Experiments with an Economic Model of the Worldwide Web*

Georgios Kouroupas[1], Elias Koutsoupias[2], Christos H. Papadimitriou[3], and Martha Sideri[1]

[1] Athens University of Economics and Business, Patission 76, 104 34 Athens, Greece
{kouroupa, sideri}@aueb.gr
[2] University of Athens, Dept of Informatics and Telecommunications, Panepistimioupolis, Ilissia, 157 84 Athens, Greece
elias@di.uoa.gr
[3] University of California, Berkeley, Soda Hall 689, Berkeley, 94720 CA, USA
christos@cs.berkeley.edu

Abstract. We present a simple model in which the worldwide web (www) is created by the interaction of selfish agents, namely document authors, users, and search engines. We show experimentally that power law statistics emerge very naturally in this context, and that the efficiency of the system has certain monotonicity properties.

1 Introduction

The worldwide web is an unstructured hypertextual corpus of exploding astronomical size and global availability. But perhaps the most fundamental, differentiating characteristic of the web is that is created, supported, used, and ran by a multitude of selfish, optimizing economic agents with various and dynamically varying degrees of competition and interest alignment. Web page authors want their pages to be visited because they benefit from such visits, either directly (e.g., in the case of e-shops) or otherwise (recognition, influence, etc.). End users seek in the web the most relevant and helpful sites for their information needs. Search engines want to improve their reputation (and therefore profits) by helping users to find the most relevant web pages. The selfish nature of the agents suggests immediately that economics and game theory can be used in modeling and understanding the web.

Incidentally, a very obvious and striking example of the game theoretic aspects of the web (not discussed further, however, in this paper) is the so-called *search engine spam*, whereby document authors attempt to deceive the ranking algorithm of the search engine in order to receive a high rank for their page, thus attracting visitors who would otherwise have no interest in them, while search engines devise and deploy countermeasures to such deception.

---

* A preliminary version of this work, without the experimental results, was presented as a poster in WWW 05 [5].

[1] Supported by 'IRAKLEITOS – Fellowships for Research of the Athens University of Economics and Business' grant.
[2] Supported in part by FP-015964 (AEOLUS).
[3] Supported by NSF ITR grant CCR-0121555, and by a grant from Microsoft Research.

X. Deng and Y. Ye (Eds.): WINE 2005, LNCS 3828, pp. 46 – 54, 2005.

The idea that economics can inform web search was first proposed by Hal Varian in [6]; however, the classical "economics of search" discussed there do not seem to apply directly to web information retrieval. Economic concepts were also used in the study of electronic markets [7] and for understanding the ranking of documents [8]. But these works do not develop an economic model of the web.

In this paper we introduce an economic model of the web. Our goals in proposing this model are: simplicity, modeling economy, and predictive power. We think of the web as a theater where three types of agents interact: document authors, users, and search engines. We postulate a utility *U(i,d),* unknown a priori to all agents, that a user *i* obtains if he or she obtains a document *d*; its main characteristics are randomness and clustering. We assume that search engines propose documents to the users, users choose and endorse those that have the highest utility for them, and then search engines make better recommendations based on these endorsements. This is how the web is created: it is the sum total of all these user endorsements.

There are several important questions: Does the structure of the web thus produced resemble that of the real web? How efficient (now in a concrete economic sense) are various search engine algorithms? And how is this efficiency affected by, say, the amount of randomness and clustering of the utility function?

In this paper we address these questions by performing experiments with our model; we hope that theorems will follow. We find that the web graph of our model has power-law distributed degrees, and that the efficiency of the process improves with clustering and endorsement intensity,

In section 2 we describe the model in detail. In section 3 the experimental results are presented. Finally in section 4 we provide some directions for future work.

2 The Model

We aim at a model of the www that captures some of the economic issues involved while at the same time being simple (not obscured by a multitude of extraneous details) and with some predictive power (it behaves, provably or experimentally, in ways consistent with observations about the www, without, of course, encoding such observations in its assumptions).

Our model consists of three types of actors: *documents*, *users* and a *search engines*. There are *m* users, indexed by *i*, that can be thought as simple queries asked by individuals, and *n* documents, indexed by *d*. The search engine, assumed to be unique at this stage of the model, provides users with document recommendations based on information it has about their preferences.

