 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. So, crawls provide a data set that is not the real web, but a sample, possibly big enough to draw valid conclusions from.

On the other hand, if a valid model of the system in question exists, many issues may be resolved through simulation. Even if various models have been proposed, 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. The approach undertaken in this paper is based on comparing model 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,b; Laura et al., 2002).

Model validation is accomplished by evaluating the capability of the model to approximate the in and out degree power law distributions, as these are key features of the web for which widely accepted results have been presented. The validation process is depicted in Figure 1 and is generic, so that it may be widely employed for web graph models.

Figure 1 Web graph model validation process



5 Simulating the EGC model

We simulated the operation of EGC as follows: the model starts with a small initial graph (number of vertices, edges), a parameter vector [p, γ, $\gamma^{\cdot}$, d] and the terminating condition N (number of vertices in graph). The initial state is epoch 1 (current epoch denoted by e). Let n denote the current number of vertices in the graph up to the last epoch and v the number of vertices to be added in the next epoch. The value of v is decided by random sampling from the binomial distribution with parameters p and n. For each vertex to arrive in the next epoch (1 to v), the algorithm creates γ self-loop edges. Then, d new edges are created and each one's head and tail are set to the selected vertices: the head of a new edge is chosen uniformly among the heads of all existing edges up to the last epoch (i.e. preferential attachment on in-degree among the n existing vertices). Then, a random variable r, sampled uniformly in (0, 1), is used to determine the tail of this edge as follows:

- if $r \in (0, \gamma^{\cdot})$, the tail is chosen randomly among the tails of all existing edges up to the last epoch (i.e. preferential attachment on out-degree among n existing vertices)
- if $r \notin (0, \gamma^{\cdot})$ the tail is chosen randomly among the (newly created) vertices of the new epoch.

These edges are added to the new epoch, until all edges have been created. Then, the new epoch 'arrives': all vertices and edges of the new epoch are added to the graph and counters e and n are suitably updated. This process keeps adding epochs until n becomes greater or equal to N, when it terminates and exports the resulting graph for further processing.

We used Matlab for the EGC model implementation and the experiments were conducted on a dual processor Pentium IV system with 1 Gb RAM. Each simulation run was made for a number of epochs until at least 1,000,000 vertices were generated in the resulting graph. Initial conditions were the same for all runs, specifically one vertex with γ self-loop edges.

According to the model definition, $0<p\leq 1$, $\gamma>1$, $\gamma^{\cdot}\in(0, 1)$ and $d>0$; we studied EGC for low, intermediate and high values of these parameters. Specifically, we examined the following characteristic values:

p: {0.05, 0.5, 0.95}
γ: {2, 5, 10}
$\gamma^{\cdot}$: {0.05, 0.5, 0.95}
d: {1, 2, 3, 4, 5}.

We thus experimented with 135 different combinations of $[p, \gamma, \gamma^{\cdot}, d]$, to explore their impact on EGC efficiency. These combinations we name parameter vectors PV. For each parameter vector $PV_i(1\leq i\leq 135)$, 30 independent replications (i.e. runs) were made, to enhance the statistical features of simulation results.

In each run $j=1, 2, \ldots, 30$ of each parameter vector $PV_k(k=1, 2, \ldots, 135)$, the algorithm exports the edge set $L_{k,j}$ and in/out degrees of all vertices are computed. Self-loop edges are removed (this was also done in the empirical analysis of web crawls) and in/out degree distributions for each run (denoted by $I_{k,j}(x)$ and $O_{k,j}(x)$ respectively, where x denotes a degree value) are computed. Dividing them by the number $n_{k,j}$ of vertices of each run, we derive the in and out degree probability distributions:

$$PI_{kj}(x) = I_{kj}(x)/n_{kj}$$
$$PO_{kj}(x) = O_{kj}(x)/n_{kj}.$$

$IN_k(x)$ and $OUT_k(x)$ are the average in and out degree probability distributions resulting from all 30 runs for each parameter vector PV_k:

$$IN_k(x) = \frac{\sum_{j=1}^{30} PI_{kj}(x)}{30}$$

$$OUT_k(x) = \frac{\sum_{j=1}^{30} PO_{kj}(x)}{30}.$$

To determine the efficiency of EGC for each [p, γ, $\gamma^{\cdot}$, d] vector, we use a simple ranking scheme, as explained in Section 4. Let $P_{in}(x)$ and $P_{out}(x)$ denote the empirical power laws for in and out degree probability on the WWW:

$$P_{in}(x) = x^{-2.1}$$
$$P_{out}(x) = x^{-2.7}.$$

Then, the in and out error distributions for all parameter vectors PV_k($\forall$ x, k=1, 2, . . . , 135) are:

$$EI_k(x) = IN_k(x) - P_{in}(x)$$
$$EO_k(x) = OUT_k(x) - P_{out}(x).$$

Let $\overline{EI}_k$ and $\overline{EO}_k$ denote the EI_k and EO_k distribution means and $\overline{EIO}_k$=$\overline{EI}_k$+$\overline{EO}_k$. We set:

$$e_{in} = \min\left(\overline{EI}_k, k = 1,2,\ldots,135\right)$$
$$e_{out} = \min\left(\overline{EO}_k, k = 1,2,\ldots,135\right)$$
$$e_{io} = \min\left(\overline{EIO}_k, k = 1,2,\ldots,135\right).$$

