STATISTICAL PHYSICS OF COMPLEX NETWORKS

By

HUAFENG XIE

A dissertation submitted to the Graduate Faculty in Physics in partial fulfillment of the requirements for the degree of Doctor of Philosophy,
The City University of New York

2008

This manuscript has been read and accepted for the
Graduate Faculty in Physics in satisfaction of the
dissertation requirement for the degree of Doctor of Philosophy.

Brian Schwartz

Date — Chair of Examining Committee

Sultan Catto

Date — Executive Officer

Sergei Maslov

Joseph Krieger

Steven A Schwarz
Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

Abstract

STATISTICAL PHYSICS OF COMPLEX NETWORKS

By

HUAFENG XIE

Adviser: Professor Brian Schwartz

We live in a connected world. It is of great practical importance and intellectual appeal to understand the networks surrounding us. In this work we study ranking of the nodes in complex networks. In large networks such as World Wide Web (WWW) and citation networks of scientific literature, searching by keywords is a common practice to retrieve useful information. On the WWW, apart from the contents of webpages, the topology of the network itself can be a rich source of information about their relative importance and relevancy to the search query. It is the effective utilization of this topological information [50] which advanced the Google search engine to its present position of the most popular tool on the WWW. The World-Wide Web (WWW) is characterized by a strong community structure in which communities of webpages are densely interconnected by hyperlinks. We study how such network architecture affects the average Google ranking of individual webpages in the community. Using a mean-field approximation, we quantify how the average Google rank of community's 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. We proceed with numerical study of simulated networks and 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.

We then extend Google's PageRank algorithm to citation networks of scientific literature. Unlike hyperlinks, citations cannot be updated after the point of publication. This results in strong aging characteristics of citation networks that affect the performance of the PageRank algorithm. To rectify this we modify the PageRank algorithm to a new ranking method, CiteRank, in which the starting point of random surfers is exponentially biased towards more recent publications. The ranking results are compared for two rather different citation networks: all American Physical Society publications between 1893 and 2003 and the set of high energy physics theory (hep-th) preprints. Despite major differences between these two networks, we find that their optimal parameters of the CiteRank algorithm are remarkably similar.

To my parents. . .

who have always loved and supported me

Acknowledgments

First and foremost I would like to express my deepest appreciation to my two advisors, Professor Brian Schwartz and Dr. Sergei Maslov. Professor Schartz has always believed in me in my pursuing of my goals. This dissertation would have been impossible without his support and encouragement. His dedication to work and care of others are the examples that I was and will be learning from and I'll never forget the valuable lessons he taught me. All the research work in this dissertation has been done under direction of Dr. Sergei Maslov. Dr. Maslov has profound knowledge of many aspects of physics and complex networks. I have always admired his passion for science, his wisdom and ingenuity, which have made working with him a joy. I feel honored to have Dr. Maslov as my advisor and have worked with him on my dissertation research. I also want to thank my group mates/collaborators and friends: Koon-kiu Yan, Dylan Walker and Dmitri Volja, for their help, inspirations and friendship. It has been one of the happiest periods of my life working with the group at the Brookhaven Lab. Last but not the least I would like to thank my parents, XIE Guanghai and LUO Huaiqin, and my wife Xiaoshan Zhu for their love and support, which are the driving forces of my moving forward in my research as well as in life.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

INTRODUCTION

1.1 History and background of networks research

We live in a connected world, where we are attached to each other through all kinds of ties: family, friendship, sexual relationship, employment etc. Many physical networks are indeed indispensable foundations of our modern civilization: transportation networks, power grid networks, telecommunication networks, the Internet, the World Wide Web and so on. In fact the human body as a biological system is a sophisticated network of biochemical reactions. Even though these networks are very different in nature, they have many common characters, such as being composed of many interacting individuals, normally large in size and growing or evolving with time, etc.

It is of great practical importance and intellectual appeal to understand these networks. Traditionally networks study is the domain of graph theory, which was born in 1736 when mathematician Leonhard Euler provided the solution to the Königsberg Bridge Problem. In the past century graph theory has developed into a full fledged field and provided insights to many practical questions such as coloring graphs, finding shortest path and finding max flow [6] [7] [8]. Social scientists are another group of pioneers in network studies. Since the early 1920s they studied social networks of relationships among social entities, for example trades among nations and communications between members of a community.

In recent years the research interests in this field has been growing very rapidly and shifting to new directions. Thanks to the advancement of computer technology, it is possible to gather and analyze data sets of large scale networks such as the Internet, email communication networks and protein interaction networks in a cell etc. Networks research naturally shifted from understanding single small graphs and local properties of individual nodes and links to understanding the large scale statistical properties of the network. In understanding systems consisting of many interacting agents, statistical physics has been of great success, which made researchers realize that tools from statistical physics can provide insights to understanding complex networks too. Meanwhile, the complex network research has raised new questions and challenges for statistical physics. The main aims of complex network research has been on the following three aspects: defining proper concepts and statistical measures to describe the feature and characters of the networks; finding the designing principles or evolutionary rules which enable the networks to become as they are; understanding the behavior of the networks as a result of the global structural features and local properties of individual agents. Exciting progresses has been made from late the 1990s to the present and a large body of literature has been accumulated [1] [2] [3] [4] [5].

1.2 Structural measures of complex networks

A network is a collection of nodes(vertices) and links(edges) between them. Links can be directed and undirected, for example, hyper links on a Web page are directed, because one can only follow a link to go to a target web page, but not the reverse. So a hyper link is a directed link. Whereas a normal network cable allows data to be transferred both ways, so a network cable is an undirected link. Another kind of link,the weighted link, is introduced to capture the characteristic

of the network in more detail, when the strength of the link is important. See Illustration 1.1. The most common mathematical representation of a network is using adjacency matrix A. For a directed network A, its element $a_{ij} = 1$, if there is a link from node i to node j; other wise $a_{ij} = 0$; for an undirected network, $a_{ij} = a_{ji} = 1$ if there's a link between i and j; for a weighted network, the magnitudes of the matrix entries represent the weights of the links.

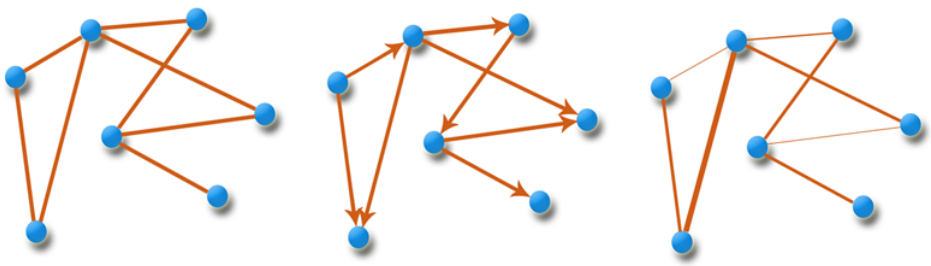

Figure 1.1: Three types of networks. A, undirected network; B, directed network; C, weighted network. The arrows in network B indicate the direction of the link. The thickness represent the different weights of the links in network C

Degree K of a node is the number of links or the immediate neighbors it has. If the total number of nodes is n,

$$K_i = \sum_{j=1}^{n} a_{ij}. \tag{1.1}$$

In a directed network the number of links a node receiving from other nodes is called in-degree K_{in}; conversely the number of links pointing from a node to other nodes is called out-degree K_{out}. An important property of a network is the degree distribution $P(k)$, which is the probability of a randomly picked node from the network to have degree K; in other words, it's the percentage of nodes with degree K among all the nodes in the network.

Distance and Diameter: Distance between two nodes in a network is the number of links in the shortest path connecting these two nodes. Diameter of a network is the largest distance between all pairs of the nodes in the network.

Betweenness: Node betweenness, also called betweenness centrality, was originally introduced in social network analysis to quantify the importance of individual nodes. Node betweenness of i is defined as the number of shortest pathes between other nodes that pass through i [9, 27]. Edge betweenness, on the other hand, is defined as the number of shortest pathes passing through an edge [11]. It indicates the traffic or load of an edge in a network.

clustering coefficient: Clustering is a common property of acquaintance networks, which represents that two people with a common friend are likely to know each other [9]. It was found later that the clustering feature presents in many other complex networks too. A widely used measure to quantify the clustering property is proposed by Watts and Strogatz [10]. Clustering Coefficient C_i of a node i is the ratio of number of links exist E_i to the total possible links among this node's K_i neighboring nodes:

$$C_i = \frac{2E_i}{K_i(K_i - 1)} \tag{1.2}$$

Clustering coefficient for the whole network is the average of all C_is of the nodes in the network.

1.3 Examples of networks in nature and society

1.3.1 Networks in Nature

- Food Web

Food Webs are used by ecologists to model the predator-prey relationships between different species [12, 13] in nature. Nodes in food webs are species and there's a directed link between two species if one is predator or prey of the other. Extensive studies of several ecological systems [14, 15, 16] have provided great data sets of these networks. Statistical studies on these networks [17, 18, 19] have shown that, with a relatively small size (the largest one contains less than 200 nodes), food webs are highly clustered and the diameter of the network is very small with a value of 3 or less. These studies also shed light on the degree distributions and robustness of the food webs.

- Gene regulatory network

  Gene regulatory networks are simplified models of the gene expression process inside living cells. Genes control the production of proteins through transcription and translation. These processes can be activated and inhibited by other proteins, which are the products of other genes. In a Gene regulatory network, nodes are proteins and directed links are the activation and inhabitation relationship between the proteins. The action of such networks is a major driving force of the animal development from a simple egg to a complex system. Statistical studies show that the degree distribution of these networks share some universal characteristics: the degree distribution scale as a power law $P(K) \sim K^{-\gamma}$, in which γ is a constant and there exists small-world effect and high clustering feature. Many efforts have been make to understand better these networks in terms of structure, evolution, robustness, etc [20, 21, 22, 23]. Techniques such as correlation profile[24] and boolean networks [25] have been applied to study these networks.

1.3.2 Social Networks

A social network consists of a set of agents that interact with each other [26, 27]. The interaction can be friendship or sexual relationship between people, intermarriages between families, trade between companies, etc. Social scientist started using networks to map and study the complex relationships since the 1920s. Many fundamental concepts and tools such as small-world property, clustering coefficient and node centrality were developed in sociology and remain useful in today's complex network research. Traditional social network studies often suffer from shortcomings of inaccuracy and small sample size due to their labor intensive data collection procedures such as using questionnaires and interviews. The tremendous advances in the technology of telecommunications and the Internet make it possible to collect more accurate and much larger social networks.

- Movie actor networks

  The Internet Movie Database (`http://www.imdb.com`) has the most extensive data of movies made around the world after the 1890's. In a movie actor network, two actors are connected if they act in the same movie. This network with number of nodes $N \sim 10^5$ is a good medium to study the collaboration in human society. Some statistical properties of these networks have been reported [28, 10, 29] that the average path length of the network is similar to a random network with the same degree distribution; it is highly clustered; the degree distribution follows a power law $P(K) \sim K^{-\gamma}$ with $\gamma \approx 2.3$.

- Scientific coauthorship networks

  Scientific coauthorship is another well documented example of collaborations among scientists [30, 31]. In the network, two scientists are connected

if they coauthored at least one manuscript. A number of studies have been focused on the coauthorship networks of physics, biomedical research, high-energy physics and computer science [32, 33, 34] ; mathematic and neuro-science [35]. Again short average path length, large clustering coefficient and power law degree distribution are found in all those networks.

1.3.3 Technological/Information Networks

- The Internet

  The internet is the physical network of routers and computers connected by network cables. Because of its large size and ever changing nature, it's mainly studied at two coarse-grained levels. One level is to consider the network of routers, which are specialized devices on the Internet to control the data transferring. Another level is to study the network of autonomous systems. An autonomous system is a group of routers and computers belonging to a same organization or under a same domain name. Early studies by Faloutsos *et al* [36]have found and later confirmed by Govindan *et al* [37] that at both levels the Internet has power law degree distributions with exponentials ranging from 2 to 2.3. Other authors have studied the clustering and betweenness [39] and degree assortivity [41]; and shed light on the modularity [40]and evolution of the Internet[38].

- The World Wide Web

  The World Wide Web (WWW) is the largest artificial network of which the topological data is accessible[42, 47]. It contains more than 10 billion web pages and still growing rapidly. While in daily life people use the Web and the Internet interchangeably, they are indeed two distinct networks. The

nodes in the Internet are hardware: computers and routers and links are undirected physical links, whereas nodes on the Web are web-pages we can browse and links are directed hyper links on the web pages. Because of the interesting properties the Web has and it's tremendous importance in delivering information, promoting business and facilitating people's daily life, WWW has drawn a lot research interest, especially after the discovery of the degree scaling property in 1999 [43, 44]. It has been reported that both the in-degree and out-degree of the web scale with power law distributions [45, 46]; the average path length is small [47]. Another important feature is that the web displays strong community structure in which groups of web pages (e.g. those devoted to a common topic or belonging to the same organization) are densely interconnected by hyper links [48].

CHAPTER 2

DYNAMIC PROCESSES TAKING PLACE ON THE NETWORKS

2.1 Introduction

The purpose of studying the structure and formation of networks is to eventually understand how the systems built upon the networks carry out their function, in other words, to understand the dynamic processes taking place on the networks. Scientists have been studying, for example, how the structure of a food web affects population dynamics, how the structure of the metabolic networks control cells' biological cycle, how the topology of the World Wide Web affect web browsing and searching; how the structure of contact networks spread disease. On the other hand, an auxiliary dynamic process set in a network can be very helpful in probing the network structure. In the following section we will review a study [49] utilizing such a diffusion process to study the modularity of the Internet. Similar diffusion processes taking place on the World Wide Web and scientific literature citation networks are considered to study the ranking in these networks in following chapters.

2.2 Diffusion on the internet

The network in question is the physical layout of the Internet on a coarse-grained level of the Autonomous Systems (AS), which are large groups of routers and servers belonging to the same organization such as a company, an Internet Service Provider (ISP) or a university. The conjecture is that this network can be divided into smaller sub-networks (modules), which interact with each other relatively weakly.

In the work of ref[49] the authors used the January 3,2000 dataset when the Internet consisted of 6474 Autonomous Systems exchanging information via 12572 undirected links. The auxiliary diffusion process describes the dynamics of a large number of random walkers moving on the network at discrete time steps. At each step, every walker moves from its current node to a neighbor of this node with equal probability among the neighbors. Denote the expected number of random walkers on node j as $\rho(j)$, the dynamics of this process can be written as:

$$\rho_i(t+1) = \sum_j T_{ij}\rho_j(t) \tag{2.1}$$

where T is the transfer matrix and its elements T_{ij} is $1/K_j$ if there is a link between node i and j, zero otherwise. K_j is the degree of node j. For $\sum_i T_{ij} = 1$, the total number of random walks is conserved at all times. Eq. 2.1 can be rewritten as $\rho_i(t+1) - \rho_i(t) = \sum_j (T_{ij} - \delta_{ij})\rho_j(t)$. Hence the diffusion equation for this process in matrix form is:

$$\Delta\rho = \mathbf{D}\rho \tag{2.2}$$

where $\mathbf{D} = \mathbf{T} - 1$ is the diffusion matrix.

