A New Paradigm for Ranking Pages on the World Wide Web

John A. Tomlin
IBM Almaden Research Center
650 Harry Road K53/80-2
San Jose, CA 95120
tomlin@almaden.ibm.com

ABSTRACT

This paper describes a new paradigm for modeling traffic levels on the world wide web (WWW) using a method of entropy maximization. This traffic is subject to the conservation conditions of a circulation flow in the entire WWW, an aggregation of the WWW, or a subgraph of the WWW (such as an intranet or extranet). We specifically apply the primal and dual solutions of this model to the (static) ranking of web sites. The first of these uses an imputed measure of total traffic through a web page, the second provides an analogy of local "temperature", allowing us to quantify the "HOTness" of a page.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Experimentation, Theory

Keywords

Search Engines, Static Ranking, Entropy, Optimization

1. INTRODUCTION

Most analyses of the world wide web have focused on the *connectivity* or *graph structure* of the web. The analysis by Broder et al. [4] , discussing the so-called "bowtie" structure is a particularly good example. However, from the point of view of understanding the net effect of the multitude of web "surfers" and the commercial potential of the web, an understanding of WWW *traffic* is even more important. This paper describes a method for modeling and projecting such traffic when the appropriate data are available, and also proposes measures for ranking the "importance" of web sites. Such information has considerable importance to the commercial value of many web sites.

We may abstract the WWW as a graph $G = (V, E)$ where V is the set of pages, corresponding to vertices or nodes and E is the set of hyperlinks (henceforth referred to simply as links) corresponding to *directed* edges in the graph, such that if page i has a link to page j then edge (i, j) exists. For convenience, we define d_i as the *out-degree* of page i; that is, the number of hyperlinks on page i. We also define a *strongly connected component* of G to be a subset $V' \subset V$ of the vertices such that for all pairs of pages $i, j \in V'$, there exists a directed path from i to j in (V', E).

Copyright is held by the author/owner(s).
WWW2003, May 20–24, 2003, Budapest, Hungary.
ACM 1-58113-680-3/03/0005.

For the moment we shall consider an "ideal" model in which the whole graph is strongly connected—that is any site (URL) can be reached by following the links from any other site.

Given such a graph, a popular model of the behavior of web surfers, and hence of the WWW traffic is a *Markov Chain* model. That is, web surfing is assumed to be a random process, where presence at a web page is viewed as a "state", and at every tick of some notional "clock" every web surfer clicks on an out-link from that page with some fixed probability, independent of the path by which the surfer arrived at the page. While this Markov Chain approach has been used in other contexts (e.g [5]), by far the best known application is to the static ranking of WWW pages known as "PageRank" (see [15]). We briefly describe this approach in the following section.

2. STATIC RANKING AND PAGERANK

Link-based ranking schemes are customarily divided into two classes— *query specific* and *static*. A query specific method such as the HITS/CLEVER approach (see [14]) builds a subgraph of the web which is relevant to the query and then uses link analysis to rank the pages of the subgraph. This form of ranking is not addressed in this paper. In a static ranking scheme all pages to be indexed are ordered once-and-for-all, from best to worst—this ordering is the "static ranking" itself. When a query arrives, and some fixed number of pages (say, 10) must be returned, the index returns the "best" 10 pages that satisfy the query, where best is determined by the static ranking. This simple scheme allows highly efficient index structures to be built. The scheme can then be augmented by incorporating other parameters.

PageRank is a static ranking of web pages initially presented in [15], and used as the core of the Google search engine. It is the most visible link-based analysis scheme, and its success has caused virtually every popular search engine to incorporate link-based components into their ranking functions. PageRank uses a Markov Chain model as described above, assuming that the probabilities of following the out-links from a page are equal.

We define a matrix P such that[1]

$$p_{ij} = \begin{cases} d_i^{-1} & \text{if } (i,j) \in E \\ 0 & \text{otherwise} \end{cases} \qquad i, j = 1, ..., n$$

[1]Some descriptions allow for multiple links between i and j by defining $p_{ij} = n_j/d_i$, where n_j is the number of links from i to j, and d_i is the "out degree" of page i, that is the number of out-links. We make the simplifying, but non-essential, assumption that such multiple links are coalesced, and that all $n_j = 1$.

p_{ij} then represents the transition probability that the surfer in state i (at page i) will move to state j (page j). If we let x_i be the probability that the surfer is in state i then by elementary theory (see e.g. [10]) for any initial vector x representing a distribution over possible locations of the surfer:

$$x^T P^k \rightarrow v^T \text{ as } k \rightarrow \infty$$

where v is the vector of probabilities representing the *steady state* of the system, with $v^T = v^T P$.