Normalising the mean values by division with the respective minimum values, we acquire the in, out and joint degree normalised mean errors:

$$N_{in}(k) = \overline{EI}_k / e_{in}$$
$$N_{out}(k) = \overline{EO}_k / e_{out}$$
$$N_{io}(k) = \overline{EIO}_k / e_{io}.$$

$N_{in}(k)$ and $N_{out}(k)$ are then sorted to implement the ranking functions $r_{in}(PV_k)$, $r_{out}(PV_k)$ and $r_{io}(PV_k)$, which return a number between 1 and 135 (the position of $N_{in}(k)$, $N_{out}(k)$ and $N_{io}(k)$ in the sorted lists), thus ranking the performance of each [p, γ, $\gamma^{\cdot}$, d] vector.

6 Simulation output analysis

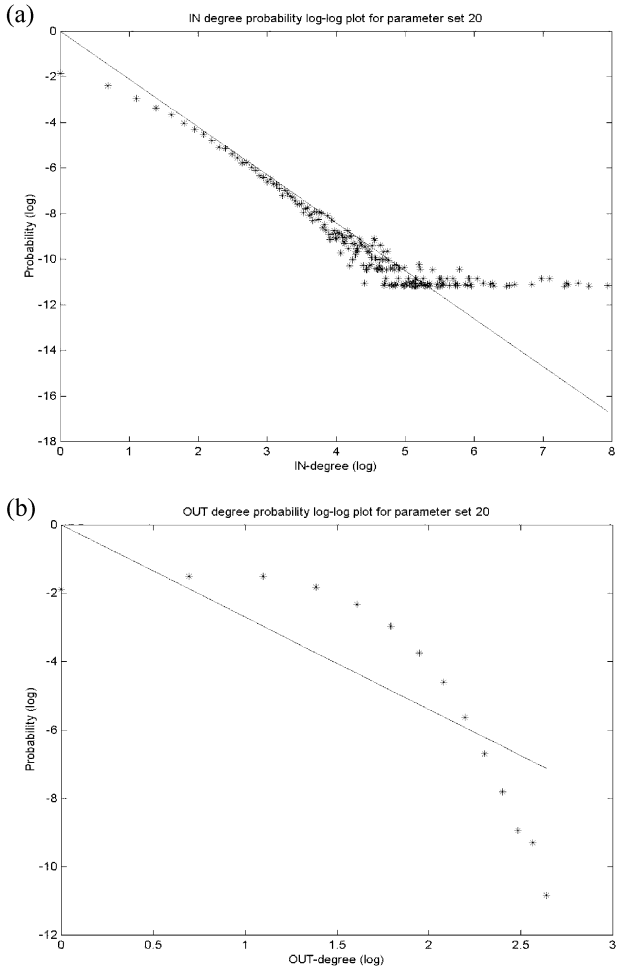
Our evaluation is oriented towards:

- validating the EGC model concerning the in and out degrees, through comparing model results with the corresponding empirical data (in/out degree power laws)
- determining the effect of individual parameters p, γ, $\gamma^{\cdot}$, d on EGC efficiency
- suggesting appropriate values for parameters to improve EGC efficiency.

6.1 In/out degree approximation

In Figures 2–4, simulation results are presented for the in and out degree distributions for three representative parameter vectors (from the 135 ones examined). In these figures, $\gamma^{\cdot}$ is variable while p=0.5, γ=2 and d=3. An in/out degree probability log–log plot is given for all test cases, to enable cross evaluation with the power laws for in and out degree. Concerning the in degree distribution, in all 135 cases, in degree probability plots keep a steady form of 'heavy tailed power law', depicted in Figures 2(a)–4(a). Concerning the out degree distribution, out degree probability plots range between 'nopower law' in Figure 2(b), to 'power law but no heavy tail' in Figure 3(b), and 'heavy tailed power law' in Figure 4(b). Output analysis of simulation results presents the following ad hoc conclusions, also briefly discussed in (Kogias and Anagnostopoulos, 2003):

Figure 2 (a) in degree probability distribution (b) out degree probability distribution for $\gamma^{\cdot}$=0.05 (p=0.5, γ=2, d=3)



- Results for the probability of vertices with a low in/out degree are significantly lower than the values predicted by the correspondent power laws, resulting in negative error means for all degree distributions. This is expected though, since power law values in the low range of degrees are too high. The same phenomenon has been reported in web crawl results.

