Yana Volkovich

Stochastic Analysis of Web Page Ranking

Yana Volkovich

Stochastic Analysis of Web Page Ranking

University of Twente, The Netherlands
CTIT PhD thesis series number 09-139
Beta dissertation series D118
ISBN 978-90-365-2823-8
ISSN 1381-3617

Yana Volkovich

Stochastic Analysis of Web Page Ranking

Stochastic Analysis of Web Page Ranking

Yana Volkovich

University of Twente, The Netherlands
CTIT PhD thesis series number 09-139
Beta dissertation series D118
ISBN 978-90-365-2823-8
ISSN 1381-3617

Stochastic Analysis of Web Page Ranking

by

Yana Volkovich

Composition of the graduation committee:

Chairman and secretary:	
prof.dr.ir. A.J. Mouthaan	University of Twente
Promoter:	
prof.dr. R.J. Boucherie	University of Twente
Assistant promoter:	
dr. N. Litvak	University of Twente
Members:	
prof.dr. W. Albers	University of Twente
dr. K.S. McCurley	Google Inc.
prof.dr. R.W. van der Hofstad	Eindhoven University of Technology
prof.dr. M.J. Uetz	University of Twente
prof.dr. A.P. Zwart	VU University Amsterdam

UT / EEMCS / AM / SOR
P.O. Box 217, 7500 AE Enschede
The Netherlands

CTIT PhD Thesis Series 09-139
Centre for Telematics and Information Technology

Beta Dissertation Series D118
BETA, Research School for Operations Management and Logistics

Part of the research in this thesis has been funded by the Dutch BSIK/BRICKS project

This thesis was edited with WinEdt and typeset with LaTeX.
Printed by Wöhrmann Print Service, Zutphen, The Netherlands.

ISSN 1381-3617
ISBN 978-90-365-2823-8
`http://dx.doi.org/10.3990/1.9789036528238`

STOCHASTIC ANALYSIS OF WEB PAGE RANKING

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van de rector magnificus,
prof.dr. H. Brinksma,
volgens besluit van het College voor Promoties
in het openbaar te verdedigen
op vrijdag 24 april 2009 om 13.15 uur

door

Yana Volkovich

geboren op 17 januari 1982
te Kohtla-Järve, Estland

Dit proefschrift is goedgekeurd door

prof.dr. R.J. Boucherie	(promotor)
dr. N. Litvak	(assistent-promotor)

to the memory of my mother

ACKNOWLEDGMENTS

This thesis is a result of my research carried out over the last four years. Looking back I see how much I have changed and how much I have learned. Hereby I would like to thank all people who were with me over these years.

First and foremost, I would like to thank Nelly Litvak who not only is a great advisor but also a great friend. Nelly, thank you for all your belief in me, for all our discussions, for all your help, and for all the fun we had outside of the work.

Second, I would like to thank my Stochastic Operation Research group that was always nice place to be. I am very thankful to Richard Boucherie for giving me the opportunity to join the group, and for continuous encouragement of my research. I am also grateful to Werner Scheinhardt, especially for taking care and advising me during the first year. Next, I would like to thank Roland de Haan for his good sense of humor, and Denis Miretskiy for being a good teaching partner of the Stochastic Processes course. Finally, I also like to thank all the current and former members of my group. Ahmad, Bas, Jan-Kees, Jasper, Judith, Maartje, Maurits, Nikky, Peter, Tom, Sing-Kong, and Thyra, – thank you!

Next, I also take a great pleasure to thank Ricardo Baeza-Yates for giving me the opportunity to escape from the November rains. I spend a wonderful time at Yahoo! Research Barcelona, and I truly enjoyed our collaboration with Aris Gionis and Debora Donato.

I am further grateful to all members of my committee. I would like to thank Bert Zwart for his ideas, inspiring discussions and joint work. I am also thankful to Wim Albers, Kevin McCurley, Remco van der Hofstad, and Marc Uetz for their thorough examination of the manuscript and for their comments.

All my successes in the Dutch language could not be done without Hester. I would like to thank her for all our amazing chats over all possible topics.

Further, I am grateful to all my friends in Enschede, Eindhoven and the Hague, who filled this time with many memorable moments. Undoubtedly, all my days in the Netherlands would be gray without continuous support from my friends in Saint

Petersburg. I would specially thank Misha, Nastia, Olga, and Yulia for their help through the hardest days in my life.

Last, I would like to thank my family. This work is dedicated to the memory of my mother for all of her support, belief and love.
Yana Volkovich.
Enschede, March 2009

CONTENTS

CHAPTER 1

INTRODUCTION

Twenty years ago, Tim Berners-Lee proposed to build a web of hypertextual pages, which today is known as the World Wide Web. The Web is an important part of our lives. Hence, understanding properties of the Web is one of the most essential research needs. In this thesis we focus on the stochastic analysis of different characteristics of the Web. In particular, we are interested in the Web properties that affect the Web page ranking, that is a measure of popularity and importance of a page in the Web. One of the most well-known and widely-used algorithms for the Web ranking is the `Google`'s *PageRank*. We focus on the asymptotic behavior of the PageRank distribution in various information networks, such as the Web and the Wikipedia. For the majority of such self-organized networks it was observed that the PageRank distribution follows a power law. One of the goals of this thesis is to define how various network characteristics influence the distribution of the PageRank. To this end, we introduce a stochastic equation that corresponds to the original definition of the PageRank, and apply the theory of regular variation to study this equation.

Further results of our work is the application of extremal dependencies and angular measure to the problem of measuring correlation between different characteristics of the power law graphs, and to the problem of rank aggregation. The angular measure has been designed for measuring correlations between power law distributed random variables, but it has never been applied to large power law graphs.

We start this chapter with a brief introduction into the Web search process in Section 1.1, and with definitions of the main Web ranking algorithms in Section 1.2. Then, in Section 1.3 we discuss the determinative properties of the Web structure. In particularly, we focus on power law distributions in Section 1.3.2. In Section 1.3.3 we provide an overview of graph models that possess various properties of the Web.

Section 1.4 briefly describes main ideas and techniques that we use in this thesis. In Section 1.4.1 we define regularly varying random variables which are natural

mathematical formalization of power laws. In Section 1.4.2 we briefly explain the idea of modeling the PageRank distribution as a solution of a stochastic equation. Moreover, we propose a generalized version of the stochastic equation of the PageRank in the way that it can be used in other real-life applications. For details on this kind of stochastic equations we refer to Section 1.4.3. Further, in Section 1.4.4 we give an introduction on applications of angular measure for evaluating dependencies between various characteristics of the Web graph, and for rank aggregation problems.

Finally, in Section 1.5 we present the outline of the thesis.

1.1 Web search

A significant role in the Web evolution was played by Web search engines. At the beginning, it was enough to have a complete list of all Web servers. However, with the increase of the number of pages this central list became not only incomplete, but too large to be of any practical use, and then the first search engines appeared. These engines were primitive, and hence they had poor performance. The returned search results were just lists of content relevant pages, whereas quality of these pages still remained to be a subject for the user to determine. Thus, to access relevant Web pages, users referred to colleagues, friends, or special web guide books.

The insufficiency of the search results was caused by the fact that the first search engines were based on the already existing techniques that were developed for document collections, in which all documents were assumed to have high quality, and to be homogeneous. This assumption holds, for example, for collections of papers or books, where the number of citations is a good measure of popularity. However, the homogeneity assumption is definitely violated in a representative collection of Web pages, where the best text match does not imply the highest relevance, and the large number of incoming links can often indicate a spam. To resolve the problem, Brin and Page with PageRank algorithm [23, 92] and Kleinberg with *HITS* algorithm [63] proposed to use link analysis for measuring importance of pages in Web search. The idea turns out to be very successful, and both of the algorithms are widely used today not only in search engines (`Google` or `Ask.com`), but in different ranking related problems. In Section 1.2 we provide formal definitions of PageRank and HITS. Now, we briefly describe how search engines work in order to define the place of the ranking in the Web search process.

Figure 1.1 shows a schematic diagram of the Web search process. At the beginning, a search engine must collect information about available Web pages. Using specially designed programs called *crawlers*, the search engines collect information about the content of the Web pages, and links between them. The crawlers need to discover new pages, and to update already visited Web pages. Here we do not focus on the design of Web crawlers. In general, it is a complicated problem, for a survey on the subject we refer to Castillo [27]. After being crawled, every page is classified. If a page is 'good' according to some rules (e.g., non-duplicate, or non-spam), then it

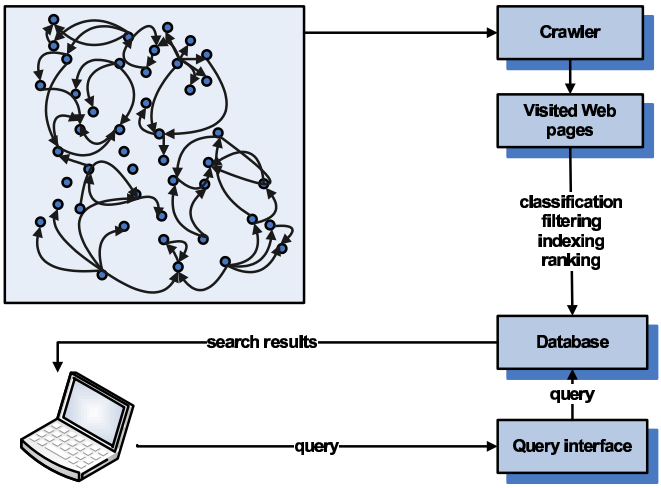


Figure 1.1: Schematic representation of Web search engine

is indexed, and stored in a database together with its rank. This rank is assigned to every page, and it is computed according to position of a page in the Web graph. The rank is pre-computed and usually query-independent. PageRank is one of examples of such ranks.

We briefly explain theoretical foundation [65, 97] for how to incorporate a probability distribution as suggested by PageRank into the overall scoring of a page for a given query. We are interested in the probability $\mathbb{P}(d|q)$ that a document d is relevant for a given query q. Using Bayes' rule we can rewrite this probability as

$$\mathbb{P}(d|q) = \frac{\mathbb{P}(d)\mathbb{P}(q|d)}{\mathbb{P}(q)}.$$

For page ranking purposes, $\mathbb{P}(q)$ is irrelevant since it does not depend on the document. The term $\mathbb{P}(q|d)$ is one of the main interests of the information retrieval community. Various heuristics are used to estimate the relevance of a query to a document. The $\mathbb{P}(d)$ term has a natural interpretation from PageRank (or similar models) as the likelihood that a document would be relevant independent of the query. One of the points of this thesis is that it provides better understanding on what the $\mathbb{P}(d)$ term might look like, and how it is distributed under the PageRank model. We note that this speaks in terms of the actual value of the PageRank and not the actual position in the ordering of documents, and therefore the value of the PageRank is important.

When a user types a query, first, the query gets translated into the search system language query (usually number code) through query interface. Second, using the modified query, search engine searches for relevant pages in the database. Returned results are listed on the screen in order of their importance. To achieve the

best performance, search engines define the importance of the page based on secret combination of rankings according to different criteria, such as content-relevance, browsing histories, search engine logs, users personal preferences, e.g., geographical locations, and positions of the pages in the Web graph.

Thus, link-related rank of a page plays an important role in the final listing of the Web search results. In the next section we define two most well-known ranking techniques that are based on the link analysis.

1.2 Web page ranking

The PageRank [23, 92], HITS [63], SALSA [70] and a number of other link-based ranking algorithms have been successfully used for evaluating the importance of a page in the Web graph. In this work we restrict our attention to the PageRank, the most popular ranking algorithm, and HITS. For surveys on other ranking schemes we refer to Langville and Meyer [69], and Berkhin [12]. Besides their primary application in the Web search, the ranking algorithms help to solve other problems of evaluating popularity of nodes in various information networks. For instance, the PageRank has been used for spam detection [52], graph partitioning [5], and finding gems in scientific citations [29], just to name a few. In the next section we start with the definition of the PageRank, the main subject of our research.

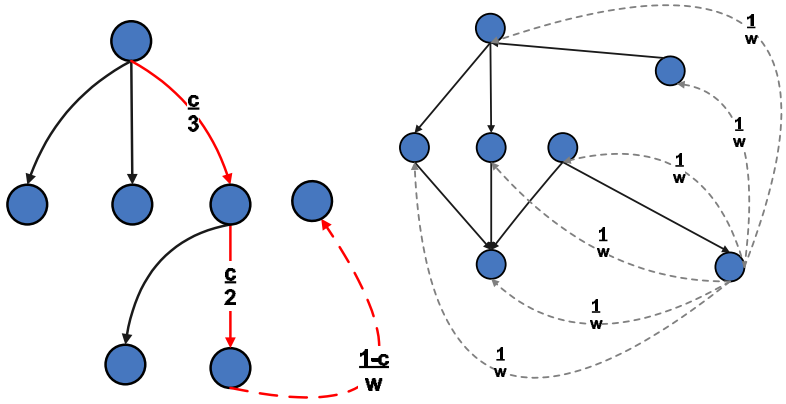
1.2.1 PageRank

The PageRank was introduced by Brin and Page [23, 92] in 1998. This was one of the ideas that brought `Google` to success. We start with the definition of the simplest version of the PageRank, so called standard PageRank. Consider the Web as a graph, where nodes are pages, and edges are links. Denote by w the number of nodes in the Web graph. We use the terms *in-degree* and *out-degree* for the number of incoming and outgoing hyperlinks of a page, respectively.

The PageRank is defined as a stationary distribution of an 'easily bored surfer' random walk on the graph (see Figure 1.2(a)). At each step, with probability c, the random walk follows a randomly chosen outgoing link of a page, and with probability $(1-c)$ the walk starts afresh from a page chosen uniformly among all pages. In other words, at each step the surfer makes a *teleportation jump* to a random page with probability $(1-c)$. The constant c is called a *damping factor*, and takes values between 0 and 1. We can summarize the PageRank definition in the next formula:

$$PR(i) = c \sum_{j \to i} \frac{1}{d_j} PR(j) + \frac{1-c}{w}, \; i = 1, \ldots, w, \tag{1.1}$$

where $PR(i)$ is the PageRank of page i, d_j is out-degree of page j, the sum is taken over all pages j that link to page i, and w is the number of pages in the Web graph.



(a) Random walk on the Web graph

(b) Dangling nodes assumption

Figure 1.2: Standard PageRank

From (1.1) it is clear that high value of PageRank of a page depends not only on quantity, but also on quality (PageRank value) of pages that links to this page. Unlike ranking by in-degree, when adding the large number of links can improve the page position, the PageRank is not easy to cheat. To achieve higher PageRank, page should receive links from important pages. Note that in-degree, as well as out-degree, is a local characteristic of the Web, whereas PageRank is a global one. Thus, adding a link affects only degrees of two pages, however adding a link can affect PageRank in many other pages [7]. The question how in-degree and PageRank are related is not trivial to answer, and it is one of the main questions of this thesis. We refer for discussion on the subject to Section 1.4.2.

If we consider PageRank of a page as a time that surfer spends on this page, then we see that *dangling nodes*, namely pages without out-going links, receive too much 'attention'. In order to solve this unfairness various approaches have been proposed. Page et al. [92] suggest to remove all dangling pages, Kamvar et al. [60] propose to add dangling nodes at the final step of the PageRank computation, and Jeh and Widom [58] modify dangling nodes by adding self-loops. In [14] and [42] authors suggest to add a sink page with self-loop, such that all dangling pages link to it. However, the most popular approach [55, 61, 68, 92] is to assume that every dangling page instead of links to nobody, links to everybody (see Figure 1.2(b)). Then we obtain that the probability to follow a particular link from such page becomes $1/w$, and it is almost zero for large w. This approach leads to the following definition of

the PageRank:

$$PR(i) = c \sum_{j \to i} \frac{1}{d_j} PR(j) + \frac{c}{w} \sum_{j \in \mathcal{D}} PR(j) + \frac{1-c}{w},\ i = 1, \ldots, w, \tag{1.2}$$

where $\mathcal{D}$ is a set of dangling nodes.

The damping factor c plays a crucial role in the definition of the PageRank. First of all, $c < 1$ insures that the PageRank is well defined. Next, presence of c makes computation of the PageRank faster [69]. Traditionally the value of c is chosen as 0.85, and it appears that this value provides reasonable ranking for the Web pages. In [8, 14, 19, 30] authors study other values of the damping factor. Avrachenkov et al. [8], and Boldi et al. [19] obtain that changing the value of c to the value close to 1 leads to distortion of highly ranked pages. Decreasing of the c factor results to more robust PageRank, i.e. the influence of outgoing links of a page on PageRanks of other pages [14], and on the PageRank of this page [7] is possible to bound. In [8] authors suggest to use $c = 0.5$ to achieve more fair ranking for central strongly connected component of the Web graph (see Section 1.3.1). In this work we mainly consider $c = 0.5$ and $c = 0.85$. Depending on the type of underlying graph, the change of the value of the damping factor can affect the top ranked pages like in the Web graph, or, in opposite, has minor influence like in the Wikipedia graph. We refer for details to Section 5.4.

It is common in the literature to rewrite (1.2) in a matrix form. To this end we introduce normalized *hyperlink matrix* H, where $H_{ij} = 1/d_j$ if there is a link from page i to page j, and $H_{ij} = 0$ otherwise. Recall that d_j is the out-degree of page j. Thus, non-zero elements of row i correspond to the outgoing links of page i, whereas non-zero elements of column j correspond to incoming links of page j. Next, we modify matrix H to S as follows: for every dangling node i, we replace corresponding zero row with $(1/w)e^T$, where e^T is a row of ones. Then PageRank vector π^T can be found as a solution of the following equations:

$$\pi^T = \pi^T \left[cS + \frac{1-c}{n} E \right],\ \pi^T e = 1.$$

It is easy to see that π_i corresponds to $PR(i)$ from (1.2). Matrix $G = cS + (1 - c)/wE$ is called Google matrix. This matrix is stochastic (each row sums to 1), irreducible (all pages are connected due to the teleportation jump), aperiodic ($G_{ii} > 0$), and primitive ($G^k > 0$), which implies that a unique positive π^T exists and *power method* guarantees to converge to this vector. Given some initial distribution $\pi^{(0)}$, e.g., $\pi^{(0)} = e$, the power method is defined as an iteration procedure:

$$\pi^{(k)T} = \pi^{(k-1)T} G,\ k \geq 1.$$

Note that uniqueness of π^T gives that the limiting distribution does not depend on the initial distribution $\pi^{(0)}$. Then, the number of iterations that is needed to

achieve ϵ-accuracy is of the order $k = \log(\epsilon)/log(c)$ independent of the underlying graph structure [14]. It possible to accelerate the power method. Kamvar et al. [61] proposed to use extrapolation methods that are based on the expansion of the result after kth iteration, $\pi^{(k)}$, into a series of eigenvectors of G. In [60] Kamvar et al. note that pages within domain are connected more frequently, than pages in different domains, and therefore they modify matrix H into block matrix. Using precomputed values of the PageRank on the relatively small blocks as initial distribution, the authors improve the speed of convergence. For more details about the PageRank computation we refer to [12, 69].

1.2.2 Non-uniform and Personalized PageRank

In the definition of standard PageRank (1.2), the distribution of the random jump, the *teleportation distribution*, is assumed to be uniform, i.e., $1/w$ for every $i = 1, \ldots, w$. In the original paper [92] authors suggest to modify PageRank by adjustment in the teleportation jumps to favor trusted nodes and be the same for all users, or to favor specific nodes for each user with respect to the individual user tastes. Then we can define the non-uniform PageRank as follows:

$$PR(i) = c\sum_{j \to i} \frac{1}{d_j} PR(j) + \frac{c}{w} \sum_{j \in \mathcal{D}} PR(j) + (1-c)T(i),\ i = 1, \ldots, w, \tag{1.3}$$

where $T(i)$ is the probability to start walk afresh in page i.

The knowledge of the user preferences can be based on the usage data, such as browsing histories, or search engine logs; and on the user data, such as information about personal characteristics of the user, e.g., name, age, or geographic location [82]. However, the individual-personalized PageRank, i.e PageRank that is personalized for every user, is computationally infeasible in practice. Then the idea is to build an approximation of such individual PageRank, that is still allows to achieve good level of personalization. Below we list several approaches for this approximation [54]. The Topic-Sensitive PageRank [53] restricts the interests of a user to the small number of topics, say $K = 20$. Then the teleportation jump can be defined as follows: $T(i) = \sum_{i \in J} p_J p_{i,J}$, where p_J is the teleportation probability to the topic J, $J = 1, \ldots, K$, and $p_{i,J}$ is a probability to teleport into particular page i within topic J. Intuitively, if some individuals like to surf for pages about sport, then their search result can be improved by enlarging the $T(i)'$s in (1.3) for the pages with sport content. Then, the Topic-Sensitive PageRank represents user preferences for the beneficial topics choice. Modular PageRank, that was proposed by Jeh and Widom in [58], is similar to the above approach. In this case the surfer teleports to the certain pages with high ranks instead of set of the topic-related pages.

In the BlockRank [60] the Web is considered to be combined from the blocks, for example, each block represents a host. Then, the teleportation jump can be defined as follows: $T(i) = p_J PR_J(i)$, where p_J is a probability to jump into block J, and $PR_J(i)$ is the 'local' PageRank of page i in block J.

We also mention next two approaches that modify the PageRank not through the teleportation. The first, the query-dependent PageRank [101], is based on the idea to replace $1/d_j$ in (1.3) with $p_q(j \to i)$, the probability that random walk follows the link to page i given that it is on page j and is searching for query q. In the second, Constantine and Gleich [30] suggest to modify the damping factor c accordingly to the user surfing properties.

With any of the above mentioned approaches, the resulting distribution of the PageRank scores for a given Web graph, depends on local graph characteristics such as in-degree and out-degree. In Sections 1.3.2 and 1.4.2 we discuss the tail behavior of the PageRank distribution, and its relations to different parameters in the Web.

1.2.3 HITS ranking scheme

Here we give brief introduction to HITS, another way of ranking Web pages. Although it is not as popular as PageRank, it plays an important role in the Web search. HITS algorithm was used in search engine `Teoma`, that is now part of `Ask.com`. The name HITS comes form Hypertext Induced Topic Search, that suggests that HITS is a query dependent algorithm unlike PageRank. The main idea of HITS is to assign for every page two scores: *authority* and *hub scores*. An authority is a page with many incoming links, while a hub is a page with many outgoing links. Then, a good authority is referred by good hubs, and a good hub has links from good authorities. To formulate it mathematically we denote by x_i and y_i authority and hub scores of page i, respectively. Given that every page has been assigned initial scores $x_i^{(0)}$ and $y_i^{(0)}$ we define an iterative procedure as follows:

$$x_i^{(k)} = \sum_{j \to i} y_j^{(k-1)}, \text{ and } y_i^{(k)} = \sum_{i \to j} x_j^{(k-1)}, \ k = 2, \ldots, \tag{1.4}$$

where $i \to j$ means that i links to j. After every iteration $x^{(k)}$ and $y^{(k)}$ need to be normalized.

If we consider *adjacency matrix* A, such as $A_{ij} = 1$ if there is a link from i to j, and $A_{ij} = 0$ otherwise, then we can rewrite (1.4) as

$$x^{(k)} = A^T y^{(k-1)}, \text{ and } y^{(k)} = A x^{(k-1)}, \tag{1.5}$$

where $x^{(k)}$ and $y^{(k)}$ are vectors of authority and hub scores after kth iteration. From (1.4) and (1.5) we obtain

$$x^{(k)} = A^T A x^{(k-1)}, \text{ and } y^{(k)} = A A^T y^{(k-1)}.$$

The matrices $A^T A$ and $A A^T$ are called authority matrix and hub matrix, respectively. The last equations define an iterative power method for computing the dominant eigenvectors for corresponding matrices. The matrices $A^T A$ and $A A^T$ are symmetric, positive semidefinite and non-negative, so their eigenvalues $\lambda_1, \ldots \lambda_w$ are necessary

real and non-negative with $\lambda_1 > \cdots > \lambda_w$. In other words, HITS with normalization always converges as $[\lambda_2(A^TA)/\lambda_1(A^TA)]^k$. Unlike power method of the PageRank, there is no better approximation to the asymptotic rate of convergence. Experiments show that around 10-15 iterations are required for a good approximation [69].

To implement HITS we build neighborhood graph Q that relates to query. To this end we add all pages that contain references to the query to this graph, and expand it by adding pages that links to, or from the pages in Q. This procedure allows to build semantic associations, for example it solves problem of synonyms. In real life such graph expansion can lead to the huge graph, so usually the number of additional pages are limited by some number, say 100 links into and 100 links out of every page. Thus, we obtain a graph that is relatively small compared to the Web graph. Then we calculate hub and authority scores on G, and list pages in two lists accordingly to the scores. Depending on search proposes, user can chose authorities (deep search on the query), or hubs (broad search).

Note that we can find eigenvector just for one of A^TA and AA^T, and then we simply obtain, for instance, hub vector from the equation $y = Ax$. The disadvantages of the HITS algorithm are that it depends on the initial vectors [70], and it is easy to spam. There are different modifications for HITS, that solve mentioned problems. We refer to [69] for details.

