An Economic Model of the Worldwide Web

George Kouroupas
Athens University of Economics and Business
+30-210-820-3149
kouroupa@aueb.gr

Elias Koutsoupias
University of Athens
+30-210-727-5122
elias@di.uoa.gr

Christos H. Papadimitriou
University of California, Berkeley
+1-510-642-1559
christos@cs.berkeley.edu

Martha Sideri
Athens University of Economics and Business
+30-210-820-3149
sideri@aueb.gr

ABSTRACT: We believe that much novel insight into the worldwide web can be obtained from taking into account the important fact that it is created, used, and run by selfish optimizing agents: users, document authors, and search engines. On-going theoretical and experimental analysis of a simple abstract model of www creation and search based on user utilities illustrates this point: We find that efficiency is higher when the utilities are more clustered, and that power-law statistics of document degrees emerge very naturally in this context. More importantly, our work sets up many more elaborate questions, related, e.g., to www search algorithms seen as author incentives, to search engine spam, and to search engine quality and competition.

Categories and Subject Descriptors: Gm, miscellaneous theory; K4, computers and society.

General Terms: Economics, Theory, Algorithms

Keywords: web search, utility function, economic model, market, price of anarchy, power laws, game theory

Motivation

What are the most important characteristics of the current global information environment? Its hypertextual nature? Its exploding astronomical size? Its lack of structure? Its global availability? These are all good answers, of course. But we believe that among the most salient, fundamental, differentiating characteristics of the worldwide web (www) is that, unlike past information systems (which were typically confined within the economic interests of a single enterprise) it is created, supported, used, and run by a multitude of selfish, optimizing economic agents with various (and dynamically varying) degrees of competition and interest alignment: Document authors (who want to be clicked and read as widely as possible), end users (who seek the most relevant and helpful information and most gainful opportunities), and search engines (who want to improve their reputation, measured perhaps in terms of user satisfaction), to name only the obvious ones. The selfish, optimizing nature of the agents suggests immediately that microeconomics and game theory may provide informative contexts and tools for studying the www. For an obvious example (which we are not exploring here), the game-theoretic nature of "google spam" is quite obvious and striking.

Recently we have seen a surge of research in a new important interface between Theoretical Computer Science and Economics/ Game Theory, motivated by the Internet (itself, like the www, the result of anarchic interaction between selfish agents), see e.g. [6]. However, this research has heretofore not touched on the www (besides occasional suggestions in that direction by one of the present authors, see e.g. [7]).

That economics can inform web search was first proposed by Hal Varian in his SIGIR keynote in 1999 [8]. However, the classical economic theories of search outlined there are of little direct applicability to www retrieval; this proposal was followed up, e.g., in the study of specific microeconomic questions about electronic markets [5] or for understanding the ranking of documents [9]. To our knowledge it has not led to the development of economic models specific to www search.

The Model

In this first step of our on-going effort we propose a very simple model of the www that we believe begins to capture the economic issues involved. What we want from such model at this stage is simplicity, elegance, and focus (so its features are not obscured by extraneous issues and details), as well some predictive power (experimental or theoretical results that confirm observations or intuition).

The main ingredients of our model are *documents* (www sites, say) and *users* (a "user" can be thought of as an individual query asked by a particular www user). We assume that there is a utility *U(i,d)* associated with each user *i* and each document *d*, capturing the satisfaction (information wealth or economic opportunity) user *i* would obtain if presented with *d*. This matrix is the basis of our model, and much depends on its quantitative features, discussed extensively below. Another important entity is the *search engine* (assumed to be unique in this simple form of the model), which provides users with document recommendations based on information it has about their preferences.

We assume that *the www is created by this interaction* of documents, users, and the search engine: Users receive the recommendations of the search engine (initially random or at best based on noisy utility statistics) and *endorse* them (click them, point hyperlinks to them, etc. in ways observable by the search engine; our model need not be specific as to the precise nature of such endorsement) depending on the utility received. We assume that, because of limited attention capacity, users can endorse a limited number of documents at any time (this capacity is another key parameter of our model). The current bipartite graph of endorsements is called the *www state*. The www state further informs the recommendations of the search engine, which in turn affect user endorsements (as users abandon earlier favorites for new, higher-utility ones), and so on. Thus, the *search algorithm* of the search engine is a function mapping a www state (bipartite graph of endorsements between users and documents) to a set of

WWW 2005, May 10-14, 2005, Chiba, Japan.
ACM 1-59593-051-5/05/0005.

recommendations. Notice that the search algorithm determines (modulo randomization, of course) the ultimate www state.

