A Maximum Entropy framework for Higher Order Link Analysis on Directed Graphs

Sreangsu Acharyya
Electrical and Computer Engineering
University of Texas at Austin
srean@lans.ece.utexas.edu

Joydeep Ghosh
Electrical and Computer Engineering
University of Texas at Austin
ghosh@lans.ece.utexas.edu

ABSTRACT

Link-analysis based techniques for ranking of the vertices of a directed graph have been widely studied in the social networks and bibliometrics communities. More recently, they have been popularized in the context of web graphs by the Pagerank [1] and HITS [2] algorithms, both of which under appropriate normalization correspond to different random walk (or "surfing") models. Pagerank is maximally local in the sense that its equivalent surfer ignores the links of the surrounding vertices, whereas the corresponding surfing model for normalized HITS is a second-order model, as its behaviour is independent of the rest of the graph, given the vertices a single hop away. In this paper we propose a way of generalizing these strategies by taking into account non-local effects of higher order, while remaining computationally efficient. The need for such an extension is motivated by the fact that Pagerank and HITS have complementary biases, Pagerank can only take advantage of direct endorsement, whereas HITS can only identify close-knit structures. The approach leads to a series of parameterized schemes, where the value of the parameter determines the weights assigned to the neighbouring vertices, connected either by forward or by backward links (edges). The parametric form allows us to select its value to optimize some desirable quality. Access to "correct" rankings would allow finding the optimal value of the parameter in a supervised learning setup. Typically, however, such data is difficult to come by, so we propose and provide solutions for two optimization criteria (i) maximum entropy, which is motivated by the desire to make minimal extra assumptions and (ii) maximum stability. The framework and techniques developed in this paper apply to a wide range of networks. We empirically illustrate its application to a large citation graph and study both the quality and the stability of the resultant rankings.

1. INTRODUCTION

Methods for establishing an ordering among the vertices V of a directed graph $G(V, E)$ by mapping them to $\mathbb{R}$, form an important class of link analysis techniques. The mapping is usually expressed in terms of the values of the connected vertices and weights on the edges or by defining an equivalent random walk on the graph. The notion of connectedness is directional and, depending on the application, forward and backward edges may have different importance. In the domain of hyperlinked documents, incoming links are deemed more valuable, whereas in a viral marketing scenario the outgoing edges might be more important. This is usually determined from domain knowledge, either explicitly, i.e. the expert assigns explicit values to forward and backward links, or implicitly by specifying the correct ordering. Such implicit information, is more difficult to handle, but can be incorporated, for instance by weighted majority type algorithms [3]. In cases where such domain knowledge does not exist, one has to use a suitable optimization principle. In this paper, an adversarial framework is used wherein one picks an option such that the maximum possible loss is minimized. The other fundamental choice that needs to be made is the "effective size" of the neighborhood. We show how this choice is related to the weightings of higher order link analysis components, and propose the use of the maximum entropy principle to determine the weighting factors.

Our framework is applicable to a broad range of directed graphs, including the World Wide Web, citation graphs and call graphs. However, we shall develop and describe it by drawing an analogy with well-known web surfing models used primarily for re-ranking the results of a keyword based web search. This is done to concretize the description and to help the reader make connections with models he is familiar with. Then we shall empirically apply the model to a citation graph of Computer Science related literature to underscore its generality, and also indicate how it may be useful in text analysis applications.

1.1 Link Analysis on the Web

Several link analysis based techniques for rank-ordering the web pages returned as search results have been effective and popular. Their success have been attributed to the view that the web graph $G(V, E)$ is a distributed recommendation system, where each web-page author confers (corrupted) recommendation to other pages through outlinks. The two most popular link analysis algorithms, Pagerank [1] and HITS [2] are examples of such models. The equivalence between some of these recommendation conferring schemes to their corresponding surfing behaviour has been known [1] for specific cases. This equivalence allows one to consider a surfing model to rank pages by the number of transitions a surfer would have made to these pages, if allowed to surf indefinitely on some subgraph of the web. In this paper, we discuss the surfing behaviours assumed in the existing models and show that they can be interpreted in terms of certain first order and second order locality characteristics, which we make explicit in the sections to follow. As a preview we mention that the pagerank surfer can be looked upon as surfing based on maximally local recommendation, whereas a HITS surfer considers recommendation from pages one hop distance away. This surfing or "flow of recommendation value" view is then naturally generalized to extend the surfing model to a broader family that takes into account such non-local effects of higher order. An upshot of this is that pages are now rewarded not only for having inlinks from influential pages, but also for providing outlinks to good pages that haven't been discovered in a massive scale by others. Pagerank and HITS have been generalized before [4] but not based on their locality of behaviour. Our approach leads to a family of algorithms controlled by a parameter α, which governs the weights apportioned to the recommendations of the experts polled in the neighborhood of the current page but at different hop distances. For random surfer based ranking schemes, a question that can be asked is, which surfing model is best suited for an application, i.e., Pagerank or HITS etc. The Maximum entropy principle provides a conservative answer by selecting one which makes the most non-commital assumption. Algorithms are provided to select appropriate values for α based on such principles. The optimal value would depend on the graph structure.

On the computational side, we note that for a given value of α, the complexity of these algorithms are exactly the same as that of existing link analysis methods, but converge faster under the proper choice of α. In other words the surfers form opinions about the pages rapidly. On graphs that are scale-free with respect to the optimal α; its calculation would be needed only sparingly. A comment is made about using similar ideas in clustering, especially in the domain of vector space modeling of documents where domain dependent distances need to be defined.

In the next section we explain the existing Pagerank and HITS algorithm in the setting of the general muti-order models to motivate the weighted higher order model introduced in Section 3. Section 4 deals with different cost functions to guide the choice of the effective neighborhood via α and devise algorithms for that purpose. The experimental results are presented in Section 5, followed by a sketch on how the model can be applied in the domain of text analysis.

2. PRELIMINARIES

The surfing models of *pagerank* and *normalized hubs & authorities* make different assumptions on the locality of their behaviour. The behaviour of the *pagerank* surfer is conditionally independent of the rest of the graph G given the current page V_i. Thus it neglects the opinion that the surrounding pages might have. Similarly one can show that for the *normalized hubs & authorities* the behaviour is conditionally independent given the current page and its parents (or children). It should be possible to gradually weaken the independence relation by expanding the Markov neighborhood. In this section we review the *pagerank* and *normalized hubs & authorities* models in some detail, to prepare for the natural generalization which follows.

2.1 Pagerank: (1^{st} order surfer)

The pagerank surfer traverses the web graph mostly by following outlinks, with occasional jumps to a random page on the web. The surfer either picks one outlink from the current page uniformly at random or resets to a random page on the web (with probabilities $1-\beta$ and β respectively). The escape probability β was introduced as a mechanism to ensure non-zero pageranks and irreducibility of the resulting Markov chain [1], which is required for the Markov chain to have an unique stationary distribution. Note however that this is just an equivalent view of considering a bounded measure of recommendation value being distributed by outlinks, proportional to the value of the page according to the same measure. The probabilistic interpretation follows as a result of L_1 normalization of this measure.

Let D_{out} be the diagonal matrix formed by the out-degrees, A the adjacency matrix of the web graph, and $N=|V|$ be the number of vertices of G. The transition matrix, T_{pr} of the pagerank surfer can then be expressed in terms of the out-degree normalized adjacency matrix ${}_nA = D_{out}^{-1}A$, and the random jump probability β as follows:

$$T_{pr} = \begin{bmatrix} {}_nA(1-\beta) & \mathbf{e}' \\ 1/N \cdots 1/N & 0 \end{bmatrix} \tag{1}$$

where the components of the vector $\mathbf{e}'$ are given by

$$e'_i = \begin{cases} \beta & \text{if } row_i \text{ of } A \neq \mathbf{0} \\ 1 & otherwise \end{cases} .$$

The page rank iteration converges to the primary eigenvector π of T_{pr}^T or, equivalently, the steady-state probabilities of the Markov chain defined by it. The convergence rate is proportional to λ_1/λ_2, the ratio of the first and second eigenvalues of T_{pr} [1].