Let us now consider the "ideal" definition of PageRank [15]—that is page i has rank x_i as a function of the rank of the pages which point to it:[2]

$$x_i = \sum_{(j,i)\in E} d_j^{-1} x_j. \tag{1}$$

This recursive definition gives each page a fraction of the rank of each page pointing to it—inversely weighted by the number of links out of that page. We may write this in matrix form as:

$$x = Ax \tag{2}$$

Note that A is the transpose of the transition probability matrix P.

Now, let us look at the ideal model (2). The PageRank vector x is clearly the *principal eigenvector* corresponding to the *principal eigenvalue* (with value 1) if this is nonzero (see e.g. [12]). Unfortunately, in the real web, many pages having zero in-degree and others have zero out-degree. Even if self loops are added to the latter, or some other means is used to preserve unit row sums of M, this means that the transition matrix will be reducible, and the eigenvector corresponding to the principal eigenvalue will contain a great many zeros.

To get around this difficulty, Page et al [15] proposed an "actual PageRank model":

$$x_i = (1-\alpha) + \alpha \sum_{(j,i)\in E'} d_j^{-1} x_j \quad \forall i \tag{3}$$

or in matrix terms:

$$x = (1-\alpha)e + \alpha A x \tag{4}$$

where e is the vector of all 1's, and α $(0 < \alpha < 1)$ is a parameter. Unless stated otherwise we use a value of 0.9 for α, but Page et al. [15] report using a value of 0.85. This modification clearly overcomes the problem of identically zero PageRank—we may think of (3) as "seeding" each page with a rank of $(1-\alpha)$.

Page et al [15] and others use this device to obtain an analogous eigenvalue problem again, as follows. Let us suppose that in addition to following links out of a page with probability p_{ij} a surfer makes a "random jump" every so often to some other page with uniform probability $1/n$. Let us suppose the surfer follows some link with probability α and makes the random jump with probability $(1-\alpha)$. The modified transition probability is then $(1-\alpha)/n + \alpha p_{ij}$ and the modified method requires us to solve

$$x = [(1-\alpha)E/n + \alpha A]x, \tag{5}$$

where E is ee^T, the matrix of all 1's.

It is easy to show that solving (4) and (5) are equivalent. If we scale the eigenvector obtained from (5) so that $e^T x = n$ we immediately obtain (4). Conversely, taking any solution x of (4) and noting that $e^T A = e^T$, we see that $e^T x = n$, and (5) follows (see [1]) .

[2]This definition is often interpreted to mean that the "importance" of a page depends on the importance of pages pointing to it.

Finally we note that the above modification is precisely equivalent to augmenting the graph G by an additional node $n+1$, and defining an augmented transition matrix:

$$\begin{pmatrix} \alpha P & (1-\alpha)e \\ e^T/n & 0 \end{pmatrix} \tag{6}$$

It is easy to verify that the stationary states of this matrix are in one-to-one correspondence with those of the modified problem above. This augmentation technique will be pursued further below.

3. A NETWORK FLOW APPROACH

In this paper we move beyond the Markov Chain model to a much richer class of models—network flow models (see e.g. [11]).

Again assuming that users click on a link at each tick of the "clock", let us define:

$$y_{ij} = \text{ the number of users following link } (i,j) \text{ per unit time}$$

then we note that

$$H_j = \sum_{i|(i,j)\in E} y_{ij}$$

is the number of "hits" per unit time at node j.

