Optimal ranking in networks with community structure

Huafeng Xie[1,3], Koon-Kiu Yan[2,3], Sergei Maslov[3] *
[1] *New Media Lab, The Graduate Center, CUNY New York, NY 10016, USA*
[2] *Department of Physics and Astronomy, Stony Brook University, Stony Brook, New York, 11794, USA*
[3] *Department of Physics, Brookhaven National Laboratory, Upton, New York 11973, USA*
(Dated: May 30, 2006)

The World-Wide Web (WWW) is characterized by a strong community structure in which groups of webpages (e.g. those devoted to a common topic or belonging to the same organization) are densely interconnected by hyperlinks. We study how such network architecture affects the average Google rank of individual communities. Using a mean-field approximation, we quantify how the average Google rank of community webpages depends on the degree to which it is isolated from the rest of the world in both incoming and outgoing directions, and α – the only intrinsic parameter of Google's PageRank algorithm. Based on this expression we introduce a concept of a web-community being decoupled or conversely coupled to the rest of the network. We proceed with empirical study of several internal web-communities within two US universities. The predictions of our mean-field treatment were qualitatively verified in those real-life networks. Furthermore, the value $\alpha = 0.15$ used by Google seems to be optimized for the degree of isolation of communities as they exist in the actual WWW.

PACS numbers: 89.20.Hh, 05.40.Fb, 89.75.Fb

The World Wide Web (WWW) – a very large ($\sim 10^{10}$ nodes) network consisting of webpages connected by hyperlinks – presents a challenge for the efficient information retrieval and ranking. Apart from the contents of webpages, the network topology around them could be a rich source of information about their relative importance and relevance to the search query. It is the effective utilization of this topological information [1] that advanced the Google search engine to its present position of the most popular tool on the WWW and a profitable company with a current market capitalization around $80 billion. As webpages can be grouped based on their textual contents, language in which they are written, the organizations to which they belong etc, it should come as no surprise that the WWW has a strong community structure [2] in which similar pages are more likely to contain hyperlinks to each other than to the outside world. Formally a web community can be defined as a collection of webpages characterized by an above-average density of links connecting them to each other.

In this study, we are going to address the following question: how does the relative isolation of community's webpages from the rest of the network affects their Google rank? In addition we would speculate the parameters of Google's PageRank algorithm were selected for its optimal performance given the extent of the community structure in the present WWW network.

In the heart of the Google search engine lies the PageRank algorithm determining the global "importance" of every web page based on the hyperlink structure of the WWW network around it. When one enters a search keyword such as e.g. "statistical physics" on the Google website the search engine first localizes the subset of webpages containing this keyword and then simply presents them in the descending order based on their PageRank values. While the details of the PageRank algorithm have undoubtedly changed since its introduction in 1997, the central "random surfer" idea first described in [1] remained essentially the same. From a statistical physics standpoint the PageRank simulates an auxiliary diffusion process taking place on the network in question. A large number of random walkers are initially randomly distributed on the network and are allowed to move along its directed links. Similar diffusion algorithms have been recently applied to study citation and metabolic networks [3] and the modularity of the Internet on the hardware level represented by an undirected network of interconnections between Autonomous Systems [4]. As in real web surfing, a random walker of the PageRank algorithm could "get bored" from following a long chain of hyperlinks. To model this scenario, the authors introduced a finite probability α for a random walker to directly jump to a randomly selected node in the network not following any hyperlinks. This leaves the probability $1-\alpha$ for it to randomly select and follow one of the hyperlinks of the current webpage. According to [5], in the real PageRank algorithm α was chosen to be 0.15. The algorithm then simulates this diffusion process until it converges to a stationary distribution. The Google rank (PageRank) $G(i)$ of a node i is proportional to the number of random walkers at this node in such a steady state, and is usu-