There are other ranking techniques, and many modifications of PageRank and HITS. In this thesis we focus only on PageRank. In Section 4.4 we mention HITS when we introduce PAR ranking scheme, that has properties of HITS and PageRank.

In the next section we consider probabilistic structure of the Web graph, in particularly, we focus on the PageRank distribution. In Section 1.4.2 we define a stochastic equation that describes relations between PageRank and other Web characteristics.

1.3 Probabilistic structure of the Web

1.3.1 Web structure

The Web has a complex structure with some notable features. Cardinally, it is huge. Recently `Google` reported that they succeeded to collect 1 trillion (10^{12}) unique URLs on the Web at once.[1] Despite the fact that unique URLs do not always identify unique pages, the obtained number still looks impressive. In 1998, Bharat and Broder [13] estimated the size of indexed Web at 200 million pages. Seven years later Gulli and Signorini in [51] claimed that indexable web is more than 11.5 billion pages. Thus, the Web is growing, and it is growing fast.

The understanding of the Web structure is an important problem that yields to better design of algorithms for crawling, searching and indexing. From a macroscopic point of view, the Web graph can be seen as a bow-tie structure. This concept was

[1] googleblog.blogspot.com/2008/07/we-knew-web-was-big.html;(Accessed in January 2009).

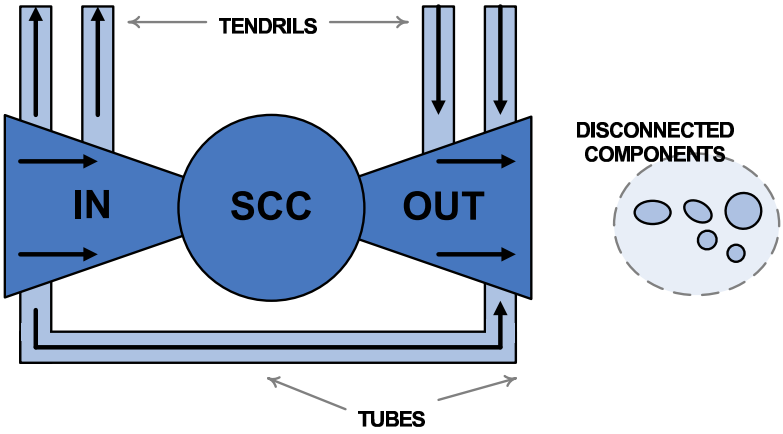


Figure 1.3: The bow-tie structure of the Web

for first time introduced by Broder et al. in [24]. We illustrate this idea in Figure 1.3. According to [24], the Web can be divided into several major components:

SCC, or Strongly Connected Component, that consists of all pages that can reach one another following directed links;

IN component combines all pages that can reach pages from SCC, however, can not be reached from it;

OUT component consists of all pages that are possible to access from SCC, and have no links back to SCC.

Moreover, there are pages that are not in SCC, however are reachable from IN, and pages that can reach OUT without passage through SCC. Such pages are called *TENDRILS. TUBES* are formed from TENDRILS that hang off from IN to hook into TENDRILS leading into OUT. We refer to the remaining parts of the Web pages as to *DISCONNECTED* components. In [24] authors report that the size of SCC (27.7%), while IN, OUT and TENDRILS components have similar sizes, and consist of 22.3%, 21.2% and 21.5% of the Web pages, respectively. Later, similar results were obtained by Donato et al. in [33], where they study another sample of the whole Web. Surprisingly different behavior were observed in [18, 56, 77]. In [18] Boldi et al. discover that half of pages in African Web are condensed into a single giant SCC pointing to several smaller components. Liu et al. [77], and later Hirate et al. [56], report that SCC in Chinese Web consists of 70% of the Web pages. In recent work by Donato et al. [35], authors study inner structure of the various components. Thus, they observed that the IN and OUT components are highly fragmented, while SCC is well interconnected. Moreover, they observed large size of the SCC component for Italian (72.3%), Indochina (51.4%) and UK (65.3%) Web samples. The large

size of SCC in the various national Web domains was also observed in [9]. There can be several explanations for phenomena. The first one is that the national Web domains should be more connected by nature. The second explanation is that the Web possibly becomes denser over time like it was observed in [71] for various social networks. The increasing of the SCC's size over time was also discovered in Wikipedia graph [25].

Assume that we know that two pages in the Web are connected, and we are interested in the length of the shortest path from one page to the other. We call the average value of such lengths as an average diameter [28]. In the Web the average diameter is surprisingly small. Thus, Broder et al. [24] find that the average path length is about 16 edges if the Web graph is directed, and 7 edges if the Web graph is undirected. Albert et al. [4] obtain that the average diameter in `nd.edu` domain equals 11.2 links. The phenomenon of the small diameter is called as small-world effect [84], and popularly known as 'six degrees of separation'. Another important observation about the Web structure is so-called self-similarity of the Web. In short, it means that the Web consists of miniature replicas of itself [32].

One of the most notable features of the Web is a presence of power laws. In the next section we discuss power laws in more details.

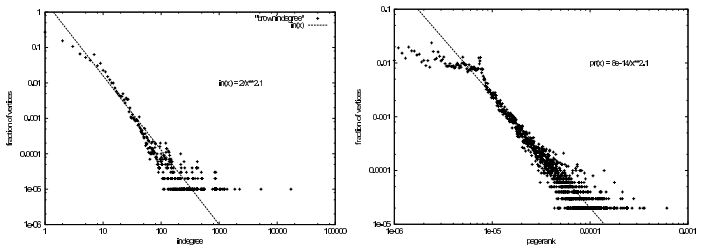
1.3.2 Power laws

In simple words, a random variable X has a *power law distribution* with exponent $\alpha > 0$, if its probability of obtaining a value grater than x is proportional to $x^{-\alpha}$. The power laws are a special family of distributions. In data analysis, many measured parameters have typical size, or scale. For instance, if we consider heights of human beings, the obtained values can deviate significantly, however can not exceed some value. Another example can be speeds of cars on the highway. However, there are some parameters that can vary over an enormously dynamic range. If we consider population of cities, size of files downloaded from the Internet, citation of scientific papers, copies of a book sold, and even diameters of the moon craters, then we can see that the obtained values can be incomparable large or small. For further reading about history and examples of the power law distributions in various research areas we refer to Mitzenmacher [86, 87], and Newman [89].

The standard strategy to reveal a presence of a power law is to plot a histogram of a quantity on log-log scale to obtain a straight line. We have $\log[\mathbb{P}(X = x)] = \log(C) + [\alpha + 1]\log(x)$, where C is some constant. However, this technique is often not efficient. In [89] Newman clearly illustrated that even for generated random numbers with a known distribution the noise in the tail region has a strong influence on the estimation of the power law parameters. Instead of the histogram, we suggest to plot the fraction of measurements that are not smaller than a given value, i.e. the complementary cumulative distribution function $\mathbb{P}(X \geq x)$. The advantage is that we obtain a less noisy plot. Additionally, this idea is consistent with our analysis for complementary cumulative distribution functions. We note that if the distribution

of X follows a power law with exponent α, then the corresponding histogram has an exponent $(\alpha + 1)$. Thus, the plot of $\mathbb{P}(X \geq x)$ on logarithmic scales has a smaller slope than the plot of the histogram. To avoid ambiguity in this work we present all results accordingly to our approach. In Section 4.1 we also discuss other techniques for power law evaluation.

In [47] Faloutsos et al. for the first time discover power law behavior of degree distribution in the undirected graph that represents paths between backbone routers (the AS graph). In the same time Albert et al. [4] observed that in-degree distribution in `nd.edu` follows power law with exponent $\alpha = 1.1$, and, Broder et al. [24] find the same exponent for the in-degree distributions in the entire Web. The next fundamental result was obtained by Pandurangan et al. in [93], where they observe that in-degree and PageRank in the Web graph have similar asymptotic behavior, namely they follow power laws with the same exponent. In Figure 1.4 we present



(a) in-degree (b) PageRank

Figure 1.4: Histogram plots from [93] for in-degree and PageRank in log-log scale.

the log-log plots for histograms for the in-degree and the PageRank from [93]. This observation is one of the results that motivate our research. Subsequent works by Donato et al. [33], and Fortunato et al. [49] confirmed the observation about similarity in tail behavior. Becchetti and Castillo [10] investigate the influence of the damping factor c on the power law behavior of PageRank. Thus, they have shown that the PageRank of the top 10% of the nodes always follows a power law with the same exponent independent of the value of the damping factor. In [26] Capocci et al., and in [25] Buriol et al. analyze in- and out-degrees distribution, and distribution of the PageRank for the Wikipedia samples, and also confirm the similarity in the power law behavior of the in-degree and the PageRank. In our works [74, 112, 113, 111, 115] this problem was studied for the different Web and Wikipedia samples. We refer for numerical results to Chapter 4.

In the next section we focus on various models that allow to achieve various properties of the Web graph, in particular power law distribution of the in-degree.

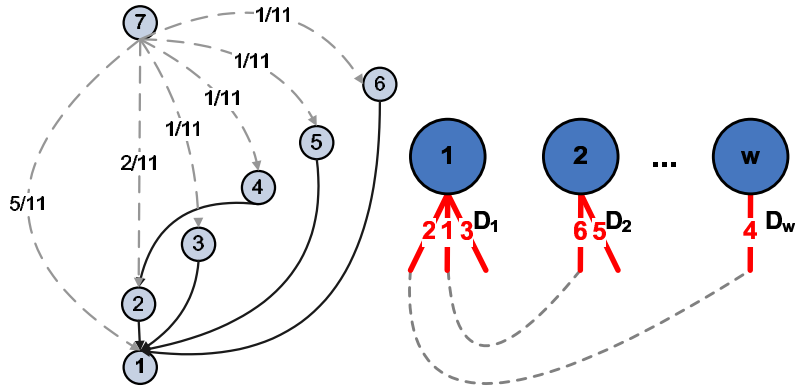
1.3.3 Web models

To better understand underlying structure and evolution of the Web graph, a convenient way to analyze the Web graph is through the random graph models. The pioneering works on the random graphs have been done by Erdös and Rényi [45, 46]. They considered a model for random graphs in which every edge between every pair of nodes is added with some fixed probability. The degree distribution in such a graph is Poisson rather than the observed power law distribution in the Web.

The dynamic *preferential attachment model* is the far-reaching approach for designing graphs with heavy tailed degree distribution. In [86] Mitzenmacher gives a survey on various version of the model arising in different contexts already since 1920s. In their seminal paper [2], Albert and Barabási developed and applied the preferential attachment model to describe the dynamics of wide range of complex networks. This approach had a major impact on studies of the Web structure.

The model is characterized by 'rich-gets-richer' approach. Informally, it means that newcomers prefer to donate their links to already popular pages then to unknown strangers. Thus, we start with d initial nodes, and then every time step we add new node, that link to d already existed nodes. These nodes are selected with probabilities proportional to their degree (see Figure 1.5(a)). In [2] authors propose a model for an undirected graph, that has been shown to have degree distribution with exponent $\alpha = 2$ [37]. Later, Bollobás and Riordan obtain the estimation for diameter at time w as $O(\log(w))$ for $d = 1$, and $O(\log(w)/\log\log(w))$ for $d \geq 2$. However, the original model has few disadvantages: it generates undirected graphs, and power law exponent for degree distribution is stuck at $\alpha = 2$. In order to model graphs with exponent that are in $(1, \infty)$, Dorogovtsev et al. [36, 37], Albert and Barabási [3], and Pennock et al. [95] proposed various modifications for the connection probability. In this thesis we mainly use model from Pennock et al. [95], where new pages connect to uniformly chosen pages with some probability δ, and with probability $(1 - \delta)$ it follows preferential attachment rule. There are also many other variations of preferential attachment models, like as copying model by Kleinberg et al [64] and Kumar et al. [67], general preferential attachment model by Aiello et al. [1], and forest fire model by Leskovec et al. [71]. We refer for a survey on the preferential attachment models to Chakrabarti and Faloustsos [28]

Configuration Model [88, 90] is a static random graph model with predescribed degree sequence. In order to build such a model we first assign degree D_j for every vertex j, and assume that $L_w = \sum_{j=1}^{w} D_j$ is even. Second, we say that page j has D_j 'stubs,' or half-edges. We number the stubs from 1 till L_w randomly, and connect the first stub to one of $L_w - 1$ remaining stubs. Later, we repeat the procedure for the second, unless it was chosen on the first step, and so on until all stubs will be connected (see Figure 1.5(b)). If power law exponent is greater than 2, which means that variance and mean of D exist, then distance between uniform pair of nodes $H_w \approx \log_\nu(w)$, where $\nu = E(D(D-1))/\mathbb{E}(D)$ [109]. In the case of degree distribution on the Web graph, such that the degree distribution has finite mean and



(a) Growing Network (b) Configuration Model

Figure 1.5: Graph models

infinite variance, the obtained distance equals to $H_w = 2 \log\log(w)/\log(\alpha - 1)$. This case was studied by van der Hofstad et al. in [110] and Reittu and Norros in [99]. Finally, for the graphs with infinite mean of degree distribution, van der Hofstad et al. proved that H_w is uniformly bounded [108].

Besides preferential attachment and configuration models there are many other interesting models. We refer to Bonato [21], Chakrabarti and Faloutsos [28], and Newman [88] for excellent surveys.

In the next section, we formalize power laws by the theory of regular variation. Later, in Section 1.4.2 we discuss how the tail distribution of the PageRank relates to the various characteristics of the Web. In Section 1.4.4 give an introduction on applications of angular measure for evaluating dependencies between various characteristics of the Web graph.

1.4 Motivation and methodology

1.4.1 Regular variation

It is difficult to overestimate an importance of study of power law distributions. A common mathematical way to analyze this kind of distributions is based on the theory of regular variations. This theory has been successfully used in many applications, such as mathematical finance [44, 83] for modeling of large insurance claims and stock market shocks; telecommunications [94, 100] for modeling of file sizes; and analysis of extremes [31] for modeling sea floods, just to name a few. Although many large self-organizing networks exhibit power laws, for example, social networks [2, 85],

epidemic networks [73], internet graph [47], or the Web graph [9, 24, 33, 93], most of the studies are restricted to only finding the presence of power laws in degree distributions. The main goal of this work is to fill this gap. We propose to use the theory of regular variation to explain similarity in the asymptotic behavior of in-degree and PageRank, the two most popular measures for page importance in the Web. Furthermore, we apply the theory of multivariate regular variation, and suggest to use the *angular measure* for measuring dependencies between different parameters of power law graphs (see Section 1.4.4). This approach is especially important in the Web, where power law exponents usually smaller than 2. In this case the second moment does not exist, and the correlation coefficient cannot be calculated.

One of the goals of this thesis is to build the correspondence between various Web characteristics and the PageRank distribution . Since the PageRank was introduced, this problem draws a lot of attention. We discuss different approaches in the next section.

To obtain the asymptotic behavior of PageRank we employ the theory of regular variation that provides natural mathematical formalism for analyzing power laws.

Definition 1.1. *A non-negative random variable X is said to be* regularly varying *with index α, if*

$$\mathbb{P}(X > x) = x^{-\alpha} L(x) \quad as \quad x \to \infty, \tag{1.6}$$

for some positive slowly varying *function $L(x)$, that is defined as follows: for every $y > 0$ we have*

$$\frac{L(yx)}{L(x)} \to 1 \text{ as } x \to \infty.$$

For more comprehensive treatment we refer to books of Bingham et al. [17], Resnick [100], and Seneta [105].

1.4.2 In-degree and PageRank

The asymptotic similarity between in-degree and PageRank was first time observed by Pandurangan *et al.* in [93]. Indeed, from the definition of the PageRank ((1.1), (1.2), and (1.3)) we can see that the PageRank should be related to the in-degree. However, as we saw above, the main idea of PageRank is that it depends not only on quantity but also on quality of incoming links of a page. Moreover, we emphasize that PageRank is a global characteristics of the Web while in-degree is a local one. Thus, the phenomena of asymptotic similarity between the in-degree and the PageRank is not trivial to justify.

One of the ways to approach this problem is to build a model of the Web, that has a power law distribution of the in-degree, and then define the PageRank distribution for this model. In [6, 50] authors verify asymptotic properties of PageRank distribution for the case of preferential attachment models.

In this thesis we characterize the power law behavior of the PageRank using the approach that we developed in our works [74, 75, 111, 112]. In the remainder of the

section we briefly describe the main ideas of the approach. The model in its most general form will be presented in Section 2.1, and the tail behavior of the PageRank will be obtained in Chapter 2 and 3.

We model the PageRank as a solution of a distributional identity, and the tail behavior of the solution is obtained under various assumptions. We note that the PageRank values in (1.3) scale as $1/w$ with the number of pages. In our analysis, it is more convenient to deal with corresponding *scale-free PageRank* scores

$$R(i) = wPR(i), \quad i = 1, \dots, w,$$

assuming that w goes to infinity. In this setting, it is easier to compare the probabilistic properties of PageRank and in- and out-degree, that are also scale-free.

We view the PageRank of a random page as a random variable R with $\mathbb{E}(R) = 1$. Our goal is to analyze to what extent the tail probability $\mathbb{P}(R > x)$ for large enough x depends on in-degree distribution N, on distribution of out-degree of a page that links to our randomly chosen page D, on teleportation distribution T, and on fraction of dangling nodes p_0. To this end, we model PageRank R as a solution of a stochastic equation involving N, T and D.

We start our analysis with simplified model in [74, 75], where we assume that all pages have constant out-degree, that equals average in- and out-degree. Then, inspired by formula (1.1), the stochastic equation for the PageRank is as follows:

$$R \stackrel{d}{=} c \sum_{j=1}^{N} \frac{1}{\mathbb{E}(N)} R_j + (1 - c), \tag{1.7}$$

where $a \stackrel{d}{=} b$ means that a and b have the same probability distribution. The relation between PageRank and in-degree is modeled through a distributional identity which is analogous to the equation for the busy period in the M/G/1 queue (see details in Section 1.4.3). We analyze (1.7) using the approach employed in [81] for studying the tail behavior of the busy period in case when the service times are regularly varying random variables.

In [75] we also consider pages without out-going links, i.e. the dangling nodes. We assume that the PageRank of a random page does not depend on whether the page is dangling, then the fraction of the total PageRank mass concentrated in dangling nodes, approximately equals the fraction of dangling nodes p_0.

In [112] we extend stochastic equation (1.7) for the case of random out-degrees. To this end we consider a random variable D, which represents the out-degree of a page that links to a particular randomly chosen page i. We note that D is not the same random variable as an out-degree of a random page since the additional information that a page has a link to i, alters the out-degree distribution. Assuming random out-degrees, in [112] we rewrite the stochastic equation for PageRank as follows:

$$R \stackrel{d}{=} c \sum_{j=1}^{N} \frac{1}{D_j} R_j + [1 - c(1 - p_0)]. \tag{1.8}$$

The solution of the last equation can be found as a limit of $R^{(k)}$'s, where $R^{(k)}$ is defined through a distributional identity

$$R^{(k)} \stackrel{d}{=} c\sum_{j=1}^{N} \frac{1}{D_j} R_j^{(k-1)} + [1 - c(1 - p_0)].$$

If $R^{(0)} \equiv 1$ then $R^{(k)}$ serves as a stochastic model for the result of the kth power iteration in standard PageRank computations. Since PageRank vector is always a result of a finite number of iterations, it follows that $R^{(k)}$ describes the distribution of PageRank if the power iteration algorithm stops after k steps. Using probabilistic techniques from Jessen and Mikosch [59], we defined asymptotical properties of $R^{(k)}$.

Finally, we combine techniques from [74, 75] and [112] in a generalization of our model for the case of non-uniform PageRank. Thus, in Chapter 2 we define asymptotics of PageRank after each iteration using probabilistic approach as in [112], and in Chapter 3 we justify the power law behavior of the PageRank using an analytical approach similar to [74]. Since the model from [112] is a generalization of the previous result, then in this thesis we consider only the last model, where we take into account many different factors affecting the PageRank, such as personalization of the PageRank, and a possible dependence between personalized preference scores and in-degrees of the Web pages. The PageRank stochastic equation can be modified as follows:

$$R \stackrel{d}{=} c\sum_{j=1}^{N} \frac{1}{D_j} R_j + cp_0 + (1 - c)wT, \tag{1.9}$$

To simplify the notation we introduce $A \stackrel{d}{=} c/D$ and $B \stackrel{d}{=} cp_0 + (1-c)nT$, and obtain the generalized stochastic equation (1.10), that is discussed in the next section.

1.4.3 Stochastic equations

From a mathematical point of view, in Chapter 2 and 3 we present the analysis of the following distributional identity

$$R \stackrel{d}{=} \sum_{j=1}^{N} A_j R_j + B, \tag{1.10}$$

where we assume that all random variables are positive; R_j's are independent and distributed as R; and A_j's are independent and distributed as some random variable A with $\mathbb{E}(A) = [1 - \mathbb{E}(B)]/\mathbb{E}(N) < 1$. We also set R_j's and A_j's to be independent, and to be independent of N and B. Moreover, it is essential that $\mathbb{E}(B) < 1$. We emphasize that N and B can be dependent.

Equations similar to (1.10) are well known in the literature. For instance, such equation can also describe the distribution of the busy period in the $M/G/1$ queue,

i.e. the queue with exponentially distributed interarrival times and an arbitrary distribution for service times:

$$R \stackrel{d}{=} \sum_{j=1}^{N(S_1)} R_j + S_1,$$

where R is the distribution of the busy period (the time interval during which the queue is non-empty), S_1 is the service time of the customer that initiated the busy period, $N(S_1)$ is the number of Poisson arrivals during this service time, and R_j's are independent and distributed as R. We refer to [81, 117] for more details on the asymptotics of a busy period in queues with heavy tails.

Another version of (1.10) arises in the theory of branching processes. For $B = 0$ we can obtain the following equation:

$$R \stackrel{d}{=} \sum_{j=1}^{N} A_j R_j,$$

that has been analyzed in detail by Liu [78, 79].

1.4.4 Dependencies and rank correlations

In order to analyze equation (1.9) we have to make assumption on the dependence of the evolved parameters. In our work [76, 114, 115] we study the question of the measuring dependencies between heavy-tailed network parameters. In particular, we focus on the relation between in-degree and PageRank. From the definition of the PageRank (1.3), it is clear that it is influenced largely by in-degree. However, there is no agreement in the literature on the dependence between these two quantities, e.g. [33, 49]. The disagreement is caused by the fact that only the value of the correlation coefficient has been considered as a dependence measure. However, the correlation coefficient is an uninformative dependence measure in heavy-tailed data [11, 28, 31, 100]. Indeed, the correlation coefficient is a 'crude summary' of dependencies that is most informative for jointly normal random variables. It is a common and simple technique but it is not subtle enough to distinguish between the dependencies in large and in small values. This becomes a problem if we want to measure the dependence between two heavy tailed network parameters, because in that case we are mainly interested in the dependence between extremely large values.

We propose to employ the extreme value theory [11] and the theory of regular variation [100] that provide a range of statistical procedures designed to deal with multivariate data of which the marginal distributions exhibit power laws. In particular, the body of statistical theory contains a well-developed notion of dependence. This notion called *extremal dependence* is characterized by *angular measure*, which is much more suitable for the power law data than standard correlation measures.

Based on the stochastic equation of the non-uniform PageRank (1.9), in Section 5.2 we characterize the tail dependence between in-degree and PageRank by

two-point angular measure. This result formalizes the common understanding of two main sources for the high PageRank: high in-degree and a high rank of one of the ancestors. In Section 5.3 we empirically compute the angular measures for the various Web characteristics. Our experimental results reveal a dramatically different correlation structure in the Web, the Wikipedia and preferential attachment graph.

The proposed dependence measure can be also used for measuring rank correlations. We refer for more details to Section 5.4. Using this approach, in Chapter 6 we define rank distance and study possible application for the rank aggregation problems.

1.5 Overview of the thesis

This section gives an overview of the results in the thesis.

In Chapter 2 we define the models for in- and out-degrees, and provide stochastic equation for PageRank in the form (1.10), where each random variable represents a certain parameter in the Web. In Section 2.2 we use a probabilistic approach to show that the proposed equation has a unique non-trivial solution with fixed finite mean. To this end, we introduce a recurrent stochastic model for the power iteration algorithm commonly used in PageRank computations. Further, in Section 2.3 we obtain the PageRank asymptotics after each iteration. In Section 2.4 we predict tail behavior of the limiting distribution of the PageRank as a convergence of the results for iterations. To show the predicted behavior we use alternative techniques in Chapter 3.