As time advances, the diffusion process approaches a steady state $\rho_i(\infty)$, in which the number of random walkers on each node doesn't change with time. This

can be satisfied with a distribution in which the number of random walkers $\rho_i(\infty)$ on a node is proportional to its degree, so that the out-going current is exactly balanced by the in-coming current of the node and equation $\rho_i(\infty) = \sum_j T_{ij}\rho_j(\infty)$ holds, which states that $\rho_i(\infty)$ is the principle eigenvector of $\mathbf{T}$ corresponding to the principle eigenvalue $\lambda^{(1)} = 1$. For networks consists of one single component such as the Internet, this steady state configuration is unique. The remaining eigenvectors which correspond to eigenvalues less than 1 represent initial configurations that decay with time towards the steady state. The characteristic decay time $\tau^{(\alpha)}$ of the αth eigenvector is given by $exp(-1/\tau^{(\alpha)}) = |\lambda^{(\alpha)}|$. The eigenvectors with eigenvalues $\sim \pm 1$ describe the slowly decaying modes, which we will see are closely related to the the modularity of the network. The authors of the aforementioned paper proposed a statistical measure Participation Ratio (PR) of the diffusion eigenvector to quantify the effective number of nodes participating in a given eigenvector with a significant weight. To calculate PR, first normalize the eigenvector by each node's degree to take into account the fact that in steady state or slowly decaying states, $\rho_i^{(\alpha)}$ is proportional to K_i: $c_i^{(\alpha)} = \rho_i^{(\alpha)}/K_i$. $c_i^{(\alpha)}$ is again normalized by $\sum c_i^2 = 1$. The Participation Ratio is defined as:

$$PR = (\sum_{i=1}^{N} c_i^4)^{-1} \tag{2.3}$$

It's straightforward to show that $PR = N$ for the steady state solution $\rho_i(\infty)$, where N is the total number of nodes in the network.

Fig.2.2 shows the PR of $\rho_i^{(\alpha)}$ and density of the eigenvalues with $-1 < \lambda^{(\alpha)} < 1$ in the Internet and its randomized counterpart, which is obtained by applying "Local Rewiring" [24] algorithm on the Internet. The resultant Random Scale Free Network (RSFN) preserves the exact degree of each node but the neighbors of a node are randomly reassigned. It is proposed [24] that RSFN is a good null model to differentiate non-random features of a complex network from random

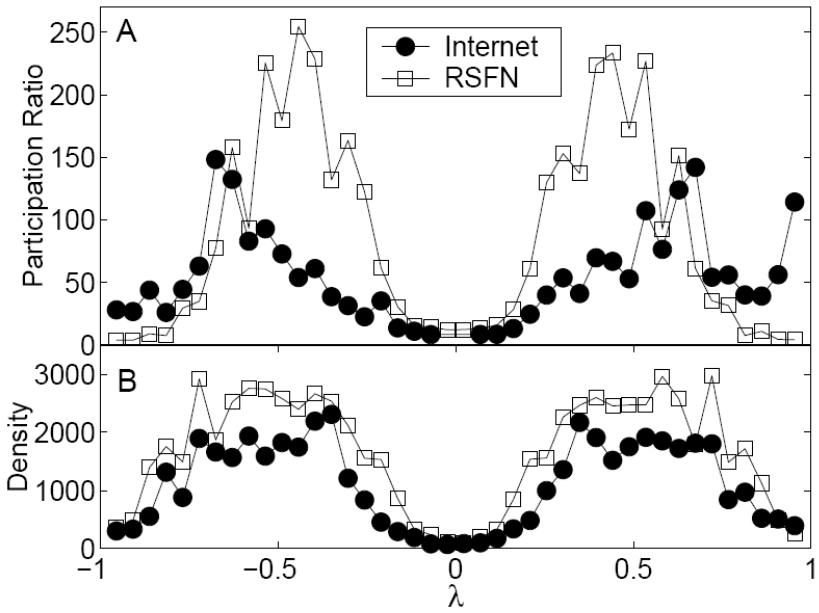

Figure 2.1: The participation ratio and the eigenvalue density as a function of the eigenvalue measured in the Internet and in its randomized counterpart - a Random Scale Free Network. Notice that for $|\lambda| \approx 1$ participation ratios in the Internet significantly exceed those in an RSFN indicating the modularity in the Internet. Image courtesy of K. A. Eriksen, I. Simonsen, S. Maslov, K. Sneppen.

ones. While the eigenvalue density of the Internet and a RSFN is similar to each other, the participation ratio for slowly decaying modes ($\lambda \approx \pm 1$) of the Internet is significantly larger than that of a RSFN. A closer look at the slowly decaying modes reveals that in those non-oscillatory modes (with $\lambda \sim 1$) the diffusion current flows from relatively isolated regions (modules) along the few links connecting them to the rest of the network. If for such modes these links would be hypothetically removed one by one, the corresponding eigenvalue would gradually increase towards 1, while the eigenvector would become more and more localized on the module. When eventually the module is completely disconnected from the rest of the network, the eigenvector evolves to the steady state solution on the module, which has PR equal to its size. This shows that the Participation Ratio of slowly decaying modes is a good quantitative estimate of the size of modules in the network. And Fig.2.2 indicates that the PR for slowly decaying modes is significantly larger than that of a RSFN, these modules in slowly decaying modes are real and not by chance.

In an effort to determining the organizing principle behind the Internet modules, [49] provided Fig.2.2, in which components of the slowest decaying mode are plotted against AS number (a numeric ID of Autonomous Systems). Autonomous Systems known to be located in Russia are marked with a circle. The total number of AS in Russia is 174, while the PR for this eigenvector is 107. The figure also shows that almost all AS that significantly participate in the eigenvector belong to Russia. Hence the slowest decaying module correspond to AS in Russia; similar study shows that other slow decaying modules are identified to be countries or organizational or geographical features within a large country.

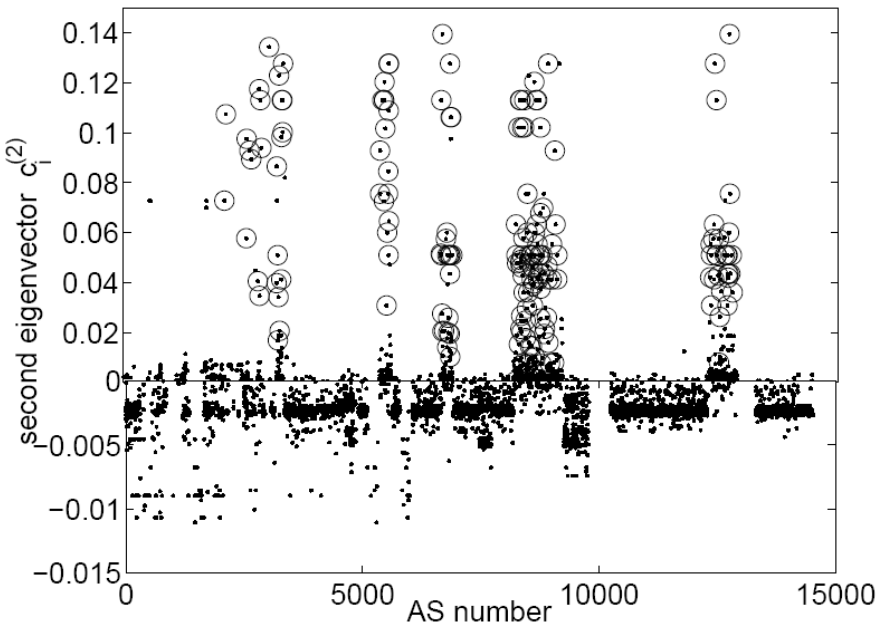

Figure 2.2: Scatter plot of the components of eigenvector $c_i^{(2)}$ corresponding to eigenvalue $\lambda^{(2)} = 0.9626$. The AS known to be located in Russia are marked with circles. Out of 100 AS with the most negative components, 23 are identified to be associated with the US Military. Image courtesy of K. A. Eriksen, I. Simonsen, S. Maslov, K. Sneppen.

CHAPTER 3

OPTIMAL RANKING IN NETWORKS WITH COMMUNITY STRUCTURE

The World Wide Web (WWW) – a very large ($\sim 10^{10}$ nodes) network consisting of web pages connected by hyper links – presents a challenge for efficient information retrieval and ranking. Apart from the contents of web pages, the topology of the network itself can 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 [50] 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.

3.1 The Google PageRank algorithm

In the heart of the Google search engine lies the PageRank algorithm determining the global "importance" of every web page based on the hyper link structure of the WWW network around it. When one enters a search keyword such as "statistical physics" on the Google website the search engine first localizes the subset of web pages 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 [50] 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 applied to study citation and metabolic networks [52] and the modularity of the Internet on the hardware level represented by an undirected network of interconnections between Autonomous Systems [40]. As in real web surfing, a random walker of the PageRank algorithm could "get bored" from following a long chain of hyper links. 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 hyper links. This leaves the probability $1 - \alpha$ for it to randomly select and follow one of the hyper links of the current web page. According to [50], 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 usually 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 is

$$G(i) = \alpha + \sum_{j \to i} (1 - \alpha) \frac{G(j)}{K_{out}(j)}. \tag{3.1}$$

Here $K_{out}(j)$ denotes the number of hyper links (the out-degree) of the node j and the summation goes over all nodes j that have a hyper link 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 [51].

3.2 Effects of community structure on ranking of the web pages

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} denote the total number of hyper links pointing from nodes in the *c*ommunity to the outside *w*orld, while E_{wc} is the total number of hyper links pointing in the opposite direction. Similarly E_{cc} and E_{ww} denote the total number of links connecting nodes within the community and, respectively, the outside world. (See Fig.3.1 for an illustration).

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 hyper links between the community and the outside world, and the current due to random jumps.

We first consider the effect of the community structure on Google ranking in the simplest and most physically transparent case of $\alpha = 0$. Consider a network in which N_c nodes form a community characterized by higher than average density of edges linking these nodes to each other. The total number of hyper links pointing to nodes inside the community is given by $E_{cc} + E_{wc} = N_c\langle K_{in}\rangle_c$ where $\langle K_{in}\rangle_c$

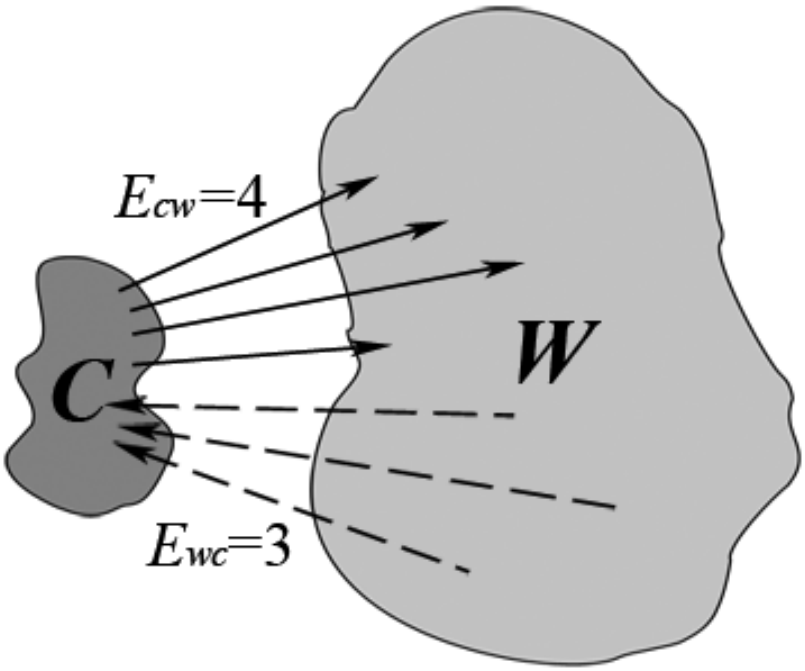

Figure 3.1: Illustration of hyper link connections between the community C and the outside world W. E_{cw} and E_{wc} are numbers of links from the community to the outside world and from the outside world to the community, respectively.

is the average in-degree of community nodes. Similarly, $E_{cc} + E_{cw} = N_c\langle K_{out}\rangle_c$, where $\langle K_{out}\rangle_c$ is the average out-degree in the community, gives the total number of hyper links originating on community nodes. The Google rank is computed in the steady state of the diffusion process where the number of random surfers currently visiting any given web page does not change with time. This means that the total current of surfers J_{cw} leaving the community for the outside world must be precisely balanced by the current J_{wc} entering the community during the same time interval.

Let $G_c = \langle G(i)\rangle_{i\in C}$ denote the average Google rank inside the community given by the average number of random surfers on its nodes. Here we introduce a mean-field approximation 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. Under this assumption, the average current flowing along any of those edges would be given by $G_c/\langle K_{out}\rangle_c$ while the total current leaving the community $J_{cw} = E_{cw}G_c/\langle K_{out}\rangle_c$. Similar analysis gives $J_{wc} = E_{wc}G_w/\langle K_{out}\rangle_w$, where $\langle K_{out}\rangle_w$ is the average out-degrees of nodes in the world outside the community.

Balancing these two currents one gets:

$$\frac{G_c}{G_w} = \frac{E_{wc}}{E_{cw}} \cdot \frac{\langle K_{out}\rangle_c}{\langle K_{out}\rangle_w} \quad . \tag{3.2}$$

The Eq. 3.2 is based on a mean-field assumption that average values of the Google rank and the out-degree on those community nodes that actually send links to the outside world are equal to their overall average values inside the community

It is tempting to assume that higher than average density of hyper links connecting nodes in the community is beneficial for the Google rank of its nodes

as it "traps" random surfers to spend more time within the community. It turned out that this naive argument is not necessarily true. In fact one is equally likely to observe an opposite effect: an excess of intra-community links could lead to a lower than average Google rank of its nodes.