- Results for the probability of vertices with in degree in the middle range are very close to the ones predicted by the in degree power law, confirming the suitability of EGC for the representation of the in degree.

- The EGC model is not as efficient when representing the out degree, as a strong variation in its behaviour is encountered for different values of [p, γ, $\gamma^{\cdot}$, d].

Figure 3 (a) in degree probability distribution (b) out degree probability distribution for $\gamma^{\cdot}$=0.5 (p=0.5, γ=2, d=3)

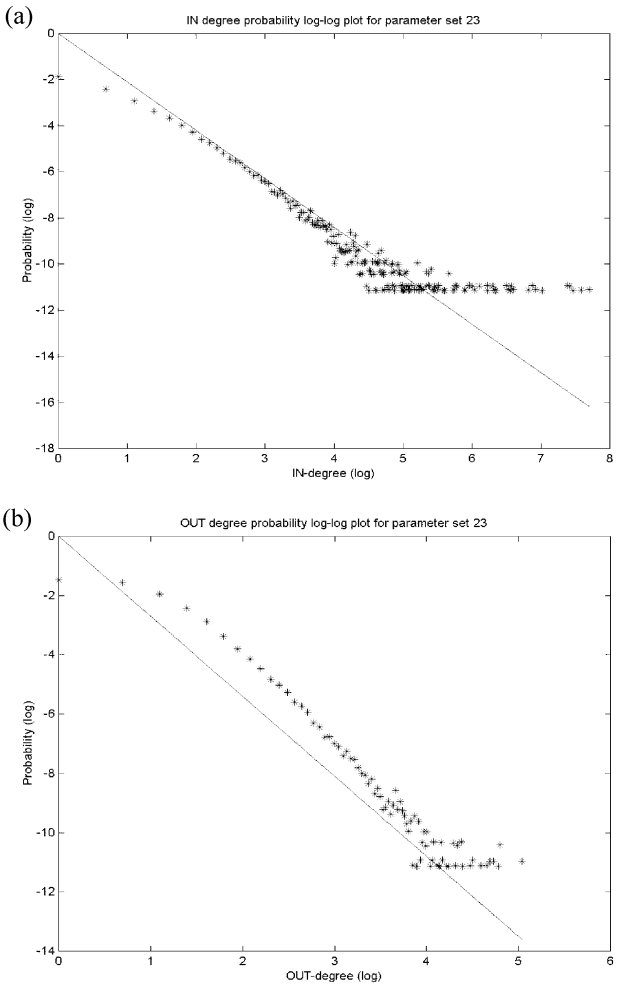
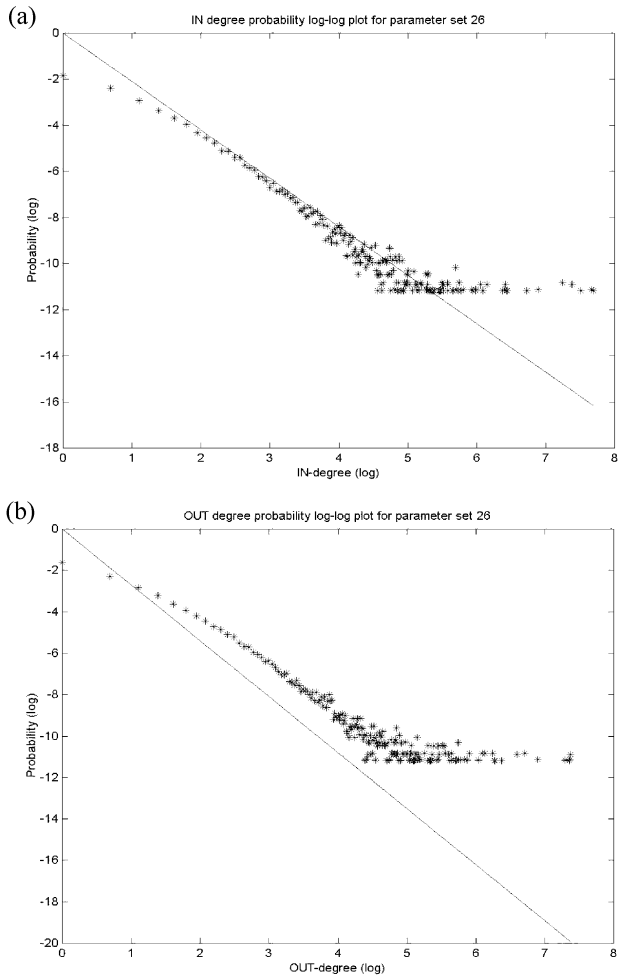