*To whom the correspondence should be addressed: maslov@bnl.gov

ally normalized by $\langle G(i)\rangle = 1$. In this normalization, the flux of walkers entering a given site due to random jump from all the other nodes is given by $\sum_{i=1}^{N} \alpha G_i/N = \alpha$. The continuity equation for this diffusion process reads $G(i) = \alpha + \sum_{j\to i}(1-\alpha)\frac{G(j)}{K_{out}(j)}$. Here $K_{out}(j)$ denotes the number of hyperlinks (the out-degree) of the node j and the summation goes over all nodes j that have a hyperlink pointing to the node i. In the matrix formalism the PageRank values are given by the components of the principal eigenvector of an asymmetric positive matrix related to the adjacency matrix of the network. Such eigenvector could be easily found using a simple iterative algorithm. To do this, all nodes must satisfy $K_{out}(i) > 0$. Practically, it is done by iteratively removing pages with zero out-degrees from the network [5]. Consider a network in which N_c nodes form a community characterized by an above-average density of edges linking these nodes to each other. Let E_{cw} to denote the total number of hyperlinks pointing from nodes in the *community* to the outside *world*, while E_{wc} - the total number of hyperlinks pointing in the opposite direction. As the Google rank is computed in the steady state of the diffusion process , the total current of surfers J_{cw} leaving the community must be precisely balanced by the opposite current J_{wc} of surfers entering the community. Note that both J_{cw} and J_{wc} consist of two contributions: the current via the direct hyperlinks between the community and the outside world, and the current due to random jumps.

In this paper we solve the problem of interplay between the community structure and the average Google rank inside the community using a mean-field approximation. This approximation holds provided that websites connecting the community to the outside world for outgoing and incoming traffic are an unbiased sample of all websites in the community and the outside world correspondingly. That is to say, we assume that their average in-degree and Google rank are approximately equal to those of other websites in their compartment. Needless to say this approximation might prove to be wrong in real life WWW networks. As we will see later this would lead to an effective renormalization of parameters in our equations, while for the most part preserving their functional form.

Let $G_c = \langle G(i)\rangle_{i\in C}$ to denote the average Google rank of webpages inside the community. Within our mean-field approximation the average Google rank of community nodes sending links to the outside world is equal to its overall average value inside the community G_c, so the average current flowing along a hyperlink pointing away from the community is given by $(1-\alpha)G_c/\langle K_{out}\rangle_c$ and the total current leaving the community along all those out-going links is $(1-\alpha)E_{cw}G_c/\langle K_{out}\rangle_c$. The total number of random walkers residing on nodes inside the community is G_cN_c and the probability of a random jump to lead to a node outside the community is $N_w/(N_c+N_w)$, which is close to 1 as $N_c \ll N_w$. The contribution to the outgoing current due to such jumps is given by αG_cN_c, and thus the total outgoing current is $J_{cw} = (1-\alpha)G_cE_{cw}/\langle K_{out}\rangle_c + \alpha G_cN_c$. Similarly the incoming current J_{wc} is given by $(1-\alpha)G_wE_{wc}/\langle K_{out}\rangle_w + \alpha G_wN_c$. Equating these two currents one gets $\dfrac{G_c}{G_w} = \dfrac{(1-\alpha)E_{wc}/(\langle K_{out}\rangle_w N_c)+\alpha}{(1-\alpha)E_{cw}/(\langle K_{out}\rangle_c N_c)+\alpha}$. One may notice that $\langle K_{out}\rangle_w N_c$ and $\langle K_{out}\rangle_c N_c$ are respectively equal to $E_{wc}^{(r)}$ and $E_{cw}^{(r)}$ – expected numbers of links connecting the community to the outside world in a random network with the same degree sequence as the network in question [6]. By approximating $G_w \approx 1$, we finally arrive at the following equation:

$$G_c = \frac{(1-\alpha)\frac{E_{wc}}{E_{wc}^{(r)}}+\alpha}{(1-\alpha)\frac{E_{cw}}{E_{cw}^{(r)}}+\alpha}. \tag{1}$$