To see it explicitly one should replace E_{wc} and E_{cw} in Eq. 3.2 with identical expressions $\langle K_{in}\rangle_c N_c - E_{cc}$ and $\langle K_{out}\rangle_c N_c - E_{cc}$ respectively:

$$\frac{G_c}{G_w} = \left(\frac{\langle K_{in}\rangle_c N_c - E_{cc}}{\langle K_{out}\rangle_c N_c - E_{cc}}\right) \cdot \frac{\langle K_{out}\rangle_c}{\langle K_{out}\rangle_w} \quad . \tag{3.3}$$

From this equation it follows that enhancing the community structure (increasing E_{cc}) while keeping other parameters such as $\langle K_{in}\rangle_c$,$\langle K_{out}\rangle_c$ and $\langle K_{out}\rangle_w$ fixed can be both good and bad for the average Google rank of the community web pages. It depends on $\langle K_{in}\rangle_c/\langle K_{out}\rangle_c$ – the ratio between average in- and out-degrees of community nodes. If the ratio is less than 1 the increase in E_{cc} leads to a further decrease of G_c/G_w below one. If the community constitutes just a small fraction of the whole network one could safely assume that G_w remains approximately constant so that the average Google rank of the community, G_c, has to decrease. Similarly if the ratio is larger than 1, G_c grows with the number of inter-community links E_{cc} (see Fig. 2 for an illustration of both cases).

3.2.1 Numerical Study

For numerical studies of networks with a community structure we propose a particular version of the Metropolis network rewiring algorithm [41]. It allows one to generate an ensemble of random networks with user-defined in- and out-degree distributions as well as with any desired density of links between pre-selected N_c "artificial community" nodes. It starts from a seed network with the preferred (scale-free in our case) distributions of in- and out-degrees [53] and proceeds by a sequence of edge-swapping steps changing a pair of randomly selected edges

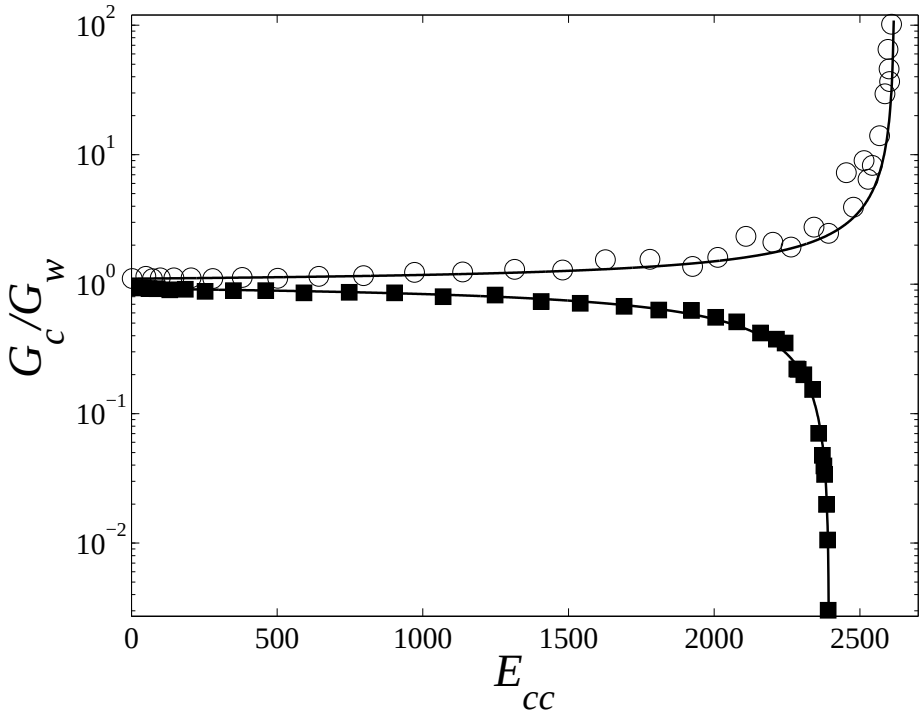

Figure 3.2: The ratio of average Google ranks in the community and the outside world G_c/G_w as a function of E_{cc} – the number of intra-community links – in two series of model networks with varying degree of community structure. Open circles correspond to a beneficial effect of the community structure on Google ranking in a scale-free network with $\langle K_{out}\rangle_c = 5.24 < \langle K_{in}\rangle_c = 5.9$. On the other hand, filled squares show a detrimental effect in another series of networks where $\langle K_{out}\rangle_c = 5.6 > \langle K_{in}\rangle_c = 4.8$. Solid lines are fits with the Eq. 3.3 with a given set of parameters for each of the networks. All networks with $10,000$ nodes have a community of 500 nodes were generated by the Metropolis rewiring algorithm described later on in the text.

A→B and C→D into A→D and C→B correspondingly. The decision on whether to accept or reject an elementary step depends on changes in the energy function H and the inverse temperature β. For our purposes of generating networks with community structure we choose $H = -E_{cc}$, where E_{cc} is the number of edges connecting pre-selected community nodes to each other. Fig. 3.3 shows the results of a numerical test of Eqs. 3.2, on simulated directed networks with scale-free distributions of in- and out-degrees: $P(K_{in}) \sim K_{in}^{-2.1}$ and $P(K_{out}) \sim K_{out}^{-2.5}$ correspondingly. The exponents were selected to be identical to their values in the actual WWW network [48, 44] and a community structure was generated by the Metropolis algorithm described above. The reciprocal temperature β used in the Metropolis algorithm indirectly determines the number of links within the community. A network without any community structure is realized at an infinite temperature ($\beta = 0$), while the algorithm run at zero temperature ($\beta = \infty$) produces a network with the largest possible number of links within the community. One could also run the algorithm with $\beta < 0$. Negative values of β generate networks with an anti-community structure in which the number of intra-community links is lower than that in a random network. The relation between E_{cc} and β for both positive and negative values of β is shown in Fig. 3.4. To analytically derive such a relation we consider the detailed balance in the steady state of the Metropolis rewiring algorithm, in which the probabilities of an increase and a decrease in E_{cc} are equal to each other. It results in the equation

$$E_{cw}E_{wc} = E_{cc}E_{ww}e^{-\beta} \qquad . \tag{3.4}$$

Additional constraints (i) $E_{cc}+E_{wc} = \langle K_{in}\rangle_c N_c$ (the sum of in-degrees of all nodes within the community), (ii) $E_{cc} + E_{cw} = \langle K_{out}\rangle_c N_c$ (the sum of out-degrees of all nodes within the community) and (iii) $E_{cc} + E_{cw} + E_{wc} = E$ (the total number of edges in the network) plugged into the Eq. (3.4) result in a quadratic equation for

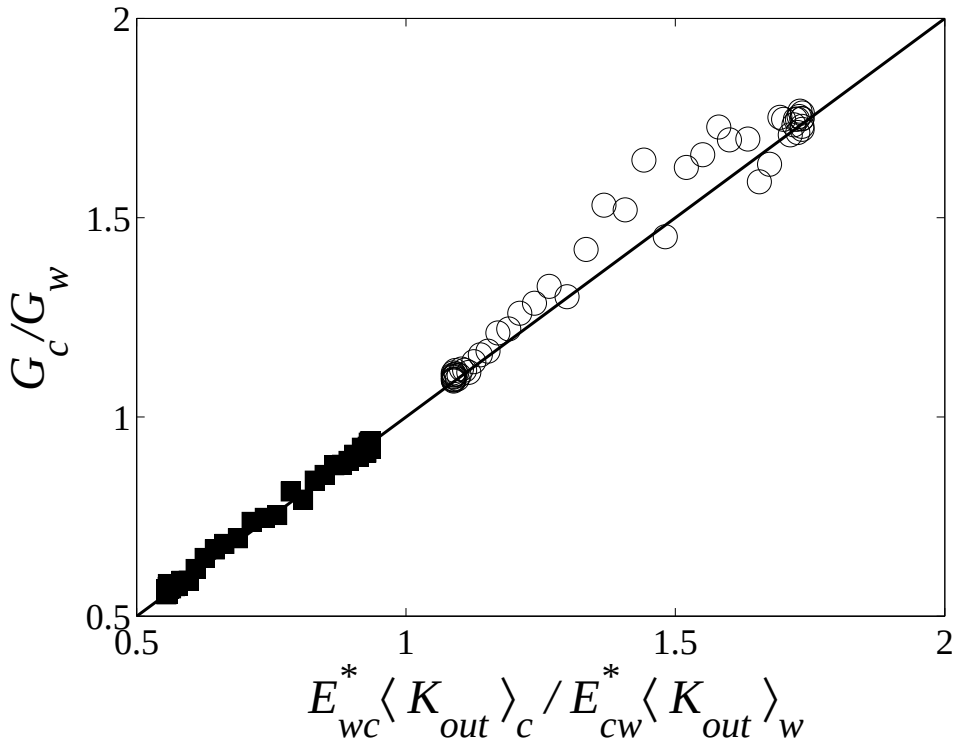

Figure 3.3: The ratio of average Google ranks in the community and the outside world G_c/G_w as a function of the ratio of effective numbers of links E^*_{wc}/E^*_{cw}. As predicted by the Eq. 3.2 these two ratios are basically equal to each other. Different symbols correspond to series of networks described in Fig.3.2.

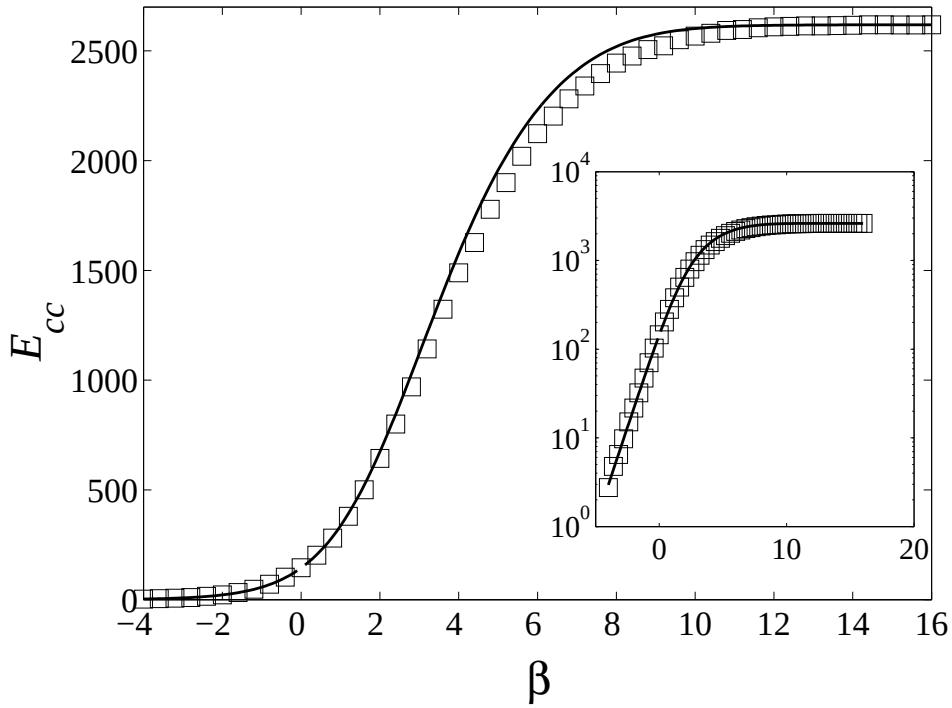

Figure 3.4: The number of intra-community links E_{cc} in networks generated by the rewiring algorithm as a function of the inverse temperature β. Negative values of β correspond to networks with anti-community structure and are generated by changing the sign in front of the Hamiltonian H. The solid line is the fit with the analytical expression obtained by solving the Eq. 3.4 for E_{cc}. The inset shows the same plot with a logarithmic scale of the Y-axis.

E_{cc} as a function of $\langle K_{in}\rangle_c$, $\langle K_{out}\rangle_c$, E, and β – the parameters strictly conserved in our rewiring algorithm. The Fig. 3.4 compares the analytical expression for $E_{cc}(\beta)$ obtained by solving the Eq. 3.4 with numerical simulations for different values of β. Clearly, E_{cc} increases with β in general accord with the Eq. 3.4. When β is sufficiently large, E_{cc} exponentially approaches a limiting value equal to $\min(\langle K_{in}\rangle_c, \langle K_{out}\rangle_c)N_c$ – the minimal number of links within a community given the set of in- and out-degrees of its nodes. The deviations between the analytical formula and numerical results visible for large values of β could be attributed to the "no multiple edges" restriction in networks generated by our rewiring algorithm. As the density of inter-community links increases more and more of rewiring steps leading to an increase of E_{cc} are aborted as the new link they are attempting to create within a community already exists. This situation is more appropriately described by the following equation: $E_{cw}E_{wc}(1-E_{cc}/E)(1-E_{ww}/E) = E_{cc}E_{ww}(1-E_{cw}/E)(1-E_{wc}/E)e^{-\beta}$, reminiscent of the detailed balance equation in two-fermion scattering (see also [55]).

3.3 Optimal ranking in networks with community structures

In this section we study how the only intrinsic parameter in the PageRank algorithm affects the effectiveness of the ranking results given the presence of community structure on the World Wide Web.

Let $G_c = \langle G(i)\rangle_{i\in C}$ denote the average Google rank of web pages 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

hyper link 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 $E^{(r)}_{wc}$ and $E^{(r)}_{cw}$ – 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 [56]. By approximating $G_w \approx 1$, we finally arrive at the following equation:

$$G_c = \frac{(1-\alpha)\frac{E_{wc}}{E^{(r)}_{wc}}+\alpha}{(1-\alpha)\frac{E_{cw}}{E^{(r)}_{cw}}+\alpha}. \tag{3.5}$$

For simplicity of notation, let us refer to the ratios $E_{wc}/E^{(r)}_{wc}$ and $E_{cw}/E^{(r)}_{cw}$ 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 [57]. One implication of the Eq.3.5 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.3.5 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 the real World-Wide Web, we downloaded [58] complete sets of hyper links contained in all web pages 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 system). (See Table 3.1 for details.) The relation between G_c and α for six such communities are shown in Fig.3.5. 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.3.5, 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 3.2). Furthermore, by looking at from which values of α, 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 exam-

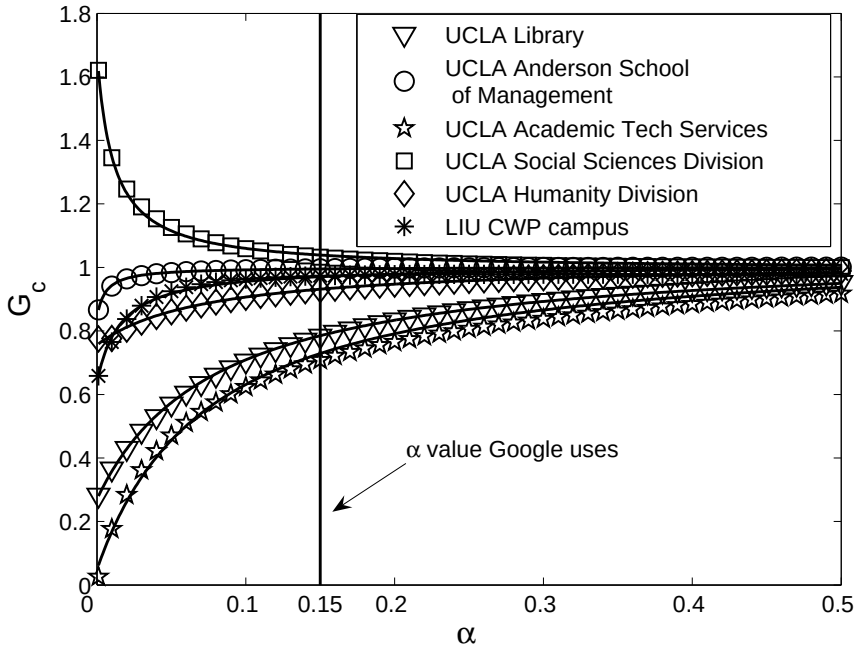

Figure 3.5: The average Google rank G_c of different communities as a function of the parameter α. The communities are within real WWW networks of two US universities (see Table 3.1 for details). The data points are obtained by running the PageRank algorithm for different values of α. Solid lines are two-parameter best fits to the data with the Eq.3.5.

ple, "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. 3.5 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.3.5 to deviate from their values prescribed by the mean-field theory provides a simple mathematical 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.3.5 (see Table 3.2.) 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 3.2 and visualized in Fig.3.6. Generally speaking, the closer

Table 3.1: The basic statistics about the academic WWW networks downloaded from Ref. [58]. We choose to study hyper link 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 [50] we iteratively removed web pages 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 3.2: 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 [56]. R^*_{cw} and R^*_{wc} are result of fitting the G_c and α dependency via Eq.3.5.