Figure 4 (a) in degree probability distribution (b) out degree probability distribution for $\gamma^{'}$=0.95 (p=0.5, γ=2, d=3)



Furthermore, in all parameter vectors, the error distribution means were negative ($\overline{EI}_k$<0, $\overline{EI}_k$<0, $\overline{EIO}_k$<0 for k=1, 2, . . ., 135). Analysis of the error distributions shows that this is due to the values predicted by power laws for very low values of degrees, especially when degree=1; both in and out degree power laws predict probability=1 for this degree. This value is evidently out of bounds as probability, but since it is a constant phenomenon in all degree distributions and it only affects a minimal subset of degrees on the lowest part of the spectrum, we choose to ignore it and focus on the main body of degree distributions. Such a deviation from the power laws has also been reported on results from web crawls, thus it is not unjustified. The minimum values of error distribution means were:

$$e_{in} = -9.993 \times 10^{-5}$$
$$e_{out} = -8.53 \times 10^{-5}.$$

As $e_{in} \approx e_{out}$, we compared the normalised values. Using ranking functions $r_{in}(PV_k)$ and $r_{out}(PV_k)$ for PV_j $j=1 \ldots 135$, we see that N_{in} climbs slowly up to about 445 times its minimum value ($min(N_{in})=1$, $max(N_{in})=445$), while N_{out} climbs faster and up to about 1055 times its respective minimum ($min(N_{out})=1$, $max(N_{out})=1055$). This leads to the following conclusions:

- out degree approximation is less efficient than in degree approximation
- out degree approximation has a significant sensitivity to [p, γ, $\gamma^{\cdot}$, d] values.

6.2 Examining individual parameters

The second activity of this study involves the examination of the impact of individual parameters p, γ, $\gamma^{\cdot}$, d on EGC results. We experimented with different values of each single parameter but in all experiments with a specific parameter value, we used the same value sets for the remaining parameters. As a result, a periodicity effect is formed in the corresponding plots for p (Figure 5), γ (Figure 6), $\gamma^{\cdot}$ (Figure 7) and d (Figure 8). Based on simulation results, we reached the following conclusions concerning the impact of parameters on the normalised mean errors N_{in} and N_{out}:

- There is no clear influence of parameter p on the normalised mean errors (Figure 5). Therefore the 'growth' factor does not seem to affect the effectiveness of EGC model.
- As parameter γ increases, the normalised mean errors tend to increase (Figure 6). Also, N_{in} tends to increase and be almost equal to N_{out}, which is undesirable, as N_{out} is always high. This imposes that low values of the 'self loop' factor should be more effective.
- As parameter $\gamma^{\cdot}$ increases, N_{out} tends to decrease and be equal with N_{in} (Figure 7). This imposes that high values of the 'tail copy' factor should be more effective.
- As parameter d increases, both N_{in} and N_{out} decrease (Figure 8). This imposes that high values of the 'natural link' factor should be more effective.