Our main economic assumption is this: We assume that there is a utility *U(i,d)* associated with user *i* and document *d*. This utility value represents the satisfaction user *i* will obtain if he or she is presented with document *d*. The quantitative features of the *m* x *n* matrix *U* are of central importance for our model, and will be discussed in greater detail below. Users know their utility values only for the documents they have been presented so far (since they cannot value something that is unknown to them). *The search engine initially has no knowledge of U,* but acquires such knowledge only by observing user endorsements. In the ideal situation in which the search engine knows *U*, it would work with perfect efficiency, recommending to each user the documents he or she likes the most.

Briefly, the model works as follows: The search engine recommends some documents to the users, initially at random. Every user reviews the documents seen so far and endorses those with the highest utility. To model the limited attention capacity of the user, we assume a bound on the number of documents he or she can endorse in total. The endorsement mechanism does not need to be specified, as soon as it is observable by the search engine. For example, endorsing a document may entail clicking it, or pointing a hyperlink to it. In our model we represent the endorsements as edges from the users to the respective documents. *A basic assumption of the model is that the www is created by this kind of interaction between users and documents.* The bipartite graph $S = ([m],[n],L)$ of document endorsements by users is called *the www state.* Furthermore, the search engine, by observing the www state, recommends new documents to users, who change their endorsements to new, higher utility documents. The search engine is using a *search algorithm*, which is a function mapping the www state to a set of *a* recommendations.

Even at this early stage, some interesting questions arise regarding the model:

1. What are the characteristics of the ultimate www state that results from this process? Do most users point to the highest-utility documents? After how many iterations does the process converge in utility achieved? Does the graph have the peculiar statistics, such as power law distribution, observed in the real web?
2. What is the efficiency or "price of anarchy" [1] of the search algorithm? In other words, which fraction of the maximum possible utility (the ideal situation where each user sees the documents of maximum utility of him or her) can be realized by a search engine?
3. What is the best search algorithm with respect to total utility? That is, which algorithm mapping www states to recommendations optimizes the price of anarchy? Note that a search engine need not be altruistic or socially conscious to strive to maximize social welfare: total user satisfaction would be a reasonable objective for a search engine in a more elaborate model in which multiple search engines compete.

In order to be able to answer to these questions we must define in more detail utility matrix *U*. In this paper we treat experimentally the questions 1 and 2. Question 3, although it is quite important is left open and maybe be the subject of a future work.

2.1 The Utility Matrix *U*

It turns out that the answers to the above questions depend heavily on the quantitative and the statistical characteristics of the utility matrix *U*. If the entries of *U* are completely random and uncorrelated the search engine will be confined to random sampling, naturally with quite poor results. But in reality, utilities are highly correlated. Documents have quality and value that make them more or less useful to users. Also documents and users (recall that by "users" we model queries) are clustered around topics.

To accommodate these characteristics, following [2], we model *U* as a *low rank matrix with added noise*. *U* is generated as follows: There are *k* topics, for some reasonable small number *k*. For each topic $t \leq k$ there is a document vector D_t of length *n*, with entries drawn independently from some distribution *Q*. The value 0 is very probable in *Q* so that about *k* - 1 of every *k* entries are 0. Also for each topic *t* there is a

user vector R_t of length m, whose entries also follow distribution Q, with about m/k non-zero entries. In other words each entry in vectors D_t and R_t represents how relative each document and user is with respect to topic t. Finally, let N be a m by n "noise" matrix with normally and independently distributed entries with mean zero and standard deviation σ. Then the utility matrix is composed as follows:

$$U = \sum_{t=1}^{k} R_t^T \cdot D_t + N \quad (1)$$

In other words, the utility Matrix U is the sum of k rank-one matrices, plus a Gaussian noise. By modeling U like this we ensure that the resulting matrix has the desired properties. So the parameters of the model so far are k, Q, and σ.

2.2 A Search Algorithm and an Endorsement Mechanism

To specify the model in full detail we need to specify a search algorithm and an endorsement mechanism. Through the endorsement mechanism users show their preference is some documents, in a way that is observable by the search engine. In our model users link to the documents they wish to endorse. As we said earlier, there is a finite number of endorsements per user, say b. That is, each user endorses the b highest utility documents he has seen so far.

The search algorithm maps the current www state to a set of recommended documents. A very simple search algorithm is to recommend to each user, at each stage, among documents, with positive utility, the a documents of highest in-degree in the www state, where a is another integer parameter. Like many successful search algorithms, this algorithm takes into account the link structure of the graph. To capture other factors affecting search, besides link structure (such as occurrences of query terms), we have assumed that the search engine has partial knowledge of utility matrix U, by knowing whether an element of it, is zero or non-zero. Also it can be shown that the "highest in-degree" heuristic is a common specialization of the well-known Pagerank [3] and HITS [4] algorithms.