The essence of a network flow model is that the flows are required to satisfy *conservation equations* at the nodes of the network. Assuming that the web traffic is in a state of equilibrium, so that the traffic out of any node is equal to the traffic in per unit time, and initially making the simplifying assumption that the network is strongly connected, these equations are:

$$\sum_{j|(i,j)\in E} y_{ij} - \sum_{j|(j,i)\in E} y_{ji} = 0 \quad (i = 1, ..., n) \tag{7}$$

We also let Y be the total traffic, which in turn equals the total number of hits per unit time, i.e.

$$Y = \sum_{i,j} y_{ij} = \sum_j H_j \tag{8}$$

Usually we prefer to work with normalized values (probabilities) $p_{ij} = y_{ij}/Y$, in which case (7) and (8) become:

$$\sum_{j|(i,j)\in E} p_{ij} - \sum_{j|(j,i)\in E} p_{ji} = 0 \quad (i = 1, ..., n) \tag{9}$$

$$\sum_{i,j} p_{ij} = 1 \tag{10}$$

It is these probabilities that we wish to use our model to estimate.

Now the ideal PageRank model specifies that traffic out of each node be split in fixed (equal) proportions:

$$p_{ij} = \frac{H_i}{Y d_i} \quad \forall (i,j) \in E \tag{11}$$

These values are readily seen to satisfy the conservation equations (9), but this is only one of many possible solutions for this much richer model. Furthermore, we have no *a priori* grounds for imposing this very restrictive solution.

It is then necessary to ask what solution *should* we propose? We may gain some guidance by looking at the models which have been used to estimate traffic patterns and flows in road networks (see [16],[23]). It turns out that both may be derived by examining a quite general class of problem.

4. ENTROPY MAXIMIZATION AND INFORMATION

Following Jaynes [13] we consider the situation where we have a random variable x which can take on values $x_1, x_2, ..., x_n$ with probabilities $p_1, p_2, ..., p_n$. These probabilities are not known. All we know are the expectations of m functions $f_r(x)$:

$$E[f_r(x)] = \sum_{i=1}^{n} p_i f_r(x_i) \quad (r = 1, ..., m) \tag{12}$$

where of course

$$\sum_{i=1}^{n} p_i = 1 \tag{13}$$

Jaynes asserts that "... our problem is that of finding a probability assignment which avoids bias, while agreeing with whatever information is given. The great advance provided by information theory lies in the discovery that there is a unique, unambiguous criterion for the "amount of uncertainty" represented by a discrete probability distribution, which agrees with our intuitive notions that a broad distribution represents more uncertainty than does a sharply peaked one, and satisfies all other conditions which make it reasonable." This measure of the uncertainty of a probability distribution was given by Shannon [20] as:

$$S(p_1, p_2, ..., p_n) = -K \sum_{i=1}^{n} p_i \log p_i \tag{14}$$

where K is a positive constant. This function is the expression for *entropy* used in statistical physics ([2], [19]), information theory and applied probability. Jaynes continues: "It is now evident how to solve our problem; in making inferences on the basis of partial information we must use the probability distribution which has maximum entropy subject to whatever is known." We therefore use the method of Lagrange multipliers to maximize (14) subject to (12) and (13).

Assigning Lagrange multipliers λ_r to the constraints (12), and λ_0 to (13), the unique maximizing solution is easily seen to be of the form:

$$p_i = \exp[-\lambda_0 - \sum_{r=1}^{m} \lambda_r f_r(x_i)] \quad (i = 1, ..., n) \tag{15}$$

and we define the *partition function* of this distribution to be:

$$Z = e^{\lambda_0} = \sum_{i=1}^{n} \exp[-\sum_{r=1}^{m} \lambda_r f_r(x_i)]. \tag{16}$$

Note that the maximum entropy can be expressed in terms of the optimal Lagrange multipliers as follows:

$$S_{max} = \lambda_0 + \sum_{r=1}^{m} \lambda_r E[f_r(x_i)] \tag{17}$$

5. MAXIMUM ENTROPY TRAFFIC DISTRIBUTION

We now apply the preceding general discussion to the specific problem of estimating a traffic distribution on the web which satisfies (9) and (10). The single-subscript probabilities p_i are replaced by the link probabilities p_{ij}, and we see that the equations (9) result if the functions f_r are specified to have the form:

$$f_r(x_{ij}) = \begin{cases} +1 & \text{for j = r, } (i,r) \in E \\ -1 & \text{for i = r, } (r,j) \in E \\ 0 & \text{otherwise} \end{cases}$$

$$E[f_r(x)] = 0$$

It then follows that the maximum entropy web traffic distribution is given by:

$$p_{ij} = \exp[-\lambda_0 - \lambda_i + \lambda_j] \quad \forall (i,j) \in E \tag{18}$$

and the *partition function* of this distribution is seen to be:

$$Z = e^{\lambda_0} = \sum_{(i,j) \in E} \exp[-\lambda_i + \lambda_j] \tag{19}$$

6. BASIC ALGORITHM

To solve this model we might view it as a nonlinear network optimization problem and solve it by application of some general method for this class of problems (see e.g. [6]). However, in view of the special form of the solution displayed in (18) above, we can also use an approach found to be appropriate with other entropy maximization applications ([16],[21], [23]) that is an *iterative scaling* or *matrix balancing* approach. Letting

$$a_i = e^{-\lambda_i} \quad , \quad \mathcal{A} = diag(a_1, ..., a_n)$$

and defining the sparse matrix $\mathcal{M}$ by

$$m_{ij} = \begin{cases} Z^{-1} & \text{for } (i,j) \in E \\ 0 & \text{otherwise} \end{cases}$$

then the solution is given by

$$\mathcal{P} = \mathcal{A}\mathcal{M}\mathcal{A}^{-1}$$

where $\mathcal{P}$ is the matrix of the solution values p_{ij}, which of course must satisfy (9) and (10).

This class of methods has received considerable attention in the literature (see e.g. [8],[17],[18]), though not for problems of web or near-web scale. In practice the following method has worked well.

Initially estimate the value of Z^{-1} and then perform the following steps:

1. Start with initial values for the a_i (e.g. 1.0, or values from a previous run if available), denoted $a_i^{(0)}$, and let $p_{ij}^{(k)} = Z^{-1} a_i^{(k)} / a_j^{(k)}$, though these values are only computed as needed.

2. Compute the row and column sums:

$$\rho_i^{(k)} = \sum_{j|(i,j) \in E} p_{ij}^{(k)} \quad , \quad \sigma_i^{(k)} = \sum_{j|(j,i) \in E} p_{ji}^{(k)}$$

3. Let $\eta_i^{(k)} = (\sigma_i^{(k)} / \rho_i^{(k)})^{1/2}$

4. If $1 - \epsilon \leq \eta_i^{(k)} \leq 1 + \epsilon$ go to step 6.

5. Update $a_i^{(k+1)} = a_i^{(k)} \eta_i^{(k)}$, for some or all of the $\eta_i^{(k)}$ not close to 1, and go to step 2.

6. Check if the final sum of the p_{ij} is sufficiently close to 1.0. If so, exit. If not, adjust the estimate of Z^{-1} and go to 1.

In general we must use a slightly more complicated model.

7. A MODIFICATION OF THE MODEL

The web (and intranets) do not satisfy the strongly connected (SC) property. There are usually many pages (nodes) with no in-links, and many others with no out-links. As in the Markov chain model, this leads to an ill-posed problem, which can be dealt with in a similar, if not identical way. Let us again add an "artificial node" $n+1$ to the model, which is connected by links to *and* from every node of the set V. Let the augmented graph be denoted $G' = (V', E')$, where

$$V' = V \cup \{n+1\}$$

$$E' = E \bigcup_{i \in V} \{(i, n+1)\} \bigcup_{j \in V} \{(n+1, j)\}$$

Variables y_{ij} , and hence p_{ij}, are defined for the new links in E' and the *total* flows in and out of $n+1$ are constrained to be a fraction $(1-\alpha)$ of the overall flow. Thus we have:

$$\begin{array}{l} \sum_{j \in V} p_{n+1,j} = (1-\alpha) \\ \sum_{i \in V} p_{i,n+1} = (1-\alpha) \end{array} \tag{20}$$

Note that this is not quite the same as the modification imposed for the Markov chain model, which fixes the proportion of the $p_{i,n+1}$ and $p_{n+1,j}$ as well as their sum, whereas here they can adopt any positive values subject to (20).