The important things to note are that the pagerank surfer considers only those pages that are linked *directly* from the current page and considers them with equal probability. Here we neglect the effect of the small escape term for simplicity of explanation. One of the drawbacks of this surfing mode is that it is oblivious to the close-knit structures in the web, for example, bipartite cliques, which have been observed to have high correlation with coherence of topics [5] and have been actively trawled for their value.

2.2 Hubs and Authorities : (2^{nd} order surfer)

Kleinberg's hubs ($\mathbf{h}$) and authorities ($\mathbf{a}$) ranking of a page is given by the eigenvectors of the normalized [1] co-occurrence and the co-reference matrices A^TA and AA^T derived from

[1]In the original paper Kleinberg [2] used L_2 normalization, however the eigenvectors themselves are independent of the normalization used.

the adjacency matrix A [2]. Bharat and Henzinger [6] suggested a per-page normalized variant of *hubs and authorities* where each row of A^TA and AA^T are normalized to sum to one, in order to limit the overly strong self feedback effect inherent in the original formulation. The *hubs and authority* rankings with this modification can be looked upon as a surfing behaviour, as described below.

The *hubs (and authorities)* model, though initially motivated from the mutually reinforced flow of recommendation on the web [2], can be obtained by considering a Markov chain, in which the transition probability $p(i,j)$ is proportional to the the number of common parents (children) of i and j, i.e. $p(i,j) \propto \sum_k A_{i,k}A_{j,k}$. This induces the transition matrix

$$T_{nh} = D_{nh}^{-1}A^TA.$$

The diagonal normalization matrix D_{nh}^{-1} is chosen so that all the rows of T_{nh} sum to one. Similarly, if we consider transitions proportional to the number of children, we get $T_{na} = D_{na}^{-1}AA^T$. Recognize that the ij^{th} entry of A^TA is the product of their inlink vectors and represents the number of common parents. Thus *hubs* surfing is parent-referred, and *authority* based surfing is child-referred.

So far we have been normalizing the product, i.e., the matrix A^TA or AA^T. However if we interleave the normalization [2] we can define two transition matrices, one for the forward direction $T_f = D_{out}^{-1}A$ as defined in the pagerank model (but with $\beta = 0$), and another for the backward direction by normalizing the transposed adjacency matrix A^T i.e. $T_b = D_{in}^{-1}A^T$. T_f thus represents following outlinks with equal probability and T_b following inlinks with equal probability. The Markov chain defined by the transition matrix below can be used to form rank ordering of search results, very similar to the *hubs and authority* rankings called SALSA [7].

$$T_{salsa} = T_bT_f = D_{in}^{-1}A^TD_{out}^{-1}A.$$

It follows trivially that if the outlink degree is constant (or sufficiently concentrated about its mean) this is the same as the hubs and authorities model. Moreover, as shown elaborately in [7] the normalized in-degree vector $\bar{\mathbf{d_{in}}}$ is the closed form solution for the stationary distribution of the above chain, here we give a short proof for completeness.

Claim: $\bar{\mathbf{d_{in}}}$ *is the stationary distribution of* T_{salsa}

Proof: $T_{salsa}^T.\bar{\mathbf{d_{in}}} = A^TD_{out}^{-1}AD_{in}^{-1}.\bar{\mathbf{d_{in}}}$

$= A^TD_{out}^{-1}A.\mathbf{1} = A^TD_{out}^{-1}.\bar{\mathbf{d_{out}}} = A^T.\mathbf{1} = \bar{\mathbf{d_{in}}}$

Though *hubs & authority*, like flow of recommendation, takes into account opinions of pages one hop distance away, it ignores the recommendation of the current page. Thus it is blind to direct recommendations. It is therefore natural to feel prompted to combine the effects of these two modes: furthermore there is no reason to be limited to a second order effect only. The resulting generalization is discussed in the next section.

[2]By interleaving L_1 normalization we obtain

$$\mathbf{a}_{t+1} \propto A^T\mathbf{h}_t \rightarrow D_{in}^{-1}A^T\mathbf{h}_t \quad (2)$$

$$\mathbf{h}_t \propto A\mathbf{a}_{t-1} \rightarrow D_{out}^{-1}A\mathbf{a}_t. \quad (3)$$

3. WEIGHTED HIGHER ORDER SURFER

In the previous section we saw two different surfing models, one in which the surfing was dictated only by the outlinks (pagerank) and the other specified only by *parental* or (*children's*) supervision (*hubs and authorities*). However, considering the sparsity of outlinks, a superposition of these strategies may be more desirable, since it would make it possible to utilize greater number of forward pointers to (possibly) related pages.

In this section we first examine the case of undirected graphs, which is considerably simpler as the higher order effects can be captured easily by the powers of the symmetric adjacency matrix (with interleaved normalization). Surprisingly, it turns out that this model is equivalent to that of a lethargic surfer discussed later. One can also show that the exponentiated matrix suggested by Miller, Rae et al., [8] as means of guarding against repetition of the principal eigenvalue can be derived from this framework.

3.1 Undirected Graph

The weighted higher order surfing model on an undirected graph is captured by

$$T_{ho} \propto \sum \alpha_i T^i$$

where α_i is a sequence such that T_{ho} converges. Note that both pagerank and *hubs and authorities* are specific instantiations of this form. Certain choices of α_i lead to analytically closed form expressions for T_{ho}. A possible sequence $\alpha_i = \frac{1}{i!}$ leads to $P = exp^T$ a very well known form in matrix analysis. The exponentiated transition matrix was also proposed by Miller, Rae et al., [8] as an ad-hoc substitute for the adjacency matrix to deal with the problem of repeated eigenvalues and reducibility. Here we can see that it can be derived from a simple surfing model. In [8] the authors do not offer an efficient means of calculating exp^T for large matrices, which is necessary for their algorithm to be applied in practice.

We choose a different sequence that is computationally tractable and at the same time captures the desired flavor and has similar guarantees of irreducibility:

$$T_{ho} \propto I + \alpha T + \alpha^2 T^2 \cdots - I = (I - \alpha T)^{-1} - I = \alpha T(I - \alpha T)^{-1} \quad (4)$$

Furthermore, we define the sequence in terms of a parameter α, the value of which can be chosen to achieve different goals. Before we progress further, we need to ensure that $(I - \alpha T)^{-1}$ in equation (4) exists, the condition for which is $||\alpha T|| < 1$ or $\alpha < 1$.

Claim: $(I - \alpha T)^{-1}$ exists for $\alpha < 1$.
Proof: Assuming $(I - \alpha T)$ to be singular we can solve $(I - \alpha T)x = 0$ for non-trivial $x = \alpha Tx$. Taking L_1 norm on both

sides we get $|x| \leq \alpha|T||x| = \alpha 1|x|$ Thus by contradiction, $\alpha < 1$.