For simplicity of notation, let us refer to the ratios $E_{wc}/E_{wc}^{(r)}$ and $E_{cw}/E_{cw}^{(r)}$ as R_{wc} and R_{cw} respectively. Roughly speaking, R_{cw} and R_{wc} quantify how isolated is a given community in both directions connecting it to the outside world. In fact, in most communities both ratios R_{wc} and R_{cw} are below 1 because E_{wc} and E_{cw} are typically less than their expected values in a randomized network [7]. One implication of the Eq.1 is that the average Google ranking of a community depends on the pattern of their connections with the outside world through the ratios R_{cw} and R_{wc}. For example if R_{wc} is close to 1 (i.e. the number of links pointing to the community is roughly the same as in a random network with the same degree distribution), G_c gets its maximum value $1/\alpha$ when $R_{cw} \ll \alpha$, which could be interpreted as the community very isolated in the out-direction. On the contrary, if the number of out-going links from the community to the outside world is roughly the same as in a corresponding randomized network, G_c attains its minimum value of α if the community is very isolated in the in-direction ($R_{wc} \ll \alpha$). ¿From Eq.1 one could easily see that the relative values of isolation ratios R_{cw}, R_{wc} and the parameter α determines the sensitivity of G_c to community's connections with the outside world. If either R_{cw} or R_{wc} is comparable to α, G_c is sensitive to the exact number of links connecting the community to the outside world in this particular direction. Conversely, if both $R_{wc}, R_{cw} \ll \alpha$ the average Google rank of community is no longer sensitive to its outside connections, and its value is close to 1 which is the overall average value of G_i for all nodes. In this case, we would refer to this community as being "decoupled" from the outside world. Of course, whether a community is decoupled or coupled depends on the value of α. A community decoupled at a particular α could become coupled if a smaller α is chosen.

To empirically investigate the interplay between G_c and α in real World-Wide Web, we downloaded [8] complete sets of hyperlinks contained in all webpages within two US universities. We then studied intra-university communities based either on common interests (like schools or departments) or common geographic locations (like individual campuses of a large university

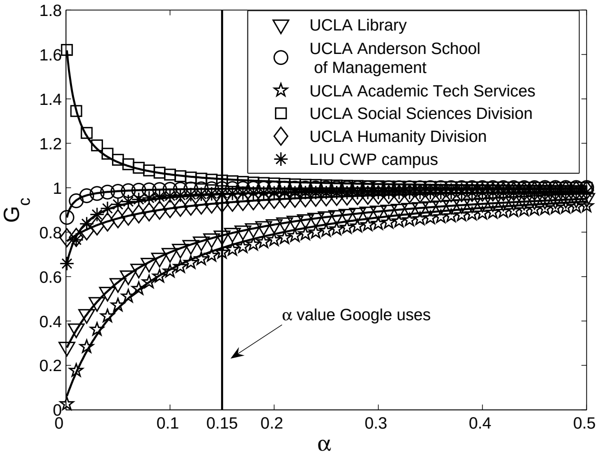

FIG. 1: The average Google rank G_c of different communities as a function of the parameter α. The communities are within real WWW networks of two US universities (see Table I for details). The data points are obtained by running the PageRank algorithm for different values of α. Solid lines are two-parameter best fits to the data with the Eq.1.

system). (See Table I for details.) The relation between G_c and α for six such communities are shown in Fig.1. As expected from our calculations, as α is lowered in all these communities G_c starts to significantly deviate from 1. Moreover, the community "UCLA social science" deviates upward while all the others deviate downward. This could be qualitatively explained by the Eq.1, with the observation that R_{wc} is greater than R_{cw} in this community, while R_{wc} is less than R_{cw} in all the others (see Table II). Furthermore, by looking at which values of α does G_c starts to significantly deviate from 1, one can see that different communities become coupled to the outside world for different α's. For example, "UCLA Library" and "UCLA Academic Tech. Service" reach the level of $G_c = 0.8$ when α is around $0.2 - 0.3$, while "UCLA Anderson School of Management" and "LIU CWP campus" reach the same level of coupling only for much lower $\alpha \approx 0.01 - 0.05$.

We would like to point out that the "mean-field" assumption we used in deriving the Eq. 1 can never be perfectly true for real web-communities. For example, a community may be linked from the outside world by a highly ranked authority page, and receive an in-coming current larger than predicted by our mean-field calculation. Conversely, it can only get links from relatively unimportant pages which would result in our mean-field model overestimating the actual current. There is no universal rule for estimating even the sign of the deviation from the mean field predictions. Thus it is impossible to calculate "corrections" to our mean-field formula. Instead those corrections have to be considered on a case-by-case basis. By allowing parameters R_{cw} and R_{wc} in the Eq.1 to deviate from their values prescribed by the mean-field theory provides a simple mathematical

TABLE I: The basic statistics about the academic WWW networks downloaded from Ref. [8]. We choose to study hyperlink networks within the Long Island University (LIU, 29476 nodes and 160457 edges) and separately within the University of California at Los Angeles (UCLA, 135533 nodes and 636595 edges). Following Google's original recipe [1] we iteratively removed webpages with zero out-degree. The resulting networks consist of 15471 nodes and 90111 edges for the LIU and 31621 nodes and 353370 edges for the UCLA. We then studied several large communities defined by the URL of their servers (e.g. .library.ucla.edu for the "UCLA Library" community.)