Already we can ask some interesting and intriguing questions:

1. What is the efficiency or "price of anarchy" (see, e.g., [6]) of this process? That is, which fraction of the maximum possible utility (the ideal situation in which each user endorses the documents of highest utility to her/him, attainable only in the unrealistic case in which the utilities *U(i,d)* were known) can be actually realized by a search algorithm?
2. What are the characteristics of the ultimate www state that results from this process? For example, does it reflect the peculiar statistics (such as power-law degree distributions) observed on the real www, see e.g. [3].
3. What is the best search algorithm in terms of total utility? That is, which algorithm mapping www states to "document ranking" optimizes the price of anarchy? Notice that a search engine need not be altruistic or socially conscious to strive 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.

It turns out that answers to these questions depend heavily on the quantitative, and indeed the statistical, characteristics of the utility matrix *U.* Intuitively, if the entries of *U* are completely random and uncorrelated, there is nothing that the search engine can do beyond random sampling (with terrible results, of course). And in reality, utilities *are* highly correlated: Documents tend to have intrinsic quality and value that make them more or less useful, queries are clustered in "topics", and a query by a user may be more or less likely to generate high utility.

To accommodate such clustering and correlations, and following the lead of [1], we model *U* as an *m* by *n low-rank matrix with added noise.* That is, we assume that *U* is generated as follows: There are *k topics*, where *k* is some reasonably small number (the rank of *U*). For each topic $t \leq k$ there is a *document vector* D_t, with entries drawn independently from some distribution *Q*; the value 0 is overwhelmingly probable in *Q,* so that about *k*-1 out of every *k* entries of these vectors are zero. Also for each topic *t* there is a *user vector* R_t related to this topic; it has *m/k* nonzero entries also drawn from *Q* (restricted to positive values); furthermore, these *k* sets of nonzero entries form a partition of the set of users 1,…,*m* into *k* sets. Finally, we let *N* be an *m* by *n* "noise" matrix with normally and independently distributed entries with mean zero and standard deviation σ. Then the utility matrix is:

$$U = \sum\nolimits^{k}_{t=1} R_t^T \cdot D_t + N$$

That is, *U* is the sum of *k* rank-one matrices, plus a Gaussian noise (notice that, as a minor point, it may have negative entries). This model is in fact a rather minimalistic way of ensuring that the resulting matrix has the desired properties, and it does so in a quantifiable way. The parameters of the model so far are *k*, *Q*, and σ.

To fully specify the model, we also need to fix (a) a search algorithm, that is, a procedure that recommends a few documents to each user depending on the www state (graph of current endorsements); and (b) the mechanism whereby users choose documents to endorse. The search algorithm is this: at each stage the search engine recommends to each user at each topic the *a* top documents with nonzero D_t entry in this topic that are endorsed by the greatest number of users. It can be shown that this "highest-indegree" heuristic is in this generic case a common specialization of both Page rank [2] and HITS [4]. We also assume that each user endorses the *b* highest-utility documents s/he has seen so far. Here *a* and *b* are the two final parameters of our model.

Within this model, preliminary experiments and theoretical analysis point to the following answers, in connection to questions (1-3) above:

1. It is verified experimentally, and can be shown theoretically for the case $a = b = 1$, that the expected efficiency of the system is a decreasing function of *k* and σ; that is, by increasing clustering and correlation between utilities the expectation of total system utility is increased.
2. We observe experimentally that, in the ultimate www state resulting from this process, document indegrees are indeed power-law distributed for a wide range of parameters. We expect to prove analytically a theorem to this effect.
3. We conjecture that the "highest endorsement" heuristic, which as we mentioned is a common specialization in this case of both Page rank and HITS, is optimal in this model. We can show an analytical lower bound on efficiency; however, it does not yet match the performance of the search engine.

Finally, a host of important questions, far too numerous to list here, are naturally suggested by our model and results so far.

Acknowledgments: Many thanks to Dimitris Achlioptas, Christos Amanatidis, Anna Karlin, Jon Kleinberg, and Milena Mihail for valuable discussions and suggestions at various stages of this work.

References

[1] D. Achlioptas, A. Fiat, A. Karlin, F. McSherry, "Web search via hub synthesis", *Proc. 2001 FOCS*

[2] S Brin, L. Page "The anatomy of a large-scale hypertextual web search engine", www-db.stanford.edu/pub/papers/google.pdf

[3] A. Broder *et al.* "Graph Structure in the web," *WWW9*, 2000

[4] J. Kleinberg "Authoritative sources in a hyperlinked environment", *JACM 46,* 5, 1999

[5] T. Koivumäki, R. Svento, J. Perttunen, H. Oinas-Kukkonen "Consumer Choice Behavior and Electronic Shopping Systems – A Theoretical Note", *Netnomics* 4, 2, 2002

[6] C. Papadimitriou "Algorithms, games, and the Internet", *Proc 2001 STOC*

[7] C. Papadimitriou "The New Problems", *Proc 2003 Workshop in Memoriam of Paris Kanellakis*

[8] H. Varian, "The Economics of Search," *Proc. SIGIR* 1999

[9] C.X. Zhai, W.W. Cohen, J. Lafferty "Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval," *Proc. SIGIR* 2003