This modification only slightly changes the algorithm described above. The matrices $\mathcal{P}$ and $\mathcal{M}$ are augmented by an $(n+1)$th row and column so that

$$\bar{\mathcal{M}} = \begin{pmatrix} \mathcal{M} & e \\ e^T & 0 \end{pmatrix} \tag{21}$$

and our iterative scaling problem is to find diagonal $\bar{\mathcal{A}}$ and $\bar{\mathcal{B}}$ such that:

$$\bar{\mathcal{P}} = \bar{\mathcal{A}} \bar{\mathcal{M}} \bar{\mathcal{B}}$$

where:

$$\bar{\mathcal{A}} = \begin{pmatrix} \mathcal{A} & 0 \\ 0 & a_{n+1} \end{pmatrix} \ , \ \bar{\mathcal{B}} = \begin{pmatrix} \mathcal{A}^{-1} & 0 \\ 0 & b_{n+1} \end{pmatrix} \tag{22}$$

and the algorithm of the previous section is adjusted so that in step 2 we also compute

$$\rho_{n+1}^{(k)} = \sum_{j \in V} p_{ij}^{(k)} \ , \ \sigma_{n+1}^{(k)} = \sum_{j \in V} p_{ji}^{(k)}$$

and in step 5 also update as follows:

$$a_{n+1}^{(k+1)} = (1-\alpha) a_{n+1}^{(k)} / \rho_{n+1}^{(k)} \ , \ b_{n+1}^{(k+1)} = (1-\alpha) b_{n+1}^{(k)} / \sigma_{n+1}^{(k)}$$

Otherwise the algorithm is unchanged. In what follows we will work only with the expanded edge set, and refer to it for convenience simply as E.

8. THE PRIMAL SOLUTION

The maximum entropy traffic model yields both a primal solution (the p_{ij}) and dual values (the λ_r). As remarked earlier, the number of "hits" at a node is given by the total traffic into (or equivalently, out of) the node, i.e.

$$H_j = \sum_{i|(i,j) \in E} y_{ij} = \sum_{i|(i,j) \in E} p_{ij} Y$$

and this value is available essentially as a byproduct of step 2 of the algorithm. It would seem reasonable to conclude that nodes with a large expected traffic through them should be regarded as "important", and use these quantities as a "traffic" ranking, or *TrafficRank*. Note, however, that the expected traffic through a node will be influenced to some extent by the links *out* of the page—unlike the situation for PageRank, which is specifically defined only in terms of the in-links.

9. THE DUAL SOLUTION

It is frequently the case in optimization models that we can gain considerable information and insight from the dual, as well as the primal solution. This true in the present case, and in a particularly interesting way. We noted earlier that the maximum entropy value is a function of the optimal Lagrange multipliers:

$$S = \lambda_0 + \sum_{r=1}^{m} \lambda_r E[f_r(x_i)] \tag{23}$$

Now varying the functions $f_r(x)$ in an arbitrary way, so that $\delta f_r(x_{ij})$ may be specified independently for each r and (i,j), and letting the expectations of the $f_r(x)$ change in a manner which is also independent, we obtain from (16)

$$\delta \lambda_0 = \delta \log Z = -\sum_r \{\delta E[f_r] + E[\delta f_r]\}$$

and it follows from (23) that

$$\delta S = \sum_r \lambda_r \{\delta E[f_r] - E[\delta f_r]\} \tag{24}$$

Let us denote $\delta Q_r = \delta E[f_r] - E[\delta f_r]$, then

$$\delta S = \sum_r \lambda_r \delta Q_r \tag{25}$$

where we *define* Q_r as the "r^{th} form of heat"(see [13], [22]). We do this by analogy with the classical thermodynamic formula where one has the following relationship between entropy and heat

$$\delta S = \frac{\delta Q}{T}$$

where Q is heat added (this defines absolute temperature T).

From (25) we see that the λ_r play the role of the *inverse* of (local) temperature. Thus by analogy we may propose a page *temperature*:

$$T_r \models 1/\lambda_r$$

and rank the pages by the values of $1/\lambda_r$ from highest to lowest. In practice we may use some function of the λ_r which preserves the same order and since the values $e^{-\lambda_r}$ have this property and fall out of the solution algorithm, we use these values to form the *Hyperlinked Object Temperature Scale* (HOTS). These values may now be used (in decreasing order) to provide a "temperature" or *HOTness* rank, in addition to the TrafficRank described above.