In Chapter 3 we define the tail behavior for the model of the PageRank distribution. To this end, we use Laplace-Stieltjes transforms and apply Tauberian theorem, see Theorem 3.2 in Section 3.1. We start with the analysis of the model for the in-degree distribution in Section 3.2. In Section 3.3 we continue with the stochastic model for the PageRank. Then, in Section 3.3.1 we derive the equation for Laplace-Stieltjes transforms, that corresponds to the general stochastic equation (1.10), and in Section 3.3.3 we obtain our main result that establishes the tail behavior of the solution of (1.10). Finally, in Section 3.3.4 we discuss asymptotics for the PageRank distribution under various assumptions on the distribution of the in-degree and the teleportation. Chapters 2 and 3 are based on Volkovich and Litvak [111].

Then, in Chapter 4 we perform a number of experiments on the Web and the Wikipedia data sets, and on preferential attachment graphs in order to justification for the results obtained in Chapters 2 and 3. The numerical results show a good agreement with our stochastic model for the PageRank distribution. Moreover, in Section 4.1 we also address the problem of evaluating power laws in the real data sets. To this end, we define several state of the art techniques from the statistical analysis of heavy tails, and provide empirical evidence on the asymptotic similarity between in-degree and PageRank. Inspired by the minor effect of the out-degree distribution on the asymptotics of the PageRank, in Section 4.4 we introduce PAR

ranking scheme, that combines features of HITS and PageRank ranking schemes. In this chapter we use results from [74, 75, 111, 112, 113]

In Chapter 5 we analyze the dependence structure in the power law graphs. In Section 5.2 we analytically define the tail dependencies between in-degree and PageRank of a one particular page by using the stochastic equation (1.10). Then, in Section 5.3 we compute the angular measures for in-degrees, out-degrees and PageRank scores in three large data sets. The analysis of extremal dependence leads us to propose a new rank correlation measure which is particularly plausible for power law data in Section 5.4. This chapter is based on [76], [114] and [115].

Finally, in Chapter 6 we apply this new rank correlation measure to various problems of rank correlation. This is work in progress that was started during a research visit at `Yahoo!Research Barcelona` in November 2008.

CHAPTER 2

PROBABILISTIC ANALYSIS OF THE PAGERANK DISTRIBUTION

In this chapter we study how asymptotical behavior of the PageRank relates to the various characteristics of the Web graph. We keep definition of the PageRank (1.3) almost unchanged but we transform it into a stochastic equation. We start with models for degree distributions in the Web. In Section 2.1.1 we model in-degree of a random page as an integer valued random variable N, and in Section 2.1.2 we introduce so-called effective out-degree D, that is out-degree of a page that points into the randomly chosen page. Then, in Section 2.1.3 we define PageRank of a random page in the network as a solution of stochastic equation.

We want to analyze to what extent the tail probability of the non-uniform PageRank depends on the distributions of the in-degree, the effective out-degree, and the teleportation jump. We note that the stochastic equation of the PageRank is a special case of the following stochastic equation:

$$R \stackrel{d}{=} \sum_{j=1}^{N} A_j R_j + B. \tag{2.1}$$

In Sections 2.2 and 2.3 as well as in Chapter 3 we consider (2.1) instead of the stochastic equation of the PageRank for the sake of simplicity in notation.

In Section 2.2 we start our analysis with showing that (2.1) has a unique solution R such that $\mathbb{E}(R) = 1$. To this end, in Section 2.2.1 we iteratively define random variables $R^{(k)}$'s, $k \geq 0$. These variables converge to the solution of (2.1) as $k \to \infty$. Next, in Section 2.2.2 we apply the results from the theory of regular variation in order to define the tail behavior of $R^{(k)}$. We state the results in Theorem 2.4, where we obtain that asymptotic of $R^{(k)}$ is determined by the asymptotics of the random

variable with the heaviest tail among N and B. Since the random variable $R^{(k)}$ can be seen as a stochastic model for the result of the kth matrix iteration in the PageRank computation, and the PageRank vector is always a result of a finite number of iterations, then we conclude that the distribution of PageRank should follow power law with exponent that is minimum of exponents of in-degree N and teleportation jump T. However, in Theorem 2.5 we note that if initial distribution $R^{(0)}$ has one of the heaviest tail among $R^{(0)}$, N and T, then the PageRank distribution after kth iteration should follow power law with exponent that is the same as exponent of $R^{(0)}$. Since the limiting distribution of $R^{(k)}$ as $k \to \infty$ does not depend on the initial distribution, then we predict that the asymptotic behavior of R should be defined as a convergence of the results of Theorem 2.4. In order to show the predicted behavior we need to use alternative technique that is based on the Laplace-Stieltjes transforms analysis, and is a subject of Chapter 3.

2.1 Model

In this section we present the models for the distributions of the in- and out-degrees, and the PageRank.

2.1.1 In-degree

We set in-degree of a randomly chosen page in the network to be an integer valued random variable N. In the Web graph as well in some other graphs, where we observe power law behavior of the in-degree distribution, we set N to be an integer valued regularly varying random variable with index $\alpha_N > 1$. One of the ways to model such N is as follows: we assume that $N = N(X)$, where X is regularly varying with index α_N and $N(x)$ is the number of Poisson arrivals during the time interval $[0, x]$, when arrival rate is 1. Then, if X is regularly varying then $N(X)$ is also regularly varying and asymptotically identical to X. In Section 3.2 we demonstrate the tail similarity between X and $N(X)$ by using the Laplace-Stieltjes transforms. Then $N(X)$ is indeed an integer and obeys the power law. We use this representation of N in Chapter 3. In this chapter we do not make any assumptions on N except we require it to be integer valued.

2.1.2 Out-degree

Next, we model the weights $1/d_j$ in the definition of the PageRank (1.3), where d_j is the out-degree of page j that has a link to page i. To this end, we consider a random variable D that represents the out-degree of a page that links to a particular randomly chosen page i. Note that D is not the same random variable as an out-degree of a random page since the additional information that a page has a link to i alters the out-degree distribution. This phenomenon is known as inspection paradox. The inspection paradox roughly states that an interval containing a random point tends

to be larger than a randomly chosen interval [102]. For instance, in [103], a number of children in a family, to which a randomly chosen child belongs, is stochastically larger than a number of children in a randomly chosen family. Likewise, a number of out-links D from a page containing a random link, should be stochastically larger than an out-degree of a random page. If p_j is a fraction of the pages with out-degree $j \geq 0$, then we can obtain

$$\lim_{w\to\infty} \mathbb{P}(D = j) = \frac{jp_j}{\mathbb{E}(N)}, \quad j \geq 1. \tag{2.2}$$

where $\mathbb{E}(N)$ is the average in/out-degree, and w is the number of pages in the Web. For sufficiently large networks, we may assume that the distribution of D is equal to its limiting distribution as defined by (2.2). We refer to D as an *effective out-degree*. The term is motivated by the fact that the distribution of D is the one that participates in the PageRank formula (1.3).

2.1.3 Stochastic equation for the PageRank

Now, we are ready to model the PageRank distribution. We view the PageRank of a random page as a random variable R with $\mathbb{E}(R) = 1$. Further, we assume that the PageRank of a random page does not depend on the fact whether the page is dangling. Indeed, it can be shown that the PageRank of a page can not be altered significantly by modifying outgoing links [7]. Moreover, experiments, e.g. in [42], show that dangling nodes are often just regular pages whose links have not been crawled. Besides, even authentically dangling pages such as `.ps`, `.jpg` or audio files, often contain important information and gain a high ranking independently of the fact that they do not have outgoing links. We note that such independence immediately implies that in large networks, the fraction of the total PageRank mass concentrated in dangling nodes is equal to the fraction of dangling nodes p_0, simply by the law of large numbers:

$$p_0 = \frac{1}{w} \sum_{j\in\mathcal{D}} R(j).$$

Our goal is to analyze to what extent the tail probability $\mathbb{P}(R > x)$ for large enough x depends on the in-degree N, the effective out-degree D, the teleportation jump T and the fraction of dangling nodes p_0. To this end, we model PageRank R as a solution of a stochastic equation involving N, T and D. Inspired by the original formula (1.3), the stochastic equation for the PageRank is as follows:

$$R \stackrel{d}{=} c \sum_{j=1}^{N} \frac{1}{D_j} R_j + cp_0 + (1 - c)wT. \tag{2.3}$$

Here R_j's and D_j's are independent and distributed as R and D, respectively. Moreover, we need to assume that R_j's and D_j's are independent and independent of N

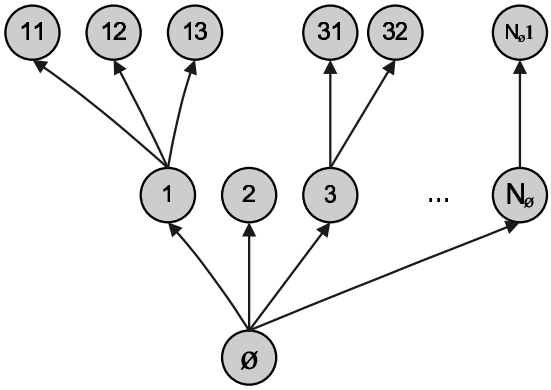


Figure 2.1: An example of Galton-Watson tree

and T. As before, $c \in (0,1)$ is a damping factor. We emphasize that N and T are allowed to be depended, that is often the case for the non-uniform PageRank.

Hence, in stochastic equation (2.3) we generalize models (1.7) and (1.8) for the case of random out-degree, and random teleportation jump. Moreover, here we allow this personalization jump to be dependent on the in-degree. In the next section we will show that (2.3) has a unique solution R such that $\mathbb{E}(R) = 1$.

2.2 Solution of stochastic equation

In the remainder of this chapter and in Chapter 3 we will analyze the following stochastic equation

$$R \stackrel{d}{=} \sum_{j=1}^{N} A_j R_j + B, \tag{2.4}$$

where we assume that all random variables are positive; R_j's are independent and distributed as R; and A_j's are independent and distributed as some random variable with $\mathbb{E}(A) = [1 - \mathbb{E}(B)]/\mathbb{E}(N)$. We also set R_j's and A_j's to be independent, and to be independent of N and B. Moreover, it is essential that $\mathbb{E}(B) < 1$. We emphasize that N and B can be dependent. It is easy to see that the above equation corresponds to (2.3) for $A \stackrel{d}{=} c/D$ and $B \stackrel{d}{=} cp_0 + (1-c)nT$.

In Sections 2.2.2 and 2.3 we establish the existence and the asymptotic properties of R in (2.4) using an iterative procedure defined in the next section.

2.2.1 Iterations

We use the following notations adopted from Liu [79]. Let $\{(N_u, A_{u_1}, A_{u_2}, \dots)\}_u$ be a family of independent copies of $(N, A_1, A_2, \dots)$ indexed by all finite sequences

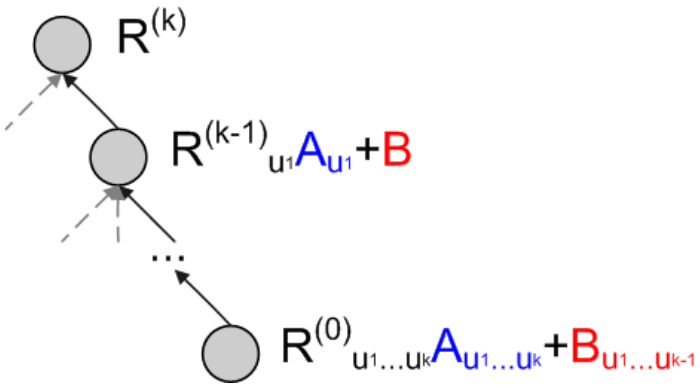


Figure 2.2: The kth iteration

$u = u_1 \dots u_i$, where $u_j \in \{1, 2, \dots\}$, $j = 1 \dots i$. Further, let $\mathbb{T}$ be *the Galton-Watson tree* with defining elements $\{N_u\}$: we have $\emptyset \in \mathbb{T}$ and, if $u \in \mathbb{T}$ and $j \in \{1, 2, \dots\}$, then concatenation $uj \in \mathbb{T}$ if and only if $1 \le j \le N_u$. In other words, we indexed the nodes of the tree with root $\emptyset$ and the first level nodes $1, 2, \dots N_\emptyset$, and at every subsequent level, the jth offspring of u is termed uj (see Figure 2.1).

We start with initial distribution $R^{(0)}$, and for every $k \ge 1$, we define the result of the kth iteration of (2.4) through a distributional identity:

$$R^{(k)} = \sum_{j=1}^{N} A_j R_j^{(k-1)} + B, \tag{2.5}$$

where $R_j^{(k-1)}$ and A_j, $j \ge 1$, are independent and distributed as $R^{(k-1)}$ and A, respectively.

Repeatedly applying (2.5), we derive the following representation for $R^{(k)}$, $k \ge 1$:

$$R^{(k)} = \sum_{u_1 \dots u_k \in \mathbb{T}} A_{u_1} \dots A_{u_1 \dots u_k} R^{(0)}_{u_1 \dots u_k} + \sum_{i=0}^{k-1} \sum_{u_1 \dots u_i \in \mathbb{T}} A_{u_1} \dots A_{u_1 \dots u_i} B_{u_1 \dots u_i}, \tag{2.6}$$

where $\mathbb{T}$ is a notation for the Galton-Watson tree. In Figure 2.2 we display the graphic interpretation of $R^{(k)}$.

2.2.2 Existence and uniqueness of solution

We start with the following definition. A stochastic process $\{Z_i,\ i \ge 1\}$ is said to be a *martingale* process if $\mathbb{E}(|Z_i|) < \infty$ for all i, and $\mathbb{E}(Z_{i+1}|Z_1, \dots, Z_i) = Z_i$.

We use the next lemma to prove the existence of the solution (2.4). This lemma is a result mentioned in [79].

Lemma 2.1. *If* $\mathbb{E}\left(\sum_{j=1}^{N} A_j\right) = 1$, *then the sequence* $\sum_{u_1 \dots u_i \in \mathbb{T}} A_{u_1} \dots A_{u_1 \dots u_i}$ *is a martingale.*

In the next theorem we show that iterations $R^{(k)}$, $k \geq 1$, converge to the unique solution of (2.4).

Theorem 2.2. *Equation (2.4) has the unique non-trivial solution with mean* 1 *given by*

$$R^{(\infty)} = \lim_{k \to \infty} R^{(k)} = \sum_{i=0}^{\infty} \sum_{u_1 \dots u_i \in \mathbb{T}} A_{u_1} \dots A_{u_1 \dots u_i} B_{u_1 \dots u_i}. \tag{2.7}$$

Proof. It is easy to verify that $R^{(\infty)}$ in (2.7) is a well-defined solution of (2.4). In particular, because all random variables are positive, we apply Fubini's theorem [15] to obtain

$$\begin{aligned}
\mathbb{E}\left(R^{(\infty)}\right) &= \mathbb{E}\left[\sum_{i=0}^{\infty} \sum_{u_1 \dots u_i \in \mathbb{T}} A_{u_1} \dots A_{u_1 \dots u_i} B_{u_1 \dots u_i}\right] \\
&= \mathbb{E}(B) \sum_{i=0}^{\infty} (1 - \mathbb{E}(B))^n \mathbb{E}\left[\sum_{u_1 \dots u_i \in \mathbb{T}} \frac{1}{1 - \mathbb{E}(B)} A_{u_1} \dots \frac{1}{1 - \mathbb{E}(B)} A_{u_1 \dots u_i}\right] = 1,
\end{aligned}$$

where the final equation holds since $\sum_{u_1 \dots u_i \in \mathbb{T}} (A_{u_1}/(1 - \mathbb{E}(B))) \dots (A_{u_1 \dots u_i}/(1 - \mathbb{E}(B)))$ is a martingale with mean 1 according to Lemma 2.1. In the second equality we can take $\mathbb{E}(B)$ outside of the summation since $B_{u_1 \dots u_i}$ comes from the $(i-1)$th step, and is independent of the number of incoming links at the level i. We refer to Figure 2.2 for illustration.

To prove the uniqueness, assume that there is another solution with mean 1 and take this solution as an initial distribution $R^{(0)}$ with $\mathbb{E}(R^{(0)}) = 1$. Consider $R^{(k)}$, then the first part of (2.6) has a mean:

$$\mathbb{E}\left(\sum_{u_1 \dots u_k \in \mathbb{T}} A_{u_1} \dots A_{u_1 \dots u_k} R^{(0)}_{u_1 \dots u_k}\right) = (\mathbb{E}(N))^k \left(\frac{(1 - \mathbb{E}(B))}{\mathbb{E}(N)}\right)^k = (1 - \mathbb{E}(B))^k,$$

and hence this part converges in probability to 0, as $k \to \infty$, because, by the Markov inequality, the probability that this term is greater than some $\epsilon > 0$ is at most $(1 - \mathbb{E}(B))^k/\epsilon \to 0$ as $k \to \infty$. Moreover, the second part of (2.6) converges a.s. to $R^{(\infty)}$ as $k \to \infty$. It follows that (2.6) converges to $R^{(\infty)}$ in probability. We conclude that there is no other fixed point of (2.4) with mean 1 except $R^{(\infty)}$. □

2.3 Asymptotics for iterations

Our main goal is to show how the asymptotics of R in (2.4) depends on the distribution of N and B. We divide this problem into three possible cases. In the first

case, we assume that N is a regularly varying random variable, and B has some distribution with lighter tail, that is, $\mathbb{P}(B > x) = o(\mathbb{P}(N > x))$ as $x \to \infty$. Then we recall that N is an integer valued regularly varying random variable

$$\mathbb{P}(N > x) \sim x^{-\alpha_N} L_N(x) \text{ as } x \to \infty, \tag{2.8}$$

where $L_N(x)$ is slowly varying function. In the second case, we take B to be regularly varying and N to have a lighter tail. Then, we have

$$\mathbb{P}(B > x) \sim x^{-\alpha_B} L_B(x) \text{ as } x \to \infty, \tag{2.9}$$

where $L_B(x)$ is a slowly varying function. In the final case, we consider both variables to be regularly varying with the same indexes.

At this point, we assume that $\mathbb{E}(N)\mathbb{E}(A^\alpha) < 1$, where $\alpha = \min(\alpha_N, \alpha_B)$.

We start with lemma that describes the asymptotic behavior of product, sum and random sums of regularly varying random variables. We use these results in Theorems 2.4 and 2.5 for defining asymptotic properties of PageRank, when the PageRank is a result of the finite number of the iteration steps. In the lemma, relation (iii) is known as Breiman's theorem (see e.g. Lemma 4.2.(1) in [59]). Properties (iv), (v), and (vi) are statements (2), (1) and (5) of Lemma 3.7 in [59], respectively. The results (i) and (ii) directly follow from Lemma 3.12 and 3.1 in [59], respectively.

Lemma 2.3. (i) *Assume that X_1 is non-negative regularly varying random variable with index $\alpha \geq 0$. If random variable $X_2 > 0$ is such that $\mathbb{P}(X_2 > x) = o(\mathbb{P}(X_1 > x))$, then*

$$\mathbb{P}(X_1 + X_2 > x) \sim \mathbb{P}(X_1 > x) \text{ as } x \to \infty.$$

(ii) *Assume that X_1 is non-negative regularly varying random variable with index $\alpha \geq 0$. If random variable $X_2 > 0$ satisfies $\mathbb{P}(X_2 > x) \sim C\, \mathbb{P}(X_1 > x)$ for some constant $C > 0$, and $\mathbb{P}(X_1 > x, X_2 > x) = o(\mathbb{P}(X_1 > x))$, then*

$$\mathbb{P}(X_1 + X_2 > x) \sim (1 + C)\mathbb{P}(X_1 > x) \text{ as } x \to \infty.$$

(iii) *Assume that X_1 and X_2 are two independent non-negative random variables such that X_1 is regularly varying with index α and that $\mathbb{E}(X_2^{\alpha+\epsilon}) < \infty$ for some $\epsilon > 0$. Then*

$$\mathbb{P}(X_1 X_2 > x) \sim \mathbb{E}(X_2^\alpha)\mathbb{P}(X_1 > x) \text{ as } x \to \infty.$$

(iv) *Assume that N is regularly varying with index $\alpha \geq 0$; if $\alpha = 1$, then assume that $\mathbb{E}(N) < \infty$. Moreover, let (X_i) be i.i.d. sequence such that $\mathbb{E}(X_1) < \infty$ and $\mathbb{P}(X_1 > x) = o(\mathbb{P}(N > x))$. Then as $x \to \infty$,*

$$\mathbb{P}\left(\sum_{i=1}^{N} X_i > x\right) \sim (\mathbb{E}(X_1))^\alpha \mathbb{P}(N > x) \text{ as } x \to \infty.$$

(*v*) *Assume* (X_i) *is i.i.d. sequence of regular varying random variables with index* $\alpha > 0$, $\mathbb{E}(N) < \infty$, *and* $\mathbb{P}(N > x) = o(\mathbb{P}(X_1 > x))$. *Then*

$$\mathbb{P}\left(\sum_{i=1}^{N} X_i > x\right) \sim \mathbb{E}(N)\mathbb{P}(X_1 > x) \text{ as } x \to \infty.$$

(*vi*) *Assume that* $\mathbb{P}(X_1 > x) \sim C\ \mathbb{P}(N > x)$ *for some constant* $C > 0$, *that* X_1 *is regularly varying with index* $\alpha \geq 1$, *and* $\mathbb{E}(X_1) < \infty$. *Then*

$$\mathbb{P}\left(\sum_{i=1}^{N} X_i > x\right) \sim (C\ \mathbb{E}(N) + (\mathbb{E}(X_1))^{\alpha})\mathbb{P}(N > x) \text{ as } x \to \infty.$$

In the next theorem we consider the case when the initial distribution $R^{(0)}$ has a lighter tail than N or B. This assumption makes sense since iterations usually start with $R^{(0)} \equiv 1$. For other types of distribution of $R^{(0)}$ we refer to Theorem 2.5. In short, the next theorem states that the tail behavior of $R^{(k)}$ is determined by the asymptotics of the random variable with the heaviest tail among N and B. Moreover, if the tails of N and B are equally heavy, then in fact we get the sum of two asymptotic expressions.

Theorem 2.4. (*i*) *If* $\mathbb{P}(B > x) = o(\mathbb{P}(N > x))$ *and* $\mathbb{P}(R^{(0)} > x) = o(\mathbb{P}(N > x))$, *then for all* $k \geq 1$:

$$\mathbb{P}(R^{(k)} > x) \sim C_N^{(k)}\mathbb{P}(N > x) \text{ as } x \to \infty,$$

where $C_N^{(k)} = (\mathbb{E}(A))^{\alpha_N} \sum_{i=0}^{k-1} \left[\mathbb{E}(N)\mathbb{E}(A^{\alpha_N})\right]^i$.

(*ii*) *If* $\mathbb{P}(N > x) = o(\mathbb{P}(B > x))$ *and* $\mathbb{P}(R^{(0)} > x) = o(\mathbb{P}(B > x))$, *then for all* $k \geq 1$,

$$\mathbb{P}(R^{(k)} > x) \sim C_B^{(k)}\mathbb{P}(B > x) \text{ as } x \to \infty,$$

where $C_B^{(k)} = \sum_{i=0}^{k-1} \left[\mathbb{E}(N)\mathbb{E}(A^{\alpha_B})\right]^i$.

(*iii*) *If* $\mathbb{P}(B > x) \sim C_{BN}\mathbb{P}(N > x)$ *for some constant* C_{BN}, $\mathbb{P}(R^{(0)} > x) = o(\mathbb{P}(N > x))$, *and* $\mathbb{P}(N > x, B > x) = o(\mathbb{P}(N > x))$, *then for all* $k \geq 1$,

$$\mathbb{P}(R^{(k)} > x) \sim C^{(k)}\mathbb{P}(N > x) \text{ as } x \to \infty,$$

where $C^{(k)} = [C_{BN} + (\mathbb{E}(A))^{\alpha_N}] \sum_{i=0}^{k-1} \left[\mathbb{E}(N)\mathbb{E}(A^{\alpha_N})\right]^i$.

Proof.