The matrix T_{ho} is term-wise positive, with row sum of $\frac{\alpha}{1-\alpha}$, thus inducing the effective transition matrix

$$T_{ho} = \frac{1-\alpha}{\alpha}[(I-\alpha T)^{-1} - I]. \quad (5)$$

Claim: The stationary probability distribution or the principal eigenvector π of T_{ho}^T is invariant of α.
Proof: $T_{ho}^T = \frac{1-\alpha}{\alpha}[(I-\alpha T^T)^{-1} - I]$ $T^T\pi = \pi \rightarrow T_{ho}^T\pi = \frac{1-\alpha}{\alpha}[(I-\alpha T^T)^{-1}\pi - \pi] = \frac{1-\alpha}{\alpha}[\frac{1}{1-\alpha\lambda}\pi - \pi]$.
The stationary probabilities are independent of the choice of the parameter α, so we can choose it to impose some desirable qualities. The new and the old eigenvalues, however, are related as $\frac{\lambda_{i\ ho}}{\lambda_{j\ ho}} = \frac{\lambda_i}{\lambda_j}\frac{1-\alpha\lambda_j}{1-\alpha\lambda_i}$, which shows us that it is order preserving and we can select α such that the eigengap between the first and the second is increased. Note that this amounts to the well known "shifted-inverse method" of finding the eigen-vectors of a matrix [9].

3.1.1 Lethargic Surfer

In this section we examine a feasible generative model, that leads to the behaviour we have discussed and indicate why the stationary distribution values may be be useful for search. The traditional models of surfing behaviour try to summarize the effect of traversing the web for an infinite amount of time according to the transition matrix induced by the web graph. Consider an impatient and lethargic surfer who is interested in finding one particular page on the web v^*, so much so that a reward of 1 is assigned to this page and 0 to others. Being lethargic, at every time step the surfer either decides to stay on the current page with probability $1-\alpha$ or check the next pages with probability α (where next page is defined in terms of the current page and the transition matrix T). Let $\mathbf{v}$ be the expected value vector [3] of the pages (the probability of being rewarded). It must satisfy the equation $\mathbf{v} = (1-\alpha)\mathbf{v} + \alpha T\mathbf{v}$ or,

$$\begin{aligned}
\mathbf{v} &= \frac{1}{1-\alpha}\mathbf{v} - \frac{\alpha}{1-\alpha}T\mathbf{v} \quad = \tfrac{1}{1-\alpha}(I-\alpha T)\mathbf{v} \\
(I-\alpha T)^{-1}\mathbf{v} &= (I + \frac{\alpha}{1-\alpha})\mathbf{v} \\
(I-\alpha T)^{-1}\mathbf{v} - I\mathbf{v} &= \frac{\alpha}{1-\alpha}\mathbf{v} \\
\frac{1-\alpha}{\alpha}((I-\alpha T)^{-1} - I)\mathbf{v} &= \mathbf{v}.
\end{aligned}$$

Thus $\mathbf{v}$ is the eigenvector of $\frac{1-\alpha}{\alpha}[[I-\alpha T]^{-1} - I]$ and therefore equivalent to the higher order surfer model described in the previous section, equation (5).

3.2 Directed Graph

The adjacency matrix of a directed graph is not symmetric and the previous derivation of the effective transition matrix do not hold. The non-commutative nature of matrix multiplication makes it possible to define several variants of higher order schemes, none of which are more fundamental than the others. We pick a few instances, that appear to be ad hoc but have a connection that they can be reduced to a mixture of forward and backward models, as explained later in this section. One particular instantiation is to consider only the forward view, as pagerank does. This gives us

$$T_{eff} = \frac{1-\alpha}{\alpha}[(I-\alpha T_f^T)^{-1} - I].$$

Alternatively, one can obtain *grandparent authorities* by considering multi-order supervision from parents, grandparents and higher ancestors.

$$T_{ga} \propto T_f + T_f(\alpha T_b) + T_f(\alpha T_b)^2 \cdots = T_f(I-\alpha T_b)^{-1}. \quad (6)$$

Similarly, multi-order *grandchildren hubs* are obtained from

$$T_{gh} \propto= T_b + T_b(\alpha T_f) + T_b(\alpha T_f)^2 \cdots = T_b(I-\alpha T_f)^{-1}. \quad (7)$$

After normalization we obtain for the authorities $T_{ga} = (1-\alpha)T_f(I-\alpha T_b)^{-1}$, the normalization constant being the same for the hubs. The *grandparent authority* weights are obtained as the eigenvectors of $T_{eff}^T = (I-\alpha T_b^T)^{-1}T_f^T$ which can be obtained from the generalized eigenvalue problem

$$(I-\alpha T_b^T)\pi_{ga} = \gamma T_f^T\pi_{ga}.$$

Other generalizations to the directed multi order surfing are also possible. *Generalized Salsa* is derived for the forward backward model (T_f, T_b), followed by the *Generalized Hubs-Authorities* model where mutual reinforcements are not restricted to binary interactions, but can take place along paths of any hop distance.

To obtain the effective transition matrix for these different modes, the basic premise remains the same i.e. of considering higher order transitions with decreasing weights. In *generalized salsa* all possible higher order effects are included, for instance for order = 2 we have to consider $\{T_b^2, T_bT_f, T_fT_b, T_f^2\}$, similarly for order n we have to consider all 2^n transitions of n steps by which we can get from page i to page j.

$$\begin{aligned}
T_{eff} \propto\ & T_b + T_f + \alpha(T_b^2 + T_fT_b + T_bTf + T_f^2) + \\
& \alpha^3(T_b^3 + T_b^2T_f + T_bT_f^2 + T_bT_fT_b + T_fT_b^2 + \\
& T_fT_bT_f + T_f^3 + T_f^2T_b) \cdots
\end{aligned}$$

$$T_{eff} \propto \quad [I - \alpha(T_b + T_f)]^{-1} - I \quad (8)$$

The stationary distribution is given by the eigenvectors of $(T_b + T_f)$. As a curiosity we observe that if the web graph is regular, the effective transitions are made symmetric and the ranks given as the eigenvector of $(A + A^T)$, which has the normalized vertex degree vector $\bar{\mathbf{d}}$ as its closed form solution.

[3] Being an expectation of an indicator variable, it is also the probability of the state. It is also equivalent to the reinforcement learning model where the reinforcements are discounted by α.

Claim: $\bar{\mathbf{d}}$ *is the stationary distribution of the Markov chain induced by* $(A^T + A)$

Proof: $(A^T + A).\bar{D}^{-1}.(\bar{\mathbf{d}}) = (A^T + A).\mathbf{1} = \bar{\mathbf{d}}$

Note that we can construct several other models by engineering the r^{th} term of the equivalent matrix. The r^{th} term is of the form $\alpha_r T_{1*}T_{2*}\cdots T_{r*}$, where the the subscripts are either f to signify the forward chain, or b to signify the backward chain. Note that in the three modes introduced so far, T_{gh} and T_{ga} correspond to cases where $1*$ is constrained to T_f and the rest to T_b and vice versa, whereas there are no such constraints for T_{gslasa}. Observation that the stationary distribution of these chains and several others in this family where the first term T_{1*} is constrained and the remaining terms may or may not be, are equivalent to the chain

$$\alpha' T_f + (1-\alpha')T_b, \tag{9}$$

where α' is some function of α. Note for $\alpha' = 1$ it reduces to the pagerank model.

Grandparent Authorities : $(I - \alpha T_b^T)\pi_{ga} \propto T_f^T \pi_{ga} \implies \pi_{ga} = \left[\alpha' T_b^T + (1-\alpha')T_f^T\right]\pi_{ga}$.

Grandchild Hubs: $(I - \alpha T_f^T)\pi_{gh} \propto T_b^T \pi_{gh} \implies \pi_{gh} = \left[\alpha' T_b^T + (1-\alpha')T_f^T\right]\pi_{gh}$.

Generalized Salsa: $\pi_{gs} = \left[\alpha' T_f^T + (1-\alpha')T_b^T\right]\pi_{gs}$ where $\alpha' = 0.5$.

Constraining $T_{1*} = T_f \implies T_{eff} \propto T_f[1-\alpha(T_f+Tb)]^{-1} \implies \pi = \left[\alpha' T_f^T + (1-\alpha')T_b^T\right]\pi$.