10. COMPUTATIONAL RESULTS

The methods described here have been implemented and tested extensively on graphs resulting from two crawls of the IBM Intranet (yielding about 19 and 17 million pages) and a partial crawl of the WWW made in 2001, yielding about about 173 million pages. In both sets of experiments, the graph is confined to those pages actually crawled. For both crawls, a large number of other pages were linked to, but not crawled. These links and pages are ignored.

We first consider the Intranet results. To test the quality of the results of the Traffic and HOTness ranking, they were compared

with PageRank both empirically, and on two specific test sets of URLs.

The test URLs were those which in the judgement of "experts" should be the primary results of a specified set of queries. Using all three ranking schemes, each page is ranked from highest to lowest (1 through n). Thus each page has a set of three ranks. To measure the "quality of the results", we took the average of the ranks for those test URLs which were covered by the crawl. Thus a low average would indicate results judged favorable by the "experts". The first test set is the smaller of the two, less than 100 pages. The second test set is somewhat larger (about 200). The averages obtained are shown in Table 1.

Test	PageRank	TrafficRank	HOTness
1	0.6443	2.275	0.4610
2	1.242	1.417	1.160

Table 1: Average Ranks of Intranet Test URLs ($\times 10^6$)

For the smaller test set 1, the average value is considerably better for HOTness than PageRank, giving greater "precision" by this measure. The TrafficRank is much worse. This is because these ranks are measuring somewhat different things. PageRank (and evidently HOTness) measures the "attractiveness" of a page, or what is sometimes referred to as *authority* (see [14]). TrafficRank measures total flow through a page. This is affected by its out-links, as well as its in-links, and indeed pages which score well on TrafficRank tend to point to many other pages. Examples are the indices of manuals, and catalogs. Thus this measure tends to capture *hubs* (see also [14]). The test set of URLs used here is intended to be a set of authorities, so the result is not surprising. A similar trend is observed for the larger test set 2. The difference in the averages is somewhat less, but they are in the same relationship.

By ordinary optimization standards, problems with a million or more variables or constraints are presently considered large, and so computing even an aproximate solution to the maximum entropy model for the WWW segment should represent a significant challenge—the associated nonlinear network model has 173 million constraints and over 2 billion variables. Gratifyingly, this very special network model ran, after calibration, in a small multiple (about 2.5) of the time required for the PageRank calculation on a desktop machine. There seems no reason why this approach should not scale in the same way as PageRank to the full web.

To evaluate results from this partial crawl of the WWW we adopt an approach similar to that used for the Intranet - the average position of humanly chosen pages. We use the pages chosen by the Open Database Project (ODP - see http://dmoz.org). The evaluation was structured as follows: Only URLs identified by the "r:resourcE' tag were considered. For all three static ranks, the average rank of such URLs was computed by "level". Thus in Table 2 only those resource URLs at the first level of the ODP hierarchy (such as /Top/Computers or /Top/Games) are considered to be at level 1. Level 2 includes these, plus those at the next level of the hierarchy (such as /Top/Computers/Education), level 3 also includes URLs in /Top/Computers/Education/Internet, etc. Level ∞ includes all resource URLs in the hierarchy. Column 2 of the table gives the number of URLs for each level found in the partial crawl.

For level 1, PageRank wins quite easily. However this very small set of pages is presumably highly authoritative, so the result is not particularly surprising or significant. For all the higher levels we see the HOTness average is better (by about 10%) than the PageRank average, with the TrafficRank inferior to both at levels greater than 3. This is consistent with the intranet results.

Level	Number	PageRank	TrafficRank	HOTness
1	27	0.753	6.404	1.656
2	4258	3.143	2.862	2.614
3	65343	4.448	4.385	3.949
4	228943	4.686	4.887	4.286
5	427578	4.817	5.127	4.438
∞	990354	5.236	5.677	4.812

Table 2: Average Ranks of WWW Test URLs ($\times 10^7$)

11. RANK AGGREGATION

While the new static ranks may be used individually, they may be even more useful when used in conjunction with other static ranks (e.g. PageRank or in-link count), or even with query-specific ranks.

As simple example of aggregation, we took the best (i.e. lowest ordinal) of the three ranks for each URL in the first test set of intranet URLs. The average of these best ranks is now 85,113 — considerably improved. Of course the rank scale has now been compressed by a factor of (at most) three. However, the score has improved by a factor of five.