(*i*) We will use induction on k. For $k = 1$ we apply Lemma 2.3 (*i*) and (*iv*) to obtain

$$\mathbb{P}\left(R^{(1)} > x\right) = \mathbb{P}\left(\sum_{j=1}^{N} A_j R_j^{(0)} + B > x\right) \sim \mathbb{P}\left(\sum_{j=1}^{N} A_j R_j^{(0)} > x\right)$$
$$\sim (\mathbb{E}(A))^{\alpha_N}\mathbb{P}(N > x) \text{ as } x \to \infty,$$

since $\mathbb{E}(N) < \infty$, $\mathbb{E}\left(A_1 R_1^{(0)}\right) = \mathbb{E}(A) \ < \infty$, and $\mathbb{P}\left(A_1 R_1^{(0)} > x\right) = o(\mathbb{P}(N > x))$. Now, assume that the result has been shown for the $(k-1)$th iteration, $k \geq 2$, then Lemma 2.3 (iii) yields

$$\mathbb{P}\left(A_1 R_1^{(k-1)} > x\right) \sim C_N^{(k-1)} \mathbb{E}\left(A^{\alpha_N}\right) \mathbb{P}(N > x), \tag{2.10}$$

Because of (2.10) and $\mathbb{E}\left(A_1 R_1^{(k-1)}\right) = \mathbb{E}(A) < \infty$, we can apply Lemma 2.3 (i), and (vi) to obtain

$$\begin{aligned}\mathbb{P}\left(R^{(k)} > x\right) &\sim \mathbb{P}\left(\sum_{j=1}^{N} A_j R_j^{(k-1)} + B > x\right) \\ &\sim \left[C_N^{(k-1)} \mathbb{E}(A^{\alpha_N}) \mathbb{E}(N) + (\mathbb{E}(A))^{\alpha_N}\right] \mathbb{P}(N > x) = C_N^{(k)} \mathbb{P}(N > x) \text{ as } x \to \infty.\end{aligned}$$

(ii) From Lemma 2.3 (i) we have that

$$\mathbb{P}\left(R^{(1)} > x\right) \sim \mathbb{P}\left(\sum_{j=1}^{N} A_j R_j^{(0)} + B > x\right) \sim \mathbb{P}(B > x) \text{ as } x \to \infty.$$

Assume that the statement holds for $(k-1)$, where $k \geq 2$. Then, from Lemma 2.3 (iii) we obtain

$$\mathbb{P}\left(A_1 R_1^{(k-1)} > x\right) \sim C_B^{(k-1)} \mathbb{E}\left(A^{\alpha_B}\right) \mathbb{P}(B > x).$$

Because $\mathbb{E}(N) < \infty$, we apply Lemma 2.3 (ii) and (v) to obtain

$$\begin{aligned}\mathbb{P}\left(R^{(k)} > x\right) &\sim \mathbb{P}\left(\sum_{j=1}^{N} A_j R_j^{(k-1)} + B > x\right) \\ &\sim \left[\mathbb{E}(N) C_B^{(k-1)} \mathbb{E}(A^{\alpha_B}) + 1\right] \mathbb{P}(B > x) = C_B^{(k)} \mathbb{P}(B > x) \text{ as } x \to \infty.\end{aligned}$$

(iii) We start the induction with $k = 1$ as follows

$$\begin{aligned}\mathbb{P}\left(R^{(1)} > x\right) &\sim \mathbb{P}\left(\sum_{j=1}^{N} A_j R_j^{(0)} + B > x\right) \sim (\mathbb{E}(A))^{\alpha_N} \mathbb{P}(N > x) \\ &+ \mathbb{P}(B > x) \sim [(\mathbb{E}(A))^{\alpha_N} + C_{BN}] \mathbb{P}(N > x) \text{ as } x \to \infty,\end{aligned}$$

where we use Lemma 2.3 (ii) and (iv). Next, from (2.10), $\mathbb{E}\left(A_1 R_1^{(k-1)}\right) = \mathbb{E}(A) <$

∞, and using of Lemma 2.3 (ii) and (vi) we obtain that for any $k \geq 2$:

$$\begin{aligned}\mathbb{P}\left(R^{(k)} > x\right) &\sim \mathbb{P}\left(\sum_{j=1}^{N} A_j R_j^{(k-1)} + B > x\right) \\ &\sim \left[\mathbb{E}(N) C^{(k-1)} \mathbb{E}(A^{\alpha_N}) + (\mathbb{E}(A))^{\alpha_N} + C_{BN}\right] \mathbb{P}(N > x) \\ &= C^{(k)} \mathbb{P}(N > x) \text{ as } x \to \infty.\end{aligned}$$

□

With $R^{(k)}$ for $A \stackrel{d}{=} c/D$ and $B \stackrel{d}{=} cp_0 + (1-c)wT$, the random variable $R^{(k)}$ serves as a stochastic model for the result of the kth power iteration in the PageRank computation (see Section 1.2.1). Since the PageRank vector is always a result of a finite number of iterations, we can conclude that the distribution of PageRank should follow power law with exponent $\alpha = \min(\alpha_N, \alpha_B)$. However, if the initial distribution $R^{(0)}$ has one of the heaviest tails, then the following results hold.

Theorem 2.5. *Let $R^{(0)}$ be a regularly varying random variable with index $\alpha_R > 0$. Then the following statements hold.*

(i) If $\mathbb{P}(N > x) = o(\mathbb{P}(R^{(0)} > x))$ and $\mathbb{P}(B > x) = o(\mathbb{P}(R^0 > x))$, then for all $k \geq 1$:

$$\mathbb{P}(R^{(k)} > x) \sim C_R^{(k)} \mathbb{P}(R^{(0)} > x) \text{ as } x \to \infty,$$

where $C_R^{(k)} = \prod_{i=0}^{k} \left[\mathbb{E}(N)\mathbb{E}(A^{\alpha_R})\right]^i$.

(ii) If $\mathbb{P}(R^0 > x) \sim C_{RN}\mathbb{P}(N > x)$, and $\mathbb{P}(B > x) = o(\mathbb{P}(R^{(0)} > x))$, then for all $k \geq 1$:

$$\mathbb{P}(R^{(k)} > x) \sim C_{RN}^{(k)} \mathbb{P}(N > x) \text{ as } x \to \infty,$$

where $C_{RN}^{(k)} = [\mathbb{E}(N)\mathbb{E}(A^{\alpha_N})]^k C_{RN} + [\mathbb{E}(A)]^{\alpha_N} \sum_{i=0}^{k-1} \left[\mathbb{E}(N)\mathbb{E}(A^{\alpha_N})\right]^i$.

(iii) If $\mathbb{P}(N > x) = o(\mathbb{P}(R^{(0)} > x))$, $\mathbb{P}(R^{(0)} > x) \sim C_{RB}\mathbb{P}(B > x)$, and $\mathbb{P}(R^{(0)} > x, B > x) = o(\mathbb{P}(B > x))$, then for all $k \geq 1$:

$$\mathbb{P}(R^{(k)} > x) \sim C_{RB}^{(k)} \mathbb{P}(B > x) \text{ as } x \to \infty,$$

where $C_{RB}^{(k)} = [\mathbb{E}(N)\mathbb{E}(A^{\alpha_B})]^k C_{RB} + \sum_{i=0}^{k-1} \left[\mathbb{E}(N)\mathbb{E}(A^{\alpha_B})\right]^i$.

(iv) If $\mathbb{P}(R^0 > x) \sim C_{RN}\mathbb{P}(N > x)$, $\mathbb{P}(B > x) \sim C_{BN}\mathbb{P}(N > x)$, $\mathbb{P}(R^{(0)} > x, N > x) = o(\mathbb{P}(N > x))$, and $\mathbb{P}(B > x, N > x) = o(\mathbb{P}(N > x))$, then for all $k \geq 1$:

$$\mathbb{P}(R^{(k)} > x) \sim C_{RBN}^{(k)} \mathbb{P}(N > x) \text{ as } x \to \infty,$$

$$C_{RBN}^{(k)} = [\mathbb{E}(N)\mathbb{E}(A^{\alpha_N})]^k C_{RN} + [C_{BN} + [\mathbb{E}(A)]^{\alpha_N}] \sum_{i=0}^{k-1} \left[\mathbb{E}(N)\mathbb{E}(A^{\alpha_N})\right]^i.$$

Proof. We again use induction. We start with $k = 1$ for which all statements are valid. Next, we assume that result has been shown for $(k-1)$th iteration, where $k > 2$. Then we consider every case respectively.

(i) We apply Lemma 2.3 (i), (iii) and (v) to obtain the following:

$$\mathbb{P}(R^{(k)} > x) = \mathbb{P}\left(\sum_{j=1}^{N} A_j R_j^{(k-1)} + B > x\right) \sim \mathbb{P}\left(\sum_{j=1}^{N} A_j R_j^{(k-1)} > x\right)$$
$$\sim \mathbb{E}(N)\mathbb{E}(A^{\alpha_R})\mathbb{P}(R^{(k-1)} > 0) = C_R^{(k)}\mathbb{P}(R^{(0)} > 0).$$

(ii) In this case we have

$$\mathbb{P}(R^{(k)} > x) = \mathbb{P}\left(\sum_{j=1}^{N} A_j R_j^{(k-1)} + B > x\right) \sim \mathbb{P}\left(\sum_{j=1}^{N} A_j R_j^{(k-1)} > x\right)$$
$$\sim \left[\mathbb{E}(A^{\alpha_N})\mathbb{E}(N)C_{RN}^{(k-1)} + (\mathbb{E}(A))^{\alpha_N}\right]\mathbb{P}(N > x) = C_{RN}^{(k)}\mathbb{P}(N > x),$$

where we use Lemma 2.3 (i), (iii) and (vi).

(iii) From Lemma 2.3 (ii), (iii) and (v) we obtain the statement:

$$\mathbb{P}(R^{(k)} > x) = \mathbb{P}\left(\sum_{j=1}^{N} A_j R_j^{(k-1)} + B > x\right) \sim \mathbb{P}\left(\sum_{j=1}^{N} A_j R_j^{(k-1)} > x\right)$$
$$+ \mathbb{P}(B > x) \sim \left[\mathbb{E}(A^{\alpha_B})\mathbb{E}(N)C_{RB}^{(k-1)} + 1\right]\mathbb{P}(B > x) = C_{RB}^{(k)}\mathbb{P}(B > x).$$

(iv) Here we use Lemma 2.3 (ii), (iii) and (vi) and get the following result:

$$\mathbb{P}(R^{(k)} > x) = \mathbb{P}\left(\sum_{j=1}^{N} A_j R_j^{(k-1)} + B > x\right) \sim \mathbb{P}\left(\sum_{j=1}^{N} A_j R_j^{(k-1)} > x\right)$$
$$+ \mathbb{P}(B > x) \sim \left[\mathbb{E}(A^{\alpha_N})\mathbb{E}(N)C_{RBN}^{(k-1)} + (\mathbb{E}(A))^{\alpha_N} + C_{BN}\right]\mathbb{P}(N > x)$$
$$= C_{RBN}^{(k)}\mathbb{P}(N > x).$$

□

Recall that for $A \stackrel{d}{=} c/D$ and $B \stackrel{d}{=} cp_0 + (1-c)wT$, equation (3.1) is a stochastic equation for the non-uniform PageRank. Then, we use $\mathbb{E}(1/D) = (1-p_0)/\mathbb{E}(N)$, $\mathbb{P}(B > x) \sim (1-c)^{\alpha_T}\mathbb{P}(wT > x)$ as $x \to \infty$, and Theorem 2.4 to obtain the following equivalence.

Corollary 2.6. *(i) If in-degree N follows power law with exponent α_N, and $\mathbb{P}(wT > x) = o(\mathbb{P}(N > x))$, then $\mathbb{P}(R^{(k)} > x) \sim C_N^{(k)}\mathbb{P}(N > x)$ as $x \to \infty$, where*

$$C_N^{(k)} = \frac{c^{\alpha_N}(1-p_0)^{\alpha_N}}{(\mathbb{E}(N))^{\alpha_N}} \sum_{i=1}^{k} \left[c^{\alpha_N}\mathbb{E}(N)\mathbb{E}(1/D^{\alpha_N})\right]^i .$$

(ii) If normalized teleportation jump wT follows power law with exponent α_T, and $\mathbb{P}(N > x) = o(\mathbb{P}(wT > x))$, then $\mathbb{P}(R^{(k)} > x) \sim C_T^{(k)}\mathbb{P}(T > x)$ as $x \to \infty$, where

$$C_T^{(k)} = (1-c)^{\alpha_T} \sum_{i=1}^{k} \left[c^{\alpha_T}\mathbb{E}(N)\mathbb{E}(1/D^{\alpha_T})\right]^i .$$

(iii) If N and wT obey power law with the same exponent α_N, and $\mathbb{P}(wT > x) \sim C_{BN}(1-c)^{-\alpha_N}\mathbb{P}(N > x)$, then $\mathbb{P}(R^{(k)} > x) \sim C^{(k)}\mathbb{P}(N > x)$ as $x \to \infty$, where

$$C^{(k)} = \left[C_{BN} + \frac{c^{\alpha_N}(1-p_0)^{\alpha_N}}{(\mathbb{E}(N))^{\alpha_N}}\right] \sum_{i=1}^{k} \left[c^{\alpha_N}\mathbb{E}(N)\mathbb{E}(1/D^{\alpha_N})\right]^i .$$

As we can see the values of the multiplicative constants $C_N^{(k)}$, $C_T^{(k)}$ and $C^{(k)}$ increase with increasing of the number of the iterations. Since $C_N^{(k)}$, $C_T^{(k)}$ and $C^{(k)}$ are always smaller than 1, we can claim that with each new iteration the log-log plot for the PageRank creeps up to the log-log plot of the in-degree distribution. For more details and empirical results we refer to Chapter 4. We also refer to discussion on the asymptotics of limiting PageRank distribution in Section 3.3.4

2.4 Asymptotics: from $R^{(k)}$ to $R^{(\infty)}$

Combining the results from Theorem 2.2 and 2.4, we can presume the following asymptotic similarities for $R^{(\infty)}$, the unique non-trivial solution of (2.4):

(i) If $\mathbb{P}(B > x) = o(\mathbb{P}(N > x))$, then $\mathbb{P}(R^{(\infty)} > x) \sim C_N\mathbb{P}(N > x)$ as $x \to \infty$, where $C_N = \lim_{k\to\infty} C_N^{(k)} = (\mathbb{E}(A))^{\alpha_N}\left[1 - \mathbb{E}(N)\mathbb{E}(A^{\alpha_N})\right]^{-1}$.

(ii) If $\mathbb{P}(N > x) = o(\mathbb{P}(B > x))$, then $\mathbb{P}(R^{(\infty)} > x) \sim C_B\mathbb{P}(B > x)$ as $x \to \infty$, where $C_B = \lim_{k\to\infty} C_B^{(k)} = \left[1 - \mathbb{E}(N)\mathbb{E}(A^{\alpha_B})\right]^{-1}$.

(iii) If $\mathbb{P}(B > x) \sim C_{BN}\mathbb{P}(N > x)$ for some constant C_{BN}, and $\mathbb{P}(N > x, B > x) = o(\mathbb{P}(N > x))$, then $\mathbb{P}(R^{(\infty)} > x) \sim C\,\mathbb{P}(N > x)$ as $x \to \infty$, where $C = \lim_{k\to\infty} C^{(k)} = \left[C_{BN} + (\mathbb{E}(A))^{\alpha_N}\right]\left[1 - \mathbb{E}(N)\mathbb{E}(A^{\alpha_N})\right]^{-1}$.

Proving these results by probabilistic methods requires an exchange of limits in x and k, which is usually a difficult technical problem. Indeed, if we assume that $\mathbb{P}(R^{(k)} > x) \sim h_k(x)$ as $x \to \infty$ for every k and some function $h_k(x)$, then $\mathbb{P}(R^{(\infty)} > x) \sim \lim_{k\to\infty} h_k(x)$ is not true in general. For instance, from Theorem 2.5 we know that the asymptotics of $R^{(k)}$ can be defined by the asymptotics of $R^{(0)}$, whereas representation (2.7) clarifies that $R^{(\infty)}$ does not depend on the distribution of $R^{(0)}$. In Section 3.3 we prove the above similarities using the Laplace-Stieltjes transforms analysis.

CHAPTER 3

LAPLACE-STIELTJES TRANSFORMS' ANALYSIS

In this chapter we define the tail behavior for our models for the distributions of the in-degree and the PageRank. To this end, we apply a technique that was used by de Meyer and Teugels [81] to obtain asymptotic behavior of the busy period in the $M/G/1$ queue. To define similarity in asymptotics between two random variables we analyze the corresponding equation for the Laplace-Stieltjes transforms. The key theorem of our analysis is a Tauberian theorem was introduced by Bingham and Doney in [16], and later in the book [17]. The theorem establishes the relation between the asymptotic behavior of a regularly varying distribution and its Laplace-Stieltjes transform. For details we refer to Section 3.1.

In Section 3.2 we start with the analysis of the model for the in-degree distribution. As it was discussed in Section 2.1.1, we define the in-degree of a random page in the Web graph as an integer random variable $N(X)$, where N is the number of Poisson(1) events on $[0, X]$, where X is a regular varying random variable. We show the asymptotic equivalence of $N(X)$ and X in two steps. First, in order to satisfy conditions of the Tauberian theorem (Theorem 3.2) we show that for some integer k : the kth moments of X exist if and only if the kth moment of $N(X)$ exist. We state this result in Lemma 3.4. Second, we define the desired tail similarity in Theorem 3.6.

In Section 3.3 we continue with stochastic model for the PageRank. As in Chapter 2, we define the distribution of the PageRank through stochastic equation (2.3). Again, instead of analyzing (2.3) we consider the general version of the stochastic equation. Here we recall (2.4):

$$R \stackrel{d}{=} \sum_{j=1}^{N} A_j R_j + B, \tag{3.1}$$

where we assume that all random variables are positive; R_j's are independent and distributed as R; and A_j's are independent and distributed as some random variable with $\mathbb{E}(A) = [1 - \mathbb{E}(B)]/\mathbb{E}(N)$. We also set R_j's and A_j's to be independent, and to be independent of N and B. Moreover, it is essential that $\mathbb{E}(B) < 1$. We emphasize that N and B can be dependent. In this chapter we need to assume that $A < 1$, and $\alpha = \min(\alpha_N, \alpha_B) > 1$ is non-integer, where α_N and α_B are power law exponents for N and B, respectively.

Based on (3.1) we define the equation for Laplace-Stieltjes transforms of N, B, and R in Section 3.3.1. To classify the asymptotic behavior of R, we first need to show that conditions of Tauberian theorem (Theorem 3.2) are satisfied. Particularly, in Lemmas 3.7 and 3.8 we justify that the existence of the kth moments of N and B implies the existence of the kth moment of R, and vice versa. Then, we define the necessary equivalences for the Laplace-Stieltjes transforms of N, B, and R in Corollary 3.9; and obtain the main results in Theorem 3.10. The theorem justifies asymptotics that are predicted in Section 2.4. Thus, the obtained tail behavior of R is determined by the asymptotics of the random variable with the heaviest tail among N and B, and is a convergence of the results of Theorem 2.4. For the case when generalized stochastic equation (3.1) servers to model the PageRank distribution we refer to Section 3.3.4.

We start with introduction of the main definitions and some facts that we use throughout this chapter.

3.1 Preliminaries

The Laplace-Stieltjes transforms analysis is one of the classical ways to study regular varying random variables. In this section we adopt definitions and result from [16]. More details can be also found in the book by Bingham et al. [17].

We denote by $f(s) = \mathbb{E}e^{-sX}$, $s > 0$, the Laplace-Stieltjes transform of X, and let $\xi_i = \int_0^\infty x^i dF_X(x)$ be the ith moment of X, where F_X is the cumulative distribution function of X. The successive moments of X can be obtained by expanding $f(s)$ in a series at $s = 0$. More precisely, we write the following:

Lemma 3.1. *The nth moment of X is finite if and only if there exist finite numbers $\xi_0 = 1$ and $\xi_1, ..., \xi_n$, such that*

$$f_n(s) = (-1)^{n+1} \left(f(s) - \sum_{i=0}^{n} \frac{\xi_i}{i!}(-s)^i \right) = o(s^n) \text{ as } s \to 0. \tag{3.2}$$

In that case, ξ_i is the ith moment of X.

The following theorem establishes a relation between the tail behavior of a regularly varying random variable and its Laplace-Stieltjes transform. We use this result in the proofs of Theorem 3.6 and 3.10.

Theorem 3.2. *(Tauberian Theorem) If* $n \in \mathbb{N}$, $\xi_n < \infty$, $\alpha \in (n, n+1)$, *then the following are equivalent*

(i) $f_n(s) \sim (-1)^n \Gamma(1-\alpha) s^\alpha L(\frac{1}{s})$ *as* $s \to 0$,

(ii) $\mathbb{P}(X > x) \sim x^{-\alpha} L(x)$ *as* $x \to \infty$.

The next lemma provides a useful bound for slowly varying functions. Here we present the version of the lemma from [117].

Lemma 3.3. *(Potter bounds) Let* L *be a slowly varying function. Then, for any fixed* $\vartheta > 1, \delta > 0$ *there exists a finite constant* $s_0 < 1$ *such that for all* $s_1, s_2 < s_0$,

$$\frac{L\left(\frac{1}{s_1}\right)}{L\left(\frac{1}{s_2}\right)} \leq \vartheta \max\left\{\left(\frac{s_1}{s_2}\right)^{\delta}, \left(\frac{s_1}{s_2}\right)^{-\delta}\right\}.$$

3.2 Asymptotic behavior of the in-degree model

We model the in-degree of a random page in the Web as an integer valued regularly varying random variable $N = N(X)$, where we assume that X is regularly varying with index α_N and $N(x)$ is the number of Poisson arrivals during the time interval $[0, x]$, when arrival rate is 1. The advantage of $N(X)$ is that we do not need to impose any restrictions on X and at the same time ensure that the in-degree is integer. We claim that the random variable $N(X)$ is regularly varying with the same index as X, or, more informally, $N(X)$ follows a power law with the same exponent. Thus, we can think of $N(X)$ as the in-degree of a random Web page. For the sake of completeness we present the formal statement and its proof in the remainder of this section.

We want to prove that $\mathbb{P}(X > x) \sim \mathbb{P}(N(X) > x)$ as $x \to \infty$. We assume that random variable X is a regularly varying with non-integer index $\alpha_X > 1$:

$$\mathbb{P}(X > x) \sim x^{-\alpha_X} L_X(x) \text{ as } x \to \infty, \tag{3.3}$$

where $L_X(x)$ is some slowly varying function. Then we show that $N(X)$ is also a regularly varying random variable with non-integer index $\alpha_N > 1$:

$$\mathbb{P}(N(X) > x) \sim x^{-\alpha_N} L_N(x) \text{ as } x \to \infty, \tag{3.4}$$

such that $\alpha_X = \alpha_N$ and $L_X(x) = L_N(x)$. We use Tauberian theorem (see Theorem 3.2), and therefore we first confirm that the corresponding moments of X and $N(X)$ always exist together.

We start with observation that $\mathbb{E}(X) = \mathbb{E}(N(X)) = \mathbb{E}(N)$. We denote by f and ϕ the Laplace-Stieltjes transforms of X and $N(X)$, respectively. Next, we consider the generating function of $N(X)$:

$$\mathbb{G}_{N(X)}(s) = \mathbb{E}s^{N(X)} = \int_0^\infty \mathbb{E}s^{N(t)} dF_X(t) = \int_0^\infty e^{-t(1-s)} dF_X(t) = f(1-s),$$

where F_X is the distribution function of X. Thus, we derive the Laplace-Stieltjes transform of $N(X)$ in terms of the Laplace-Stieltjes transform of X:

$$\phi(s) = \mathbb{E}e^{-sN(X)} = f(1 - e^{-s}). \tag{3.5}$$

Denote by $\nu_0 = 1,\ \nu_1 = \mathbb{E}(N), \ldots, \nu_n$ the first n moments of $N(X)$. Provided that ν_n are finite we define

$$\phi_n(s) = (-1)^{n+1}\left(\phi(s) - \sum_{i=0}^{n} \frac{\nu_i}{i!}(-s)^i\right)$$

as in Lemma 3.1. Then, we can prove the following lemma.

Lemma 3.4. *For $n \geq 1$, the following are equivalent*

(i) $\xi_n < \infty$,

(ii) $\nu_n < \infty$.

Proof.

$(i) \to (ii)$ From Lemma 3.1 we know that $\xi_n < \infty$ implies $f_n(y) = o\left(y^n\right)$. Consider

$$y(s) = 1 - e^{-s} = \sum_{i=1}^{n+1}(-1)^{i+1}\frac{s^i}{i!} + o(s^{n+1}),$$

then we can find $y^i(s)$:

$$y^i(s) = \sum_{j=i}^{n+i} \mu_{i,j} s^j + o\left(s^{n+i}\right)$$

for $i \geq 1$ and some appropriate constants $\mu_{i,j}, j = i, \ldots, n+i$. Thus, we easily obtain

$$\begin{aligned} f_n(y(s)) &= (-1)^{n+1}\left(f(y(s)) - \sum_{i=0}^{n}\frac{\xi_i}{i!}(-y(s))^i\right) \\ &= (-1)^{n+1}\left(\phi(s) - 1 - \sum_{i=1}^{n}\frac{\xi_i}{i!}(-1)^i\left(\sum_{j=i}^{n+i}\mu_{i,j}s^j + o\left(s^{n+i}\right)\right)\right) \\ &= (-1)^{n+1}\left(\phi(s) - \sum_{i=0}^{n}\frac{\hat{\nu}_i}{i!}(-s)^i + O\left(s^{n+1}\right)\right), \end{aligned}$$

for some finite constants $\hat{\nu}_0 = 1$ and $\hat{\nu}_1, \ldots \hat{\nu}_n$, that can be expressed in terms of $\xi_1, \ldots, \xi_n$. Thus, we find

$$\phi(s) = \sum_{i=0}^{n}\frac{\hat{\nu}_i}{i!}(-s)^i + (-1)^{n+1}f_n(y(s)) + O(s^{n+1}) = \sum_{i=0}^{n}\frac{\hat{\nu}_i}{i!}(-s)^i + o(s^n),$$

since $y(s) = s + o(s)$. By uniqueness of the power series expansion we have $\nu_i = \hat{\nu}_i$, $i = 0 \ldots n-1$, and then by Lemma 3.1 we have $\nu_n = \hat{\nu}_n < \infty$.