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

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 Google [51], 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.3.6 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.3.5 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. 3.6 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.3.5.

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 hyper link structure of client's web pages 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 hyper links 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 hyper links 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 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 hyper links 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

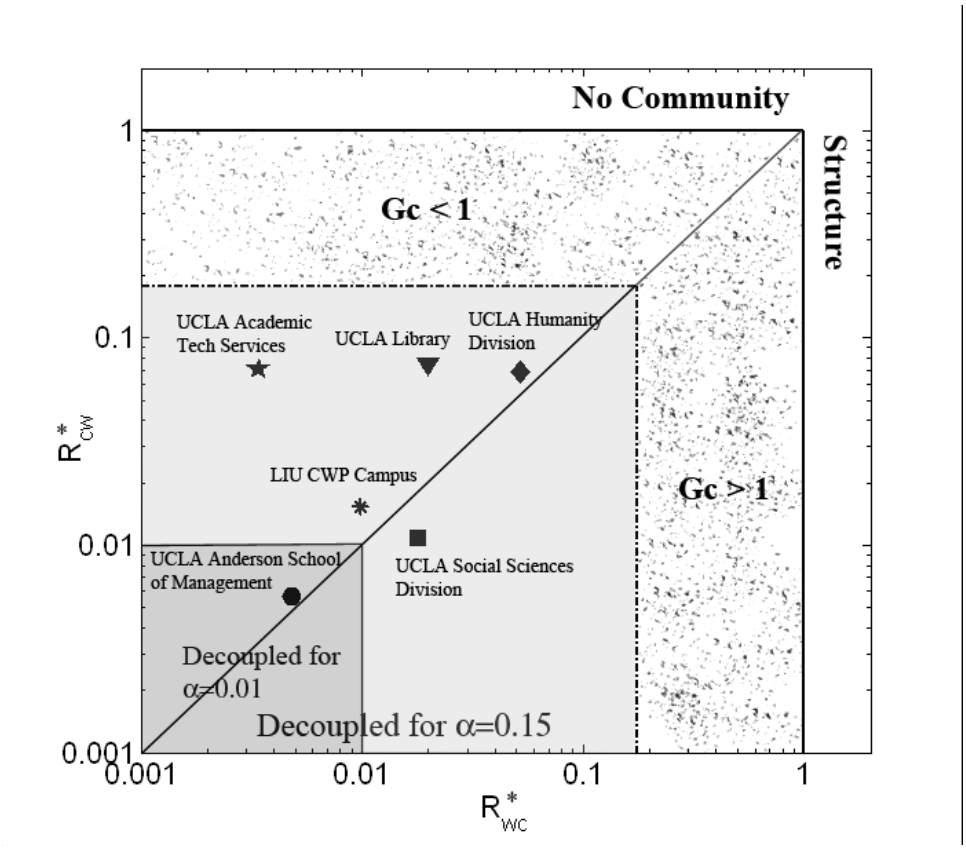

Figure 3.6: R^*_{cw} and R^*_{wc} for different communities. For communities above the diagonal, $R^*_{cw} > R^*_{wc}$ which implies that the average PageRank G_c is less than 1; for communities beneath the diagonal, $R^*_{cw} < R^*_{wc}$ which implies that the average PageRank G_c is greater than 1. 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 $\alpha = 0.01$. Google's choice of $\alpha = 0.15$ makes all communities we study decoupled from the rest of the world.

Fig.3.6). Thus, for our sample of web-communities, $\alpha = 0.15$ proposed in [50] indeed strikes the best possible balance between the opposing demands on the value of α.

CHAPTER 4

NEW WAYS OF RANKING SCIENTIFIC LITERATURE

4.1 Using Google's pagerank on citation networks

Due to their rapid growth, many information networks have become untenable to navigate without some sort of ranking scheme. This is particularly evident in the example of the World Wide Web, a network of pages connected by hyperlinks. A successful solution to the problem of ranking the Web is Google's PageRank algorithm [50]. Another class of information networks that could benefit from such a ranking method are citation networks.

With the advancement of information technology, large electronic citation data base became available. It enables relatively comprehensive study of scientific citations on certain disciplines using a network model. The citation networks are comprised of scientific publications as nodes and directed links from a citing article to a cited article. The network we study contains citation data between journals published by the American Physical Society [59]. This dataset contains around 380,000 papers and 3,100,000 citation links. We know only the year in which each paper was published and it ranges from 1893 to 2003. Current methods of measuring the influence of an article are based on the total number of citations it receives. In citation networks this measure is the in-degree of a

node. It has been reported that the in-degree in various citation data sets has a broad distribution and fits well with a power law, as shown in Fig.4.1.

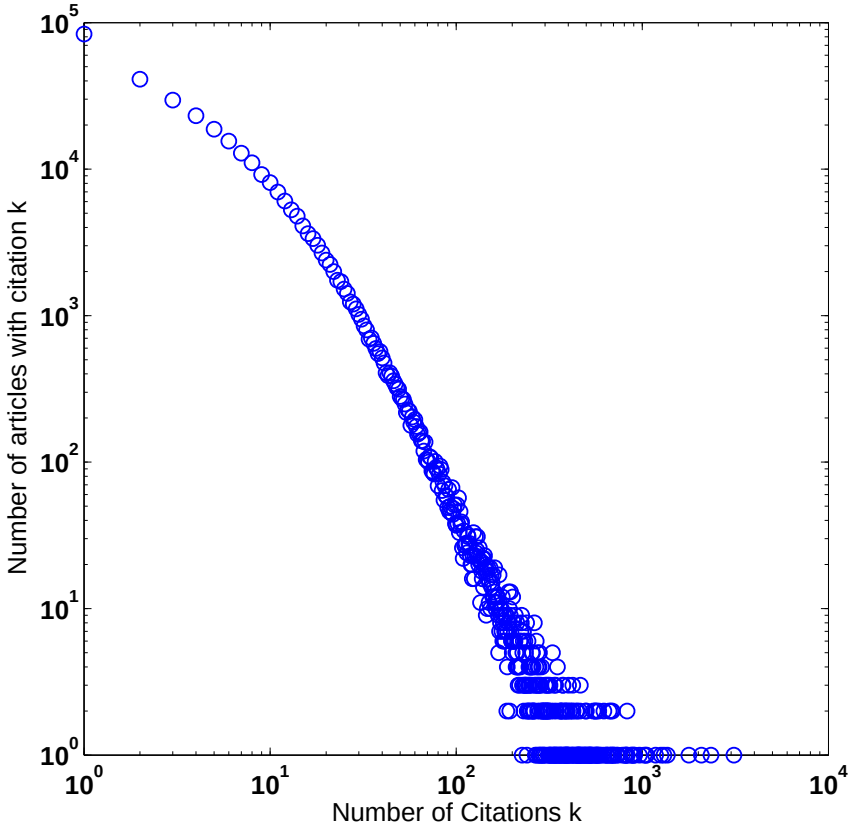

Figure 4.1: In-degree distribution of the citation network. x-axis is the number of citations an article has and y-axis is the number of articles having the particular number of citations

While it is a natural way to use total number of citations to rank articles, it is a rather crude measure for it's too "democratic" in treating all citations as equal and ignoring differences in importance and the number of citations of the citing papers. Intuitively a citation from an influential article should weight more than a citation from a less influential article. The Google PageRank behind

Google search engine is a good metric in ranking the web-pages on WWW. It can also provide a more accurate measure of the impact of an article when applied to the citation networks. As discussed in the previous chapter, PageRank is calculated by simulating a random walk on a network and the PageRank value of a node is the number of random walkers on the node in steady state. One of the advantages of Google's PageRank algorithm is that it implicitly accounts for the importance of the citing article in a self-consistent fashion. And it takes into account the effect that citation coming from an article with a large number of references should count less than those coming from one with less references. In other words, all other conditions the same, paper A should be considered more influential if it is amongst a few ones that inspired another paper B than if A is one of hundreds citations of paper B.

Authors of [60] proposed using the PageRank algorithm to improve the formula used to calculate the impact factor of scientific journals.

4.2 Optimal prameter of PageRank algorithm for citation networks

As reviewed in the previous chapter, the PageRank value $G(i)$ of a web-page i^{th} is defined by the recursion formula 3.1: $G(i) = \alpha + \sum_{j \to i}(1-\alpha)\frac{G(j)}{K_{out}(j)}$. Here the sum is over the neighboring web-pages j that link to i. The first term describes the behavior of web surfers randomly jump to another web-page with probability α while browsing on the web. The second term describes propagation of the other random surfers along hyper links with even probability $1/k_j^{out}$, where k_j^{out} is the out-degree of web-page j.

It is worthwhile to note that PageRank after all utilizes only the topological

information to rank the articles. Quality of the articles is not directly considered in the method. Whereas PageRank can be a good metric for the importance of the papers because the structure of the citation network is the result of the collective behavior of many scientists: they study scientific literature, conduct research and make new citations in their publications. Thus the quality of the articles are "encoded" in the network topology. Using in-degree, which is the number of citation an article receives, as a measure of quality can be thought as a first order approximation, whereas PageRank is a more sophisticated way to utilize the structural information.

In equation 3.1, α is a free parameter that controls the performance of the PageRank algorithm. The prefactor $(1-\alpha)$ in the second term gives the fraction of random walks that continue to propagate along the links; a complementary fraction α is uniformly re-injected into the network, as embodied by the first term. In the original Google PageRank algorithm of Brin and Page [50], the parameter α was chosen to be 0.15. It is undisclosed to the public how exactly Google has chosen this value, whereas we concluded this value is optimized for WWW in previous chapter. Choosing $\alpha = 0.15$ corresponds to that when surfing the web one typically follows of the order of $L = 1/0.15 \simeq 7$ hyperlinks before becoming either bored or frustrated with this line of browsing and jumps to a new web-page. In the context of citations, we conjecture that entries in the reference list of a typical paper are collected following somewhat shorter paths of average length 2, making the choice $\alpha = 0.5$ more appropriate for a similar algorithm applied to the citation network. The empirical observation justifying this choice is that approximately 50% of the articles [61] in the reference list of a given paper A have at least one citation to another article that is also in the reference list of A. Assuming that such "feed-forward" loops result from authors of paper A following references of paper B, we estimate the probability $1-\alpha$ to follow

this indirect citation path to be close to 0.5. To implement the Google PageRank algorithm for the citation network, we start with a uniform distribution of random walkers by placing 1 random walk on each node of the network and then iterate Eq. 3.1. Eventually a steady state set of number of random walkers for each node is reached. Then the PageRank value of each node is given by the number of random walkers residing on it. The PageRank value obtained with this method is proportional to the occupation probability at each node for the random walk process defined by Eq. 3.1. Finally, we sort the PageRank values to determine the rank of each node. It is both informative and entertaining to compare the Google rank with the citation (in-degree) rank of typical and the most important publications in Physical Review.

4.3 Google's PageRank for physical review publications

Fig. 4.2 shows the average PageRank value $\langle G(k)\rangle$ for publications with k citations as a function of k. For small k, there are many publications with the same number of citations and the dispersion in $G(k)$ is small. Correspondingly, the plot of $\langle G(k)\rangle$ versus k is smooth and increases linearly with k for $k \geq 50$ Thus the average PageRank value and the number of citations represent similar measures of popularity, a result that has been observed previously [62, 63]. In fact, the citation and PageRank value distributions are qualitatively similar, further indicating that citations and PageRank value, on the average, are similar measures of importance.

However, for large k, much more interesting behavior occurs. When k is sufficiently large, there is typically only one publication with k citations. Thus instead of an average value, individual PageRank values are plotted as a function of the number of citations in Fig. 4.3. The extreme outliers with respect to the

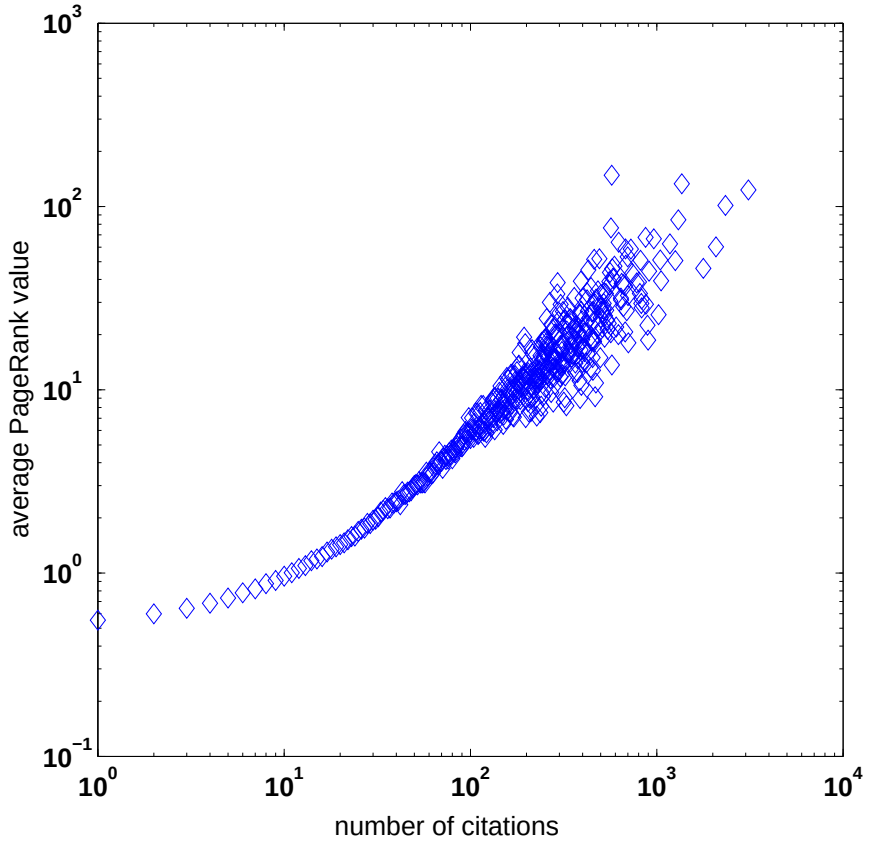

Figure 4.2: Average PageRank value $\langle G(k) \rangle$ versus number of citations k. For small k, there are many articles with the same number of citations, which results in small fluctuation in $\langle G(k) \rangle$ and PageRank value grows linearly with k.