These preliminary results encourage further experiments with aggregation of static ranks. The use of rank aggregation methods ([7],[9]) has become of growing importance, but its application to static ranking schemes for web pages is in its infancy. Obviously, such schemes need ranks to aggregate, and the two new measures defined here are a significant addition to the link-based schemes available.

12. SPAM

The problem of "spamming" search engines continues to grow. PageRank, which depends only on in-links to confer importance, is thought to be relatively resistant to spam, but there is a cottage industry which attempts to do just that.

Ranking schemes which involve hubs, such as HITS, are more vulnerable to spam, since it is easy to create many out-links, and thus create a hub. Clearly TrafficRank will be vulnerable to spamming in the same way. However, it is not at all obvious how to spam HOTness. Malicious manipulation of the dual values of a large scale nonlinear network optimization model is a problem which has not been studied, to our knowledge. Clearly, this would be an interesting topic for further research.

13. CONCLUSION

In the absence of other information the traffic on the WWW can only rigorously be modeled by use of an entropy maximization procedure. Such a model can be constructed on a large scale and there exists a computationally feasible algorithm for its solution. As a by product of this algorithm two sets of quantities—"traffic" and a local "temperature" are obtained which may be used for ranking pages. This model has the further advantage that it can be adapted to employ such data as may become available on the actual traffic and network behavior.

14. ACKNOWLEDGEMENTS

I would like to acknowledge the help of my IBM colleagues Reiner Kraft, Kevin McCurley and Andrei Broder for their work on quality measure and web crawling, and Michael Saunders and Danzhu Shi of Stanford University for helpful discussions of the algorithms.

15. REFERENCES

[1] A. Arasu, J. Novak, A. Tomkins and J. Tomlin, "PageRank Computation and the Structure of the Web: Experiments and Algorithms", Poster Proc. WWW2002, Hawaii, May 2002. http://www2002.org/CDROM/poster/173.pdf

[2] R. Balescu, "Equilibrium and Nonequilibrium Statistical Mechanics", Wiley, NY (1975).

[3] S. Brin and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine", Proc. of WWW7, Brisbane, Australia, June 1998. See: http://www7.scu.edu.au/programme/fullpapers/1921/com1921.htm

[4] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins and J. Wiener, "Graph Structure in the Web", Proc. WWW9 conference, 309-320, May 2000. See also: http://www9.org/w9cdrom/160/160.html

[5] M. Charikar, R. Kumar, P. Raghavan, S. Rajagopalan and A. Tomkins, "On Targeting Markov Segments", in *Proceedings of the ACM Symposium on Theory of Computing*, ACM Press (1999).

[6] R.S. Dembo, J.M. Mulvey and S.A. Zenios, "Large-Scale Nonlinear Network Models and Their Application", *Operations Research* **37**, 353–372 (1989).

[7] C. Dwork, R. Kumar, M. Naor, D. Sivakumar,"Rank Aggregation Methods for the Web", Proc WWW10 conference, Hong Kong, May 2001. See: http://www10.org/cdrom/papers/577/index.html

[8] B.C. Eaves, A.J. Hoffman, U.G. Rothblum and H. Schneider, "Line-sum-symmetric Scalings of Square Non-negative Matrices", *Math. Prog. Studies* **25**, 124–141 (1985).

[9] R. Fagin, "Combining fuzzy information: an overview", *SIGMOD Record* **31**, 109-118 , June 2002.

[10] W. Feller, *An Introduction to Probability Theory and its Applications, Vol 1 (3rd edition)*, Wiley, NY (1968)

[11] L.R. Ford, Jr. and D.R. Fulkerson, *Flows in Networks*, Princeton University Press, Princeton, NJ, (1962).

[12] G.H. Golub and C.F. Van Loan, *Matrix Computations (3rd edition)*, Johns Hopkins University Press, Baltimore and London (1996).

[13] E. Jaynes, "Information Theory and Statistical Mechanics", *Physical Review* **106**, 620–630 (1957).

[14] J. Kleinberg, "Authoritative Sources in a Hyperlinked Environment", *JACM* **46**, (1999).