$(ii) \to (i)$ Similar to the first part of the proof, where we use $s(y) = -\ln(1-y)$. □

Remark 3.5. *It follows from the proof of Lemma 3.4 that if $\xi_n < \infty$, then*

$$f_n\left(1-e^{-s}\right) = \phi_n(s) + O(s^{n+1}).$$

Now, we use Theorem 3.2 to prove that (3.3) implies (3.4), and vice versa.

Theorem 3.6. *The following are equivalent*

(i) $\mathbb{P}(X > x) \sim x^{-\alpha_N} L(x) \quad as \quad x \to \infty,$

(ii) $\mathbb{P}(N(X) > x) \sim x^{-\alpha_N} L(x) \quad as \quad x \to \infty.$

Proof.

$(i) \to (ii)$ From Theorem 3.2 and (i) we know that

$$f_n(s) \sim (-1)^n \Gamma(1-\alpha_N) s^{\alpha_N} L_X\left(\frac{1}{s}\right) \quad \text{as} \quad s \to 0, \tag{3.6}$$

where $\alpha_N > 1$ is not integer and n is the largest integer smaller than α_N. From $1 - e^{-s} \sim s$ as $s \to 0$, (3.6) and Lemma 3.3 we obtain that

$$f_n(s) \sim f_n(1-e^{-s}) \text{ as } s \to 0.$$

Then, we use (3.6) and Remark 3.5 to obtain

$$\phi_n(s) \sim (-1)^n \Gamma(1-\alpha) s^{\alpha} L\left(\frac{1}{s}\right) \quad \text{as} \quad s \to 0.$$

Now we again apply Theorem 3.2 to conclude that

$$1 - F_{N(X)} \sim x^{-\alpha} L(x) \quad \text{as} \quad x \to \infty.$$

$(ii) \to (i)$ Similar to the first part of the proof. □

Thus, our model for the number of incoming links properly describes the in-degree distribution that follows a power law with finite expectation and a non-integer exponent. We use $N = N(X)$ throughout the rest of this chapter.

3.3 General stochastic equation

In this section we consider the general stochastic equation (3.1). We start with deriving of the equation for Laplace-Stieltjes transforms, that corresponds to (3.1).

3.3.1 Equation for Laplace-Stieltjes transforms

We denote the first m moments of B by $\beta_1, \beta_2, \ldots, \beta_m$, and $\beta_0 = 1$. Then, provided that β_m is finite, we define

$$b_m(s) = (-1)^{m+1} \left(b(s) - \sum_{i=0}^{m} \frac{\beta_i}{i!} (-s)^i \right), \tag{3.7}$$

where $b(s)$ is the Laplace-Stieltjes transform of B.

We introduce the following notation:

$$G(t,s) = \mathbb{E}\left(e^{-tX} e^{-sB}\right), \tag{3.8}$$

where it is easy to see that $G(t,0) = f(t)$ and $G(0,s) = b(s)$. Moreover, if X and B are independent, implying that N and B are independent, then we have

$$G(t,s) = f(t)b(s).$$

Let $r(s)$ be the Laplace-Stieltjes transform of R. Then, by (3.1) and (3.5) the following holds:

$$\begin{aligned} r(s) &= \mathbb{E}\left(e^{-sR}\right) = \mathbb{E}\left[\exp\left(-s \sum_{j=1}^{N} A_j R_j \right) e^{-sB} \right] \\ &= \mathbb{E}\left[\mathbb{E}\left(\exp\left(-s \sum_{j=1}^{N} A_j R_j \right) e^{-sB} \middle| N, B \right) \right] = G\left[1 - \mathbb{E}\left(r\left(As\right)\right), s\right]. \end{aligned}$$

Thus, we derive the next equation:

$$r(s) = G\left[1 - \mathbb{E}\left(r\left(As\right)\right), s\right]. \tag{3.9}$$

Denoting

$$t(s) = 1 - \mathbb{E}\left(r\left(As\right)\right), \tag{3.10}$$

we write (3.9) as

$$r(s) = G(t(s), s). \tag{3.11}$$

3.3.2 Auxiliary results

We define $\rho_1, \ldots, \rho_k$ to be the first k moments of R. If $\rho_k < \infty$, and we write

$$r_k(s) = (-1)^{k+1} \left(r(s) - \sum_{i=0}^{k} \frac{\rho_i}{i!} (-s)^i \right), \tag{3.12}$$

as in Lemma 3.1.

Next, we denote $k = \min(m, n)$, where m and n are integer, and such that $\beta_m = \mathbb{E}(B^m) < \infty$ and $\xi_n = \mathbb{E}(X^n) < \infty$. Further, we assume that $\mathbb{E}(X^j B^{k+1-j}) < \infty$ for all $1 \le j \le k$. We note that because of $\mathbb{E}(X) < \infty$ and $\mathbb{E}(B) < \infty$ this assumption is always true in the case of the independent N and B. However, this assumption is much weaker than independence. Then we can prove the following lemma.

Lemma 3.7. *If $\xi_n < \infty$ and $\beta_m < \infty$ for some integer $m, n \ge 1$, and $\mathbb{E}(X^j B^{k+1-j}) < \infty$ for all $1 \le j \le k$, where $k = \min(m, n)$, then $\rho_k < \infty$.*

Proof. We use induction, starting from $k = 1$ for which the statement is valid. Assume that for $i = 1, 2, \ldots, k-1$, lemma has been proved, so we can use the following expansion:

$$r(s) = 1 - s + \sum_{i=2}^{k-1} \frac{\rho_i}{i!}(-s)^i + o(s^{k-1}),$$

to present $t(s)$ as a sum

$$t(s) = -\mathbb{E}\left(\sum_{i=1}^{k-1} \frac{\rho_i}{i!} A^i (-s)^i + o(s^{k-1})\right) = -\sum_{i=1}^{k-1} \frac{\rho_i}{i!}\mathbb{E}(A^i)(-s)^i + o(s^{k-1}).$$

As the result of this, we can actually obtain $t^i(s)$:

$$t^i(s) = \sum_{j=i}^{k+i-2} \zeta_{i,j} s^j + o(s^{k+i-2}), \tag{3.13}$$

for $i \ge 1$ and some appropriate constants $\zeta_{i,j},\ j = i, \ldots, k+i-2$.

Now, we consider the Taylor expansion of $G(t(s), s)$:

$$\begin{aligned} G(t(s), s) &= \left[\sum_{i=0}^{k} \frac{\xi_i}{i!}(-t(s))^i + (-1)^{k+1} f_k\left(t(s)\right)\right] \\ &+ \left[\sum_{i=0}^{k} \frac{\beta_i}{i!}(-s)^i + (-1)^{k+1} b_k(s)\right] \\ &- 1 + \sum_{i=0}^{k+1} \frac{(-1)^i}{i!} \sum_{j=1}^{i-1} \binom{i}{j} \mathbb{E}\left(X^j B^{i-j}\right) t^j(s) s^{i-j} + o(s^{k+1}), \end{aligned} \tag{3.14}$$

where $t(s) \sim \mathbb{E}(A)s$. Here we use that $G'_{t^j s^{i-j}}(0,0) = (-1)^i \mathbb{E}\left(X^j B^{i-j}\right) < \infty$ for all $0 \le i \le k+1$ and $0 < j < k+1$. Then, from (3.10), (3.11), and (3.14), we obtain

the following:

$$r(s) = 1 - \mathbb{E}(N)t(s) + \left[\sum_{i=2}^{k} \frac{\xi_i}{i!}(-t(s))^i + (-1)^{k+1} f_k\left(t(s)\right)\right] + \left[\sum_{i=0}^{k} \frac{\beta_i}{i!}(-s)^i \right.$$
$$\left. +(-1)^{k+1} b_k(s)\right] - 1 + \sum_{i=0}^{k+1} \frac{(-1)^i}{i!} \sum_{j=1}^{i-1} \binom{i}{j} \mathbb{E}\left(X^j B^{i-j}\right) t^j(s) s^{i-j} + o(s^{k+1})$$
$$= 1 - \mathbb{E}(N)\left[1 - \mathbb{E}(r(As))\right] + \sum_{i=1}^{k} \eta_i s^i + o(s^k),$$

where we use (3.13), $f_k(t(s)) = o(s^k)$, and $b_k(s) = o(s^k)$ to find the appropriate constants $\eta_1, \ldots, \eta_k$. Next, we rewrite the last equation

$$r(s) - \mathbb{E}(N)\mathbb{E}(r(As)) = 1 - \mathbb{E}(N) + \sum_{i=1}^{k} \eta_i s^i + o(s^k),$$

and apply (3.12) to obtain the following:

$$r_{k-1}(s) - \mathbb{E}(N)\mathbb{E}\left(r_{k-1}\left(As\right)\right) + (-1)^k \sum_{i=0}^{k-1} \frac{\rho_i}{i!}(1 - \mathbb{E}(A^i))(-s)^i = 1 - \mathbb{E}(N)$$
$$+ \sum_{i=0}^{k} \eta_i s^i + o(s^k).$$

Because $r_{k-1}(s) = o(s^{k-1})$, $\mathbb{E}(r_{k-1}(As)) = o(s^{k-1})$ and the uniqueness of the series expansion, we can remove all powers up to k:

$$r_{k-1}(s) - \mathbb{E}(N)\mathbb{E}\left(r_{k-1}\left(As\right)\right) = \eta_k s^k + o(s^k). \tag{3.15}$$

Now, we let $A_1, A_2 \ldots$ be independent and distributed as A. We consider the following partial sums

$$\sum_{j=0}^{M} (\mathbb{E}(N))^j \left[\mathbb{E}\left(r_{k-1}(A_1 \ldots A_j s)\right) - \mathbb{E}(N)\, \mathbb{E}\left(r_{k-1}(A_1 \ldots A_{j+1} s)\right)\right]$$
$$= r_{k-1}(s) - (\mathbb{E}(N))^{M+1} \mathbb{E}(r_{k-1}(A_1 \ldots A_{M+1} s))$$

We claim that the second term converges to 0 as $M \to \infty$. From induction hypothesis and the definition of $o(s^{k-1})$, for all $\varepsilon > 0$, there exists a $\delta = \delta(\varepsilon)$ such that $|r_{k-1}(s)| < \varepsilon s^{k-1}$ whenever $0 < s \leq \delta$. Fix some ε and take $\delta = \delta(\varepsilon)$. Then the following holds:

$$\mathbb{E}\left|r_{k-1}\left(A_1 \ldots A_{M+1} s\right)\right| < \varepsilon s^{k-1} \mathbb{E}\left(A_1^{k-1} \ldots A_{M+1}^{k-1}\right) = \varepsilon s^{k-1} \left(\mathbb{E}(A^{k-1})\right)^{M+1},$$

where final equation holds because of the independence of the A's. Taking the limit as $M \to \infty$, since $\mathbb{E}(B) < 1$, $A < 1$, $\mathbb{E}(A) = (1 - \mathbb{E}(B))/\mathbb{E}(N)$ and $\mathbb{E}(A^{n-1}) \leq \mathbb{E}(A)$ we have $\lim_{M\to\infty} \mathbb{E}(N)^{M+1}\mathbb{E}(r_{k-1}(A_1 \dots A_{M+1}s)) = 0$. It follows that we can express $r_{k-1}(s)$ as an infinite sum:

$$r_{k-1}(s) = \sum_{j=0}^{\infty} (\mathbb{E}(N))^j \left[\mathbb{E}\left(r_{k-1}(A_1 \dots A_j s)\right) - \mathbb{E}(N)\,\mathbb{E}\left(r_{k-1}(A_1 \dots A_{j+1}s)\right)\right], \quad (3.16)$$

where we can apply (3.15) to each of the terms. Form the definition of $o(s^k)$, for every $\varepsilon > 0$, there exists a $\delta = \delta(\varepsilon)$ such that

$$\left|r_{k-1}(s) - \mathbb{E}(N)\mathbb{E}\left(r_{k-1}\left(As\right)\right) - \eta_k s^k\right| < \varepsilon s^k$$

whenever $0 < s \leq \delta$. Moreover, for this ε and $0 < s \leq \delta$, we also have

$$\begin{aligned}
&\left|\mathbb{E}\left(r_{k-1}\left(A_1 \dots A_j s\right)\right) - \mathbb{E}(N)\mathbb{E}\left(r_{k-1}\left(A_1 \dots A_{j+1}s\right)\right) - \eta_k s^k \mathbb{E}\left(A_1^k \dots A_j^k\right)\right| \\
&\leq \mathbb{E}\left|\mathbb{E}\left[r_{k-1}\left(A_1 \dots A_j s\right) - \mathbb{E}(N) r_{k-1}\left(A_1 \dots A_{j+1}s\right) - \eta_k s^k A_1^k \dots A_j^k\right.\right. \\
&\left.\left.\mid A_1, \dots, A_j\right]\right| < \varepsilon s^k (\mathbb{E}(A^k))^j,
\end{aligned}$$

for every $j \geq 0$ and $A_1, \dots, A_{j+1}$, which are independent and distributed as A. Here the last inequality holds because $A < 1$, and then $0 < A_1 \dots A_{j+1}s \leq s < \delta$ for every $j \geq 0$. Using the representation of $r_{k-1}(s)$ as an infinite sum, (3.16), we obtain

$$\begin{aligned}
&\left|r_{k-1}(s) - \eta_k \sum_{j=0}^{\infty} (\mathbb{E}(N))^j \mathbb{E}\left(A_1^k \dots A_j^k\right) s^k\right| \\
&= \left|\sum_{j=0}^{\infty} (\mathbb{E}(N))^j \left[\mathbb{E}\left(r_{k-1}(A_1 \dots A_j s)\right) - \mathbb{E}(N)\mathbb{E}\left(r_{k-1}(A_1 \dots A_{j+1}s)\right)\right]\right. \\
&\left. -\eta_k \sum_{j=0}^{\infty} (\mathbb{E}(N))^j \mathbb{E}\left(A_1^k \dots A_j^k\right) s^k\right| \leq \varepsilon s^k \sum_{j=1}^{\infty} (\mathbb{E}(N)\mathbb{E}(A^k))^j \\
&= \varepsilon [1 - \mathbb{E}(N)\mathbb{E}(A^k)]^{-1} s^k.
\end{aligned}$$

Thus, we have shown that $r_{k-1}(s) - \eta_k[1 - \mathbb{E}(N)\mathbb{E}(A^k)]^{-1}s^k = o(s^k)$. Taking $\rho_k = -\eta_k[1 - \mathbb{E}(N)\mathbb{E}(A^k)]^{-1}$, from Lemma 3.1 and the last equation we conclude that ρ_k is the kth moment of R and it is finite. □

We can also proof the conversed lemma.

Lemma 3.8. *If $\rho_k < \infty$, $k \geq 1$, then $\xi_k < \infty$ and $\beta_k < \infty$.*

Proof. Let R be non-negative random variable, that satisfies (3.1) and has finite kth moment. Equation (3.1) implies that R is stochastically greater than B, and thus R is also stochastically greater than $B(AN(X)+1)$. Hence, the existence of the kth moment of R ensures the existence of the kth moment of B and $N(X)$, which in turn ensures the existence of the kth moment of X. □

The next Corollary follows from the proof of Lemma 3.7.

Corollary 3.9. *It follows from Lemma 3.7 that*

(*i*) *if* $n < m$, *then* $r_n(s) - \mathbb{E}(N)\mathbb{E}\left(r_n\left(As\right)\right) = f_n(t(s)) + O(s^{n+1})$.

(*ii*) *if* $n > m$, *then* $r_m(s) - \mathbb{E}(N)\mathbb{E}\left(r_m\left(As\right)\right) = b_m(s) + O(s^{m+1})$.

(*iii*) *if* $n = m$, *then* $r_n(s) - \mathbb{E}(N)\mathbb{E}\left(r_n\left(As\right)\right) = f_n(t(s)) + b_n(s) + O(s^{n+1})$.

Proof. Recall k to be $\min(m,n)$. Because $r_k(s) = o(s^k)$ we can consider the following expansion of (3.13):

$$t^i(s) = \sum_{j=l}^{k+i-1} \zeta_{i,j} s^j + o(s^{k+i-1}), \tag{3.17}$$

for $i \geq 1$ and appropriate constants $\zeta_{i,j}, j = i, \ldots, k+i-1$.

From (3.11), (3.14), (3.17), the definitions of $r_k(s)$, $b_k(t)$, $t(s)$, Lemma 3.7, it follows that

$$\begin{aligned}
&(-1)^{k+1} r_k(s) + \sum_{i=0}^{k} \frac{\rho_i}{i!}(-s)^i = \left[(-1)^{k+1} f_k\left(t(s)\right) + \sum_{i=2}^{k} \frac{\xi_i}{i!}(-t(s))^i + 1\right. \\
&\left. -\mathbb{E}(N)\left[1 - \mathbb{E}\left((-1)^{k+1} r_k(As) + \sum_{i=0}^{k} \frac{\rho_i}{i!}(-As)^i\right)\right]\right] - 1 + \left[\sum_{i=0}^{k} \frac{\beta_i}{i!}(-s)^i\right. \\
&\left. +(-1)^{k+1} b_k(s)\right] + \sum_{i=0}^{k+1} \frac{(-1)^i}{i!} \sum_{j=1}^{i-1} \binom{i}{j} \mathbb{E}\left(X^j B^{i-j}\right) t^j(s) s^{i-j} + o(s^{k+1}) \\
&= (-1)^{k+1}[b_k(s) + f_k(t) + \mathbb{E}(N)\mathbb{E}(r_k(As))] + \sum_{i=0}^{k+1} \varsigma_i s^i + o(s^{k+1}),
\end{aligned}$$

where $\varsigma_0, \ldots, \varsigma_{k+1}$ are appropriate constants. Due to the uniqueness of the series expansion, we can reduce the above formula to

$$r_k(s) = b_k(s) + f_k(t) + \mathbb{E}(N)\mathbb{E}(r_k(As)) + (-1)^{k+1}\varsigma_{k+1} s^{k+1} + o(s^{k+1}).$$

The corollary follows because $t(s) \sim \mathbb{E}(A)s$ as $s \to 0$. □

Now we are ready to prove our main result.

3.3.3 Main theorem

In the next theorem we obtain our main result that establishes the tail behavior of the solution of the general stochastic equation (3.1). In particular, for $A \stackrel{d}{=} c/D$ and $B \stackrel{d}{=} cp_0 + (1-c)wT$, we can derive asymptotics for the PageRank distribution under various assumptions on the distribution of the in-degree and the teleportation (see Section 3.3.4).

Theorem 3.10. *(i) if* $\mathbb{P}(B > x) = o(\mathbb{P}(N > x))$, *then the following are equivalent:*

(i.1) $\mathbb{P}(N > x) \sim x^{-\alpha_N} L_N(x)$ *as* $x \to \infty$,

(i.2) $\mathbb{P}(R > x) \sim C_N x^{-\alpha_N} L_N(x)$ *as* $x \to \infty$, *where* $C_N = (E(A))^{\alpha_N}[1 - \mathbb{E}(N)\mathbb{E}(A^{\alpha_N})]^{-1}$;

(ii) if $\mathbb{P}(N > x) = o(\mathbb{P}(B > x))$, *then the following are equivalent:*

(ii.1) $\mathbb{P}(B > x) \sim x^{-\alpha_B} L_B(x)$ *as* $x \to \infty$,

(ii.2) $\mathbb{P}(R > x) \sim C_B x^{-\alpha_B} L_B(x)$ *as* $x \to \infty$, *where* $C_B = [1 - \mathbb{E}(N)\mathbb{E}(A^{\alpha_B})]^{-1}$;

(iii) if $\mathbb{P}(B > x) \sim C_{BN}\mathbb{P}(N > x)$, *then the following are equivalent:*

(iii.1) $\mathbb{P}(N > x) \sim x^{-\alpha_N} L_N(x)$, *and* $\mathbb{P}(B > x) \sim x^{-\alpha_N} L_B(x)$ $\sim C_{BN} x^{-\alpha_N} L_N(x)$ *as* $x \to \infty$,

(iii.2) $\mathbb{P}(R > x) \sim C x^{-\alpha_N} L_N(x)$ *as* $x \to \infty$, *where* $C = [C_{BN} + (\mathbb{E}(A))^{\alpha_N}] \times [1 - \mathbb{E}(N)\mathbb{E}(A^{\alpha_N})]^{-1}$.

The results of Theorem 3.10 describe the tail behavior of R under various assumptions on the distribution of N and B. First of all, we observe that the power law exponent is defined by the random variable with the heaviest tail among N and B, representing the in-degree and the user preference, respectively. Next, we see that the obtained multiplicative constants agree with the results of Section 2.4. When B has a lighter tail than N, we observe that the distribution of B has no influence on the asymptotics of R. In the next case we find that C_B only depends on the distribution of N only through its mean, and in the case of the similar tails of N and B we have the effects from both of them.

Proof of Theorem 3.10.

$(i, ii, iii.1) \Rightarrow (i, ii, iii.2)$ It follows from $(i, ii, iii.1)$ and Theorem 3.2 that

(i) $f_n(t) \sim (-1)^n \Gamma(1-\alpha_N) t^{\alpha_N} L_N\left(\frac{1}{t}\right)$ as $t \to 0$;

(ii) $b_m(s) \sim (-1)^m \Gamma(1-\alpha_B) s^{\alpha_B} L_B\left(\frac{1}{s}\right)$ as $s \to 0$;

(iii) both previous equivalences,

where m and n are the largest integer values not exceeding α_B and α_N, respectively.

Recall that $t(s) \sim \mathbb{E}(A)s$ as $s \to 0$, because of (3.10) and $r(s) = 1 - s + o(s)$. Then, by applying Corollary 3.9 we can obtain as $s \to 0$:

$$(i)\ \ r_n(s) - \mathbb{E}(N)\mathbb{E}\left(r_n\left(As\right)\right) \sim (-1)^n\Gamma(1-\alpha_N)(\mathbb{E}(A))^{\alpha_N}L_N\left(\tfrac{1}{s}\right)s^{\alpha_N}$$

$$(ii)\ \ r_m(s) - \mathbb{E}(N)\mathbb{E}\left(r_m\left(As\right)\right) \sim (-1)^m\Gamma(1-\alpha_B)L_B\left(\tfrac{1}{s}\right)s^{\alpha_B}$$

$$(iii)\ \ r_n(s) - \mathbb{E}(N)\mathbb{E}\left(r_n\left(As\right)\right) \sim (-1)^n\Gamma(1-\alpha_N)\left[(\mathbb{E}(A))^{\alpha_N}L_N\left(\tfrac{1}{s}\right) + L_B\left(\tfrac{1}{s}\right)\right]s^{\alpha_N}.$$

Let V_N and V_B be constants that are defined as follows:
(i) $V_N = (\mathbb{E}(A))^{\alpha}$ and $V_B = 0$;
(ii) $V_N = 0$ and $V_B = 1$;
(iii) $V_N = (\mathbb{E}(A))^{\alpha}$ and $V_B = 1$.
Next, we denote

$$Z(s) = r_k(s) - \mathbb{E}(N)\mathbb{E}\left(r_k\left(As\right)\right),$$
$$Y(s) = (-1)^k\Gamma(1-\alpha)\left[V_N L_N\left(\frac{1}{s}\right) + V_B L_B\left(\frac{1}{s}\right)\right]s^{\alpha},$$

where $\alpha = \min(\alpha_N, \alpha_B)$, and $k = \min(n, m)$. We note that $Y(s) \geq 0$ for every $s > 0$.

We prove the statement of the theorem in two steps. First, we use the representation (3.16) for $r_k(s)$, and show that the following asymptotic similarity holds:

$$\sum_{i=0}^{\infty}(\mathbb{E}(N))^i\mathbb{E}\left(Z\left(A_1\ldots A_i s\right)\right) \sim \sum_{i=0}^{\infty}(\mathbb{E}(N))^i\mathbb{E}\left(Y\left(A_1\ldots A_i s\right)\right), \tag{3.18}$$

as $s \to 0$. Second, we demonstrate that the right-hand side of (3.18) has the desired asymptotics.