Figure 5 N_{in} and N_{out} versus p

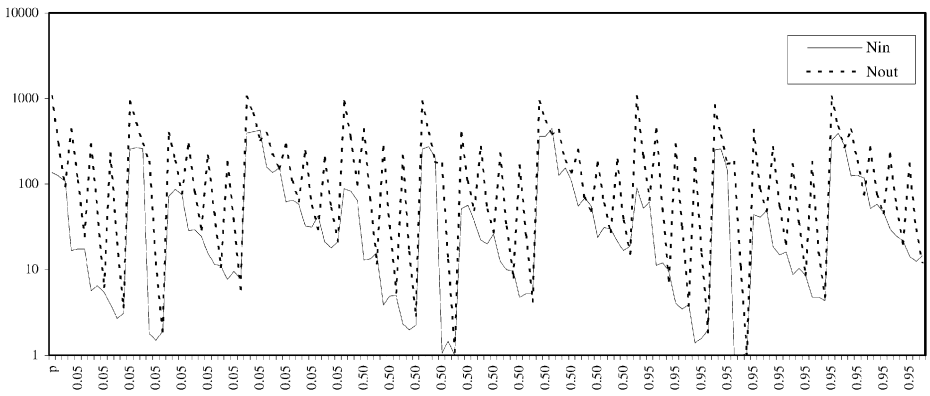


Figure 6 N_{in} and N_{out} versus γ

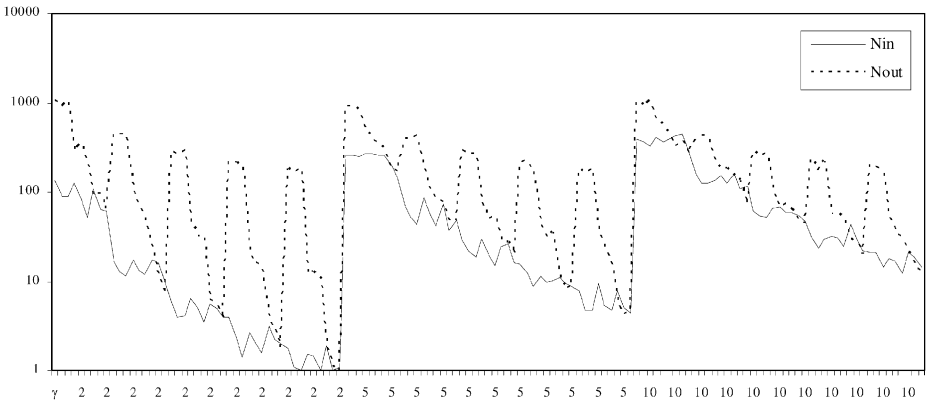
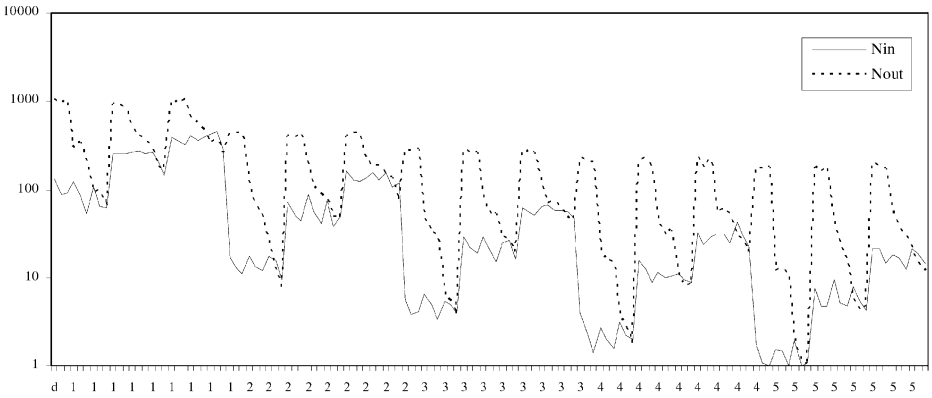


Figure 7 N_{in} and N_{out} versus γ'

Figure 8 N_{in} and N_{out} versus d



6.3 Refining the EGC model

The last activity of the study involves suggesting appropriate values for [p, γ, $\gamma^{\cdot}$, d] to improve EGC effectiveness. Towards this objective, we use our ranking scheme $r_{in}(PV_j)$ and $r_{out}(PV_k)$ and determine which values of j satisfy $1 \leq r_{in}(PV_j) \leq 35$ and which values of k satisfy $1 \leq r_{in}(PV_k) \leq 35$; i.e. we choose the 35 optimal parameter vectors and perform an analysis concerning the values of each individual parameter being used in them (i.e. we analyse the top 25% of our results). Our analysis is summarised in Figure 9 for N_{in} and Figure 10 for N_{out}.

As depicted in Figure 9, concerning the in degree approximation:

- p and $\gamma^{\cdot}$ have no impact on EGC performance, as depicted in Figure 9(a) and (c) respectively
- values of γ close to 2 and values of d equal or greater than 5 are more effective, as depicted in Figure 9(b) and (d) respectively
- overall, there is no strong impact of [p, γ, $\gamma^{\cdot}$, d] values on the performance of EGC.

On the other hand, as depicted in Figure 10, concerning the out degree approximation:

- p has no impact on EGC performance, as depicted in Figure 10(a)
- values of $\gamma^{\cdot}$ close to 0.95 are more effective, as depicted in Figure 10(c)
- values of γ close to 2 and values of d equal or greater than 5 are more effective, as depicted in Figure 10(b) and (d) respectively
- Overall, there is a strong impact of [p, γ, $\gamma^{\cdot}$, d] values on the performance of EGC.

Based on the conclusions about the general behaviour and effectiveness of the EGC model, we extended our research for producing WWW-like representations.

A significant feature of the model is its independence of the growth rate. In contrast to other models, EGC provides a convenient way of adapting to real-time evolution studies of the WWW. Other models grow by one vertex at each time-step, a fact that doesn't help when one must define the real time passing at each time step. Even if that was possible, since the WWW grows approximately exponentially, the real-time duration of each time step must be continuously adjusted to reflect the onslaught of new vertices arriving at incrementally smaller real-time intervals. The EGC model suffers from no such drawback. All one has to do is decide the real-time equivalent of the model's 'epoch' and adjust the 'growth' factor p analogously.