Constraining $T_{1*} = T_b \implies T_{eff} \propto T_b[1-\alpha(T_f+Tb)]^{-1} \implies \pi = \left[\alpha' T_f^T + (1-\alpha')T_b^T\right]\pi$.

A related family closely connected to these can be obtained by constraining the last term T_{r*} of $\alpha_r T_{1*}T_{2*}\cdots T_{r*}$. It can be shown that the stationary distributions of this family are related to the ones shown above as $\pi_{\text{``}dual\text{''}} = (I - \alpha T_*^T)\pi_{\text{``}primal\text{''}}$. We omit the simple demonstration.

The linear interpolation of forward and backward chains do not subsume all possible effects, for instance, the *generalized hubs and authorities* from the mutually re-inforcing principle. Given a hub weight vector $\mathbf{h_t}$, the authority weight of a page p can be calculated as the sum of the hub weights of all pages from which the the page p can be reached in the 0^{th} step, first step, and higher steps with appropriate discounting. The hub vectors in turn can be similarly calculated using the authority vector $\mathbf{a_t}$.

$$\begin{aligned}
\mathbf{a_{t+1}} &= \mathbf{h_t} + \alpha A^T \mathbf{h_t} + \alpha^2 (A^T)^2\, \mathbf{h_t} \cdots = [I - \alpha A^T]^{-1}\mathbf{h_t} \\
\mathbf{h_t} &= \mathbf{a_t} + \alpha A \mathbf{a_t} + \alpha^2 A^2 \mathbf{a_t}\ \cdots = [I - \alpha A]^{-1}\mathbf{a_t} \\
\mathbf{a_t} &= [I - \alpha(A^T + A) + \alpha^2 A A^T]^{-t}\mathbf{h_0}
\end{aligned}$$

Thus the *generalized hubs and authorities* are given as the principal eigenvectors of $[A^T + A - \alpha A A^T]$ and $[A^T + A - \alpha A^T A]$; which can be looked upon as $\mathcal{O}(\alpha A A^T)$ or $\mathcal{O}(\alpha A^T A)$ perturbation of a symmetric multi-order surfing model.

However, since the linear interpolation form characterizes a significant subset of the possible higher order effects, we restrict our attention to this family only and focus on the means of choosing an appropriate value for α. The question about the choice of an appropriate α may also be posed in a different way as follows: Assuming that the surfer chooses the forward and reverse jumps stochastically, find the parameter that optimizes a desirable quality, for example how fast his walk converges, how fast he explores all the vertices, etc.

4. CHOICE OF α

The maximum entropy (maxent) principle [10, 11] provides a way of parameter selection when no other guiding principle can be found. The maxent principle ensures that minimum extra assumptions are made and is thus maximally non-committal. When applied to our domain, maximizing the entropy tries to decorrelate the consecutive states of the Markov chain, thereby reducing the memory of the chain and promoting rapid mixing and consequently stability. Another beneficial consequence is that the resulting ranking scheme is conservative i.e. the ranks are as close to the uniform distribution as possible on the induced graph. This ranks a particular page higher than some other only if there are compelling reasons to do so.

Another criteria that we consider to guide our choice of α is to select a value that generates maximally rapidly mixing Markov chains. This can be motivated both from the point of view of accelerated convergence of the algorithm and also the fact that rapid mixing Markov chains have high eigen-gap and as will be shown shortly, are more resistant to perturbations.

In order to design algorithms to optimize the cost functions mentioned above, we will make use of some results relating to perturbation of Markov chains and properties of its second distinct eigenvalue. We introduced these simple results here, before discussing the algorithms in a later section. We refer to the corresponding literature rather than including the proofs, unless they are short enough to fit in this paper.

Lemma 1. *Given a discrete Markov chain X with initial distribution μ_0, stationary distribution π and a diagonalizable transition probability matrix T, its distribution μ_k at step k can be bounded from stationarity as $||\mu_k - \pi|| \le C\lambda_2^k$ where λ_2 is the second distinct highest eigenvalue of the Markov kernel T and C is some constant. Any other bound of the form $||\mu_k - \pi|| \le C'\eta^k$ implies $\eta \ge \lambda_2$.*

Next we introduce the interesting concept of a fundamental matrix of a Markov chain [12], that we use to establish some simple bounds. Let T_1 be the transition matrix induced by the adjacency matrix A. Suppose that the original matrix T_1 is perturbed to T_2 as a result of which the stationary probability shifts from π_1 to π_2. In order to quantify the effect of this perturbation consider the matrix $M = [I - T_1^T + T_1^{T^\infty}]^4$ called the fundamental matrix. M has the following

[4] $T^\infty = Lt_{n\to\infty} T^n$

property that will be used in the analysis of the algorithms presented later.

Lemma 2. *The change in the stationary distribution π can be expressed in terms of its inverse fundamental matrix as $\Delta\pi = M^{-1}\Delta T\pi_2$, whose L_2 operator norm $||M^{-1}||_2 = (\frac{1}{1-\lambda_2})$ bounds perturbation norm of stationary distribution by $(\frac{1}{1-\lambda_2})E[||\Delta T||_2]$*

Proof.

$$M(\pi_2 - \pi_1) = \pi_2 - T_1\pi_2 + T_1^{\infty}\pi_2 -$$

$$\pi_1 + T_1\pi_1 - T_1^{T^{\infty}}\pi_1 = (T_2^T - T_1^T)\pi_2$$

Let v be an eigenvector of the matrix T^T with eigenvalue $\lambda < 1$. $Mv = [I - T^T + T^{T^{\infty}}]v = v - \lambda v + T^{T^{\infty}}v = (1-\lambda)v$. Thus the eigenvalues of $M^{-1} = \frac{1}{1-\lambda}$. For $1 > \lambda_2 \geq \cdots \lambda_n$, the largest eigenvalue of M^{-1} is $\frac{1}{1-\lambda_2}$ implying that $||M^{-1}||_2 = (\frac{1}{1-\lambda_2})$.

$$E[||\Delta\pi||_2] \leq ||M^{-1}||_2 E[||\Delta T||_2]||\pi_2||_2$$

$$\leq (\tfrac{1}{1-\lambda_2})E[||\Delta T||_2]$$

$$E[||\Delta\pi||_1] \leq (\tfrac{\sqrt{2}}{1-\lambda_2})E[||\Delta T||_1]$$

□

4.0.1 Coupling inequality

The coupling inequality that we state below offers perhaps the most elegant and concise way of bounding the mixing rate of a Markov chain. The argument here follows that of Rosenthal in [13], but is simpler since here we are interested only in minimizing the mixing time and not in proving rapid mixing.

Lemma 3. *Given a discrete Markov Chain taking values in the set V with a subset $v_s \in V$, initial distribution μ_0 and a stationary distribution π, the distribution at step k can be bounded as $||\mu_k - \pi|| \leq (1-\gamma\eta)^k$ where γ and η are defined as below.*

Proof. The total variational distance between the distributions of two discrete random variables X, Y taking values in the set $\mathcal{A}$ is bounded as $sup_{A\in\mathcal{A}}||P(X \in A) - P(Y \in A)|| \leq P(X \neq Y)$ by the coupling lemma [13].

Consider two Markov chains X and Y defined over the state space (of vertices) V with transition probabilities T and initial distributions μ_0 and π, respectively. Let $v_s \subset V$ be a subset such that $Pr((X,Y) \notin (\{v_s\} \times \{v_s\})) = \gamma$ and $\forall x \in v_s T(x,V) \geq \eta\xi$, allowing us to simulate the Markov chain with a biased coin ($P(heads) = \eta$) and the distribution ξ on V as follows:

$$T = \gamma T + (1-\gamma)\left[\eta\xi + (1-\eta)\left[\frac{T-\eta\xi}{1-\eta}\right]\right]$$

If $(X_{i-1}, Y_{i-1}) \notin (\{v_s\}X\{v_s\})$ choose both X_i, Y_i independently from T, else if $(X_i, Y_i) \in (\{v_s\}X\{v_s\})$ and the coin comes heads, we sample according to the distribution ξ and set $X_i = Y_i$, else both of them are selected independently according to $\frac{T-\eta\xi}{1-\eta}$. Clearly X_k and Y_k can be made equal after the first head, thus we can bound μ_k from stationarity as follows $||\mu_k - \pi|| \leq (1-\gamma\eta)^k$. Moreover from the relation $\eta\xi(i) \leq Min_{x\in v_s}(T(x,i))$, we get the highest value of $\eta = \sum_V Min_{x\in v_s}(T(x,i))$.