So a and b are two final parameters of our model.

3 Experiments

To validate our model, we ran several experiments for various values of the parameters. The model was implemented as an iterative process. In each iteration, the search engine observed the *www state* and proposed a documents to the users according to the highest in-degree heuristic. After that, users made their decisions, endorsing the b highest utility documents they had seen so far. This algorithm was iterated to convergence, that is, until very few changes were observed in the *www state*. The number of iterations needed for convergence was between 8 and 10. The utility matrix U was constructed according to Eq. 1 and the non-zero elements of distribution Q were chosen randomly from the uniform distribution.

In subsection 3.1 we study the degree distribution of the www state and show that for a wide range of the parameters values follows a power law distribution. In subsection 3.2 we study the efficiency or price of anarchy of the search algorithm.

In the experiments reported here we have assumed that there is no noise, that is, N in Eq. 1 is the zero matrix.

3.1 The Characteristics of the www State

It has been pointed out in several studies [9,10,11,12] that the degrees of the web graph follow power law distributions, that is, the fraction of pages with in-degree *i* is proportional to i^{-x} for some *x>1*, where the value for the parameter *x* is about 2.1 for in-degrees and 2.7 for out-degrees. In our case, only the in-degrees are significant, since out-degrees are, by definition, all equal to *b*.

We find experimentally that, for a wide range of values of the parameters *m, n, k, a, b*, the in-degree of the documents seem to be clearly power-law distributed. This is shown in Figs. 1 and 2. (We have not included in the plot the documents with zero in-degree, which are the majority. In other words, endorsements are concentrated in a very small subset of the documents.) In fact, the exponent in Figure 2 (but not of Figure 1) is quite close to the observed exponent of the in-degree power law in the real www.

More work is needed in order to define how the various parameters affect the exponent of the distribution.

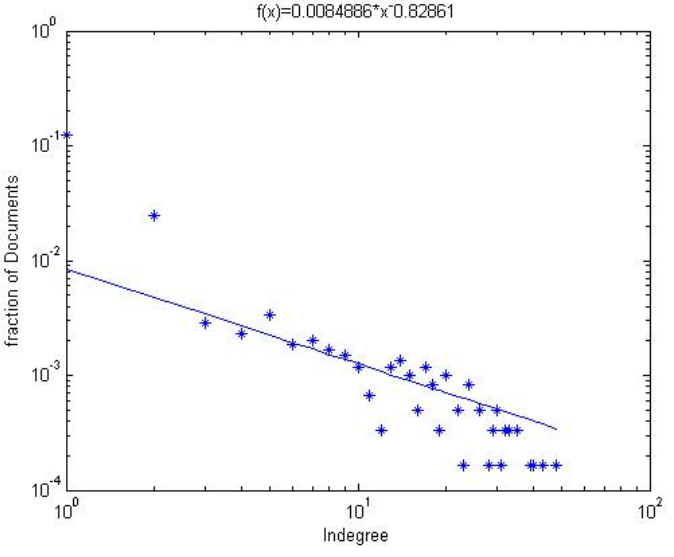

Fig. 1. Log-log plot of the in-degree distribution of an instance of the model with m=1000, n=6000, k=80, a=5, b=5

3.2 The Price of Anarchy

The second question we pose is how efficient the search algorithm can be, meaning which fraction of the maximum total utility can realize during its operation. In this situation the quantity of interest is the price of anarchy as a function of the number of iterations of the algorithm. In all experiments we made, the price of anarchy improved radically during the first 2-3 iterations and later the improvement had a slower rate. This is shown in Fig. 3.

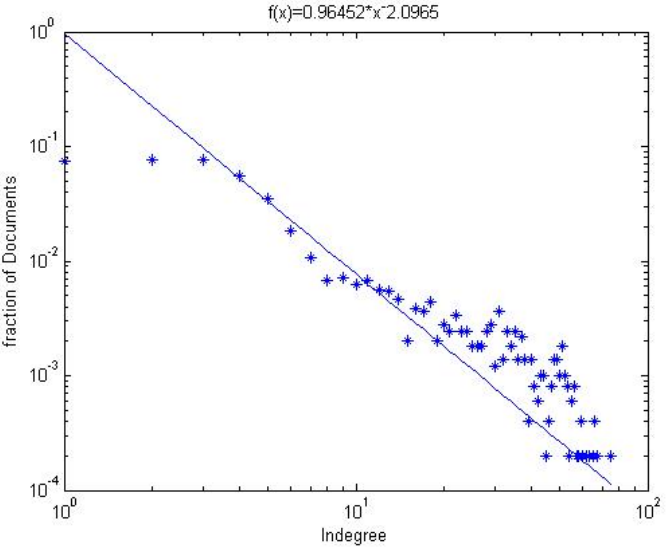

Fig. 2. Log-log plot of the in-degree distribution of an instance of the model with m=3000, n=5000, k=150, a=10, b=10