As we saw above, $Z(s) \sim Y(s)$ as $s \to 0$. Then, for every $\varepsilon > 0$, there exists a $\delta = \delta(\varepsilon)$ such that $|Z(s)/Y(s) - 1| < \varepsilon$ whenever $0 < s \leq \delta$. We fix some ε and take $\delta = \delta(\varepsilon)$. Now again let $A_1, A_2, \ldots$ be independent random variables, which are distributed as A. Because $A < 1$, and then $0 < A_1 \ldots A_i s \leq s \leq \delta$, for every $i \geq 0$ we have

$$\left|\frac{Z(A_1\ldots A_i s)}{Y(A_1\ldots A_i s)} - 1\right| < \varepsilon. \tag{3.19}$$

From (3.19) we obtain the following:

$$\begin{aligned}
&\left|\frac{\sum_{i=0}^{\infty}(\mathbb{E}(N))^i\mathbb{E}\left(Z\left(A_1\dots A_is\right)\right)}{\sum_{i=0}^{\infty}(\mathbb{E}(N))^i\mathbb{E}\left(Y\left(A_1\dots A_is\right)\right)}-1\right|\\
&\leq\frac{\sum_{i=0}^{\infty}(\mathbb{E}(N))^i\left|\mathbb{E}\left[Z\left(A_1\dots A_is\right)-Y\left(A_1\dots A_is\right)\right]\right|}{\left|\sum_{i=0}^{\infty}(\mathbb{E}(N))^i\mathbb{E}\left(Y\left(A_1\dots A_is\right)\right)\right|}\\
&\leq\frac{\sum_{i=0}^{\infty}(\mathbb{E}(N))^i\mathbb{E}\left[\left|\frac{Z(A_1\dots A_is)}{Y(A_1\dots A_is)}-1\right|Y\left(A_1\dots A_is\right)\right]}{\sum_{i=0}^{\infty}(\mathbb{E}(N))^i\mathbb{E}\left(Y\left(A_1\dots A_is\right)\right)}\\
&<\frac{\varepsilon\sum_{i=0}^{\infty}(\mathbb{E}(N))^i\mathbb{E}\left(Y\left(A_1\dots A_is\right)\right)}{\sum_{i=0}^{\infty}(\mathbb{E}(N))^i\mathbb{E}\left(Y\left(A_1\dots A_is\right)\right)}=\varepsilon,
\end{aligned}$$

which implies (3.18).

Next, we use Lemma 3.3, and then for every $\vartheta>1$ and $\delta>0$ we can find finite constants s_B and s_N such that for all $i>0$ and $0<s<\min(s_B,s_N)$,

$$\begin{aligned}
\vartheta^{-1}\left(A_1\dots A_i\right)^{\delta}&\leq\frac{L_B\left(\frac{1}{A_1\dots A_is}\right)}{L_B\left(\frac{1}{s}\right)}\leq\vartheta\left(A_1\dots A_i\right)^{-\delta},\text{ and}\\
\vartheta^{-1}\left(A_1\dots A_i\right)^{\delta}&\leq\frac{L_N\left(\frac{1}{A_1\dots A_is}\right)}{L_N\left(\frac{1}{s}\right)}\leq\vartheta\left(A_1\dots A_i\right)^{-\delta}.
\end{aligned}\tag{3.20}$$

We divide the right-hand side of (3.18) by $L_B\left(\frac{1}{s}\right)L_N\left(\frac{1}{s}\right)$, and apply (3.20) to $Y(A_1\dots A_is)/L_B\left(\frac{1}{s}\right)L_N\left(\frac{1}{s}\right)$ to obtain the following:

$$\begin{aligned}
&\vartheta^{-1}(-1)^k\Gamma(1-\alpha)\left(\frac{V_N}{L_B(\frac{1}{s})}+\frac{V_B}{L_N(\frac{1}{s})}\right)s^{\alpha}\sum_{i=0}^{\infty}(\mathbb{E}(N))^i\mathbb{E}\left(\left(A_1\dots A_i\right)^{\alpha+\delta}\right)\\
&\quad\leq\frac{\sum_{i=0}^{\infty}(\mathbb{E}(N))^i\mathbb{E}\left(Y\left(A_1\dots A_is\right)\right)}{L_B\left(\frac{1}{s}\right)L_N\left(\frac{1}{s}\right)}\\
&\quad\leq\vartheta(-1)^k\Gamma(1-\alpha)\left(\frac{V_N}{L_B(\frac{1}{s})}+\frac{V_B}{L_N(\frac{1}{s})}\right)s^{\alpha}\sum_{i=0}^{\infty}(\mathbb{E}(N))^i\mathbb{E}\left(\left(A_1\dots A_i\right)^{\alpha-\delta}\right).
\end{aligned}$$

Because $A_1,A_2\dots$ are independent and identically distributed as A we can conclude the following:

$$\begin{aligned}
&\vartheta^{-1}(-1)^k\Gamma(1-\alpha)\left(\frac{V_N}{L_B(\frac{1}{s})}+\frac{V_B}{L_N(\frac{1}{s})}\right)s^{\alpha}\frac{1}{1-\mathbb{E}(N)\mathbb{E}\left(A^{\alpha+\delta}\right)}\\
&\quad\leq\frac{\sum_{i=0}^{\infty}(\mathbb{E}(N))^i\mathbb{E}\left(Y\left(A_1\dots A_is\right)\right)}{L_B\left(\frac{1}{s}\right)L_N\left(\frac{1}{s}\right)}\\
&\quad\leq\vartheta(-1)^k\Gamma(1-\alpha)\left(\frac{V_N}{L_B(\frac{1}{s})}+\frac{V_B}{L_N(\frac{1}{s})}\right)s^{\alpha}\frac{1}{1-\mathbb{E}(N)\mathbb{E}\left(A^{\alpha-\delta}\right)}.
\end{aligned}$$

Taking $\vartheta \to 1$ and $\delta \to 0$ by the dominated convergence we obtain

$$\sum_{i=0}^{\infty}(\mathbb{E}(N))^i\mathbb{E}\left(Y\left(A_1\dots A_i s\right)\right) \sim (-1)^k\Gamma(1-\alpha)[1-\mathbb{E}(N)\mathbb{E}(A^\alpha)]^{-1}$$
$$\times\left(\frac{V_N}{L_B(\frac{1}{s})}+\frac{V_B}{L_N(\frac{1}{s})}\right)L_B\left(\frac{1}{s}\right)L_N\left(\frac{1}{s}\right)s^\alpha \text{ as } s\to 0.$$

Combining the last equivalence, (3.18), and the infinite-sum representation (3.16) for $r_k(s)$:

$$r_k(s)=\sum_{i=0}^{\infty}(\mathbb{E}(N))^i\left[\mathbb{E}\left(r_k\left(A_1\dots A_i s\right)\right)-\mathbb{E}(N)\mathbb{E}\left(r_k\left(A_1\dots A_{i+1}\right)\right)\right], \tag{3.21}$$

we then obtain

$$r_k(s)\sim(-1)^k\Gamma(1-\alpha)\left[V_N L_N\left(\frac{1}{s}\right)+V_B L_B\left(\frac{1}{s}\right)\right][1-\mathbb{E}(N)\mathbb{E}(A^\alpha)]^{-1}s^\alpha \tag{3.22}$$

as $s\to 0$. Now, we again apply Theorem 3.2 that leads to the statement of the theorem.

$(i,ii,iii.1)\Leftarrow(i,ii,iii.2)$ We denote V_N and V_B, $k=\min(n,m)$, and $\alpha\in(k,k+1)$, as before. Then, from $(i,ii,iii.2)$, and Theorem 3.2 we can obtain (3.22), that leads to the asymptotic equivalence:

$$r_k(s)-\mathbb{E}(N)\mathbb{E}(r_k(As))\sim(-1)^k\Gamma(1-\alpha)L\left(\frac{1}{s}\right)[1-\mathbb{E}(N)\mathbb{E}(A^\alpha)]^{-1}s^\alpha, \tag{3.23}$$

as $s\to 0$, where we denote

$$L\left(\frac{1}{s}\right)=V_N\left[L_N\left(\frac{1}{s}\right)-\mathbb{E}(N)\mathbb{E}\left(A^\alpha L_N\left(\frac{1}{As}\right)\right)\right]$$
$$+V_B\left[L_B\left(\frac{1}{s}\right)-\mathbb{E}(N)\mathbb{E}\left(A^\alpha L_B\left(\frac{1}{As}\right)\right)\right]$$

Next, we again use bounds (3.20) to obtain

$$\left[\frac{V_N}{L_B\left(\frac{1}{s}\right)}+\frac{V_B}{L_N\left(\frac{1}{s}\right)}\right]\left[1-\vartheta^{-1}\mathbb{E}(N)\mathbb{E}(A^{\alpha+\delta})\right]\le\frac{L\left(\frac{1}{s}\right)}{L_N\left(\frac{1}{s}\right)L_B\left(\frac{1}{s}\right)}$$
$$\le\left[\frac{V_N}{L_B\left(\frac{1}{s}\right)}+\frac{V_B}{L_N\left(\frac{1}{s}\right)}\right]\left[1-\vartheta\mathbb{E}(N)\mathbb{E}(A^{\alpha-\delta})\right]$$

Thus, by the dominated convergence for $\vartheta\to 1$ and $\delta\to 0$ we have

$$L\left(\frac{1}{s}\right)\sim[1-\mathbb{E}(N)\mathbb{E}(A^\alpha)]\left[C_N L_N\left(\frac{1}{s}\right)+C_B L_B\left(\frac{1}{s}\right)\right].$$

From last similarity and (3.23) we obtain

$$r_k(s) - \mathbb{E}(N)\mathbb{E}(r(As)) \sim (-1)^k \Gamma(1-\alpha) \left[V_N L_N\left(\frac{1}{s}\right) + V_B L_B\left(\frac{1}{s}\right)\right] s^{\alpha},$$

as $s \to 0$, from where by applying Corollary 3.9 we show $(i, ii, iii.1)$. □

3.3.4 Tail behavior of the PageRank distribution

Recall that for $A \stackrel{d}{=} c/D$ and $B \stackrel{d}{=} cp_0 + (1-c)wT$, where w is the number of pages, equation (3.1) serves a stochastic model for the non-uniform PageRank. Then, from Theorem 3.10 we obtain the following equivalences.

Corollary 3.11. *(i) If in-degree N follows power law with exponent α_N, and $\mathbb{P}(wT > x) = o(\mathbb{P}(N > x))$, then $\mathbb{P}(R > x) \sim C_N \mathbb{P}(N > x)$ as $x \to \infty$, where*

$$C_N = \frac{c^{\alpha_N}(1-p_0)^{\alpha_N}}{(\mathbb{E}(N))^{\alpha_N}(1 - c^{\alpha_N}\mathbb{E}(N)\mathbb{E}(1/D^{\alpha_N}))}.$$

(ii) If normalized teleportation jump wT follows power law with exponent α_T, and $\mathbb{P}(N > x) = o(\mathbb{P}(wT > x))$, then $\mathbb{P}(R > x) \sim C_T \mathbb{P}(wT > x)$ as $x \to \infty$, where

$$C_T = \frac{(1-c)^{\alpha_T}}{(1 - c^{\alpha_T}\mathbb{E}(N)\mathbb{E}(1/D^{\alpha_T}))}.$$

(iii) If N and wT follow power law with the same exponent α_N, and $\mathbb{P}(wT > x) \sim C_{BN}(1-c)^{-\alpha_N}\mathbb{P}(N > x)$, then $\mathbb{P}(R > x) \sim C\mathbb{P}(N > x)$ as $x \to \infty$, where

$$C = \left[\frac{(\mathbb{E}(N))^{\alpha_N} C_{BN} + c^{\alpha_N}(1-p_0)^{\alpha_N}}{(\mathbb{E}(N))^{\alpha_N}(1 - c^{\alpha_N}\mathbb{E}(N)\mathbb{E}(1/D^{\alpha_N}))} \right].$$

Consider the case when normalized teleportation has a lighter tail than the in-degree. From Corollary 3.11(i) we note that the teleportation distribution has no influence on the PageRank distribution. Then, in this case we claim that tail behavior of the non-uniform PageRank (1.3) is the same as the tail behavior of the standard PageRank (1.2). Furthermore, from the Jensen's inequality $\mathbb{E}(1/D^{\alpha_N}) \geq (\mathbb{E}(1/D))^{\alpha_N} = [(1-p_0)/\mathbb{E}(N)]^{\alpha_N}$, it follows that

$$C_N \geq \frac{c^{\alpha_N}(1-p_0)^{\alpha_N}}{(\mathbb{E}(N))^{\alpha_N}[1 - c^{\alpha_N}(1-p_0)^{\alpha_N}(\mathbb{E}(N))^{1-\alpha_N}]} = C'_N. \tag{3.24}$$

The last expression is the value of C_N in case when the out-degree of all non-dangling nodes is a constant $\mathbb{E}(N)/(1-p_0)$ as in (1.7). If $\alpha_N = 1.1$, then the difference

between the left- and the right-hand sides of (3.24) is really small for any reasonable out-degree distribution. If we also ignore the term $c^{\alpha_N}(1-p_0)^{\alpha_N}(\mathbb{E}(N))^{1-\alpha_N}$ in Corollary 3.11(i), then C_N can be approximated from above as follows

$$C_N \geq \frac{c^{\alpha_N}(1-p_0)^{\alpha_N}}{(\mathbb{E}(N))^{\alpha_N}} = c^{\alpha_N}\left[\mathbb{E}\left(\frac{1}{D}\right)\right]^{\alpha_N} = C_N^{''}.$$

Note that the asymptotic equivalence $\mathbb{P}(R > x) \sim C_N^{''}\mathbb{P}(N > x)$ as $x \to \infty$ holds if we assume that the values of the PageRank R can be approximated by $cN\mathbb{E}(1/D)$ as proposed in Fortunato et al. [49]. Furthermore, we can repeat a similar reasoning for (iii) to obtain

$$C \geq \frac{(\mathbb{E}(N))^{\alpha_N}C_{NB} + c^{\alpha_N}(1-p_0)^{\alpha_N}}{(\mathbb{E}(N))^{\alpha_N}[1 - c^{\alpha_N}(1-p_0)^{\alpha_N}(\mathbb{E}(N))^{1-\alpha_N}]} \geq C_{NB} + c^{\alpha_N}\left[\mathbb{E}\left(\frac{1}{D}\right)\right]^{\alpha_N}.$$

In Section 4.2 and 4.3 we verify the obtained asymptotics for the PageRank in several sets of Web graph data.

Since the in-degree distribution follows power law with $\alpha_N = 1.1$, the examples of the case when teleportation has the heaviest tail, are difficult to examine because it is indistinguishable from the case (iii), or because the first moment of wT does not exist. In Section 4.2 we provide an example for the non-uniform PageRank with teleportation that follows power law with exponent $\alpha_T = 0.5$ (see Figure 4.5(d)), where we can clearly see that the PageRank tends to follow a power law with the same exponent as the teleportation distribution.

CHAPTER 4

NUMERICAL RESULTS AND SPECIAL CASES

In this chapter we provide empirical justification for the results obtained in Chapters 2 and 3. To this end, we perform a number of experiments on the Web and the Wikipedia data sets, and on preferential attachment graphs, that are commonly used for modeling graphs with power law degree distribution (see details in Section 1.3.3).

We start with evaluation of power laws in the Web graph data. Despite that the power law behavior is well studied in many real-life networks, e.g. Internet graph [47], the World Wide Web [24], and citation graphs [98], the conclusion on whether or not the data follows a power law is often seem to be made purely by determining whether or not the log-log plot resembles the signature straight line. However this can be misleading especially when a size-frequency plot is used [72]. Although one may agree with Li *et al.* [72] that a cumulative (size-rank) plot is enough to reveal a power law to an experienced eye, for more reliable conclusions on realistic noisy data, we need more than just a glance at the log-log plots. Chakrabarti and Faloutsos [28] mention two goodness-of-fit methods for Pareto distribution and suggest that such methods should be applied more often. In Section 4.1 we aim at resolving these issues by using several state of the art techniques from the statistical analysis of heavy tails, cf. the recent book of Resnick [100]. Using *QQ plots*, *Hill* and *altHill plots,* and *Pickands plots* we evaluate that in-degree and PageRank follow power laws with similar exponents for various data sets.

Next, we study how well the tail behavior of the PageRank is predicted by our stochastic model. In Section 4.2 we compute the non-uniform PageRank (1.3) for the various distributions of the teleportation jump, and compare it with the results of Section 3.3.4. We confirm that the PageRank distribution tends to follow a power law with the same exponent as the exponent of the heaviest distribution among in-degree and teleportation. Moreover, we also perform experiments for some special cases. In Section 4.3 we consider distribution of the standard PageRank (1.2) on large

data sets. In particularly, we study the asymptotic behavior of the PageRank after the first, the second, and the last iterations, and observe that our model correctly captures the dynamics of the PageRank distribution in successive of power iterations.

The numerical results show a good agreement with our stochastic model for the PageRank distribution.

Finally, inspired by the minor effect of the out-degree distribution on the asymptotics of the PageRank, we propose a new ranking scheme in Section 4.4. We call this scheme as *Pure Authority Rank (PAR),* and define it as a modification of the PageRank (1.2) where we assume that the number of outgoing links of all pages is a constant, and equals to the average in- and out-degree. Note that PAR rank can be modeled by stochastic equation (1.6). We compute the PAR rank for the Wikipedia and preferential attachment graph, and again observe the similarity in the asymptotics of the PAR rank and the in-degree. Moreover, we also compare the PAR rank with PageRank, and discover that the PAR rank behaves similarly to the PageRank computed for a higher value of c, and converges faster.

4.1 Evaluation of power laws

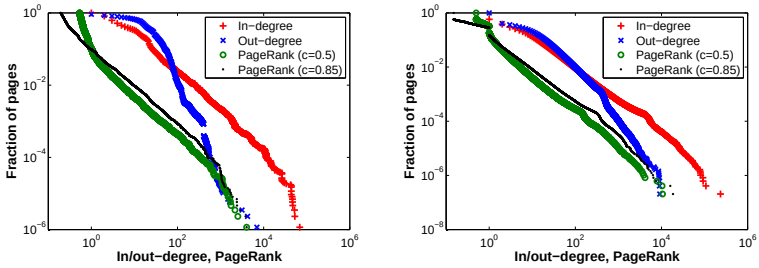
In this section we use various statistical techniques to reveal and evaluate the power laws. To this end, we chose three data sets that represent different network structures. As the Web sample, we used the EU-2005 data set with 862.664 nodes and 19.235.140 links [20]. We also performed experiments on the Wikipedia (English) data, whose structure is known to be different from the Web graph [26]. This data set contains 4.881.983 nodes and 42.062.836 links. Finally, we simulated a Growing Network by using preferential attachment rule for 90% of new links. The graph consists of 10.000 nodes with constant out-degree $d = 8$. In Figure 4.1 we show the cumulative log-log plots for in-degrees, out-degrees and PageRank scores in all data sets. The PageRank scores are computed according to definition of the standard PageRank (1.2).

The log-log plots for the in-degree and the PageRank in Figure 4.1 resemble the signature straight line indicating power laws. However, several techniques should be combined in order to establish the presence of heavy tails and to evaluate the power law exponent. We use QQ plots, Hill and altHill plots as well as Pickands plots to confirm that the in-degree and the PageRank follow power laws with similar exponents for all three data sets. We will also conclude that the out-degree can be modeled reasonably well as a power law with exponent around 2.5-3.

Denote by $X_1, \ldots, X_w$ non-negative observations of node's characteristic (e.g. in-degree) on the Web graph. We write $X_{(i)}$ for the ith largest value of $X_1, \ldots, X_w$, where $1 \leq i \leq w$:

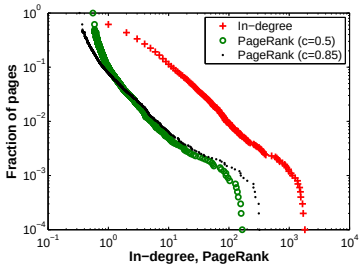
$$X_{(1)} \geq X_{(2)} \geq \ldots \geq X_{(w)}.$$

In the next sections we will provide a review of estimation techniques designed under assumption that $X_1, \ldots X_w$ are independent random variables having an identical



(a) EU-2005

(b) Wikipedia



(c) Growing Network

Figure 4.1: Cumulative log-log plots for in/(out)-degree, PageRank (c=0.5) and PageRank (c=0.85)

regularly varying distribution with tail index α. The idea is to apply several different procedures in order to make the final conclusion.

4.1.1 Hill plot

The Hill's estimator $H_{k,w}$ is a widely used estimator of $1/\alpha$, that is based on k upper order statistics:

$$H_{k,w} = \frac{1}{k}\sum_{i=1}^{k} \log\left(\frac{X_{(i)}}{X_{(k+1)}}\right). \tag{4.1}$$

It was proved (see e.g. [100]) that $H_{k,w}$ converges in probability to $1/\alpha$ as $w, k \to \infty$, $k/w \to 0$. An obvious problem with the Hill estimator is choosing the value k so that $X_{(k)}$ corresponds to a 'beginning' of the power law tail. This can be mitigated by constructing a so-called Hill plot.

To make a Hill plot for α we graph $\{(k, H_{k,w}^{-1}), 1 \leq k \leq w\}$ and if the plot looks stable around a certain horizontal line, we can pick the corresponding value of α. This sometimes works beautifully, especially for data close to pure Pareto tails. However, if $L(x)$ in the definition of regular varying random variable (1.6) deviates considerably from a constant there may be enormous errors. The Hill plot, as well as the Hill estimator, is also not location invariant. Theoretically, a shift does not affect the power law exponent, however it drastically distorts the Hill plot. Clearly, in case when the Hill plot does not look stable, the Hill estimator can not be used for the evaluation of α.

To construct confidence intervals for the Hill estimator, Newman [89] suggests to use a bootstrap method for estimating the variance of $H_{k,w}^{-1}$. A simpler way is to use the convergence of $\sqrt{k}H_{k,w}$ to a normal random variable with mean $1/\alpha$ and variance $1/\alpha^2$ as $w, k \to \infty$, $k/w \to 0$ (see [100, p.304]). Thus, one can obtain confidence intervals based on the quantiles of the standard normal distribution.

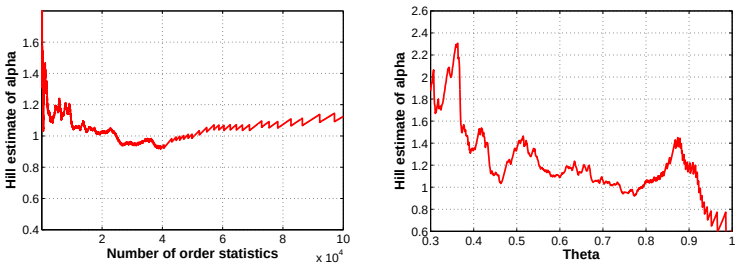
One can also display the Hill plot in the alternative form $\{(\theta, H_{\lceil w^\theta \rceil, w}^{-1}), 0 \leq \theta \leq 1\}$, where $\lceil x \rceil$ is the smallest integer greater or equal to $x \geq 0$. This plot is called the alternative Hill plot, altHill. Compared to the Hill plot, the altHill shows the largest order statistics more prominently. According to [100], if the distribution is not exactly Pareto, then the altHill spends more time in the small neighborhood of α than the Hill plot.

Here we only display Hill and altHill plots for in-degree and PageRank (c=0.85) in the Web graph (see Figure 4.2). For the various plots for the Growing Network (Figure 4.18), and the Wikipedia (Figure 4.15 and 4.17), as well for plots of out-degree and PageRank(c=0.5) in the EU-2005 data set (Figure 4.14 and 4.16) we refer to Section 4.6. We note that the saw-type picture for in-degrees and out-degrees reflects the fact that we deal with integer values that are the same for quite large groups of nodes.

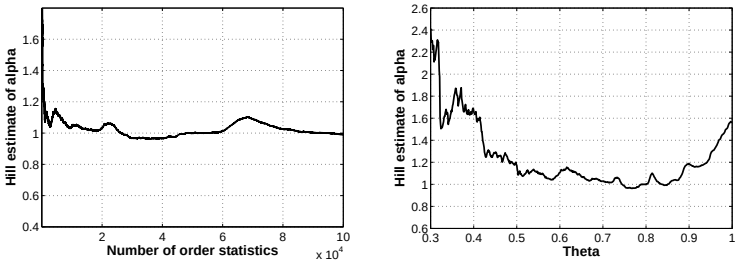
In the Web data, the Hill plots confirm the power law tail of in-degree and PageRank (c=0.85). The exponent α seems to be the same in both cases. However, it looks like the estimation 1.1 is, on average, on a higher side. Again, oscillations between 0.9 and 1.2 are essential since $\alpha = 0.9$ implies infinite mean. The altHill is stable for θ between 0.4 and 0.9. The beginning of the plot is most probably distorted by the well-known exponential cut-off of the real-life data [28], and for $\theta > 0.9$ the number of used order statistics is too large.

In the Growing Networks, the Hill plots behave reasonably nice. The plot for in-degree (Figure 4.18(a)) is more stable as it spends significant time around the line $\alpha = 1.1$. In Figure 4.18(c), the plot for PageRank (c=0.85) also behaves well and seems to suggest a slightly smaller tail index, around 1.05. From the plots we see that the estimator for α is very sensitive to the choice of k. Thus, constructing a Hill plot is a helpful step when applying a Hill estimator.