□

Having introduced the mathematical tools that will be used, we describe the algorithms in the next section.

4.1 Maximum Entropy Rate

The intent of link-analysis can be posed as modeling the behaviour of an actual surfer and identifying the page she would visit most frequently. In absence of any further information, the most natural and principled selection of α is one that is maximally non-committal about the assumptions made on the surfing behaviour. This can be obtained by Maximizing the entropy rate [14] of the Markov chain used.

The entropy rate of a Markov chain [11] is the minimum number of bits that would be required on average to index the visited states of the Markov chain. In the worst case this equals $log|V|$, but may be lower, depending on how skewed the limiting distribution is. Maxent may be motivated for this application both from the view that it favours uniform, i.e., conservative, rankings and also from the fact that it makes minimum extra assumptions about the surfing behaviour.

We will develop some intuition about this cost function, based on the fact that it is also the conditional entropy of state X_t, conditioned on the previous state X_{t-1}, demonstrated as follows. The entropy rate of a Markov chain is given by $\underset{n\to\infty}{Lt}\frac{H(X_1,X_2,\cdots,X_n)}{n} = \underset{n\to\infty}{Lt}\sum_t \frac{H(X_t|X_1,\cdots,X_{t-1})}{n}$ which equals $H(X_t|X_{t-1})$. Here we have used the independence and stationarity property of the Markov Chain. Each row of the transition matrix T of a Markov chain represents a conditional probability distribution. The expectation of the entropies of the conditional distributions are taken over the stationary probability distribution π. Deviating slightly from customary notation, we represent the entropy of the conditional distributions for each row by the vector $h(\alpha) = [h_1(\alpha), \cdots, h_N(\alpha)]^T$, which allows us to rewrite the cost function as

$$H(\alpha) = \underset{X_{t-1}}{E}\left[\!\!\left[-\underset{X_t|X_{t-1}}{E}[\![logP(X_t|X_{t-1})]\!]\right]\!\!\right] = \pi^T[h_i(\alpha)].$$

In spite of the similarity of the above cost function to $\underset{X_{t-1}}{E}[\underset{X_t|X_{t-1}}{E}[logP(X_t, X_{t-1})]]$, (the cost function maximized by the EM algorithm) it cannot be maximized by a similar process of estimating the distribution and maximizing the expectation as followed in the EM family of algorithms. This inconvenience follows from the fact that unlike likelihood the conditional entropy of the Markov chain is not bounded by its values calculated from incorrect distributions, thereby allowing the cost function to decrease when the limiting distribution is re-estimated. We provide two iterative algorithms to address the cost function above.

4.1.1 Max Min Entropy

The analytical difficulty of maximizing the entropy rate of the Markov chain, brought about by the complicated dependence of the limiting distribution to α can be mitigated, if we are ready to sacrifice the above cost function for its lower bound $Max_\alpha Min_\pi H(\alpha) = Max_\alpha Min_\pi \pi^T h(\alpha)$. We proceed as if the dependence of the stationary distribution to π is unknown and prepare for the worst. Concavity of entropy in the distribution ensures $h(\alpha x + (1-\alpha)y) \geq \alpha h(x) + (1-\alpha)h(y)$, allowing us to maximize the lower bound as shown below. The linearity of the cost function in the parameter α ensures that it is concave and possesses a global maximum.

$$Max_\alpha Min_\pi \pi^T(\alpha T_f(i) + (1-\alpha)T_b(i)) \ s.t, \ \pi_j \geq 0, \ \sum \pi_j = 1$$

This formulation can be perceived as a zero-sum game between two players, one choosing the value of α and the other a distribution over the row indices. The payoffs are given by $\alpha h(x) + (1-\alpha)h(y)$. As shown later in this paper this objective function can be attacked efficiently by primal dual linear programming methods and constitutes the first algorithm.

Algorithm:1

Solve by Linear Programming the following problem:

$$Max_{V,\alpha} V$$

s.t.

$$V - \alpha(h_i(0) - h_i(1)) \geq h_i(1).$$

Lower bounding by a linear function as above will generally result in extreme underestimations. This prompts us to look for closer approximations, that are computationally as simple. This is provided by applying Jensen's inequality to the entropy function to transform it into a related, weighted L_2 norm minimization of discrete probability distributions.

Lemma 4. *Given the stationary distribution π, the maximum entropy $H(\alpha^*) > H(\alpha'')$, where $\alpha' = \frac{\pi^T[q_i^T.\delta_i]}{\pi^T[\delta_i^T \delta_i]}$ with q_i defined as the i^{th} row of the matrix T_b and δ_i the vector difference of the i^{th} rows of the matrices T_f and T_b.*

Proof.

$$H(\alpha) = \pi^T[h_i(\alpha)] = -\sum_i \pi_i \sum_j T(i,j)\log(T(i,j)) \geq$$

$$-\sum_i \pi_i \log(\sum_j T(i,j)^2) \geq -\log(\sum_i \pi_i \sum_j T(i,j)^2).$$

Letting $\alpha'' = Argmax_\alpha - \log\left(\sum_i \pi_i \sum_j T(i,j)^2\right) = Argmin_\alpha \sum_i \pi_i \sum_j T(i,j)^2$, we obtain

$2\sum_i \pi_i \sum_j \delta_i(j)(\alpha\delta_i(j) + q_i(j)) = 0$ or $\alpha'' = \frac{\pi^T[q_i^T.\delta_i]}{\pi^T[\delta_i^T \delta_i]}$ □

Apart from being a strictly convex optimization problem, the associated weighted L_2 norm minimization problem enjoys the nice property that it has a minima close to the optimal value of α for the entropy maximization problem. This can be shown by first order Taylor approximation of the necessary and sufficient conditions for maximizing $\pi^T[h_i(\alpha)]$. Obtaining bounds from the Taylor residue using the fact that all entries in the transition matrix are bounded away from zero is tedious, but otherwise simple. We show only the first order approximation property.

Proof. $h_i'(\alpha) = \frac{\partial}{\partial\alpha}\sum_j (q_i(j) + \alpha\delta_i(j))\log(q_i(j) + \alpha\delta_i(j)) = \sum_j \delta_i(j) + \sum_j \delta_i(j)\log(q_i(j) + \alpha\delta_i(j))$

$= \sum_j \delta_i(j)\log(q_i(j) + \alpha\delta_i(j))$. From the necessary and sufficient conditions, we have $\pi^T[h'(\alpha)] = \sum_i \pi_i \sum_j \delta_i(j)\log(q_i(j) + \alpha\delta_i(j)) = 0$. Taking the Taylor series expansion around $log(1)$ we obtain $\sum_i \pi_i \sum_j \delta_i(j)(log(1) + 1.(1 - q_i(j) - \alpha\delta_i(j)) = 0$, or, $\sum_i \pi_i \sum_j \delta_i(j)(\alpha\delta_i(j) + q_i(j)) = 0$. Here we have used that $\sum_j \delta_i(j) = 0$. □

Armed with the above lower bound of the cost function we can now pose the following Mini-Max problem, which like before may be posed as a game between players choosing α and a distribution over the rows of the transition matrix. $Min_\alpha Max_\pi \sum_i \pi_i \sum_j T(i,j)^2$.