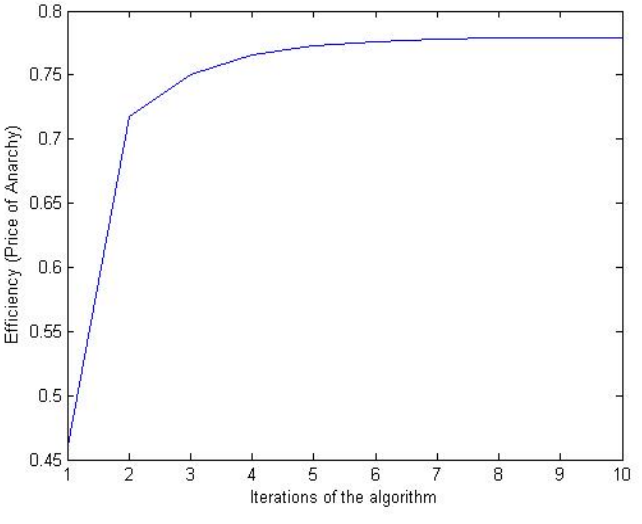

Fig. 3. Efficiency of the Search Algorithm

As we can see the algorithm performs very well since already after iteration 3 has attain more than 75% of the total attainable utility. Of course, adding noise to the model may deteriorate the efficiency.

Another interesting issue is how the efficiency of the search algorithm is affected when we vary the values of *k, a* and *b*. When the number of topics *k* increases the efficiency of the algorithm increases. This fact is clearly shown in Fig. 4.

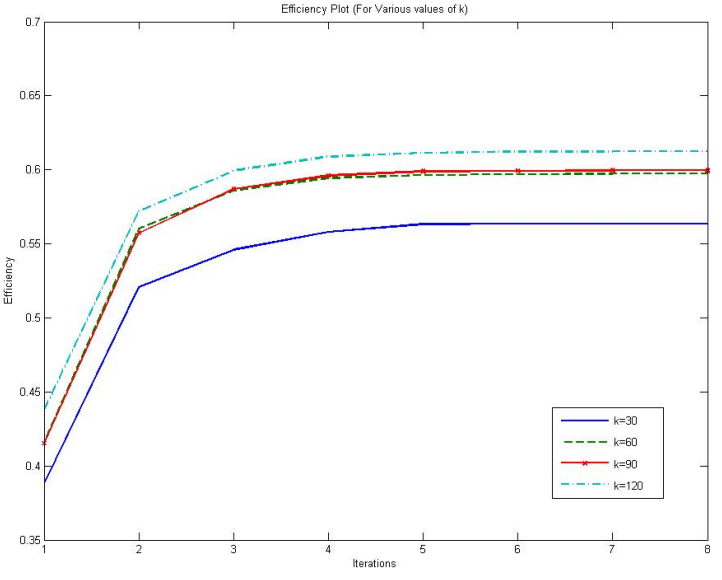

Fig. 4. Efficiency plot for various values of k (number of topics) for an instance of 1000 users, 6000 documents, *a=1, b=1*

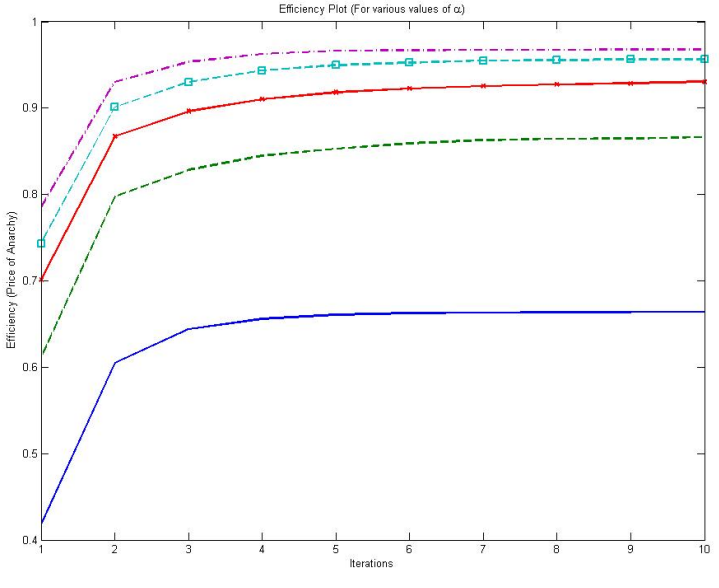

Fig. 5. Efficiency plot for various values of *a* (number of recommended documents) for an instance of 1000 users, 6000 documents, *a=2,4,6,8,10, b=2*