Community	N_c	E_{cc}	$E_{cc}^{(r)}$	E_{wc}	E_{cw}
UCLA Library	2028	23062	1699	755	2141
UCLA School of Management	1340	15983	739	175	169
UCLA Academic Tech. Services	1907	26597	2248	139	3113
UCLA Social Science Division	626	3986	50	258	142
UCLA Humanity Division	864	4846	79	397	445
LIU CWP Campus	2756	18376	4105	336	1393

TABLE II: R_{cw}, R_{wc}, R^*_{cw} and R^*_{wc} for different communities. R_{cw} and R_{wc} are obtained by counting the links from the community to the world and vice versa, divided by the corresponding number of links in a random network with the same degree distribution [6]. R^*_{cw} and R^*_{wc} are result of fitting the G_c and α dependency via Eq.1.

Community	R_{wc}	R_{cw}	R^*_{wc}	R^*_{cw}
UCLA Library	0.04	0.09	0.02	0.07
UCLA School of Management	0.01	0.01	0.005	0.006
UCLA Academic Tech. Services	0.007	0.1	0.003	0.07
UCLA Social Science Division	0.04	0.03	0.02	0.01
UCLA Humanity Division	0.04	0.08	0.05	0.07
LIU CWP Campus	0.03	0.09	0.01	0.02

formalism to quantify those corrections for real communities. We define R^*_{cw} and R^*_{wc} from the two-parameter best fit of the actual $G_c(\alpha)$ dependence in a given community with the Eq.1 (see Table II.) One may regard R^*_{cw} and R^*_{wc} as effective parameters, which in addition to simple geometrical properties of the community such as numbers of links connecting it to the outside world, take into account Google ranks of actual pages sending those links. These "renormalized" ratios R^*_{cw} and R^*_{wc} would be more accurate than their "raw" counterparts (R_{cw} and R_{wc}) in determining whether a particular web-community is coupled to or decoupled from the outside world at a given value of α.

The effective ratios R^*_{cw} and R^*_{wc} for the six communities used in our study are listed in the Table II and visualized in Fig.2. Generally speaking, the closer to the origin is a community in this figure, the lower is the value of α at which it first becomes coupled to the outside world. One could see that for $\alpha = 0.15$, which is the actual value used by the Google [5], all of our six communities are essentially decoupled from the outside world. However, if a much smaller value of α (say 0.01) is chosen, 5 out of 6 of our communities (all except for the "UCLA Anderson School of Management") would become sensitive to their connections with the outside world. In principle, Fig.2 might be extended to include the region where R^*_{cw} and R^*_{wc} are above one, but by definition those points are not referring to well-defined communities. From Eq.1 it follows that it is the asymmetry between R_{cw} and R_{wc} which determines whether G_c is greater than or less than 1. Thus the diagonal in Fig. 2 separates communities with $G_c > 1$ from those with $G_c < 1$. The ratio between the x- and y-coordinates of the community in this plot determines the asymptotic value of its Google rank G_c for α close to zero. Thus the two communities: "UCLA Academic Tech. Service" and "UCLA Social Science", whose ratios between their $x-$ and $y-$ coordinates in this plot are respectively the smallest and the largest in our set deviate the most from $G_c = 1$ as shown in Fig.1.