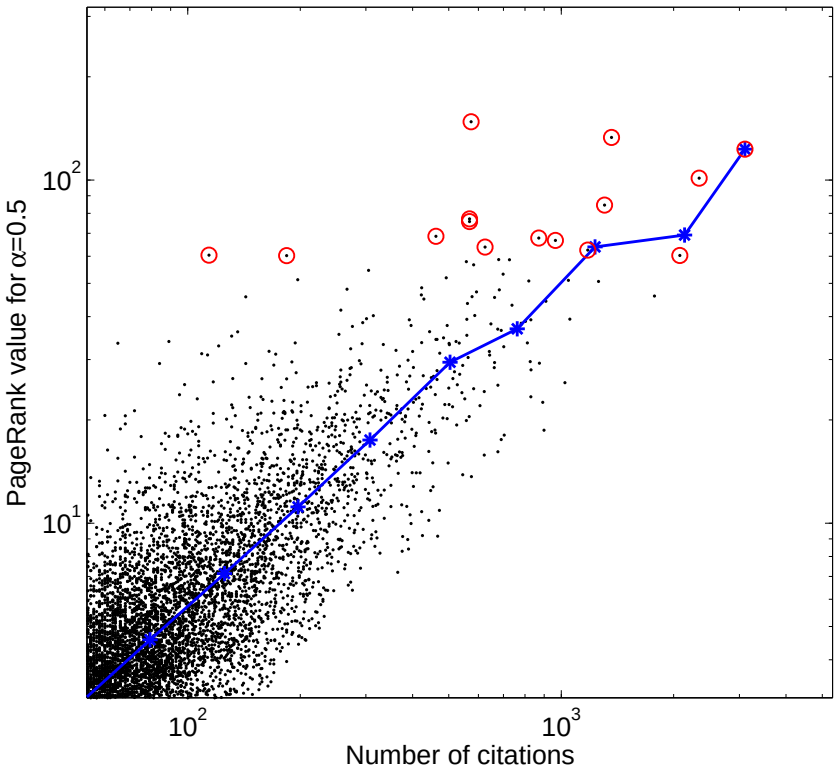

Figure 4.3: Individual outlier publications. The scatter plot of the PageRank value vs the number of citations. The top-15 Google-ranked papers are identified by author(s) initials (see Table 4.1). As a guide to the eye, the solid curve is logarithmically binned average of the data of $\langle G(k) \rangle$ versus k in Fig. 4.2.

linear behavior of Fig. 4.2 are the ones with relatively low citations but high PageRank value. To directly compare an article's rank by PageRank and by pure number of citations, we plot each article's PageRank against its rank by citation in Fig.4.4. The fifteen articles with the highest PageRank value are marked with large circles in Fig.4.3 and Fig.4.4. Some of these fifteen articles are ranked even beyond top 1000 by citations.

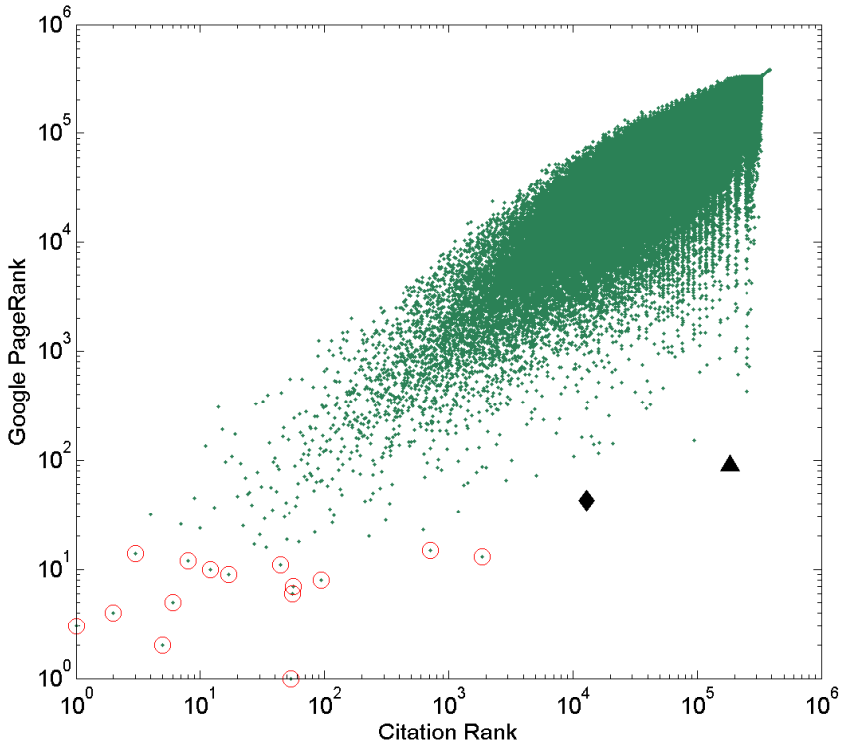

Figure 4.4: PageRank of each article against rank by number of citations. Articles that are located below the diagonal have higher PageRanks than their citation ranks. Top-15 PageRank papers are marked with red circles. Some of these fifteen articles are ranked even beyond top 1000 by citations. The black triangle and black diamond show two examples where PageRank is inaccurate.

The top-15 papers are listed Table4.1 with number of citations and the citation

rank of these publications. While several of the highest-cited Physical Review papers appear on this list, there are also several more modestly-cited papers that are highly ranked according to the PageRank algorithm. The disparity between the Google rank and citation rank arises because, as mentioned in the previous section, the former involves both in-degree as well as PageRank of the neighboring nodes. According to the Google algorithm of Eq. 3.1, an article j contributes a factor $G(j)/k(j)$ to the PageRank value of its reference i, where $k(j)$ is the out-degree of article j. Thus for a paper to have a large PageRank value, the articles citing it should be important (large $G(j)$), and have a small number of citations (small out-degree $k(j)$). The latter ensures that the contribution of a citing article is not strongly diluted.

With this perspective, let us compare the statistical measures of the two articles "Unitary Symmetry and Leptonic Decays", Phys. Rev. Lett. **10**, 531 (1963) by N. Cabibbo (C) and "Self-Consistent Equations Including Exchange and Correlation Effects", Phys. Rev. **1**40, A1133 (1965) by W. Kohn & L. J. Sham (KS). The former has the highest PageRank value of all Physical Review publications, while the latter is the most cited. The high PageRank of C stems from the fact that that value of $\langle G(j)/k(j)\rangle = 0.524$ for the children of C is an order of magnitude larger than the corresponding value $\langle G(j)/k(j)\rangle = 0.079$ for the children of KS. This difference more than compensates for the factor 5.4 difference in the number of citations to these two articles (3104 for KS and 574 for C as of June 2003). Looking a little deeper, the difference in $\langle G(j)/k(j)\rangle$ for C and KS stems from the denominator; the children of C (papers that cite C) have 15.4 citations an average, while the children of KS are slightly "better" and have 18.4 citations on average. However, the typical child of C has fewer references than a child of KS and a correspondingly larger contribution to the PageRank of C.

Table 4.1: Top-15 PageRank articles.

P.R[1]	P.R Value[2]	C.R.[3]	# of Cites[4]	Publication	Title	Author(s)
1	147.98	54	574	PRL, **10**, 531, (1963)	Unitary Symmetry and Leptonic...	N. Cabibbo
2	133.17	5	1364	PR, **108**, 1175, (1957)	Theory of Superconductivity	J. Bardeen, L. Cooper, et al.
3	123.04	1	3104	PR, **140**, A1133, (1965)	Self-Consistent Equations...	W. Kohn & L. J. Sham
4	101.29	2	2340	PR, **136** B864, (1964)	Inhomogeneous Electron Gas	P. Hohenberg & W. Kohn
5	84.57	6	1306	PRL, **19** 1264, (1967)	A Model of Leptons	S. Weinberg
6	77.12	55	568	RMP, **15** 1, (1943)	Stochastic Problems in physics...	S. Chandrasekhar
7	75.69	56	568	PR, **65** 117, (1944)	Crystal Statistics	L. Onsager
8	68.57	94	462	PR, **109** 193, (1958)	Theory of the Fermi Interaction	R. P. Feynman & M. Gell-Mann
9	67.81	17	871	PR, **109** 1492, (1958)	Absence of Diffusion in Certain...	P. W. Anderson
10	66.68	12	954	PRL, **42** 673, (1979)	Scaling Theory of Localization...	E.Abrahams, P. W Anderson et al.

Continued on next page

Table 4.1 – continued from previous page

P.R[1]	P.R Value[2]	C.R.[3]	# of Cites[4]	Publication	Title	Author(s)
11	63.78	44	625	PRL, **58** 908, (1987)	Superconductivity at 93 K...	M. K. Wu, J. R. Ashburn et al.
12	62.55	8	1178	PR, **124** 1866, (1961)	Theory of the Fano Resonance...	O. Újsághy, J. Kroha
13	60.45	1848	114	PR, **34** 1293, (1929)	The Theory of Complex Spectra	J. C. Slater
14	60.29	3	2079	PRB, **23** 5048, (1981)	Self-interaction correction to...	J. P. Perdew, Alex Zunger
15	60.22	712	184	PR, **43** 804, (1933)	On the Constitution of Metallic Sodium	E. Wigner and F. Seitz

Table 4.1: 1,PageRank; 2, PageRank value; 3, citation rank; 4, number of citations. $\alpha = 0.5$ is used to calculate PageRank

Other research articles on the top-15 PageRank list but outside the top-15 citation list are easily recognizable as seminal publications. For example, Onsager's 1944 paper presents the exact solution of the two-dimensional Ising model; both a calculational *tour de force*, as well as a central development in the theory of critical phenomena. The paper by Feynman and Gell-Mann introduced the $V - A$ theory of weak interactions that incorporated parity non-conservation and be-

came the "standard model" of weak interactions. Anderson's paper, "Absence of Diffusion in Certain Random Lattices" gave birth to the field of localization and is cited by the Nobel prize committee for the 1977 Nobel prize in physics.

Table 4.2: Top 100 PageRank articles whose citation rank are more than 10 times of their PageRank.

P.R.[1]	P.R. Value[2]	C.R.[3]	# of Cites[4]	Publication	Title	Author(s)
1	147.98	54	574	PRL, **10**, 531, (1963)	Unitary Symmetry and Leptonic...	N. Cabibbo
8	68.57	94	462	PR, **109** 193, (1958)	Theory of the Fermi Interaction	R. P. Feynman & M. Gell-Mann
13	60.45	1848	114	PR, **34** 1293, (1929)	The Theory of Complex Spectra	J. C. Slater
15	60.22	712	184	PR, **43** 804, (1933)	On the Constitution of Metallic Sodium	E. Wigner and F. Seitz
20	54.56	227	305	PR0, **106**, 364, (1957)	Correlation Energy of an ...	M. Gell-Mann, K. Brueckner
23	51.26	619	197	PRL, **58**, 408, (1987)	Bulk superconductivity at ...	R. J. Cava et al.
28	40.11	310	267	PRL, **58**, 405, (1987)	Evidence for superconductivity ...	C. W. Chu et al.
34	45.70	1192	143	PRL, **10**, 84, (1963)	Photon Correlations	R. J. Glauber

Continued on next page

Table 4.2 – continued from previous page

P.R.[1]	P.R. Value[2]	C.R.[3]	# of Cites[4]	Publication	Title	Author(s)
43	39.85	12825	38	PR0, **35**, 509, (1930)	Cohesion in Monovalent Metals	J. C. Slater
59	36.14	1326	136	PR0, **60**, 252, (1941)	Statistics of the Two-Dimensional...	H. A. Kramers, G. H. Wannier
62	35.10	1425	131	PR0, **81**, 440, (1951)	Interaction Between the ...	C. Zener
66	33.86	2906	89	PRB, **28**, 4227, (1983)	Electronic structure of $BaPb_{1-x}Bi_xO_3$	L. F. Mattheiss, D. R. Hamann
72	33.46	5048	65	PR0, **45**, 794, (1934)	Electronic Energy Bands in ...	J. C. Slater
77	32.86	1674	121	PRL, **10**, 518, (1963)	Classification of Two-Electron ...	J.Cooper, U. Fano, F. Prats
81	31.68	881	165	PR0, **75**, 486, (1949)	The Radiation Theories of ...	F. J. Dyson
85	31.21	1990	109	PR0, **109**, 1860, (1958)	Chirality Invariance and ...	E. Sudarshan & R. Marshak
87	30.93	1872	113	PR0, **46**, 509, (1934)	On the Constitution of Metallic Sodium.	II, E. Wigner, F. Seitz
90	29.91	184066	3	PRB, **22**, 5797, (1980)	Cluster formation in...	H. Rosenstock, C. Marquardt

Continued on next page

Table 4.2 – continued from previous page

P.R.[1]	P.R. Value[2]	C.R.[3]	# of Cites[4]	Publication	Title	Author(s)
98	29.21	1190	143	PR0, **76**,749, (1949)	The Theory of Positrons	R. P. Feynman
99	29.20	3193	84	PR0, **79**, 350, (1950)	Antiferromagnetism..	P. W. Anderson

Table 4.2: 1,PageRank; 2, PageRank value; 3, citation rank; 4, number of citations. $\alpha = 0.5$ is used to calculate PageRank

The striking ability of the PageRank algorithm to identify influential papers can be seen when we consider the top-100 PageRank papers. Table 4.2 shows the subset of publications on the top-100 PageRank in which the citation rank is more than 10 times lower than PageRank; that is, publications with anomalously high Google rank compared to their citation rank. This list contains many easily-recognizable papers for an average physicist. For example, the paper by Gell-Mann and Brueckner, "Correlation Energy of an Electron Gas at High Density" is a seminal publication in many-body theory. The publication by Glauber, "Photon Correlations", was recognized for the 2005 Nobel prize in physics. The Kramers-Wannier article, "Statistics of the Two-Dimensional Ferromagnet. Part I", showed that a phase transition occurs in two dimensions, contradicting the common belief at the time. The article by Dyson, "The Radiation Theories of Tomonaga, Schwinger, and Feynman", unified the leading formalisms for quan-

tum electrodynamics and it is plausible that this publication would have earned Dyson the Nobel prize if it could have been shared among four individuals. One can offer similar rationalizations for most of the remaining articles in this table.