[15] L. Page, S. Brin, R. Motwani and T. Winograd "The PageRank Citation Ranking: Bringing Order to the Web", Stanford Digital Library working paper SIDL-WP-1999-0120 (version of 11/11/1999). See: http://dbpubs.stanford.edu/pub/1999-66

[16] R.B. Potts and R.M. Oliver, *Flows in Transportation Networks*, Academic Press, New York (1972).

[17] M.H. Schneider, "Matrix Scaling, Entropy Minimization and Conjugate Duality (II): The Dual Problem", *Math. Prog.* **48**, 103–124 (1990).

[18] M.H. Schneider and S.A. Zenios, "A Comparative Study of Algorithms for Matrix Balancing", *Operations Research* **38**, 439-455 (1990).

[19] E. Schrödinger, *Statistical Thermodynamics*, Dover edition, Mineola, NY (1989).

[20] C.E. Shannon, "A Mathematical Theory of Communication", *Bell Systems Tech. J.* **27**, 379, 623 (1948)

[21] J.A. Tomlin, "An Entropy Approach to Unintrusive Targeted Advertising on the Web", Proc. WWW9 conference, 767-774, May 2000. See also: http://www9.org/w9cdrom/214/214.html

[22] A.G. Wilson, "Notes on Some Concepts in Social Physics", Regional Science Association: Papers, XXII, Budapest Conference, 1968.

[23] A.G. Wilson, *Entropy in Urban and Regional Modeling*, Pion Press, London (1970).

APPENDIX

A. GENERALIZED MODEL

So far we have only assumed that we know the structure of the graph G and the value of the parameter α. The model may be generalized to include other sets of data, or partial data, if they are available. Firstly, there may be a "prior" distribution ω_{ij} of the probabilities p_{ij} postulated. In this case, we may modify the objective function to maximize the "cross-entropy":

$$\text{Max} \quad -\sum_{(i,j)\in E} p_{ij}\log(\frac{p_{ij}}{\omega_{ij}}) \;=\; -\sum_{(i,j)\in E} p_{ij}(\log\omega_{ij} - \log p_{ij}) \tag{26}$$

When there is no such information, the ω_{ij} are assumed equal, and we revert to the original model form (14).

The second set (or sets) of data which might be exploited are those assigning some cost (e.g. congestion) or benefit (e.g. relevance to the current page) to following a link. If we assume that there is some total cost or benefit to be obtained we can add a constraint

$$\sum_{(i,j)\in E} c_{ij}p_{ij} = C \tag{27}$$

Assigning a Lagrange multiplier β to this constraint, and using the a priori probabilities ω_{ij}, the solution of our extended model now has the form:

$$p_{ij} = \omega_{ij}\exp[-\lambda_0 - \lambda_i + \lambda_j - \beta c_{ij}] \quad \forall (i,j)\in E \tag{28}$$

and the partition function is of the form:

$$Z = e^{\lambda_0} = \sum_{(i,j)\in E} \omega_{ij}\exp[-\lambda_i + \lambda_j - \beta c_{ij}]. \tag{29}$$

Computationally the algorithm need only be modified in the obvious way, in the definition of the $\mathcal{M}$ matrix,

$$m_{ij} = \begin{cases} Z^{-1}\omega_{ij}e^{-\beta c_{ij}} & \text{for } (i,j)\in E \\ 0 & \text{otherwise} \end{cases}$$

requiring the initial estimation of β as well as λ_0 and a two dimensional interpolation in the space of these parameters to find the optimal feasible solution. Multiple such constraints can in principle be added, but a Lagrange multiplier must be estimated, or interpolated, for each one. In this situation it may become desirable to consider more general nonlinear network optimization methods.

Finally, we point out that the entropy maximization formalism gives the most likely distribution subject to "whatever is known". If we know the total traffic through a node (i.e. hits H_i for a page) we can incorporate this in the model in the same way as we deal with the "artificial" page $n+1$ — that is by replacing the single conservation equation for that page (7) by the pair:

$$\begin{array}{l} \sum_{(j,i)\in E} p_{ji} = H_i/Y \\ \sum_{(i,j)\in E} p_{ij} = H_i/Y \end{array} \tag{30}$$

The computational algorithm is trivially modified to treat these in the same way as the artificial page.