Having already focused on the behaviour exhibited by EGC for various values of the 'natural link' parameter d (which is directly related to the WWW growth), we have so far concluded that it gives a – more or less – realistic representation for various values of d. As the WWW evolves, the value of the 'natural link' parameter d was found to be approximately 7 in 1999 (Broder et al., 2000; Kleinberg et al., 1999; Kumar et al., 2000). We thus endeavour to achieve a more precise approximation of the WWW using the other model parameters (especially the 'tail copy' factor $\gamma^{\cdot}$), for the currently acknowledged value of d (other suggestions have also appeared in the literature, but have not so far been substantiated (Broder et al., 2003; Fetterly et al., 2003).

Figure 9 Analysis of [p, γ, γ', d] values for the 35 optimal vectors for N_{in}

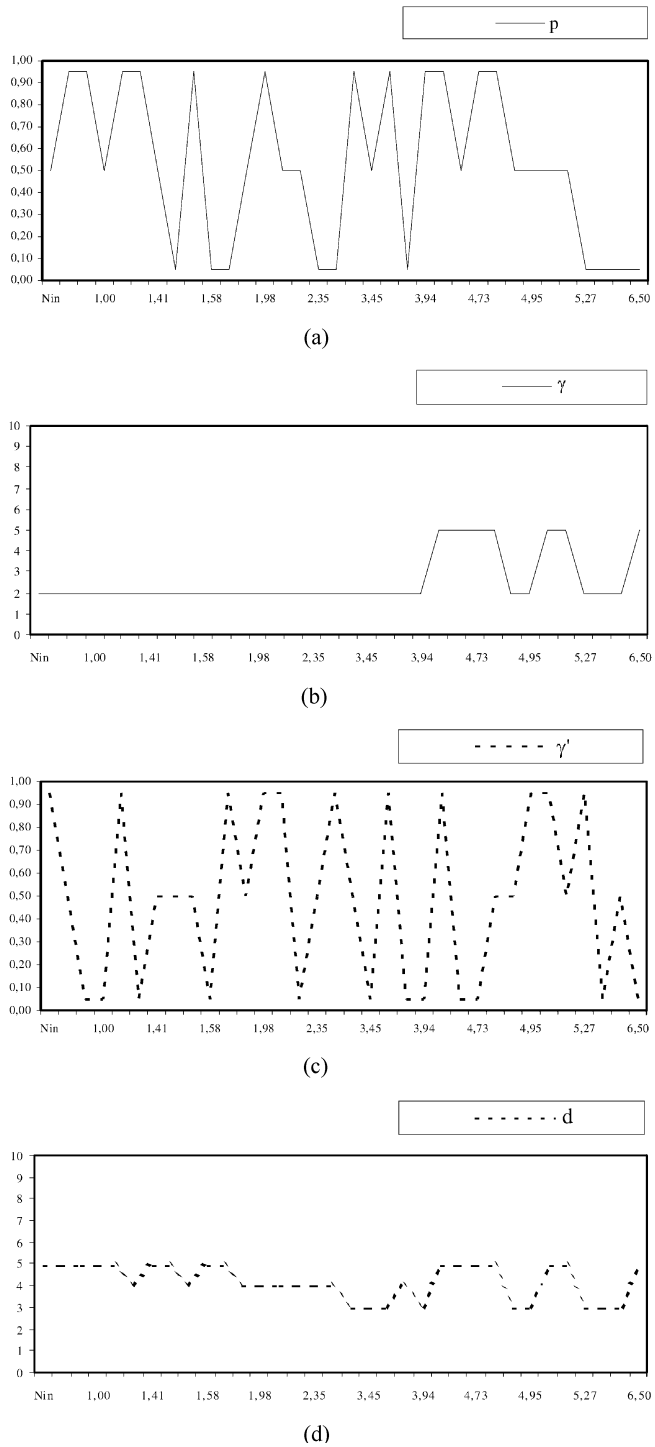
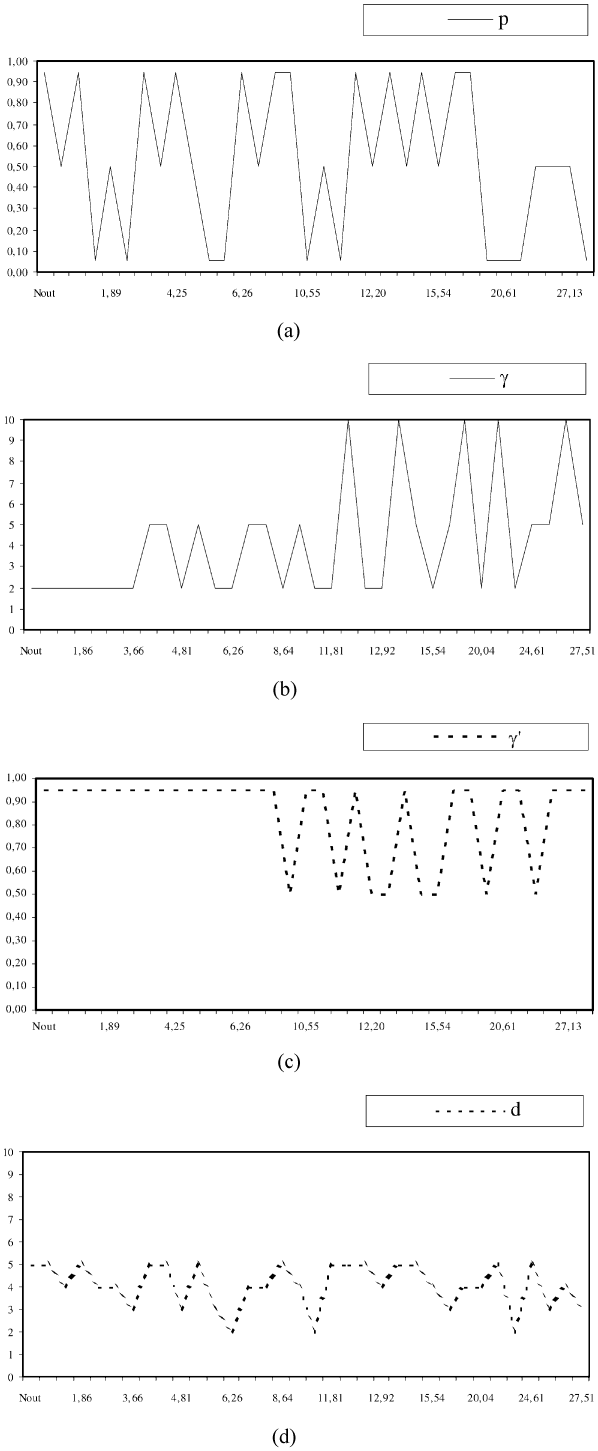