When *a* increases (the number of recommended documents by the search engine) the efficiency of the algorithm also increases. This is shown in Fig. 5 and is quite expected because the users choose from a wider collection of documents.

Increasing *b* (number of endorsed documents per user) causes the efficiency of the algorithm to decrease. This is quite unexpected, since more user endorsements mean more complete information and more effective operation of the search engine. But the opposite happens: more endorsements per user seem to confuse the search engine (see Fig. 6).

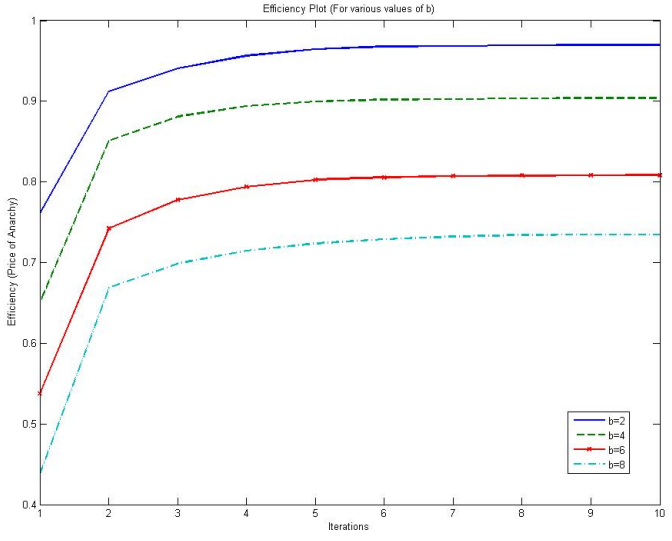

Fig. 6. Efficiency plot for various values of b (number of endorsed documents) for an instance of 1000 users, 6000 documents, *a=8, b=2,4,6,8*

4 Conclusion and Future Work

In this paper we propose a very simple economic model of the worldwide web and present some promising and interesting initial experimental results. Much more comprehensive experiments need to be conducted most notably adding noise to *U*.

Naturally, our ambition is to rigorously *prove* the power law phenomena we observed experimentally, as well as the monotonicity properties of the efficiency. Finally, we hope to use this model to approach the intriguing subject of the optimally efficient search algorithm.

References

1. C. Papadimitriou: Algorithms, Games and the Internet: Proc 2001 STOC
2. D. Achlioptas, A. Fiat, A. Karlin, F. McSherry: Web Search via Hub Synthesis: Proc 2001 FOCS

3. S. Brin, L. Page: The Anatomy of a Large Scale Hypertextual Web Search Engine
4. J. Kleinberg: Authoritative Sources in a Hyperlinked Environment: JACM 46, 5, 1999
5. G. Kouroupas, C. Papadimitriou, E. Koutsoupias, M. Sideri: An Economic Model of the Worldwide Web: Poster, 14th WWW Conference, 2005
6. H. Varian: The Economics of Search: Proc SIGIR 1999
7. T. Koivumaki, S. Svento, J. Pertunen, H. Oinas-Kokkonen: Consumer Choice Behavior and Electronic Shopping Systems – A Theoretical Note: Netnomics 4, 2, 2002
8. C.X. Zhai, W. W. Cohen, J. Lafferty: Beyond Independent Relevance. Methods and Evaluation Topics for Subtopic Retrieval: Proc SIGIR 2003.
9. L. Adamic, B. Huberman: Power Law distribution of the World Wide Web: Science Mag. 287, 2000
10. A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Weiner: Graph Structure in the Web: Proc 9th WWW Conference, 2000
11. A. Barabasi, R. Albert: Emergence of scaling in Random Networks: Science Mag, 286, 1999
12. R Kumar, P. Raghavan, S Rajagopalan, A. Tomkins: Trawling the Web For Emerging Cyber Communities. Proc 8th WWW Conference, 1999