The nonlinearity in the payoffs prevents a solution using linear programming methods directly as before, and would require discretization of the α to pose this as a finite matrix game, or else solve the semi-infinite matrix game [15]. We take a different approach and use the *Hedge* algorithm of Freund and Schapire [16] to solve the game asymptotically, such that $\overline{\pi}^T h(\alpha_T) \leq Max_\alpha Min_\pi \pi^T h(\alpha) + \mathcal{O}\left(\sqrt{\frac{\log N}{T}}\right)$. For proof please refer to [16].

Algorithm: 2(Hedging)

- Assign initial weights $w(0)_i = 1/N$ to each row (strategy) i.
- Form the initial distribution on strategies $\pi_i(0) = \frac{w_i(0)}{\sum w_i(0)}$
- Repeat T times
  - $\alpha(t) = Argmax_\alpha \pi(t)^T h(\alpha)$.
  - Calculate losses $l_i(t) = -h_i(\alpha(t))$
  - Update weights $w_i(t+1) = w_i(t) * e^{l_i(t)}$
  - Update $\pi_i(t+1) = \frac{w_i(t+1)}{\sum w_i(t+1)}$
- $\overline{\pi_i} = \frac{1}{T}\sum_j \pi_i(j)$
- $\alpha_T = Argmax_\alpha \overline{\pi}^T h(\alpha)$.

4.2 Maximization by Auxiliary Function

The main difficulty in maximizing the cost function above is that we cannot follow a sequential scheme where, for a given distribution π, we update the the parameter α to maximize $\pi(\alpha)^T h(\alpha)$, followed by updating π. The reason for the difficulty is that when α is updated so as to maximize $\pi(\alpha)^T h(\alpha)$, its equivalent $\pi(\alpha)$ may reduce the cost function. To remedy this we formulate a bound on the amount the cost function can decrease as a result of changes in π, followed by bounds on the change in π. This information allows us to construct an auxiliary function that lower bounds the cost function we want to maximize, but equals it at some value of α.

Lemma 5. *Perturbing the mixing weight α by $\Delta\alpha$ induces a perturbation in the stationary distribution π such that $||\Delta\pi|| \leq 2\kappa\Delta\alpha$. where κ is selected such that $\kappa \geq \frac{1}{1-\lambda_2}$.*

Proof. $||\Delta T_{eff}|| = ||\Delta\alpha T_f - \Delta\alpha T_b|| \leq |\Delta\alpha|||T_f|| + |-\Delta\alpha|||T_b|| = 2\Delta\alpha$.

Invoking lemma (2) completes the proof. Note also we can always compute an appropriate κ using the bounds formulated in lemma (3), since the entries are bounded away from zero because of the random jump factor β. □

Lemma 6. *The entropy rate at $\alpha+\Delta\alpha$, $H(\alpha+\Delta\alpha)$ is lower bounded by an auxiliary function $f(\Delta\alpha) = \pi^T h(\alpha + \Delta\alpha) - 2\kappa\Delta\alpha||h^*||_2$ concave in the parameter $\Delta\alpha$.*

Proof. $\pi(\alpha+\Delta\alpha)^T h(\alpha+\Delta\alpha) \geq \pi(\alpha)^T h(\alpha+\Delta\alpha) - Max\{\Delta\pi^T h(\alpha+\Delta\alpha)\}$

$\Delta\pi^T h(\alpha + \Delta\alpha) \leq ||\Delta\pi||_2 ||h(\alpha + \Delta\alpha)||_2 \leq 2\kappa\Delta\alpha||h^*||_2$. where $h^* = Max_\alpha h(\alpha)$.

Combining, we obtain $\pi(\alpha + \Delta\alpha)^T h(\alpha + \Delta\alpha) \geq f(\Delta\alpha)$ where, $f(\Delta\alpha) = \pi(\alpha)^T h(\alpha + \Delta\alpha) - 2\kappa\Delta\alpha||h^*||_2$

and $f(\Delta\alpha) = H(\alpha)$. Concavity of $f()$ follows from concavity of $h()$. □

Algorithm:3

- Assign initial $\alpha_0 = 0.5$
- Repeat untill convergence
  - Find stationary distribution π_t of $\alpha T_f + (1-\alpha)T_b$.
  - $\Delta\alpha = ArgMax_{\Delta\alpha} f_t(\Delta\alpha)$.
  - Update $\alpha_{t+1} = \alpha_t + \Delta\alpha$

The lemma above ensures that updating the parameter $\Delta\alpha$ according to first order necessary conditions for maxima, converges to the global maxima of the required cost function. The computation above involves the term κ, which needs to be estimated. One can re-estimate it at each α by solving a linear programming problem or work with a global value of κ.

4.3 Maximum Stability

Note that the term T_b, which is necessary to incorporate some of the higher order effects also makes the scheme prone to link spam. This detrimental effect is mitigated somewhat by of the normalization used, but minimization of it seems to be a valid quality to demand. In this section we first show that the stability of the chain to perturbations can be bounded in terms of the eigen-gap, as can be the rapid mixing property. This allows any bounds on mixing rate to be translated into equivalent statements about eigen-gap.

Since the second eigenvalue of a mixture of Markov chains is not a very well behaved function, we operate on an upper bound of it characterized by equivalent upper bounds on its mixing rates, which we then try to minimize. For this reason it would be helpful to obtain upper bounds that are convex in order to avoid local minima problems. These are provided by the lemmas (2, 3) established in the previous section.

Proposition 1. *The second distinct eigenvalue of mixtures Markov chain can be upper bounded by a function convex in the mixing parameter α.*

The proof is constructive and explained below, we use the upper-bound the mixing rate of the chain by coupling inequalities, which also suggests a method to maximize the eigengap.

4.3.1 Algorithm

The upper bound to the mixing rate in terms of the factor $\eta = \sum_{v_s} Min_{x\in V}(T(x,i))$ as derived in (4.0.1) suggests an approach where α is chosen so as to maximize the above quantity, i.e., $\alpha^* = ArgMax_\alpha Min_{x\in V}(T(x,i))$ where $T = \alpha T_f + (1-\alpha)T_b$. The concavity of this maximization problem proves the theorem stated before. The problem can be solved numerically by the following linear programming formulation or its corresponding dual.

$$Max \sum W_j$$

s.t.

$$\forall_{\{i\in v_s\}}\ \alpha T_f(i,j) + (1-\alpha)T_b(i,j) \leq W_j$$

For sparse matrix T the number of variables in the primal problem can be reduced by neglecting any column j such that $\exists_{\{x\in v_s\}} T(x,j) = 0$. The number of constraints can be controlled by the choice of the set v_s.

5. EXPERIMENTAL RESULTS

We carried out experiments on a subset of the cora dataset consisting of the citation graph of 11754 research papers in the "Artificial Intelligence" category. Cora was a popular search engine for computer science research papers, and has previously been used by Ng et al. in [17] to examine the properties of HITS and Pagerank. The objective of the experiments was to investigate the quality of the rankings produced by the maxent and max-stability solutions. Since there is no unique "correct" ranking of the papers, we identified papers which were ranked in the top 200 consistently, irrespective of the value of α' as defined in equation (9), and examined their ranks under the maxent scheme. We also explore the stability of the rankings as follows: Many of the papers in the dataset had self-referring links. These were used to simulate a deliberate act of increasing their rank. We created two sets of citation graphs: (I) With self references removed (33274 links) and (II) with self citations (34458 links) and observed the differences in the rankings as a result of the perturbation imposed.

The dependence of the entropy rate and convergence time as a function of α' is shown in figure 1. For dataset I the entropy rate is maximized at $\alpha' = 0.43$ and for dataset II at 0.78.