Figure 10 Analysis of [p, γ, γ', d] values for the 35 optimal vectors for N_{out}



- We concluded that the 'self loop' parameter γ should be kept as low as possible; as this is an integer number and $\gamma>1$, its further studying is restricted to the values 2 and 3. Value 1 degenerates the model to a simple preferential attachment approach, but we choose to also use it (although out-of-bounds) to examine interesting features at the degenerate case. We added values 4 and 5 for completeness and validation purposes.
- We concluded that values of the 'tail copy' parameter γ' close to 0.95 are more effective, but this conclusion stands when compared to values 0.05 and 0.5. We now try to pinpoint the area between 0.5 and 0.95, where model effectiveness peaks, using values 0.05, 0.25, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9 and 0.95. This fine-grained approach provides an increased degree of accuracy.
- We use consistently the same values for the 'growth' parameter p.

For each combination of the above values for p, γ, γ' (keeping a constant d=7), we made additional simulation experiments, consisting of 30 runs for 1,000,000 vertices. Results from all experiments were processed with Matlab. We collected the in- and out-degree distributions for each run and computed the average in- and out-degree distributions. We then employed linear regression to find the best power law fit for these distributions on the log-log scale, looking for power law exponents close to the real WWW, where values -2.1 (for the in-degree) and -2.7 (for the out-degree) have been reported (Broder et al., 2000; Kleinberg et al., 1999; Kumar et al., 1999a).

Among the 180 additional experiments carried out (for all parameter value combinations), we present the most promising (where both in- and out-power law fit exponents are close to the desired values) in Table 1.

Table 1 Parameter values of the most promising experiments (d=7)

p	γ	γ'	*IN exponent*	*OUT exponent*
0.05	2	0.7	−2.08	−2.73
0.05	2	0.75	−2.09	−2.60
0.05	3	0.80	−2.24	−2.63
0.5	2	0.65	−2.03	−2.73
0.5	3	0.7	−2.16	−2.73
0.95	2	0.6	−1.99	−2.78
0.95	3	0.7	−2.12	−2.66

Based on these results, we make the following remarks:

- approximations of both the in- and out-degree power law exponents for values of the 'self loop' parameter γ=1, 4, 5 are not satisfactory
- there are effective approximations of both the in- and out-degree power law exponents for low and high values of the 'growth' parameter p
- the best approximations of both in- and out-degree power law exponents appear for values of γ' in the range 0.6–0.8; the best values of γ' seem to be slightly dependent on the choice of the 'self loop' parameter γ: if γ=2, γ' should be close to 0.65; if γ=3, γ' should be close to 0.75.