On the other hand, there are some cases PageRank gives inaccurate ranks. One apparent example is for the paper "Cluster formation in two-dimensional random walks: Application to photolysis of silver halides" by Rosenstock and Marquardt (RM). Notice that this article has only 3 citations and is ranked No.184066 by citation, yet it has a high PageRank value of 29.91 and is ranked No.90 by PageRank. This paper is marked with black triangle in Fig.4.4. Why does RM appear among the top-100 Google-ranked publications? We found that one of the three publications citing RM is a famous paper by T. Witten and L. Sander : "Diffusion-Limited Aggregation, a Kinetic Critical Phenomenon" Phys. Rev. Lett. **47**, 1400 (1981) (WS). It has 680 citations and a high PageRank of No. 16. Unexpectedly WS has only 1 reference which is RM. Thus $1 - \alpha = 0.5$ of its popularity is passed to RM in the PageRank algorithm. The appearance of RM on the list of top-100 Google-ranked papers occurs precisely because of the mechanics of the PageRank algorithm in which being one of the few references of a famous paper makes a huge contribution to the PageRank value. But why this intuitively reasonable mechanism results in such rather bizarre rank in this particular case? We discover that it is the incompleteness of the dataset that caused the problem. WS itself has 9 citations instead of 1 as indicated in our dataset. Other references are not included because the dataset only contains citations to other articles published in APS journals. This makes the PageRank value of RM about an order of magnitude higher than what it should be. Similar situation occurs for paper "Cohesion in Monovalent Metals" by J. C. Slater (marked with black diamond in Fig.4.4). It is one of the two "inside" reference of the influential paper "On the Constitution of Metallic Sodium. II" by E. Wigner and F. Seitz. 15 of this

paper's references point to articles outside APS journals which are not included in the dataset.

Besides incompleteness of the dataset, one other aspect to worry about is if the PageRank algorithm gives an "unfair" advantage to older papers. Indeed long random walks on time-directed networks inevitably drift towards older papers. Since the average length of the walk in the PageRank algorithm is given by $1/\alpha$, this effect is especially pronounced for small values of α. To further investigate this question with our selection of $\alpha = 0.5$, we plot the average PageRank value of publications (curve A) as well as the average in-degree, ie, number of citations of a publication (curve B) against its age as of 2003 in Fig.4.5. The plot is binned to 8 data points per decade. In curve B, average number of citations of an article increases for ages less than roughly 10 years and saturates for greater ages. Due to lack of statistics and other artifacts[65], the curve drops sharply after roughly 50 years old and will be neglected in our discussion. Whereas in curve A the average PageRank of a paper increase with age until 50 years old without saturation. It shows that old papers do get advantages in PageRank algorithm, which is reasonable from a historical point of view, because publications need time to gain recognition and sometimes it takes decades. So we suggest that PageRank is a good measure of a paper's influence in the scope of the whole scientific literature in question. On the other hand, if we look for publications to inspire our current research, on average it is unreasonable to value a paper from 50 years ago more than one from recent years. To provide inspiration for current research, one needs to take into account this "aging" effect and rank the publications accordingly. We tackle this exact problem in the next section.

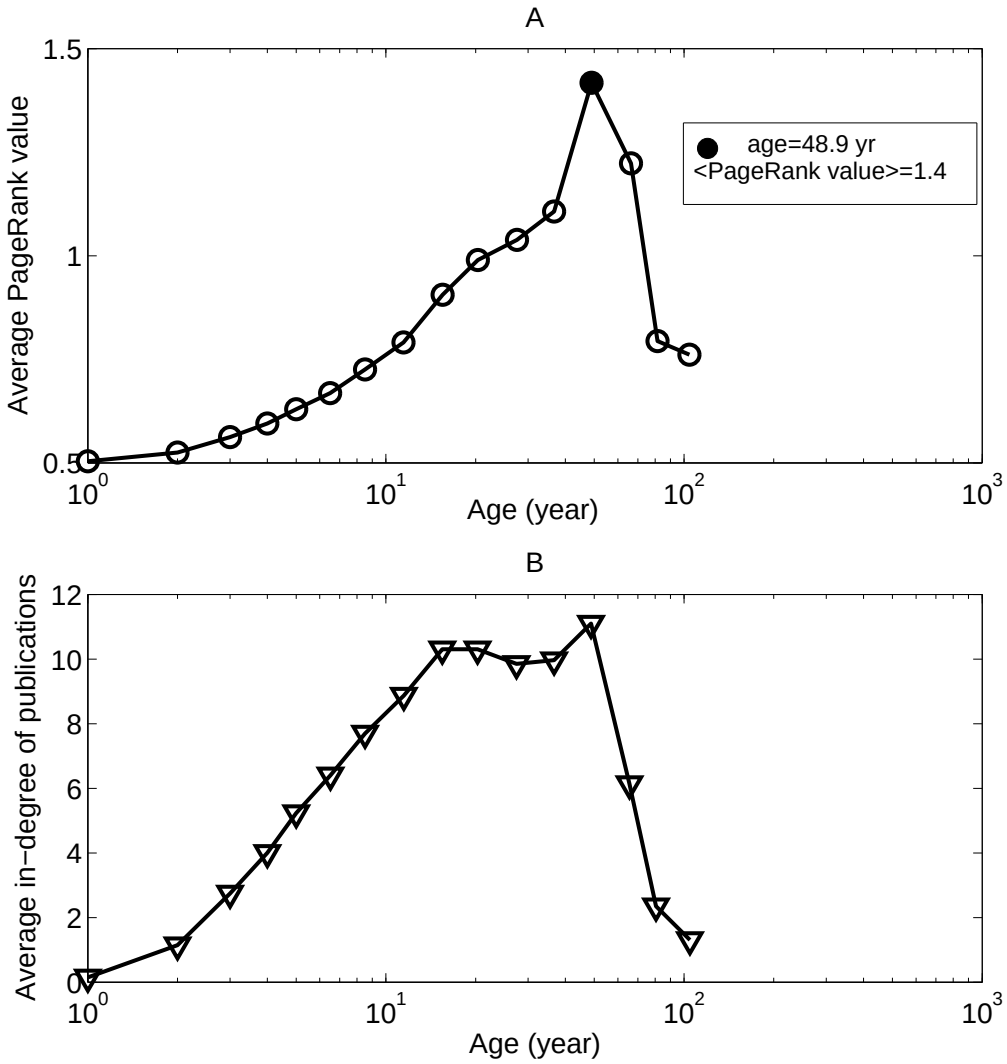

Figure 4.5: Average PageRank and average in-degree(number of citations) of a publication versus its age. Data points are log-binned to 8 points per decade. The figure shows that the average PageRank value and average number of citations of a paper grow with age for papers less than roughly 50 years old and drop sharply for old papers.

4.4 CiteRank: adapted PageRank algorithm for citation networks

Directly applying PageRank algorithm to citation networks allows us to discover a set of highly influential papers that would be undervalued based on just their number of citations. However, there exist significant differences between the World Wide Web and citation networks to make one want to modify the original PageRank algorithm [50]. The most important difference is that, unlike hyper-links, citations cannot be updated after publication. This makes aging effects [64, 65] in citation networks much more pronounced than in the WWW. The other consequence is the inherent time-arrow present in the topology of citation networks. Indeed, a paper could only cite earlier works on the subject. This significantly alters the spectral properties of the adjacency matrix which lie at the heart of the PageRank algorithm. In the extreme case of $\alpha = 0$, the absence of directed loops means that the adjacency matrix could have only zero eigenvalues.

The success of the PageRank algorithm can be attributed, in part, to its ability to capture the behavior of people randomly browsing the network of web pages. Indeed, the PageRank of a given web page can be interpreted as the predicted traffic for that page if every WWW user would follows a random path of (on average) $1/\alpha$ hyper-links starting from a randomly selected web-page. The assumption that a typical web-surfing starts at a randomly selected web-page might be not completely unreasonable for the WWW but needs to be modified for citation networks. As all of us know researchers typically start "surfing" scientific publications from a rather *r*ecent publication which caught their attention on a daily update of a preprint archive, a recent volume of a journal, or, perhaps, which was featured in a news article in the popular media. Thus a more realistic model for the traffic (quantified e.g. by the rate of downloads) along the citation

network should take into account that researchers "surfing" the citation network preferentially start their quests from recent papers and progressively get to older and older papers with every step.

In this work we introduce the CiteRank algorithm, an adaptation of the PageRank algorithm to citation networks. Our algorithm simulates the dynamics of a large number of researchers looking for new information. Every researcher independent of each other is assumed to start his/her search from a *recent* paper or review and then to follow a chain of citations until satisfied or just saturated with information. Explicitly, we define the following two-parameter CiteRank model of such process allowing one to estimate the traffic $T_i(\tau_d, \alpha)$ to a given paper i. A recent paper is selected randomly from the whole population with a probability that is exponentially discounted according to the age of the paper, with a characteristic decay time of τ_d. At every step of the path with probability α the researcher is satisfied/saturated and halts his/her line of inquiry. With probability $(1 - \alpha)$ a random citation to an adjacent paper is followed. The predicted traffic, $T_i(\tau_d, \alpha)$ to a paper is proportional to the rate at which it is visited (downloaded) if a large number of researchers independently follow such a simple-minded process.

While we interpret the output of the CiteRank algorithm as the traffic, its utility ultimately lies in the ability to successfully rank publications. High CiteRank number of a publication denotes its high relevance in the context of currently popular research directions, while the PageRank number is more of a "lifetime achievement award" [66]. It is fruitful to compare the CiteRank of a paper, T_i, with the more traditional method of ranking publications, c_i, the number of citations received. Indeed, the two are highly correlated; a result easily understood on the basis that the larger the number of citations a paper has, the more likely

it will be visited by a researcher via one of the incoming links.

However, the more refined CiteRank algorithm surpasses the conventional ranking, by number of citations, in its characterization of relevancy on two accounts. Like the original PageRank algorithm [50][60], in CiteRank, the popularity of papers is calculated in a self-consistent fashion: The effect of a citation from a more popular paper is greater than that of a less popular one. A citation from a paper that is "highly visible" will contribute more to the visibility of the cited paper. Furthermore, the age of a citing paper is intrinsically accounted for. The effect of a recent citation to a paper is greater than that of an older citation to the same paper. New citations indicate the relevancy of a paper in the context of current lines of research.

An algorithmic description of the aforementioned model can be understood as follows. The transfer matrix associated with the citation network is $W_{ij} = 1/k_j^{out}$ if j cites i and 0 otherwise, where k_j^{out} is the out-degree of the j^{th} paper. Let ρ_i, the probability of initially selecting the i^{th} paper in a citation network, be given by $\rho_i = e^{-age_i/\tau_d}$. The probability that the researcher will encounter a paper by initial selection alone is given by $\vec{\rho}$. Similarly, the probability of encountering the paper after following one link is $(1-\alpha)W \cdot \vec{\rho}$. The CiteRank traffic of the paper is then defined as the probability of encountering it via paths of any length. That is, given an initial distribution of new papers, $\vec{\rho}$, and transfer matrix, W, the CiteRank traffic is given by:

$$\vec{T} = I \cdot \vec{\rho} + (1-\alpha)W \cdot \vec{\rho} + (1-\alpha)^2 W^2 \cdot \vec{\rho} + \cdots \tag{4.1}$$

Practically we calculate the CiteRank traffic on all papers in our dataset by taking successive terms in the above expansion to sufficient convergence ($< 10^{-10}$ of the average value).

In order to assess the viability of this ranking scheme and to select optimal parameters (τ_{dir}, α), we need a quantitative measure of its performance on real citation networks. Besides the Physical Review dataset, we evaluate another dataset, **Hep-th**: An archive snapshot of the "high energy physics theory" archive (http://arxiv.org/archive/hep-th) from April 2003 (preprints ranging from 1992 to 2003). This dataset, containing around 28,000 papers and 350,000 citation links, was downloaded from [70]. We know the actual date of appearance of each of the entries in the preprint archive and thus the age of each node is known with the resolution of 1 day.

The optimal parameters are those that yield the best correlation between a predicted traffic, $T_i(\tau_d, \alpha)$ and the actual traffic (e.g. the rate at which individual papers are downloaded). Unfortunately, the actual traffic data for scientific publications are not readily available for these networks. However, it is reasonable to assume that traffic to a paper is positively correlated with the number of new citations it accrues over a recent time interval, Δk_{in}. For lack of better intuition we first assume a linear relationship between actual traffic and number of recent citations accrued. This corresponds to a simple-minded scenario in which every researcher downloading a paper with a certain small probability could add it to the citation list of the manuscript he/she is currently writing. In order to compare CiteRank with actual citation accrual, we constructed an historical snapshot of both networks used in this study. In both cases, the most recent 10 percent of papers are pruned from the network. This corresponds to the last 4 years (2000-2003) in the Physrev network and last 1 years in the Hep-th network. The CiteRank T_i of the remaining 90 percent of papers is then evaluated and correlated with their actual accrual of new citations Δk_{in} originating at the most recent 10 percent of papers.