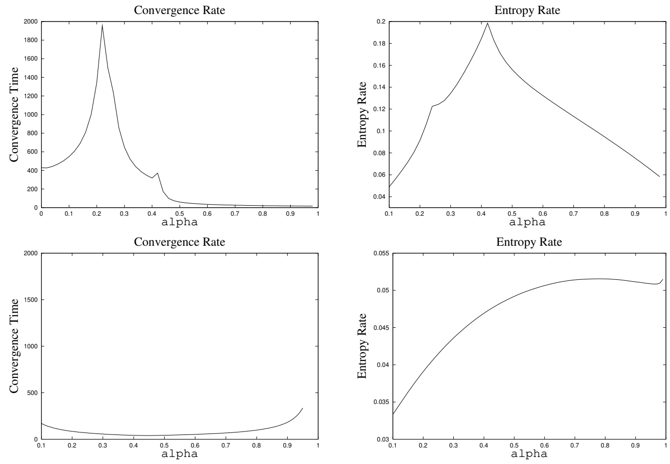

Table 1: Entropy and Convergence rates for data set I(top) and II(bottom).

In order to investigate the nature of the maximum entropy solution we carried out the rankings for α' ranging from 0.1 to 1 at increments of 0.1 and kept top 200 papers for each. From this set of rankings, we identified the papers which occurred frequently in the ranked lists. We compare the ranks assigned to these papers by the traditional pagerank algorithm ($\alpha' = 1.00$ and where α' was chosen to optimize the entropy in table 2, where we indicate, in left to right order, the number of lists a paper occurred in (out of 9), the ranks for dataset I (both maxent and normal pagerank), the ranks for dataset II (both maxent and normal pagerank). The important thing to note is that $\alpha' = 1.00$ misses many of the papers (denoted as "X" in the table) which are ranked high consistently at different values of α' as compared to maxent value of α'. It is also an indication of the fact that forward links, though not as valuable as inlinks carry some value in the setting of citation graphs. This can be understood from the fact that recent papers are not cited as frequently as older papers, but they might still be identified from their quality of outlinks. Such temporal difference would also exist for the preferential attachment based models for webgraphs [18], [19].

Next we examine the stability criteria for the choice of α'. Although $\alpha' = 1$ is sub-par from quality of citation point of view, it is a point of high stability as indicated by its fast mixing time and therefore is potentially stable and spam-resistant. We found that to be indeed the case. The stability characteristic is examined in table 3. The columns shows the ranks for the datasets I and II. We examine the top 10 papers ranked by each of these cases and observe the difference in their ranks after perturbation. Note that the differences are small at values of $\alpha' = 1.00$ (a fact also observed in the experiments in [17]), but wildly different at $\alpha' = 0.2$ where the mixing time is high.

Citations	**n**	**I 0.43**	**I 1.00**	**II 0.78**	**II 1.00**
Bernardo, Berger, Dawid, et al. Eds. Bayesian Statistics 4	9	37	104	32	123
Goldberg. Genetic Algorithms in Search Optimization and Machine Learning.	9	9	3	3	7
13195 Kaelbling, Littman, Cassandra	9	38	160	73	174
Langdon. Data Structures and Genetic Programming	8	21	X	18	X
Latombe, Robot Motion Planning	8	102	X	93	X
Langdon,Poli, Fitness causes bloat: Mutation	8	18	X	45	X
Schapire, Freund, Bartlett, Lee, Boosting the Margin: A New Explanation for the Effectiveness of Voting Methods	8	33	X	65	X
Garey, Johnson, Computers and Intractibility: A Guide to the Theory of NP-Completeness	8	17	4	4	8
Jarmo. An indexed bibliography of genetic algorithms:	8	14	X	9	X
Parekh, Yang,Honavar. Constructive neural network learning algorithms for multi-category classification.	8	8	X	49	X
1103693	8	114	X	25	X
Sutton, Precup, Singh, S. Between MDPs and semi-MDPs: learning, planning, and representing knowledge at multiple temporalscales	7	55	X	79	X
Dietterich. Machine learning research: Four current directions.	7	62	X	138	X
Stone, Veloso. The cmunited\'97 simulator team	7	48	X	156	X
Sutton. Learning to predict by the methods of temporal differences	7	30	6	6	102
Han, Towards on-line analytical mining in large databases	7	20	X	103	X
Jordan, Ghahramani, Jaakkola, Saul.An introduction to variational methods for graphical models	7	40	X	105	X
Valente, Gil, Swartout. INSPECT: an intelligent system for air campaign plan evaluation based on EXPECT	7	1	1	10	5
Valente, Swartout, Gil, A Representation and Library for Objectives in Air Campaign Plans	7	2	2	11	6
Cassandra. Exact and Approximate Algorithms for Partially Observable Markov Decision Problems	7	39	X	180	X
Srikant, Agrawal, Mining generalized association rules.	7	70	24	24	26
Roth, Applying Winnow to Context Sensitive Spelling Correction	7	24	X	66	X
Krogh, Hertz, Palmer, Introduction to the Theory of Neural Computation	7	46	7	8	11
Freund, Schapire, Experiments with a New Boosting Algorithm	7	92	43	41	50
Freund, Schapire, A Decision-theoretic Generalization of On-line Learning and an Application to Boosting	7	71	30	36	38
Guler, Generalized linear complementarity problems and interior point algorithms for their solutions	7	60	X	104	X
Simmons, A robust layered control system for a mobile robot	7	43	5	5	9
Oates,Prasad, Lesser, Cooperative Information Gathering: A Distributed Problem Solving Approach	7	99	90	53	94
Yang, Honavar, Feature subset selection using a genetic algorithm	7	19	X	81	X
Howard, Dynamic Programming and Markov Processes	7	61	25	31	29
Barto, Bradtke, Singh, Learning to Act using Real- Time Dynamic Programming	7	32	11	157	29

Table 2: Ranks occupied by the papers in dataset I and II calculated for α' at maxent and $\alpha' = 1.00$(pagerank). "X" signifies that the paper was not present in the ranked list

Paper id	I 0.2	II 0.2	Paper id	I 1.00	II 1.00
175355	1	61	571354	1	5
86515	2	51	538295	2	6
108351	3	X	35	3	7
108350	4	X	2	4	8
156935	5	53	1388	5	9
156936	6	40	6213	6	10
156934	7	X	3229	7	11
243802	8	186	4584	8	12
243808	9	X	2027	9	13
22876	10	X	1365	10	14
237404	X	1	21366	X	1
713268	X	2	628859	X	2
178718	59	3	237404	X	3
3108	X	4	713268	X	4
118873	18	5	571354	1	5
980151	X	6	538295	2	6
762584	X	7	35	3	7
705482	X	8	2	4	8
69296	15	9	1388	5	9
125090	X	10	6213	6	10

Table 3: Variation in the Ranks of the top 10 papers due to perturbations, evaluated at $\alpha' = 1.00$(pagerank) and 0.02.

6. APPLICATIONS IN SIMILARITY ANALYSIS

In this section we explore an application in word co-occurrence graphs, as a demonstration of the range of applicability of the higher order link analysis model.

The idea of incorporating higher order effects turns out to be helpful for vector space modeling of documents, where a document $\mathbf{d_i}$ is represented as a high dimensional vector $\mathbf{d_i} = \{d_{i1}, d_{i2} \cdots d_{iN}\}$ of word occurrences. For simple models, each component is binary and denotes the presence or absence of a distinct word in the dictionary $\mathcal{D}$. But better results are obtained by *tfidf* weightings or using lower dimensional projections, where each component denotes the presence or absence of synonyms or related words. In this case, identifying the groups of related words by feature extraction such as LSI [20] become important.