The Hill and altHill plots suggest that the in-degree and PageRank in the Web (Figure 4.2) and in the Growing Networks (Figure 4.18) are heavy-tailed but not exactly a Pareto. Indeed, the plots look relatively stable but it is difficult to single



(a) in-degree



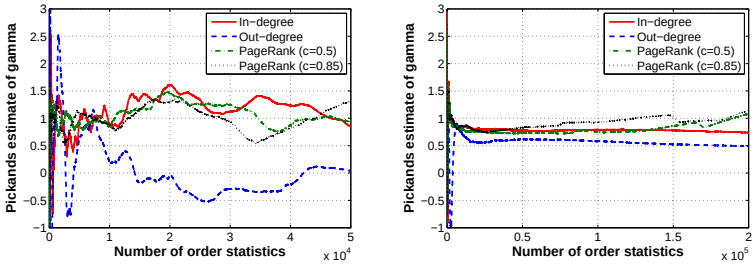
(b) PageRank (c=0.85)

Figure 4.2: Hill (*left*) and altHill (*right*) plots for the EU-2005 data set.

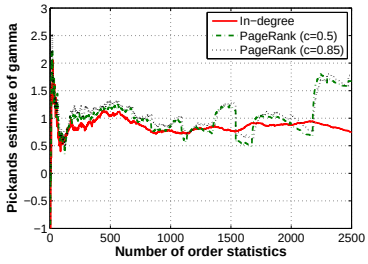
out α.

For the out-degree in the Web data (Figure 4.16(a)) the altHill plot oscillates considerably. However, the Hill plot (Figure 4.14(a)) does not behave nearly as badly as it would, for instance, for the exponential distribution (see example in [100, p.96]). Based on the Hill plot, one may therefore conclude that the out-degree has a power law.

Finally, Wikipedia turns out to be an example of perfect Hill plots (Figure 4.15) whereas altHill (Figure 4.17) shows large oscillations. We conclude that in-degree and PageRank (c=0.85) in Wikipedia follow closely a Pareto distribution with index 1.2. The index of PageRank (c=0.5) distribution is around 1.4. The out-degree is also Pareto, with index about 1.6.



(a) EU-2005 (b) Wikipedia



(c) Growing Network

Figure 4.3: Pickands plots for in/(out)-degrees and PageRanks.

4.1.2 Pickands plot

A Pickands estimator as presented in [100], is another way to evaluate α and reveal the presence of power laws. We first introduce the *extreme-value* distributions, defined as

$$G_\gamma = \exp\left(-(1+\gamma x)^{-1/\gamma}\right),\ \gamma \in \mathbb{R},\ 1+\gamma x > 0.$$

The power law case corresponds to $\gamma > 0$ and then $\gamma = 1/\alpha$.

The Pickands estimator of γ uses differences of quantiles, where the latter are estimated by means of three upper statistics, $X_{(k)}$, $X_{(2k)}$, $X_{(4k)}$, from a sample size w. The estimator is defined as

$$\hat{\gamma}_{k,w}^{(Pickands)} = \frac{1}{\log 2} \log\left(\frac{X_{(k)} - X_{(2k)}}{X_{(2k)} - X_{(4k)}}\right).$$

Determining an appropriate of k is again an important issue. Unlike the Hill estimator, the Pickands estimator is both location and scale invariant.

Similarly to the Hill plot, a *Pickands plot* consists of the points

$$\left\{\left(k, \hat{\gamma}_{k,w}^{(Pickands)}\right),\ 1 \le k < w/4\right\}.$$

A difficulty in constructing Pickands plots for integer-valued observations such as in-degrees and out-degrees in the networks, is that the values of order statistics might be identical, resulting in division by zero. To fix this problem we introduce a randomization of the data by adding uniformly $(0, 1)$ distributed random variables to each of the observations.

The Pickands plots for our data sets are presented in Figure 4.3. We note that we plot the values of $\hat{\gamma}_{k,w}^{(Pickands)}$ that estimates $1/\alpha$.

The results for in-degree and PageRank in all three data sets are in good agreement with Hill plots. The new information we find by looking at the plot for out-degree in the Web data. In Figure 4.3(a) a large part of the Pickands plot shows $\gamma < 0$ which signals light tails. This is in agreement with Donato *et al.* [33] and other papers that claim that the out-degree data does *not* follow a power law. On the other hand, the Pickands plot goes below zero only for quite large values of k, so we still can not exclude the power law tail.

4.1.3 QQ plot

Suppose we have a hypothesis that the true distribution function producing the data is $F(x)$. A goodness of fit test provides the rigorous way to verify such hypothesis, whereas the *QQ plot* is a more informal but convenient alternative. To construct a QQ plot we graph the theoretical quantiles of F versus the sample quantiles:

$$\left\{\left(F^{\leftarrow}\left(\frac{i}{w+1}\right), X_{(w-i+1)}\right),\ 1 \le i \le w\right\},$$

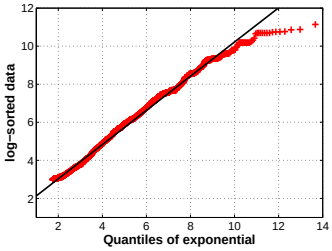
where $F^{\leftarrow}(y) = \inf\{x :\ F(x) \ge y\}$ is the inverse of distribution function F. If our hypothesis is true then the result should fall roughly on the straight line $\{(x, x),\ x > 0\}$. One potential problem is how to decide what we consider 'close enough' to linear.

To apply QQ plots to power laws, suppose that our null hypothesis is that for some $x_0 > 0$, a distribution of a random variable X satisfies

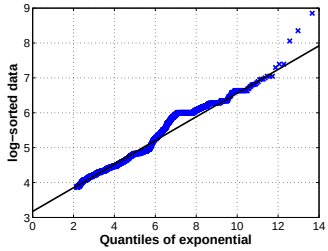
$$\mathbb{P}(X > x) = \left(\frac{x}{x_0}\right)^{-\alpha},$$

so it follows that $\mathbb{P}(\log X > y) = \mathrm{e}^{-\alpha(y - \log x_0)}$. Hence, using quantiles of exponential distribution we plot

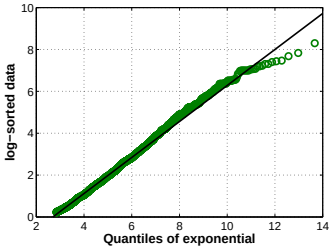
$$\left\{\left(-\log\left(1 - \frac{i}{w+1}\right), \log X_{(w-i+1)}\right), 1 \le i \le w\right\}.$$



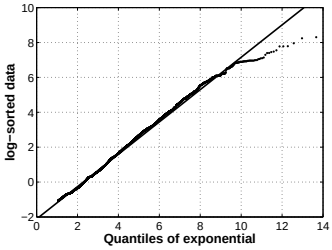
(a) in-degree: 150.000 (α = 1.11) upper statistics



(b) out-degree: 100.000 (α = 2.95) upper statistics



(c) PageRank (c=0.5): 50.000 ($\alpha = 1.16$) upper statistics



(d) PageRank (c=0.85): 300.000 ($\alpha = 1.08$) upper statistics

Figure 4.4: QQ lines for the EU-2005 data set.

The slope of the least-squared line fitted to the QQ plot is an estimate of $1/\alpha$. Thus, if $\{(x_i, y_i), 1 \leq i \leq n\}$ are n points on the plane, we can calculate the slope in standard way

$$\mathrm{SL}\{(x_i, y_i), 1 \leq i \leq w\} = S_{xy}/S_{xx},$$

where $S_{xy} = \sum_{i=1}^{w}(x_i - \bar{x})(y_i - \bar{y})$, $S_{xx} = \sum_{i=1}^{w}(x_i - \bar{x})^2$ and $\bar{x}$ means mean value of x. Now we can define the QQ estimator for $1/\alpha$ based on k upper order statistics as

$$\mathrm{SL}\left\{\left(-\log\left(1 - i(w+1)\right), \log X_{(w-i+1)}\right), w - k + 1 \leq i \leq w\right\}.$$

Clearly, there remains the problem of choosing k.

In Figure 4.4 we present the QQ plots for EU-2005 data set for good choices of k. In Section 4.6 we provide the remaining plots. Again, the data on the in-degree and the PageRank resulted in QQ plots similar to straight lines, and the estimates

for α are close to what we expected. Thus, in this case all techniques point to the same result.

With a certain amount of tolerance, we can accept that the QQ plot for out-degrees in the Web data in Figure 4.4(b) is close enough to a straight line. Moreover, the estimated $\alpha = 2.95$ is in good agreement with the Hill plot. We also note that $\alpha > 2$ implies a finite variance while power law models are especially important in case when the variance is infinite, reflecting high variability [72, 94]. Hence, in case of a finite variance, it is not really crucial whether the data obeys a power law. To exclude the possibility of exponential tail of out-degree, we also constructed a QQ plot with exponential quantiles by plotting $-\log(1 - i/(w+1))$ against $X_{(w-i+1)}$. The result that we do not present here is not any close to a straight line. To summarize, the out-degree has a finite variance and a tail heavier than exponential, so it can be modeled reasonably well as a power law with exponent around 2.5-3, according to our estimates.

4.2 Asymptotics for non-uniform PageRank

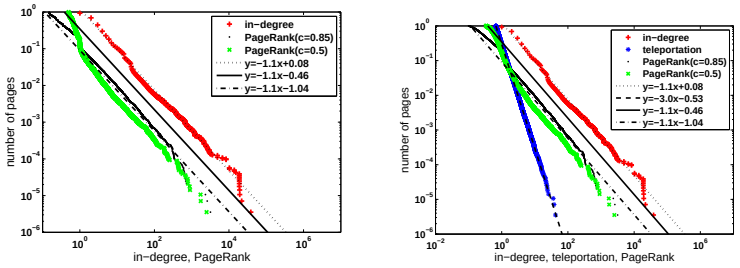
In this section we study tail behavior of non-uniform PageRank (1.3) in relation to various characteristic of the Web graph, and to distribution of teleportation jump. Thus, we want to illustrate the results of Chapters 2, and 3, in particularly, we justify asymptotical equivalences obtained in Corollaries 2.6, and 3.11. We start with the case of the non-uniform PageRank.

We use Stanford data set[1] with $w = 281.903$ pages and 2.312.497 links. It is a relatively small Web sample, however, it is known to possess basic properties of the Web. In particular, in this data set, the in-degree shows typical power law behavior with exponent $\alpha_N = 1.1$. In the next section we present more numerical results for simpler model of the standard PageRank with uniform teleportation.

We create the teleportation distribution by using the inverse transformation method. First, we generate random numbers $u_1, \ldots, u_w$ from the standard uniform distribution, and then we set $t_i = (1 - u_i)^{-1/\alpha_T}$, where $i = 1, \ldots, w$. These t_i's are random numbers that are Pareto distributed with exponent α_T. We choose $\alpha_T = 0.5$, 1.1 and 3.0. Second, we denote $\bar{t}$ as the mean value of $t_1, \ldots, t_w$, and define the teleportation probability of a jump to page i as $T(i) = t_i/(w\bar{t})$. Next, we use formula (1.3) to obtain the non-uniform PageRanks. We also compute the PageRank with uniform teleportation jumps. The computation of the PageRank is done by applying the matrix power iteration method (see [69] for more details).

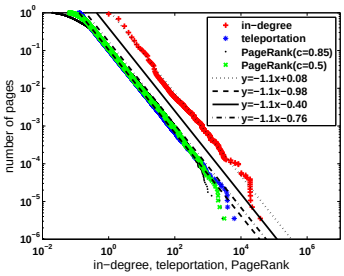
In Figure 4.5(a)-(d) we present cumulative log-log plots for in-degree, teleportation and PageRanks for damping factors $c = 0.5$ and $c = 0.85$. Here we consider a scale-free teleportation, so we plot complementary cumulative distribution function $\mathbb{P}(wT > x) = (\bar{t}x)^{-\alpha_T}$. Then, $y = -\alpha_T x - \alpha_T * \log_{10}(\bar{t})$ is the straight line that

[1] www.kamvar.org/personalization; (Accessed in April 2009).

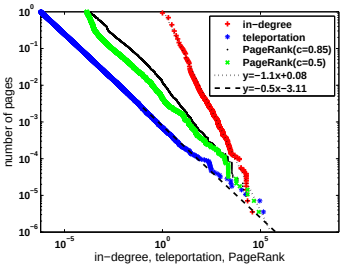


(a) uniform

(b) $\alpha_B = 3.0$



(c) $\alpha_B = 1.1$



(d) $\alpha_B = 0.5$

Figure 4.5: Cumulative log-log plots for in-degree, teleportation and PageRank.

corresponds to the teleportation log-log plot. We also fit the in-degree plot with the straight line $y = -1.1x + 0.08$.

First, we consider the log-log plots of the standard PageRank with uniform teleportation (see Figure 4.5(a)). In this case we use Corollary 3.11(i) to obtain the distance between in-degree and PageRank log-log plots as

$$\log_{10}(C_N) = \log_{10}\left[\frac{c^{\alpha_N}(1-p_0)^{\alpha_N}}{(\mathbb{E}(N))^{\alpha_N}(1-c^{\alpha_N}\mathbb{E}(N)\mathbb{E}(1/D^{\alpha_N}))}\right], \tag{4.2}$$

where, as before, N is the in-degree, D is the effective out-degree, and p_0 is the fraction of the dangling nodes. From $\mathbb{E}(N) = 8.2032$, $p_0 = 0.006$ and $\mathbb{E}(1/D^{1.1}) = 0.1043$, we predict the PageRank log-log plots: $y = -1.1x - 0.46$ for $c = 0.85$, and $y = -1.1x - 1.04$ for $c = 0.5$. In the plot we show these theoretically predicted lines and experimental PageRank log-log plots. We see that both lines perfectly match the slopes of the PageRanks, and they trace the direction of changes in the PageRank distribution in respect with changes of the damping factor. Indeed, the plot of the

PageRank with $c = 0.5$ is further from the in-degree log-log plot, then the plot of the PageRank with $c = 0.85$. We note that we underestimate the predicted distance in the case of $c = 0.85$, that can be caused by some assumptions of our model. We refer to Section 4.5 for discussion.

We again use Corollary 3.11(i) for the case of the PageRank with teleportation that follows power law with exponent $\alpha_T = 3.0$. Then we end up with the same constant as in (4.2), and therefore we get the same predicted lines for the PageRank log-log plots: $y = -1.1x - 0.46$ for $c = 0.85$, and $y = -1.1x - 1.04$ for $c = 0.5$. In Figure 4.5(b) we plot the distributions of the teleportation and the PageRanks along with the predicted straight lines. The results are similar to the previous case. Thus, we can see that the distribution of the teleportation has no influence on the tail behavior of the PageRank in case when the teleportation has a lighter tail than the in-degree.

Next, we consider the $T(i)$'s with $\alpha_T = 1.1$. In Remark 3.11(iii) we need to know C_{NB} from $\mathbb{P}(wT > x) \sim (1-c)^{-\alpha_T} C_{NB} \mathbb{P}(N > x)$ as $x \to \infty$. Since $y = -1.1x + 0.08$ and $y = -1.1x - 0.98$ are the fitted lines for log-log plots for in-degree and teleportation, respectively, we can find that $C_{NB} = 0.0108$ for $c = 0.85$, and $C_{NB} = 0.4063$ for $c = 0.5$. Thus, in the case when the in-degree and the teleportation are regular varying with the same index $\alpha_N = \alpha_T = 1.1$, we can define the distance in the following way:

$$\log_{10}(C) = \log_{10}\left[\frac{(\mathbb{E}(N))^{\alpha_N} C_{NB} + c^{\alpha_N}(1-p_0)^{\alpha_N}}{(\mathbb{E}(N))^{\alpha_N}(1 - c^{\alpha_N}\mathbb{E}(N)\mathbb{E}(1/D^{\alpha_N}))}\right]. \tag{4.3}$$

We apply these constants in the above formula to obtain $y = -1.1x - 0.41$ and $y = -1.1x - 0.76$ for PageRank plots for $c = 0.85$ and $c = 0.5$, respectively. We plot these lines in Figure 4.5(c). Compared to Figures 4.5(a) and (b), here the teleportation distribution smoothens the log-log plots of the PageRanks. Thus, we can hardly see the difference between the plots for $c = 0.5$ and $c = 0.85$. The slopes of the experimental PageRanks again correspond to the predicted power law exponent 1.1. The differences between the log-log plots of the in-degree and the PageRanks agree better than in the previous cases.

Finally, we present results for the teleportation with power law exponent $\alpha_T = 0.5$ in Figure 4.5(d). Note that we can not find the distance in this case, because the first moment of T does not exist. However, we can clearly see that the PageRank tends to follow a power law with the same exponent as the teleportation distribution.

From Section 3.3.4 we know that

$$C_N \geq \frac{c^{\alpha_N}(1-p_0)^{\alpha_N}}{(\mathbb{E}(N))^{\alpha_N}[1 - c^{\alpha_N}(1-p_0)^{\alpha_N}(\mathbb{E}(N))^{1-\alpha_N}]} = C'_N. \tag{4.4}$$

The last expression is the value of C_N in case when the out-degree of all non-dangling nodes is a constant $\mathbb{E}(N)/(1-p_0)$. If $\alpha_N = 1.1$, then the difference between the left- and the right-hand sides of (4.4) is really small for any reasonable out-degree distribution. We test the prediction of multiplicative constant C'_N from (4.4). To

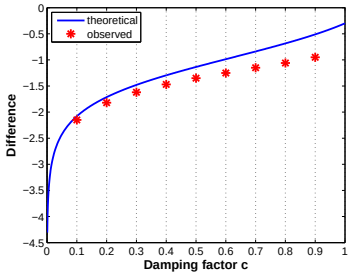
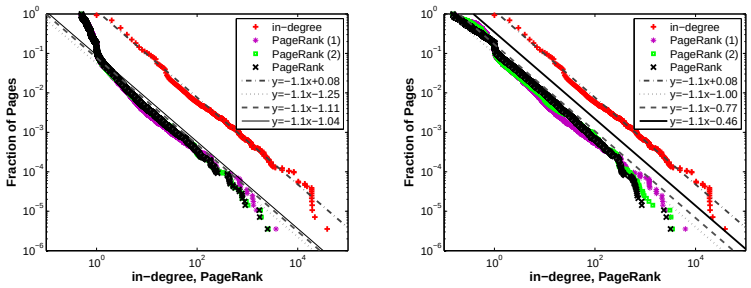


Figure 4.6: The theoretical and observed differences between asymptotics of in-degree and PageRank.

this end, we calculate the differences between the logarithms of the complementary cumulative distribution functions of PageRank and in-degree for different values of the damping factor. In Figure 4.6 we plot $\log_{10}(C'_N)$ together with the observed differences. As it can be seen, the theoretical and observed values are quite close. E.g., for typical values of c between 0.8 and 0.9, the difference is 0.41, resulting in a factor C'_N that is only a factor 2.57 larger than in the observed data. Thus, this is a good approximation for the difference between the two distributions.



(a) c=0.5 (b) c=0.85

Figure 4.7: Cumulative log-log plots for in-degree and PageRank. The straight lines for the PageRank plots are predicted by the model for the 1st, the 2nd, and the last power iterations.

Finally, we verify out model for power iterations. In Figure 4.7 we show the cumulative log-log plot of in-degree and standard PageRank after the 1st, the 2nd, and the last power iterations for the damping factors $c = 0.5$ and $c = 0.85$. The results again confirm the similarity in the asymptotic behavior of the in-degree and

the PageRank. Using Corollary 2.6(i) we calculate the difference between log-log plots of the in-degree and the PageRank after the kth iteration:

$$\log_{10}\left(C_N^{(k)}\right) = \log_{10}\left[\frac{c^{\alpha_N}(1-p_0)^{\alpha_N}}{(\mathbb{E}(N))^{\alpha_N}}\sum_{i=1}^{k}\left[c^{\alpha_N}\mathbb{E}(N)\mathbb{E}(1/D^{\alpha_N})\right]^i\right]. \tag{4.5}$$

In the case of $c = 0.85$, we predict PageRank after the 1st, the 2nd, and the last power iterations with the straight lines: $y = -1.1x - 1.00$, $y = -1.1x - 0.77$, and $y = -1.1x - 0.46$, respectively. From Figure 4.7 we see that our model correctly predicts the PageRank distributions for $c = 0.5$, and captures the dynamics of the iterations for $c = 0.85$.

4.3 Asymptotics for the standard PageRank

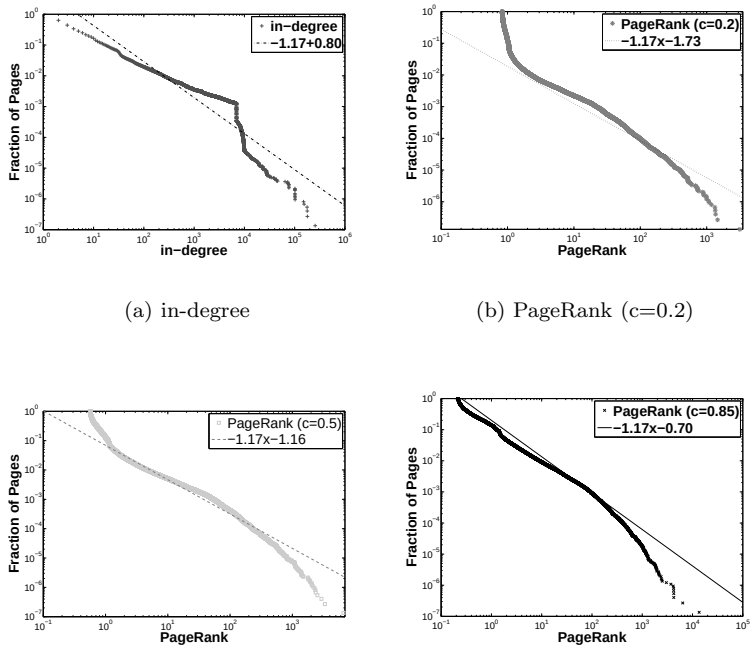
In this section we analyze distribution of the standard PageRank (1.2) in the Web data set, the Wikipedia data set and preferential attachment graphs. Then, we apply Corollaries 3.11(i) and 3.11(i), that give us the asymptotic similarity between in-degree and PageRank and the multiplicative constants as in (4.2) and (4.5).

4.3.1 Web data

We performed experiments on Indochina-2004 and EU-2005 Web samples [20]. In Figures 4.9 and 4.8 we present cumulative log-log plots for in-degree and PageRanks. We fit the straight line for in-degree accordingly to the evaluated power law exponent. For the PageRank, we plot the theoretically predicted straight lines obtained from Corollary 3.11(i).

The Indochina set contains 7.414.866 nodes and 194.109.311 links. The results are presented in Figure 4.8. The in-degree plot resembles a power law except for the excessively large fraction of pages with in-degree about 10^4. The presence of *bump* was observed also in other data samples in the past [24, 34]. In [34], the authors suggested that it could be probably due to a huge clique created by a single spammer. For more detail on this data set see [9]. For Indochina, we obtain a power law exponent $\alpha_N = 1.17$ for cumulative plot, which is quite different from the result in [9]. This demonstrates the sensitivity of estimators for the power law exponent. Indeed, the exponent 0.6 in [9] reflects the behavior in the first part of the plot, whereas 1.17 gives more weight on the tail of the in-degree distribution.

We fit the straight line $y = -1.17x + 0.8$ into the in-degree plot and then compute the distance according to (4.2) for $c = 0.2, 0.5$, and 0.85. With $\mathbb{E}(N) = 26.17$, $p_0 = 0.18$, and $\mathbb{E}(1/D^{1.17}) = 0.0248$, we obtain the following prediction for the PageRank log-log plot: $y = -1.17x - 1.73$ for $c = 0.2$, $y = -1.17x - 1.16$ for $c = 0.5$, and $y = -1.17x - 0.70$ for $c = 0.85$. In Figure 4.8(b)-(c) we show these theoretically predicted lines along with the experimental PageRank log-log plots. We see that for this data set, our model provides the linear fit with a striking accuracy.



(a) in-degree

(b) PageRank (c=0.2)

(c) PageRank (c=0.5)

(d) PageRank (c=0.85)

Figure 4.8: Indochina data set: cumulative log-log plots for in-degree and PageRank. The straight lines for the PageRank plots are predicted by the model.

We performed the same experiment for EU-2005 data set defined in Section 4.1. Then, in-degree log-log plot can be fitted perfectly by $y = -1.1x + 0.61$. We use the same approaches to calculate the difference between the in-degree and PageRank plots for $\mathbb{E}(N) = 22.3$, $p_0 = 0.08$, $\mathbb{E}(1/D^{1.1}) = 0.0314$. Thus, the theoretical prediction for the PageRank are $y = -1.1x - 1.63$, $y = -1.1x - 1.07$, and $y = -1.1x - 0.60$ for $c = 0.2, 0.5$, and 0.85, respectively. The log-log plots for experimental data, the fitted straight line for in-degree, and corresponding theoretical straight lines for PageRank, are presented in Figure 4.9.