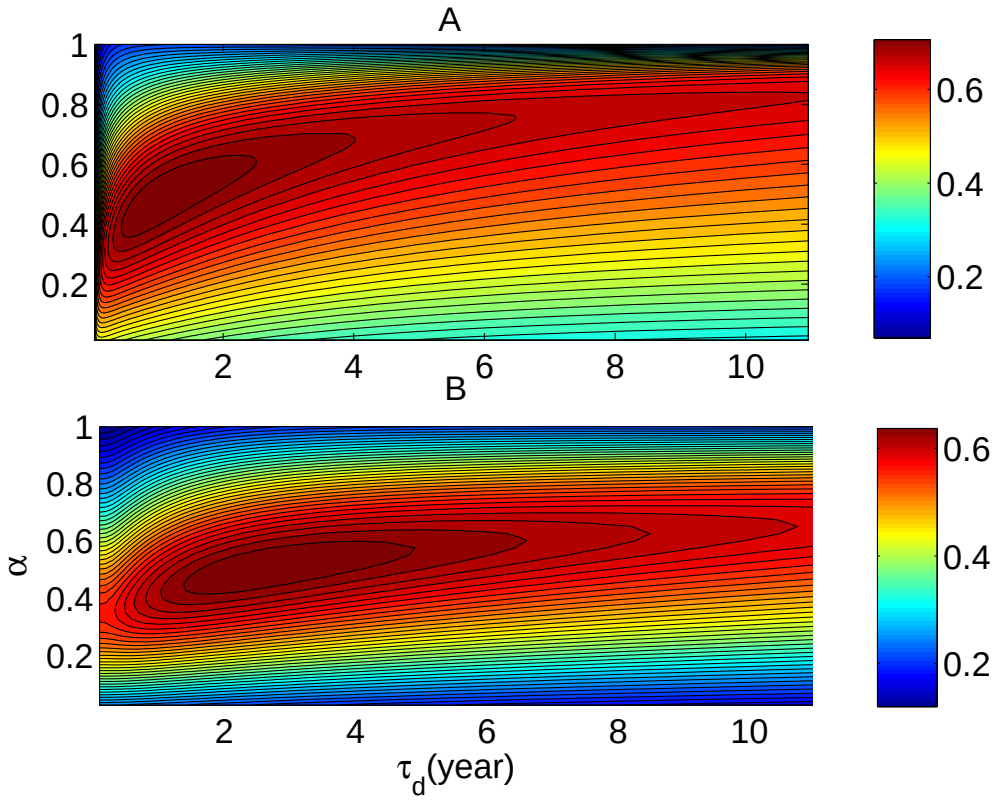

Figure 4.6: The Pearson (linear) correlation coefficient between the number of recent citations accrued (Δk_{in}) and CiteRank traffic (T_i) is calculated over the parameter space of the CiteRank model for the hep-th (A) and physrev (B) network. Both networks exhibit peaks in correlation coefficient in the α-τ_{dir} plane. The highest correlation is achieved for $\alpha = 0.48$, $\tau_{dir} = 1$ year in the hep-th network and $\alpha = 0.50$, $\tau_{dir} = 2.6$ years, in the physrev network.

It is important to note the qualitative as well as quantitative differences between the two citation networks considered. The Physical Review citation network (physrev) is comprised of a large number ($\sim$ 400,000) of peer-reviewed publications acquired over a period close to a hundred years. The high-energy physics archive citation network (hep-th) is comprised completely of a much smaller number ($\sim$ 28000) of electronically submitted publication preprints, with no associated form of peer review. Despite these significant differences in the nature of the networks considered, the general features of their correlation contours are remarkably similar. In both cases, a single sharp peak in correlation is evident for particular values of the parameters. The value of the optimal parameters for both networks are:

hep-th: $\alpha = 0.48$, $\tau_d = 1$ year

physrev: $\alpha = 0.50$, $\tau_d = 2.6$ years

Remarkably, the value of α is nearly the same for two rather different networks considered here and is in agreement with that proposed in [66] on purely empirical grounds. The difference in optimal parameter τ_{dir} for these networks is in agreement with the common-sense expectation of faster response time (and hence faster aging of citations) in preprint archives compared to peer-reviewed publications.

Another feature of Fig.4.4 is that in both networks large values of the correlation coefficient are concentrated along a diagonally-positioned ridge. For a broad range of τ_d, in either network, there appears to be linear relationship between the two model parameters, as indicated by the slope of maximal correlation. In other words, the best choice of α for a given τ_d seems to rise linearly with τ_d, a behavior that will be revisited later.

While the correlation contour plots shown in Fig.4.4 are a promising indication that the CiteRank model of traffic with optimized parameters provides a good zero-order approximation to the actual traffic along citation network, they are to some extent predicated on the assumption of a linear relationship between actual traffic and Δk_{in}. One might readily ask how this model fares in the absence of such an assumption. While the assumption of a *linear* relationship may be unreasonable, a positive, monotonic relationship between these quantities is certainly expected. There is a statistical correlation method precisely adapted for such a situation, namely, the Spearman rank correlation. Under this relaxed correlation measure, only the rank of T_i are correlated with the rank of Δk_{in}. Numerical changes in T_i that do not lead to reordering have no effect on the value of the rank correlation coefficient. Another rationale for using rank correlations is that our ultimate goal is ranking publications instead of modelling the traffic. Thus we are currently not interested in individual T_i's but only in their relative values. Spearman correlation contour plots are constructed for both networks and shown in Fig.4.4. The optimal values for both networks are:

hep-th: $\alpha = 0.31$, $\tau_{dir} = 1.6$ year

physrev: $\alpha = 0.55$, $\tau_{dir} = 8$ years

These results roughly confirm the prediction of $\alpha \sim 0.5$ from Fig.4.4, however there is a more appreciable discrepancy in τ_{dir} between linear and rank correlation for both networks.

In both panels of Fig.4.4, over a broad range of parameters the optimal value of $\alpha(\tau_d)$ for a given value of τ_d is positively correlated with τ_d. This is an indication that these two parameters are entangled. In fact, this is to be expected as it is some admixture of the two parameters which leads to the exposure of

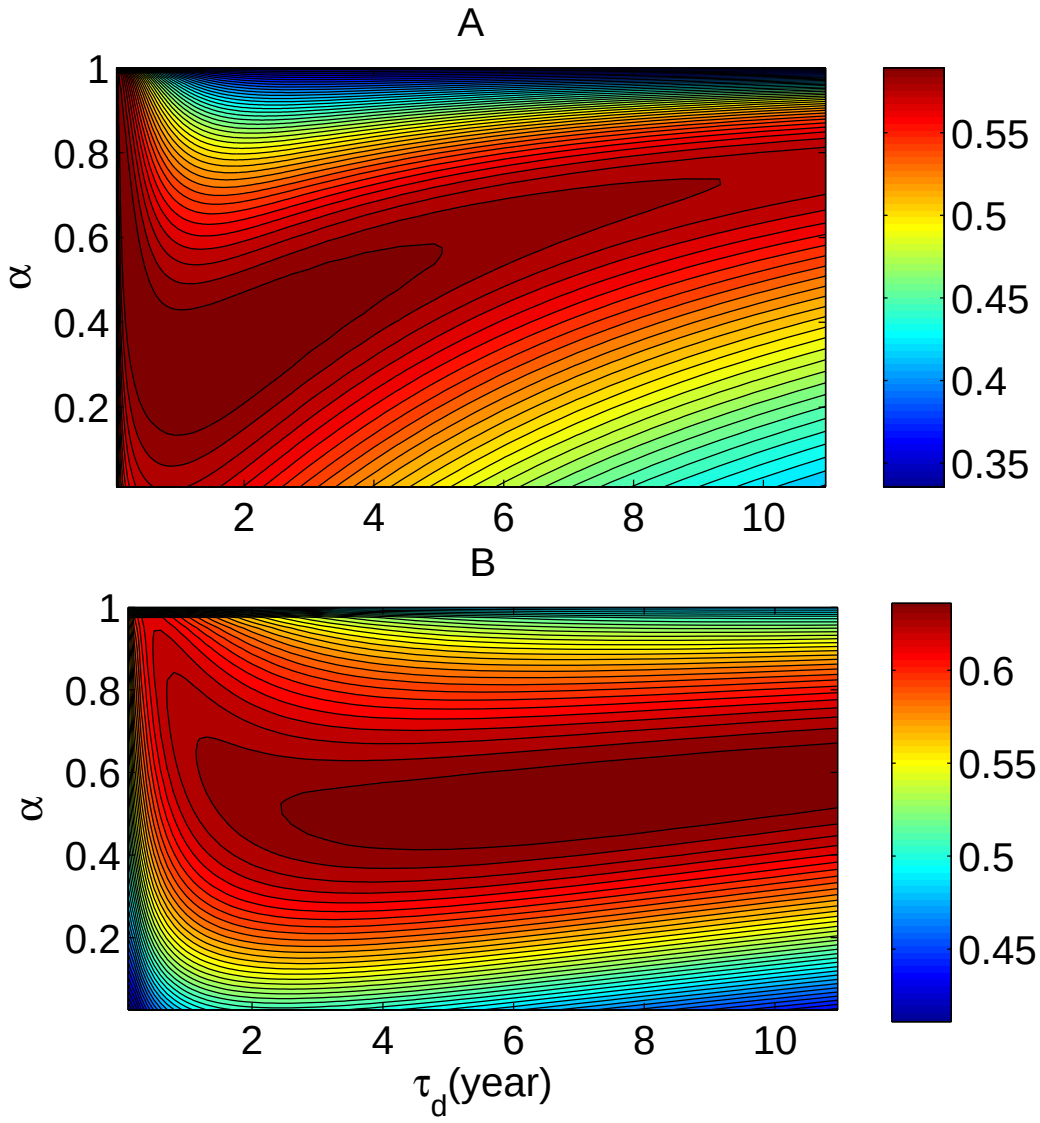

Figure 4.7: The Spearman rank correlation coefficient between recent citations accrued (Δk_{in}) and CiteRank traffic (T_i) for hep-th network (A) and physrev (B). Both networks exhibit similar behavior. There are more extended regions of good correlation relative to the linear correlation contours of fig. 1. This broadening is expected as a consequence of the more relaxed correlation measure. The highest rank correlation occurs for $\alpha = 0.31$, $\tau_d = 1.6$ years (575 days), in the hep-th network and $\alpha = 0.55$, $\tau_d = 8$ years, in the physrev network. This seems to confirm the prediction of $\alpha \sim 0.5$ from Fig.4.4, however there is a more appreciable discrepancy in τ_d between linear and rank correlation for both networks.

a given paper to the researcher. An intuitive picture of this entanglement can be understood in terms of the penetration depth, which is a measure of how far back in time a random surfer following rules of the CiteRank algorithm is likely to get. The penetration depth is affected by both τ_d - the average age of the initial paper at which he/she started following the chain of citations, and $1/\alpha$ - the mean number of steps on this chain of citations. For small τ_d and large α the penetration depth is small implying that only very recent papers receive traffic. On the other hand, for large τ_d and small α the penetration depth is very large meaning that most of the traffic would be redirected towards older papers.

To better understand how α and τ_{dir} influence the age distribution of CiteRank traffic, we performed the following quantitative analysis. Let $T_{tot}(t)$ denote the total CiteRank model traffic to papers written exactly t years ago. As described by Eq. 4.1, two distinct processes contribute to $T_{tot}(t)$. The first is the "direct" traffic $T_d(t)$ due to the initial selection of papers in this age group, which is proportional to $\exp(-t/\tau_{dir})$ [61]. The second is the "indirect" traffic $T_{ind}(t)$ arriving via one of the incoming citation links. The latter is given by $T_{ind}(t) = (1-\alpha)\int_t^\infty T_{tot}(t')P_c(t',t)dt'$, where $P_c(t',t)$ is the fraction of citations originating from papers of age t' that cite papers of age t. It should be noted that $P_c(t',t)$ is an *empirical* distribution and, as such, is a *measured* property of the citation network under consideration. According to [65] and our own findings, $P_c(t',t)$ is reasonably well approximated by the exponential form $\frac{1}{\tau_c}\exp(-(t'-t)/\tau_c)$. Taking the Fourier transform of the equation $T_{tot}(t) = T_d(t) + T_{ind}(t)$, we have

$$T_{tot}(\omega) = T_d(\omega) + (1-\alpha)T_{tot}(\omega)P_c(\omega). \tag{4.2}$$

This equation is similar, in spirit, to the well-known random phase approximation [67]. Solving Eq. 4.2 for $T_{tot}(\omega)$ and taking the inverse Fourier transform, yields

$$T_{tot}(t) \sim (\tau_c - \tau_{dir})\exp(-t/\tau_{dir}) + (1-\alpha)\tau_{dir}\exp(-\alpha t/\tau_c). \tag{4.3}$$

Thus, the traffic arriving at the subset of papers of age t is given by the superposition of two exponential functions.

Having an approximate analytical expression for $T_{tot}(t)$, we are now in a position to better understand what determines the optimal values of α and τ_{dir}. Open circles in Fig.4.4 show the age distribution of the number of recently acquired citations, Δk_{in}, for papers in the physrev dataset. The approximate CiteRank traffic, given by Eq. 4.3, is also displayed. It is calculated using the empirically determined value $\tau_c = 8$ years, optimal $\tau_{dir} = 2.6$ years and three values of $\alpha = 0.2, 0.5$ and 0.9. As one would expect, the profile of $\langle \Delta k_{in} \rangle$ vs t best agrees with the CiteRank plot for the optimal value $\alpha = 0.5$ [69]. Fig.4.4 also provides some clues to the positive correlation between near-optimal choices of α and τ_{dir}, visible as diagonal "ridges" in Fig.4.4A and B. Indeed, if the value of α is chosen to be large, the contribution from the second term is diminished; the use of a larger value of τ_{dir} could partially compensate for the loss of CiteRank traffic to older papers, and would thus be in reasonably good agreement with the Δk_{in} data.

Another encouraging observation is that, like Eq. 4.3, the age distribution of recently acquired citations shown in Fig.4.4 has two regimes characterized by two different decay constants of about 5 and 16 years, with a crossover point around $t = 15$ years. Our interpretation of this fact is that papers are found and cited via two distinct mechanisms: researchers can either find a paper directly or by following citation links from earlier papers. For each of these mechanisms, the probability that a given paper is found decays with its age but the characteristic decay time for the direct discovery is shorter. While very recent papers, especially the ones altogether lacking citations, are for the most part discovered directly, older papers are mostly discovered by following citation links.