In the vector space model, most of the processing revolves around computing the similarity $sim(\mathbf{q}, \mathbf{d_i})$ between a query $\mathbf{q}$ (typically a binary vector, as in the simple vector space model) and a document $\mathbf{d_i}$, as a dot product $\mathbf{q}^T\mathbf{d_i}$. Note that for the simple binary vector space model, $\mathbf{q}^T\mathbf{d_i}$ is just the number of common words between the query and the document. Such a definition of similarity will have difficulty in identifying documents describing identical concepts but with different words.

Popular techniques applied to alleviate this problem include LSI [20] and query expansion [21, 22]. In LSI both the query $\mathbf{q}$ and the document $\mathbf{d_i}$ are projected on to the subspace of the primary principal vectors U of the document-word matrix W^T. Thus $sim_{LSI}(\mathbf{q}, \mathbf{d_i}) = \mathbf{q}^T U^T U \mathbf{d_i}$. Note that a byproduct of LSI is equalizing the variance along the basis dimensions of the subspace, which takes care of the fact that the principal theme of the corpus does not swamp the sub-principal themes.

Query expansion works by modifying the query $\mathbf{q}$ to include the synonyms and related words. Given the word-document matrix W, one way of constructing the expansion is to find the set of documents containing the words in $\mathbf{q}$. This is given by word vector $\mathbf{w} = W^T\mathbf{q}$. All non-zero components of $\mathbf{w}$ indicate that the corresponding document contained at least one word from the query $\mathbf{q}$, while the value indicates the total number of words common to the document and the query. To expand the query we merge all the words occurring in the documents appearing with non-zero weights, $\mathbf{q}' = W\mathbf{w} = WW^T\mathbf{q}$, so the modified similarity $sim_{expanded} = \mathbf{d_i}^T(WW^T)\mathbf{q}$. The (WW^T) matrix also denotes the similarity between a pair of words as computed from their co-occurrence [23].

Based on arguments similar to higher order surfing models, we can consider higher order co-occurrences. N^{th} order co-occurrence based query expansion would yields the similarity. As $sim_{expanded^N} = \mathbf{d_i}^T(WW^T)^N\mathbf{q}$ and as $N \to \infty$ this converges to LSI using only the first principal vector u_1 i.e $sim_{expanded^\infty} = sim_{LSI} = \mathbf{d_i}^T u_i u_i^T \mathbf{q}$.

However, simply taking the higher order query expansion suffers from what is known as query drift. So similar to the link analysis theme, we can combine the higher order componenets with decreasing weights,

$$K = I + \alpha WW^T + \alpha^2(WW^T)^2 \cdots = (I - \alpha WW^T)^{-1}$$

following which the similarity may be defined as $sim' = \mathbf{d_i}^T K\mathbf{q}$.

7. CONCLUDING REMARKS

In this paper we generalize the first and second order link-analysis models represented by Pagerank and HITS respectively, to a family of models that can control the strength of their locality behaviour through a single parameter, α. As a method for choosing the parameter value we investigated two approcaches (i) Maxent and (ii) Max-stability. Both Maxent and Max-stability approaches form an alternative when expert knowledge or training examples is not provided. Our framework is applicable to a variety of undirected and directed graphs - the empirical study on a citation graph is just one illustration.

As an extension of the Maxent formulation presented here we would like to examine a broader range of entropy constrained solutions. For example, Levene *et al.* [24] have discussed an algorithm for calculating the entropy of surfing from actual data. So, instead of maximizing the entropy, we can also select the value of α to minimize the difference between the actual entropy and the entropy of the effective transition matrix. This may be computed by a procedure of alternate projection between the space of transition matrices parameterized by α and the space of transition matrices of specified entropy rates, defined on the same state space. We intend to follow up on this approach in the future.

8. ACKNOWLEDGEMENT

This research was supported by Intel Corporation under grant #8032.

9. REFERENCES

[1] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," 1998.

[2] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM*, vol. 46, no. 5, pp. 604–632, 1999.

[3] W. W. Cohen, R. E. Schapire, and Y. Singer, "Learning to order things," in *Advances in Neural Information Processing Systems* (M. I. Jordan, M. J. Kearns, and S. A. Solla, eds.), vol. 10, The MIT Press, 1998.

[4] C. Ding, X. He, P. Husbands, H. Zha, and H. Simon, "Pagerank, hits and a unified framework for link analysis," Tech. Rep. 49372, LBNL, 2002.

[5] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, "Trawling the Web for emerging cyber-communities," *Computer Networks (Amsterdam, Netherlands: 1999)*, vol. 31, no. 11–16, pp. 1481–1493, 1999.

[6] K. Bharat and M. R. Henzinger, "Improved algorithms for topic distillation in a hyperlinked environment," in *Proceedings of SIGIR-98, 21st ACM*

International Conference on Research and Development in Information Retrieval, (Melbourne, AU), pp. 104–111, 1998.

[7] R. Lempel and S. Moran, "The stochastic approach for link-structure analysis (SALSA) and the TKC effect," *Computer Networks (Amsterdam, Netherlands: 1999)*, vol. 33, no. 1–6, pp. 387–401, 2000.

[8] J. C.Miller, G. Rae, and F. Schaefer, "Modification of kleinberg's hits algorithm using matrix exponentiation and web log records," in *SIGIR'01*, (New Orleans), pp. 444 – 445, sept 2001.

[9] G. H. Golub and C. F. V. Loan, *Matrix Computations*. Baltimore, MD, USA and Oxford, England: The Johns Hopkins University Press and North Oxford Academic, 1993.

[10] E. T. Jaynes, "On the rationale of maximum-entropy methods," *Proc. IEEE*, vol. 70, pp. 939–952, 1982.

[11] T. Cover and J. Thomas, *Elements of Information Theory*. Wiley, New York, 1991.

[12] J. G. Kemeny and J. L. Snell, *Finite Markov Chains*. New York: Van Nostrand Reinhold, 1960.

[13] J. S. Rosenthal, "Minorization conditions and convergence rates for Markov chain Monte Carlo," *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 558–566, 1995.

[14] J. N. Kapur and H. K. Kesavan, *Entropy optimization principles with applications*. Academic Press, 1992.

[15] T. Basar and G. Olsder, *Dynamic Noncooperative Game Theory*. Academic Press, 1995.

[16] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," in *European Conference on Computational Learning Theory*, pp. 23–37, 1995.

[17] A. Y. Ng, A. X. Zheng, and M. I. Jordan, "Link analysis, eigenvectors and stability," in *IJCAI-01*, pp. 903–910, 2001.

[18] R. A. Barabasi and H. Jeong, "Mean-field theory for scale-free random graphs," *Physica*, vol. 272, no. A, pp. 173–187, 1999.

[19] G. Pandurangan, P. Raghavan, and E. Upfal, "Using PageRank to Characterize Web Structure," in *8th Annual International Computing and Combinatorics Conference (COCOON)*, 2002.

[20] G. Salton, *Automatic information organization and retrieval*. New York: McGraw Hill, 1968.

[21] F. Jiang and M. L. Littman, "Approximate dimension equalization in vector-based information retrieval," in *Proc. 17th International Conf. on Machine Learning*, pp. 423–430, Morgan Kaufmann, San Francisco, CA, 2000.

[22] S. K. M. Wong, W. Ziarko, and P. C. N. Wong, "Generalized vector space model in information retrieval," in *Proceedings of the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 18–25, 1985.

[23] N. Cristianini, J. Shawe-Taylor, and H. Lodhi, "Latent semantic kernels," in *Proceedings of ICML-01, 18th International Conference on Machine Learning* (C. Brodley and A. Danyluk, eds.), (Williams College, US), pp. 66–73, Morgan Kaufmann Publishers, San Francisco, US, 2001.

[24] M. Levene and G. Loizou, "Computing the entropy of user navigation in the web," Tech. Rep. Research Note RN/99/42, Department of Computer Science, University College London, 1999.