These results are consistent with the conclusions previously reached on refining the EGC model for producing WWW-like graphs.

7 Conclusions

We presented a simulation-based evaluation of the EGC model for the web graph. Validation was accomplished by measuring the capability of the model to approximate the in/out degree, as this is a key feature of the web for which widely accepted results have been presented. Weak points of EGC were indicated and appropriate parameter values for delivering a realistic model of the web graph were suggested. Overall, the EGC model provides a very good approximation of the in degree distribution but it is not analogously efficient when approximating the out degree. It is eligible for time-driven simulation, since the degree effects are irrelevant of the 'growth' factor p. For d=7, the model parameters must either be γ=2 and $\gamma^{'}\approx$0,65 or γ=3 and $\gamma^{'}\approx$0,75. The in and out degree distributions produced with these values follow power laws with exponents close enough to the desired ones.

Acknowledgements

We thank the anonymous reviewers for their valuable comments on the paper. This research was supported in part by Pythagoras program (MIS 89198) co-funded by the Greek Government and the European Union.

References

Adler, M. and Mitzenmacher, M. (2001) 'Towards compressing web graphs', *Proc. IEEE Data Compression Conference*, pp.203–212.

Aiello, W., Chung, F. and Lu, L. (2001) 'Random evolution in massive graphs', *Proc. 42nd Annual IEEE Symposium on Foundations of Computer Science*, pp.510–519.

Barabasi, A. and Albert, R. (1999) 'Emergence of scaling in random networks', *Science*, Vol. 286, pp.509–512.

Brin, S. and Page, L. (1998) 'The anatomy of a large-scale hypertextual web search engine', *Proc. 7th WWW Conference*. Available from: http://www7.scu.edu.au/programme/fullpapers/1921/com1921.htm

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', *Proc. 9th WWW Conference*, pp.309–320.

Broder, A., Najork, M. and Wiener, J. (2003) ' Efficient URL caching for world wide web crawling', *Proc. 12th WWW Conference*, pp.679–689.

Callaway, D., Hopcroft, J., Kleinberg, J., Newman, M. and Strogatz, S. (2001) 'Are randomly grown graphs really random? ', *Technical Report con-mat/0104546, LANL ArXiV, 2001*. Available from: http://www.arxiv.org/abs/cond-mat/0104546/

Dill, S., Kumar, R., McCurley, K., Rajagopalan, S., Sivakumar, D. and Tomkins, A. (2001) 'Self-similarity in the web', *Proc. International Conference VLDB*, pp.69–78.

Eppstein, D. and Wang, J. (2002) 'A steady state model for graph power law', *Proc. 2nd International Workshop on Web Dynamics*.

Erdös, P. and Rényi, A. (1959) 'On random graphs I', *Publ. Math. Dececen*, Vol. 6, pp.290–297.

Fetterly, D., Manasse, M., Najork, M. and Wiener, J. (2003) 'A large-scale study of the evolution of web pages', *Proc. 12th WWW Conference*, pp.669–678.

Hawking, D., Voorhees, E., Craswell, N. and Bailey, P. (2000) 'Overview of the TREC-8 web track', *Proc. 8th Text Retrieval Conf.*. Available from: http://trec.nist.gov/pubs/trec8/t8_proceedings.html

Kleinberg, J. (1998) 'Authoritative sources in a hyperlinked environment', *Proc. 9th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp.668–677.

Kleinberg, J., Kumar, R., Raghavan, P., Rajagopalan, S. and Tomkins, A. (1999) 'The web as a graph: measurements, models and methods', *Proc. International Conference on Combinatorics and Computing*, pp.1–18.

Kogias, A. and Anagnostopoulos, D. (2003) 'A simulation-based evaluation of the exponential growth copying model for the web graph', *ACM/IEEE WWW2003* (short paper), available from: http://www2003.org.

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

Kumar, R., Raghavan, P., Rajagopalan, S. and Tomkins, A. (1999b) 'Extracting large-scale knowledge bases from the web', *Proc. 25th VLDB Conference*, pp.639–650.

Kumar, R., Raghavan, P., Rajagopalan, S., Sivakumar, D., Tomkins, A. and Upfal, E. (2000) 'Stochastic models for the web graph', *Proc. 41st Annual Symposium on Foundations of Computer Science*, pp.57–65.

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

Tadic, B. (2001) 'Dynamics of directed graphs: the world-wide web', *Physica A*, Vol. 293, pp.273–284.