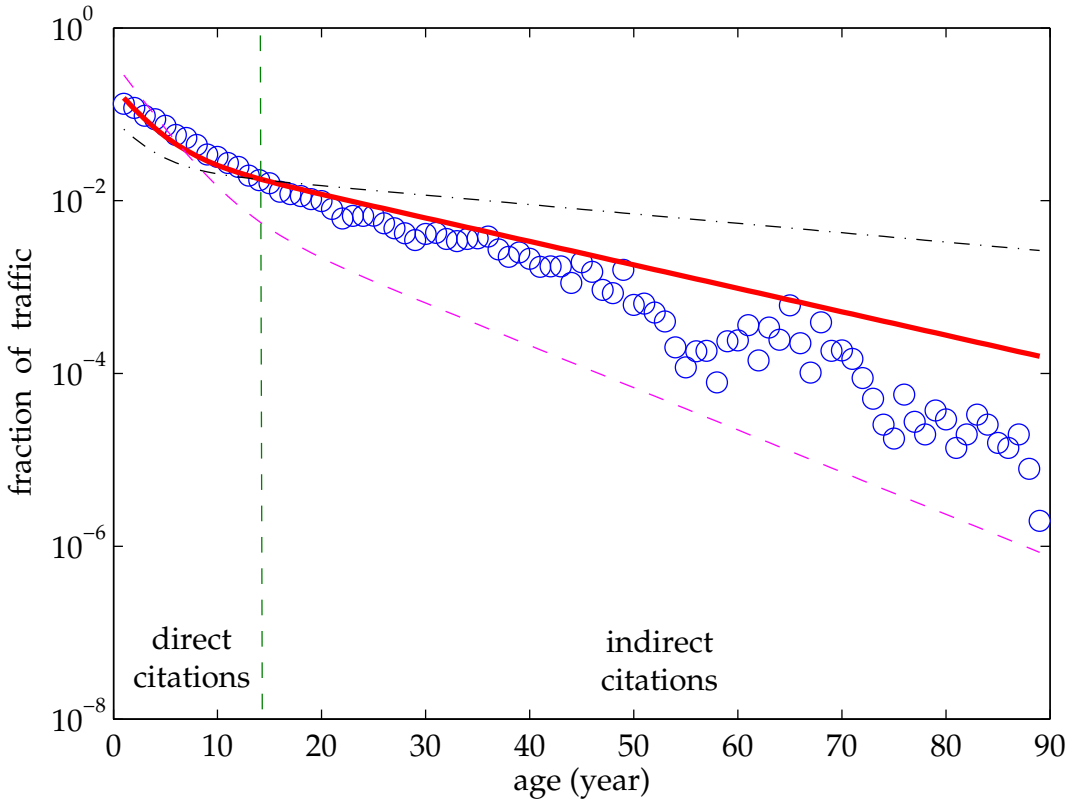

Figure 4.8: The age distribution of newly accrued citations Δk_{in} (blue) for the physrev network. Theoretical predictions [4.3] for the CiteRank traffic are calculated for the optimal $\tau_{dir} = 2.6$ and three values of $\alpha = 0.2$ (dot-dashed line), 0.5 (thick solid line), and 0.9 (dashed line). In agreement with Fig.4.4, the optimal value, $\alpha = 0.5$, provides the best agreement with Δk_{in}. All curves are normalized so that the sum of all data points is equal to 1.

The optimal values of α in the two very different citation networks considered are remarkably close to each other. In both cases it appears that, on average, the length of chains of citations followed by a typical researcher is close to $1/\alpha \simeq 2$. Since this chain includes the original starting point (the paper from which the researcher started his quest), the length of around 2 means that the average cited paper is just one link away from the starting point. This raises the disconcerting possibility that many citations may be copied directly from the initially discovered reference. Such citation copying was recently proven to be a very common scenario [68].

CHAPTER 5

Conclusions and directions for future work

We employed a "mean-field" approximation to investigat how the WWW community structure affects the Google Rank of web-pages belonging to a given community. We have shown that the only free parameter α in PageRank algorithm controls to what degree a community with given number of links entering (E_{wc}) and leaving (E_{cw}) the community isolated from the rest of the world. When a community is completely isolated, changing E_{wc} and E_{cw} will not affect its average Google rank; on the contrary, if a community is "coupled" with the world, its average PageRank value is determined by the exact balance of E_{wc} and E_{cw}, therefor webmasters will be able to manipulate the rank of their website by altering the number of hyper links between the community and the world. Since such ability of webmasters to artificially boost their ranking is undesirable for search engines, as large as possible α should be used to make most communities isolated from the world. On the other hand, for very large α the algorithm fails to take into account the topological property of the network. Our empirical study has shown that Google's choice of 0.15 does strike an optimal balance between these two opposing demands on α.

We explored the application of PageRank algorithm on citation networks. With an empirically derived choice of $\alpha = 0.5$, we calculated the PageRank for a citation network consisting of all the publications of Physical Review journals up to 2003. Comparing to ranking by number of citations, PageRank provides a

more sophisticated measure of a paper's historical influence in the scientific literature in question. Some undervalued moderately cited papaer are "discovered" by the algorithm, which turn out to be seminal papers in the literature. To provide an importance measure with respect to current research, we proposed CiteRank algorithm based on PageRank. CiteRank models the traffic (downloads) of existing publications while giving more weight to the recent ones. We found the optimal parameters through comparing model provided traffic with empirically derived traffic measure and confirm by theoretical analysis.

Being rather powerful tools to rank scientific literature, both algorithms suffer from incomplete dataset issue. It would be interesting and practically useful to introduce some correction, for instance, to include the immediate citations of Physical Review publications. This would ensure the out-degree to outside papers, which could be majority for some papers, won't be discarded. Another possible direction could be to model the growth of citation networks.

References

[1] Réka Albert and Albert-László Barabási, Statistical mechanics of complex networks, *Rev. Mod. Phys.* **74**,47-97 (2002).

[2] S. Strogatz, Exploring complex networks. *Nature* **410**:268–276 (2001).

[3] S. N. Dorogovtsev and J. F. F. Mendes, Evolution of networks, *Adv. Phys.*, **51**, 1079 - 1187 (2002).

[4] M. E. J. Newman, The structure and function of complex networks, *SIAM Review* **45**, 167-256 (2003)

[5] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.U. Hwang, Complex networks:Structure and dynamics, *Physics Reports* **424**, 175-308 (2006).

[6] B.Bollobás, *Random Graph*, Academic Press, Longdon, (1985)

[7] B.Bollobás, *Modern Random Graph*, Springer, New York, (1998).

[8] S. Janson, T. Luczak, and A. Rucinski, *Random Graphs* Willey, New Yor, (2000).

[9] S. Wasserman, K. Faust, *Social Network Analysis*, Cambridge University Press, Cambridge, (1994).

[10] D.J. Watts, S.H. Strogatz, Collective dynamics of s̀mall-worldńetworks, *Nature*, **393**,440-442 (1998).

[11] Girvan, M. and Newman, M. E. J., Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA 99, 8271C8276 (2002).

[12] Cohen, J. E., Briand, F., and Newman, C. M., *Community food webs: Data and theory*, Springer, New York (1990).

[13] S.L. Pimm, *The Balance of Nature*, Chicago Press, Chicago (1991).

[14] Baird, D. and Ulanowicz, R. E., The seasonal dynamics of the Chesapeake Bay ecosystem, *Ecological Monographs* **59**,329C364 (1989).

[15] Goldwasser, L. and Roughgarden, J., Construction and analysis of a large Caribbean food web, *Ecology* **74**, 1216C1233 (1993).

[16] Martinez, N. D., Artifacts or attributes? Effects of resolution on the Little Rock Lake food web, *Ecological Monographs* **61**, 367C392 (1991).

[17] Montoya, J. M. and Solé, R. V., Small world patterns in food webs, *J. Theor. Bio.* **214**, 405C412 (2002).

[18] Camacho, J., Guimerá, R., and Amaral, L. A. N., Robust patterns in food web structure, *Phys. Rev. Lett.* **88**, 228102 (2002).

[19] Dunne, J. A., Williams, R. J., and Martinez, N. D., Network structure and biodiversity loss in food webs: Robustness increases with connectance, *Ecology Letters* **5**, 558C567 (2002).

[20] Farkas, I. J., Jeong, H., Vicsek, T., Barabási, A.-L., and Oltvai, Z. N., The topology of the transcription regulatory network in the yeast, Saccharomyces cerevisiae, *Physica A* **381**, 601C612 (2003).

[21] Jeong, H., Mason, S., Barabasi, A.-L., and Oltvai, Z. N., Lethality and centrality in protein networks, *Nature* **411**, 41C42 (2001).

[22] Guelzim, N., Bottani, S., Bourgine, P., and Kepes, F., Topological and causal structure of the yeast transcriptional regulatory network, *Nature Genetics* **31**, 60C63 (2002).

[23] Shen-Orr, S., Milo, R., Mangan, S., and Alon, U., Network motifs in the transcriptional regulation network of Escherichia coli, Nature Genetics 31, 64C68 (2002).

[24] Maslov, S.; Sneppen, K. Specificity and stability in topology of protein networks. *Science* **296**, 910 –913 (2002).

[25] Kauffman, S. A., *The Origins of Order*, Oxford University Press, Oxford (1993).

[26] S. Wasserman, K. Faust, *Social Networks Analysis*, Cambridge University Press, Cambridge, 1994.

[27] Scott, J., *Social Network Analysis: A Handbook*, Sage Publications, London, 2nd ed. (2000).

[28] Newman, M. E. J., Strogatz, S. H., and Watts, D. J., Random graphs with arbitrary degree distributions and their applications, *Phys. Rev. E* **64**, 026118 (2001).

[29] AMARAL, L. A. N., A. SCALA, M. BARTHELEMY, AND H. E. STANLEY, Classes of Small-World Networks, *Proc. Natl. Acad. Sci.* USA, **97**, 11149-11152, (2000).

[30] Barabási, A.-L., Jeong, H., Ravasz, E., Néda, Z., Schuberts, A., and Vicsek, T., Evolution of the social network of scientific collaborations, *Physica A* **311**, 590C 614 (2002).

[31] Batagelj, V. and Mrvar, A., Some analyses of Erdős collaboration graph, Social Networks 22, 173-186 (2000).

[32] Newman, M. E. J., Scientific collaboration networks: I. Network construction and fundamental results, *Phys. Rev. E* **64**, 016131 (2001).

[33] Newman, M. E. J., Scientific collaboration networks: II. Shortest paths, weighted networks, and centrality, *Phys. Rev. E* **64**, 016132 (2001).

[34] Newman, M. E. J., The structure of scientific collaboration networks, *Proc. Natl. Acad. Sci. USA* **98**, 404C409 (2001).

[35] A. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T.Vicsek, Evolution of the social network of scientific collaborations. *Physica A*, **311**(3–4):590–614, (2002).

[36] Faloutsos, M., Faloutsos, P., Faloutsos, C., On power-law relationships of the internet topology, *Proc. ACM SIGCOMM.* (1999).

[37] Govindan, R. and H. Tangmunarunkit, Heuristics for internet map discovery, *Proc. IEEE Infocom*, (2000).

[38] Yook, S.H., Jeong, H., Barabasi, A.L., Modeling the Internet's large-scale topology, *Proc. of the National Academy of Sciences* **99** 13382–13386, (2002).

[39] A. Vázquez, R. Pastor-Satorras, and A. Vespignani. Large-scale topological and dynamical properties of the internet. *Physical Review E*, **65**(6):066130, 2002.

[40] K.A. Eriksen, I. Simonsen, S. Maslov, K. Sneppen, Modularity and Extreme Edges of the Internet, *Phys. Rev. Lett.* **90** (2003) 148701.

[41] S. Maslov, K. Sneppen, A. Zaliznyak, Detection of Topological Patterns in Complex Networks: Correlation Profile of the Internet, *Physica A*, **333**, 529-540 (2004).

[42] Huberman, B. A., *The Laws of the Web*, MIT Press, Cambridge, MA (2001).

[43] Lawrence, S. and C. L. Giles, Accessibility of information on the web, Nature 400, 107 (1999).

[44] A. Albert, H. Jeong, and A.-L. Barabasi, Diameter of the World Wide Web, *Nature*, **401**, 130, (1999).

[45] Kumar, R., P. Raghavan, S. Rajalopagan, and A. Tomkins, The Web as a Graph, *Proceedings of the 9th ACM Symposium on Principles of Database* (2000).

[46] Broder, A., R. Kumar, F. Maghoul, P. Raghavan, S. Rajalopagan, R. Stata, A. Tomkins and J. Wiener, Graph structure in the Web telecommunications networking, *Comput. Netw.* **33**, 309 (2000).

[47] Kumar, R., Raghavan, P., Rajagopalan, S., Sivakumar, D., Tomkins, A. S., and Upfal, E., Stochastic models for the Web graph, in *Proceedings of the 42st Annual IEEE Symposium on the Foundations of Computer Science*, pp. 57C65, Institute of Electrical and Electronics Engineers, New York (2000).

[48] R. Kumar, P. Raghavan, S. Rajagopalan, and A.Tomkins, Trawling the web for emerging cyber-comnities, *Computer Networks* **31**, (11-16):1481– 1493 (1999).

[49] Kasper Astrup Eriksen, Ingve Simonsen, Sergei Maslov, Kim Sneppen, Modularity and Extreme Edges of the Internet, *Phys. Rev. Lett.* **90**, 148701 (2003).

[50] S. Brin and L. Page, The anatomy of a large-scale hypertextual Web search engine, *Computer Networks and ISDN Systems*, **30**, 107 (1998).

[51] L. Page, S. Brin, R. Motwani and T. Winograd, The PageRank Citation Ranking: Bringing Order to the Web, *Stanford Digital Library Technologies Project* (1998).

[52] S. Bilke and C. Peterson, Topological properties of citation and metabolic networks, *Physical Review E* **64**, 036106 (2001).

[53] Such a seed network can be created e.g. using a stub reconnection procedure described in M. E. J. Newman, S. H. Strogatz, and D. J. Watts, *Phys. Rev. E* **64**, 026118 (2001).

[54] S. Maslov and K. Sneppen, Specificity and stability in topology of protein networks, *Science*, **296**, 910 (2002).

[55] J. Park and M. E. J. Newman, Origin of degree correlations in the Internet and other networks, *Phys. Rev. E* **68**, 026112 (2003).

[56] 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.

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

[58] Thelwall, M, A Free Database of University Web Links: Data Collection Issues, *Cybermetrics*, Vol **6/7**, Issue 1. Paper 2 (2002-3).

[59] The APS journals include Phys. Rev. Series I (1893-1912), Phys. Rev. Series II (1913- 1969), and Phys. Rev. Series III (1970-present). This latter series includes the five topical sections: Phys. Rev. A, B, C, D, and E (the latter from 1990-present). Also included are Phys. Rev. Lett., Rev. Mod. Phys., and Phys. Rev. Special Topics, Accelerators and Beams (1998-present).

[60] J. Bollen, M. A. Rodriguez, and H. Van de Sompel cs.DL/0601030

[61] The actual fraction of "followed citations" is 48% for the entire dataset and 55% for papers published during the last 4 years.

[62] S. Fortunato, M. Boguna, A. Flammini, and F. Menczer, How to make the top ten: Approximating PageRank from in-degree, cs.IR/0511016.

[63] S. Fortunato, A. Flammini, and F. Menczer, Scale-free network growth by ranking, cond-mat/0602081.

[64] D. J. De Solla Price, Networks of Scientific Papers, *Science*, **149**, 510 (1965).

[65] S. Redner, Citation Statistics from 110 Years of Physical Review, *Physics Today*, **58**, 49 (2005)

[66] P. Chen, H. Xie, S. Maslov, S. Redner, Finding Scientific Gems with Google, physics/0604130

[67] J. Jensen, A. Mackintosh, *Rare Earth Magnetism: Structures and Excitations*, Clarendon Press, Oxford, 1991.

[68] M.V. Simkin, V.P. Roychowdhury, Read before you cite!, *Complex Syst.*, **14**, 269 (2003) (cond-mat/0212043).

[69] The apparent disagreement in the tail involves profound dips due to the World War II and I [65], which of course cannot be explained by any theoretical model.

[70] This hep-th dataset was used in the KDD Cup 2003, http://www.cs.cornell.edu/projects/kddcup/.