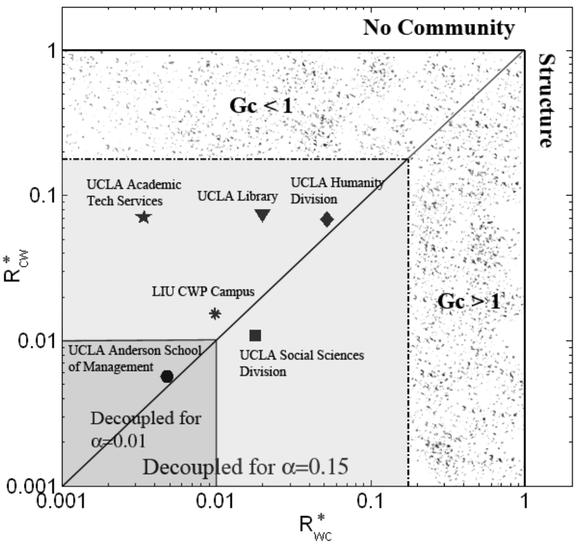

FIG. 2: R^*_{cw} and R^*_{wc} for different communities. Communities inside the lightly shaded square are decoupled from the rest of the world for $\alpha = 0.15$, while the ones inside the dark shaded square are decoupled for $\alpha = 0.01$.

The dominance of Google and the all-important role of its ranking led to the appearance of services offering "search engine optimization" to their clients. They promise to modify the content and the hyperlink structure of client's webpages to improve their Google rank. Our findings suggest one obvious way how such an "optimization" could be achieved: the number of links pointing to the outside world should be reduced to the minimum while the number of intra-community hyperlinks is kept at the maximum. However, as we demonstrated above the success of such a strategy depends on whether or not the community in question is coupled to the outside world. Indeed, the average Google rank of a decoupled community is virtually insensitive to the exact balance of hyperlinks connecting it to the outside world .

Since coupling of web-communities to the outside world and the resulting ability of their webmasters to artificially boost the ranking is undesirable for a search engine, it should come as no surprise that the internal parameter α chosen by the Google's team is carefully selected to minimize this effect. To make most of the communities decoupled the value of α in the PageRank algorithm should be as large as possible. On the other hand, for very large α the algorithm does not take into account also the relevant network properties of the WWW. Indeed for α close to 1, random surfers rarely follow hyperlinks and thus nearly all topological information about the network is lost. Therefore, the optimal value of α should be chosen based on the realistic values of isolation parameters R_{cw} and R_{wc}. In our study we found all the communities to be effectively decoupled at $\alpha = 0.15$ but not at smaller values of α (e.g $\alpha = 0.01$ shown as a dark shaded square in Fig.2). Thus, for our sample of web-communities, $\alpha = 0.15$ proposed in [1] indeed strikes the best possible balance between the opposing demands on the value of α.

Work at Brookhaven National Laboratory was carried out under Contract No. DE-AC02-98CH10886, Division of Material Science, U.S. Department of Energy.

---

[1] S. Brin and L. Page, Computer Networks and ISDN Systems, **30**, 107 (1998).

[2] R. Kumar, P. Raghavan, S. Rajagopalan, and A.Tomkins, Computer Networks **31**, 11 (1999).

[3] S. Bilke and C. Peterson Physical Review E **64**, 036106 (2001).

[4] K. A. Eriksen, I. Simonsen, S. Maslov and K. Sneppen, Phys. Rev. Lett. **90**, 148701 (2003).

[5] L. Page, S. Brin, R. Motwani and T. Winograd, Stanford Digital Library Technologies Project (1998).

[6] Indeed, in a random network out of $\langle K_{out}\rangle_w N_w$ hyperlinks starting at nodes outside the community $\langle K_{out}\rangle_w N_w N_c/(N_w + N_c) \simeq \langle K_{out}\rangle_w N_c$ would end up pointing to community nodes. Similarly, out of $\langle K_{out}\rangle_c N_c$ hyperlinks starting at community nodes $\langle K_{out}\rangle_c N_c N_w/(N_w + N_c) \simeq \langle K_{out}\rangle_c N_c$ would point to nodes in the outside world.

[7] Ususally communities have higher than expected number of intra-community links: $E_{cc} > E^{(r)}_{cc}$. Since $E^{(r)}_{cc} + E^{(r)}_{wc} = E_{cc} + E_{wc} = N_c\langle K_{in}\rangle_c$ and $E^{(r)}_{cc} + E^{(r)}_{cw} = E_{cc} + E_{cw} = N_c\langle K_{out}\rangle_c$, this automatically implies that s $E_{wc} < E^{(r)}_{wc}$ and $E_{cw} < E^{(r)}_{cw}$.

[8] Thelwall, M. Cybermetrics, Vol **6/7**, Issue 1. Paper 2 (2002-3).