4.3.2 Wikipedia

In order to further verify our results, we performed the experiments on the Wikipedia data set from Section 4.1. The structure of Wikipedia is believed to be slightly different from the Web graph. In Figure 4.10 we show the in-degree and PageRank plots, with fitted straight line $y = -1.18x + 0.30$ for the in-degree and predicted

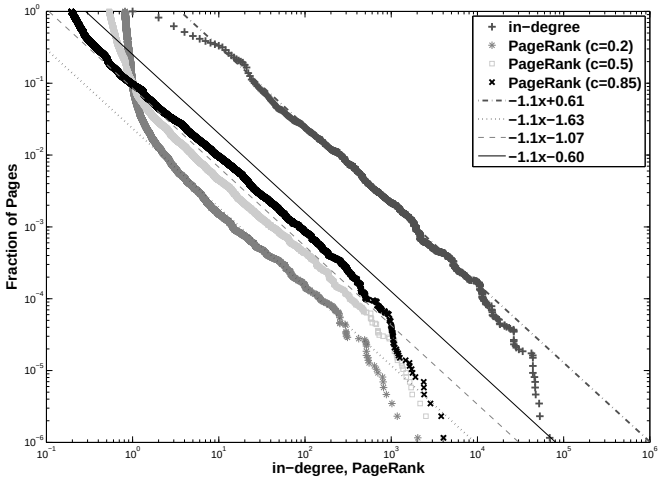


Figure 4.9: EU-2005 data set: cumulative log-log plots for in-degree and PageRank. The straight lines for the PageRank plots are predicted by the model.

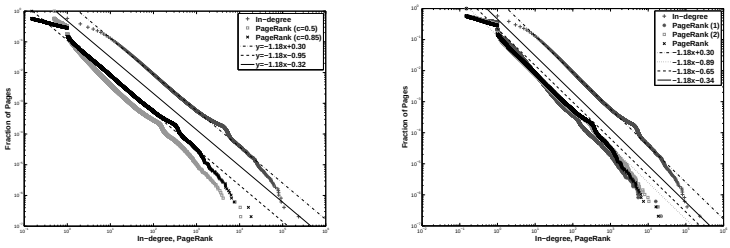
lines for the PageRank. In this data set we have $\mathbb{E}(N) = 8.6159$, $p_0 = 0$ and $\mathbb{E}(1/D^{1.18}) = 0.1006$. Figure 4.10(a) shows the PageRank plot for $c = 0.5$ and $c = 0.85$. In Figure 4.10(b) we depict the PageRank plots after the first, the second and the last iterations for $c = 0.85$.

Importantly, we observe that the PageRank for Wikipedia retains its power law distribution, and the exponent is again the same as the one for in-degree. Moreover, we see that our model correctly captures the dynamics of the PageRank distribution in successive power iterations and for different values of c.

4.3.3 Synthetic graphs

Next, we performed the experiments on a synthetic graph with out-degree close to constant. The graph of 5.000.000 nodes and 41.577.523 links was generated using preferential attachment rule (see Section 1.3.3). Further, 30% of the links were redirected in order to make the graph more realistic and comparable to Wikipedia. The original out-degree was 9, however, due to the duplicated edges, the average out-degree became 8.3155. This data set is different from the Growing Network in Section 4.1.

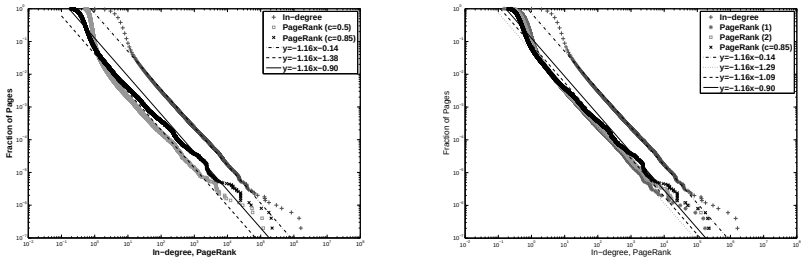
The results on the PageRank distribution are presented in Figure 4.11. For the in-degree, we computed $\alpha = 1.14$. The predicted lines for the PageRank are obtained with $\mathbb{E}(N) = 8.3155$, $p_0 = 0$, $\mathbb{E}(1/D^{1.14}) = 0.0858$ ($\approx (\mathbb{E}(N))^{\alpha}$). In Figure 4.11(a) we show the PageRank plot for $c = 0.5$ and $c = 0.85$, and Figure 4.11(b) displays



(a) different values of c (b) different number of iterations

Figure 4.10: Wikipedia data set: cumulative log-log plots for in-degree and PageRank. The straight lines for the PageRank plots are predicted by the model.

the PageRank plots after the first, the second and the last iterations for $c = 0.85$. One can see that our model provides a good estimation for the difference between



(a) different values of c (b) different number of iterations

Figure 4.11: Synthetic data: cumulative log-log plots for in-degree and PageRank. The straight lines for the PageRank plots are predicted by the model.

the graphs. Furthermore, the lines look parallel as before, although in the growing network models, the PageRank power law exponent is proved to depend on the damping factor [6]. Here we clearly face the fact that the nuances of the 'real' slope are hard to capture on the data. Consequently, our model works well in this case.

Here we also consider another version of the growing network graph that showed a quite different behavior. The difference is caused by the simulation procedure. To ensure the same number of outgoing links for all pages, we link the first $\mathbb{E}(N)$ nodes to randomly chosen pages at the end of the simulation. We simulate Growing

Network by using preferential attachment rule for 80% of new links. The graph consist of 50.000 nodes with constant out-degree $\mathbb{E}(N) = 8$. In Figure 4.12 we present cumulative log-log plots for in-degree and PageRanks. Clearly, the PageRank for $c = 0.85$ does not show good power law behavior. The observed bumps can be explained by the presence of loops with highly ranked initial nodes.

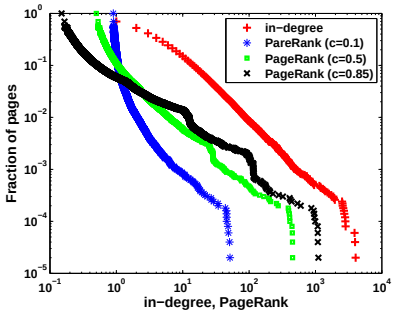


Figure 4.12: Growing network: cumulative log-log plots for in-degree and PageRank.

4.4 PAR ranking scheme

The negligible effect of out-degree distribution on the PageRank behavior made us wonder about the role of out-degrees in link-based ranking in general. In HITS (see Section 1.2.3), the ranking of a page i is determined by its authority score, which in turn depends on hub scores of pages linking to i. Furthermore, a hub score is high for pages with high out-degree, and thus getting a link from such a page is *advantageous* in HITS whereas it is *disadvantageous* in PageRank according to (1.3). Since both HITS and PageRank work well in practice, one may try to think of some ranking scheme where out-degree *does not play a role* at all.

We propose one such ranking scheme that we call a Pure Authority Rank (PAR). This algorithm is a mixture between HITS and PageRank. The PAR is defined iteratively. The initial score of each page $i = 1, \ldots, w$ is $s_i^{(0)} = 1/w$, and the results of successive iterations are computed as

$$s_i^{(k)} = \frac{c}{\mathbb{E}(N)} \sum_{j \to i} s_j^{(k-1)} + \frac{1-c}{w}, \quad k \geq 1, \tag{4.6}$$

and then normalized so that $\sum_{i=1}^{w} s_i^{(k)} = 1$. Here the summation is over all pages j that link to i.

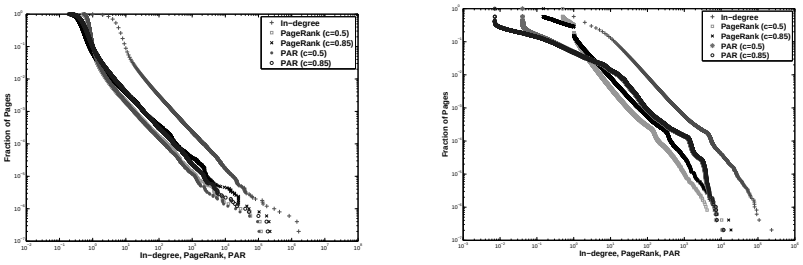
Now, let A be an adjacency matrix of the Web Graph. Then denoting $\tilde{M} = (c/\mathbb{E}(N))A + (1-c)E/w$, where E is the matrix of ones, we can write (4.6) with the subsequent normalization in a matrix-vector form as $\mathbf{s}^{(k)} = \mathbf{s}^{(k-1)}\tilde{M}/||\mathbf{s}^{(k-1)}\tilde{M}||$,

		Scores	Ranks		Iterations	
Data	c	Correlation coefficient	Kendall's τ	Spearman's ρ	PR	PAR
Synthetic graph	0.5	0.8112	0.1234	0.1827	8	7
	0.85	0.9753	0.1002	0.1488	13	9
Wikipedia	0.5	0.2474	0.3510	0.4304	8	17
	0.85	0.4675	0.3629	0.4422	29	18

Table 4.1: Comparison of PageRank and PAR.

where $\mathbf{s}^{(k)} = (s_1^{(k)}, \ldots, s_w^{(k)})$ and $||\cdot||$ is the L_1 norm. Since $\tilde{M}$ is a positive matrix, the convergence and uniqueness of the PAR scores are guaranteed by the Perron-Frobenius theorems. If we take $c = 1$ we obtain an algorithm close to HITS but without the hub-iteration. In this case, the algorithm will converge but the resulting vector might depend on the initial vector, as in HITS and SALSA [70]. We refer to [48] for the detailed uniqueness analysis of link-based ranking schemes.

We computed the PAR scores for Wikipedia and the synthetic graph. The algorithm converges fast and, remarkably, the speed of convergence does not depend on c (see Table 4.1). In Figure 4.13 we present the log-log plots for PAR and PageRank. Since the two methods are similar, it is not surprising that the PAR distribution seems to follow a power law with the same exponent as in-degree. A more interesting observation is that the PAR plot for Wikipedia in Figure 4.13(b) behaves similar to a PageRank plot computed for a higher value of c.



(a) synthetic data (b) Wikipedia

Figure 4.13: PageRank and PAR log-log plots.

Finally, we computed the Kendall's tau [62] and the Spearman's rho [107] (for definitions see (5.23) in Chapter 5), as well as correlation coefficient between PAR and PageRank scores for the top 1% pages. The results are presented in Table 4.1.

The high correlation between the scores for synthetic graph is expected since in this case the difference between PAR and PageRank is minimal. On the other hand, the correlation between the ranks is on a lower side. Similar to [22] we observe that

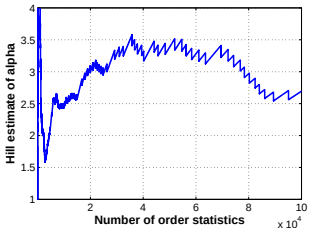
the ranking order is very sensitive to the nuances of the algorithm. We note that the ordering of the PageRank and the PAR values may be less significant when they are incorporated into an information retrieval model as described in Section 1.1. For a more fair comparison of the two algorithms, future research should reveal which pages were demoted and which were promoted. We believe that the advantages of the PAR algorithm, such as fast convergence and insensitivity to out-degrees, should definitely attract more studies.

4.5 Discussion

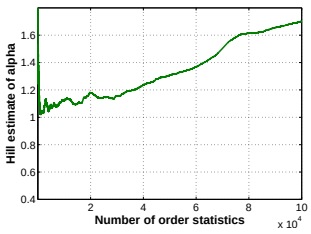
Our results are in a good agreement with the Web data. The differences between the model and the data depend on many factors, in particular, on the choice of a data set. Furthermore, equation (2.3) implicitly involves the assumption of the branching structure of the Web. Although hierarchal structure is present in the Web [40, 41], this assumption is an obvious simplification of the realistic Web structure. Future work could try to investigate how to improve the model in that respect, mainly by studying the dependencies amongst the R_i's in (2.3), or between the R_i's on the one hand and N on the other.

The Growing Network models may provide an alternative explanation [6, 50]. For instance, in [6] it was shown that the *expected* PageRank in Growing Networks follows a power law with an exponent which does depend on the damping factor. The reason for such discrepancy with our results could be that we focus only on asymptotics, whereas [6] employs a mean-field approximation. Indeed, experiments show that the shape of the PageRank distribution does depend on the damping factor, and thus, it may affect the average values while the tail behavior remains the same for all values of c.

4.6 Additional plots



(a) out-degree

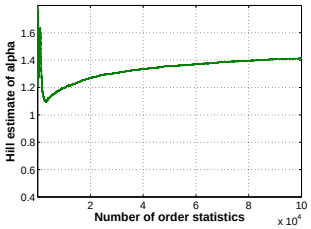


(b) PageRank (c=0.5)

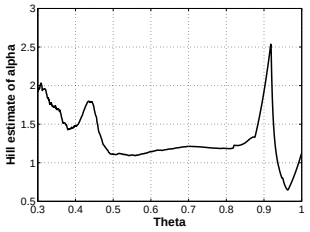
Figure 4.14: Hill plots for the EU-2005 data set.

(a) in-degree

(b) out-degree

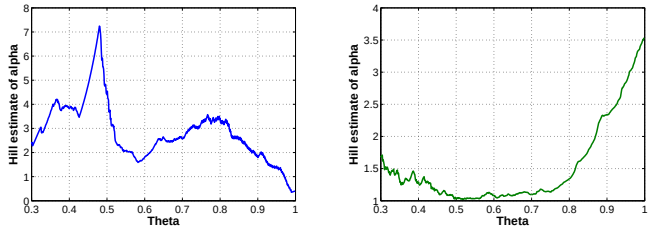


(c) PageRank (c=0.5)



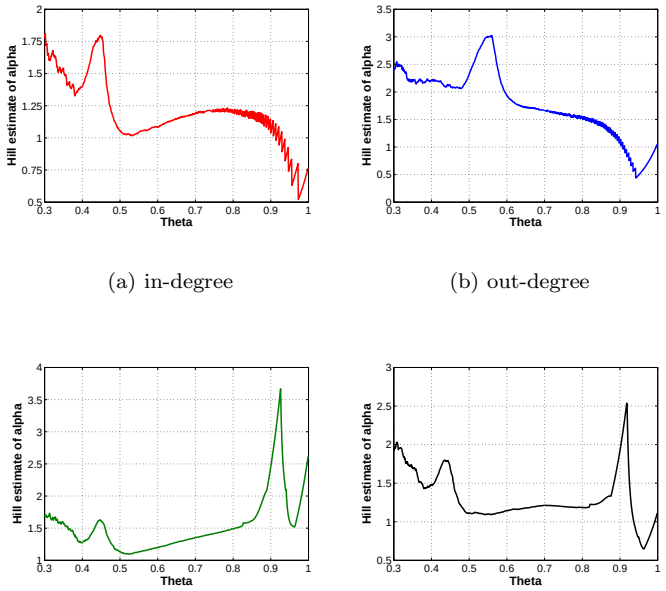
(d) PageRank (c=0.85)

Figure 4.15: Hill plots for the Wikipedia data set.



(a) out-degree (b) PageRank (c=0.5)

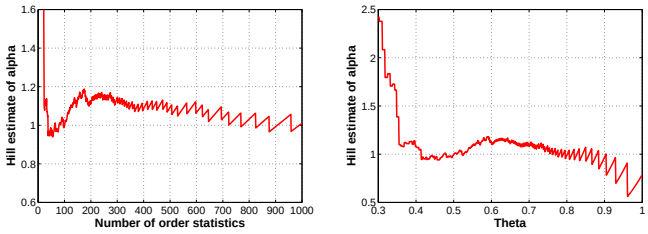
Figure 4.16: altHill plots for the EU-2005 data set.



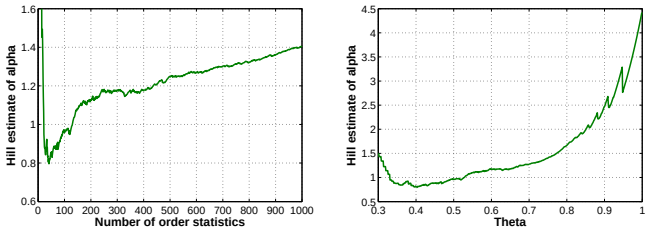
(a) in-degree (b) out-degree

(c) PageRank (c=0.5) (d) PageRank (c=0.85)

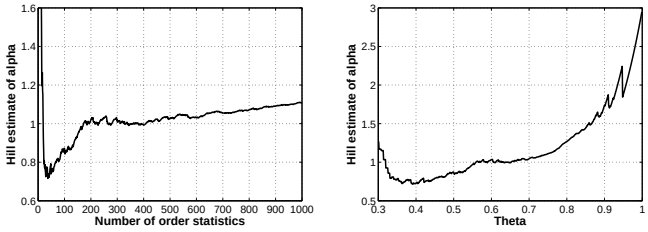
Figure 4.17: altHill plots for the Wikipedia data set.



(a) in-degree

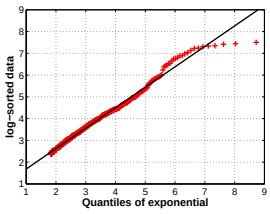


(b) PageRank (c=0.5)

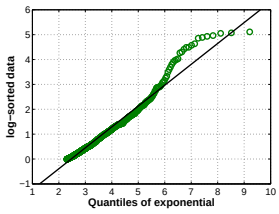


(c) PageRank (c=0.85)

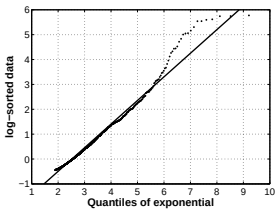
Figure 4.18: Hill ($left$) and altHill ($right$) plots for the Growing Network data set.



(a) in-degree: 1000 (α = 1.06) upper-statistics

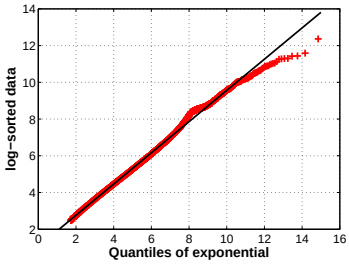


(b) PageRank (c=0.5): 1000 (α = 1.19) upper statistics

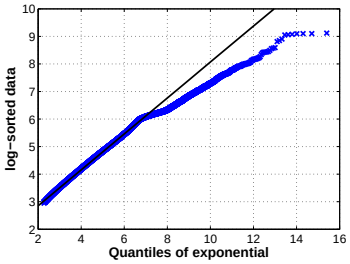


(c) PageRank (c=0.85): 1500 (α = 1.05) upper statistics

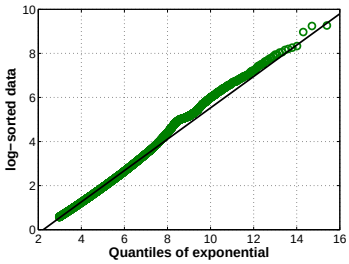
Figure 4.19: QQ lines for the Growing Network data set.



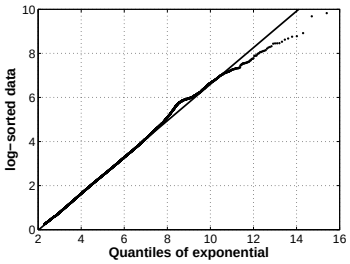
(a) in-degree: 500.000 (α = 1.18) upper statistics



(b) out-degree: 300.000 (α = 1.59) upper statistics



(c) PageRank (c=0.5): 250.000 (α = 1.40) upper statistics



(d) PageRank (c=0.85): 500.000 (α = 1.20) upper statistics

Figure 4.20: QQ lines for the Wikipedia data sets

CHAPTER 5

EXTREMAL DEPENDENCIES

5.1 Introduction

In this chapter we focus on dependence structure in power law graphs. In particular, we consider relation between two important power law characteristics: in-degree and PageRank.

We propose to employ the extreme value theory [11] and the theory of regular variation [100] that provide a range of statistical procedures designed to deal with multivariate data of which the marginal distributions exhibit power laws. We operate with the notion of *tail dependence* for a random vector (X, Y), the dependence between extremely large values of X and Y. Such tail dependence is characterized by an angular measure on $[0, 1]$, or $[0, \pi/2]$ (see Section 5.2.2 for formal definition) depending on the chosen norm. Informally, a concentration of the angular measure around 0 and/or 1 signals independence, while concentration around some other number $a \in (0, 1)$ suggests that a certain fraction of large values of Y comes together with the large values of X.

In Section 5.2 we start with an analytical approach for computing the tail dependencies between the in-degree and the PageRank. To this end, we again consider the PageRank as a solution of the general stochastic equation (2.4), and compute the angular measure analytically. The resulting angular measure is concentrated in points 0 and $a \in (1/2, 1)$, and the mass distribution depends on the network parameters. Such angular measure is a formalization of the common understanding that there are two main sources for high ranking: high in-degree and a high rank of one of the ancestors. Furthermore, the fraction of the measure mass in 0 stands for the proportion of highly ranked nodes that have a low in-degree. Thus, we obtain the description of the dependence structure, that is more informative and better relates

to reality that the correlation coefficient. We derive the results on tail dependence in Sections 5.2.1 and 5.2.2. In Section 5.2.4 we discuss the results and compare our findings to the graph data

Then, in Section 5.3 we compute the angular measures for in-degrees, out-degrees and PageRank scores in three large data sets from Section 4.1. Our experimental results reveal a dramatically different correlation structure in the three data sets. For instance, the results for in-degree and PageRank in Wikipedia strongly suggest an independence between these two parameters. Similar analysis for the Web graph reveals a non-trivial dependence structure. Finally, a preferential attachment graph shows a very strong dependence between in-degree and PageRank.

The analysis of extremal dependence leads us to propose a new rank correlation measure which is particularly plausible for power law data. The measure has the appealing property that it is especially sensitive to rank permutations for top-ranked nodes. Using the new correlation measure, we demonstrate that the PageRank ranking is not sensitive to moderate changes in the damping factor. The rank correlation measure is presented in Section 5.4.

Further, in next chapter we apply the proposed rank correlation measure to rank aggregation problems.

5.2 Characterization of tail dependence for in-degree and PageRank

As in Chapters 2 and 3, we model the distribution of the non-uniform PageRank (1.3) through the stochastic equation. Here we again study the general stochastic equation (2.4):

$$R \stackrel{d}{=} \sum_{i=1}^{N} A_i R_i + B, \tag{5.1}$$

where A_i's are independent and distributed as some random variable $A < 1$, and $B > 0$ is independent of the A_i's. Note that in this chapter we assume that N and B are independent. Next, we define

$$\bar{F}_1(u) := \mathbb{P}(N > u) \qquad \text{and} \qquad \bar{F}_2(u) := \mathbb{P}(R > u), \qquad u > 0,$$

and assume that $\bar{F}_1(u)$ is regularly varying with non-integer index $\alpha > 1$. We also assume that B in (5.1) has a lighter tail than N, that is, $\mathbb{P}(B > u) = o(\mathbb{P}(N > u))$ as $u \to \infty$. As a result, $\bar{F}_2(u)$ is also regularly varying. In fact, from Theorem 3.10(i) we have that

$$\bar{F}_2(u) \sim K \bar{F}_1(u) \qquad \text{as} \qquad u \to \infty, \tag{5.2}$$

where $K = (E(A))^{\alpha_N}[1 - \mathbb{E}(N)\mathbb{E}(A^{\alpha_N})]^{-1}$. Moreover, if we consider stochastic equation (1.7) for the standard PageRank (1.1), then we obtain that

$$K = \frac{c^\alpha}{d^\alpha - dc^\alpha}, \tag{5.3}$$

for $\mathbb{E}N = d$, $A = c/d$ and $B = 1 - c$. In the sequel we will only use the specific form (5.3) in Corollary 5.5 and Section 5.2.4. We also note that within same model (5.1), we could assume that the distribution of the R_i's is different from the one of R. In this case, if the tail of the R_i's is not heavier than the one of N, similarity (5.2) still holds (see Lemma 2.3(iv) and (vi) in Chapter 2).

We need to deal with a minor complication because $\bar{F}_1$ is not strictly decreasing, and we will in the sequel need to consider the behavior of its inverse function for small arguments. Instead of working with the generalized inverse $\bar{F}_1^{-1}(v) = \inf\{u > 0 | \bar{F}_1(u) \leq v\}$, which would make the proofs more involved, we prefer to simply work with some function that is strictly decreasing and asymptotically equivalent to $\bar{F}_1(u)$. Such a function can e.g. be defined as $f_1(u) := (1 + e^{-u})\bar{F}_1(u)$, for which the inverse function is well-defined. Thus, we arrive at the following:

$$\begin{aligned} \bar{F}_1(u) := \mathbb{P}(N > u) &\sim f_1(u), \quad \text{as } u \to \infty \\ \bar{F}_2(u) := \mathbb{P}(R > u) &\sim f_2(u), \quad \text{as } u \to \infty, \end{aligned} \tag{5.4}$$

where

$$f_1(u) = u^{-\alpha} L(u), \qquad f_2(u) = K u^{-\alpha} L(u) = K f_1(u),$$

for some slowly varying function $L(\cdot)$.

5.2.1 Tail dependence

Let us introduce two functions that are defined on $\mathbb{R}_+^2$, namely the *stable tail dependence function* [11],

$$\ell(x, y) = \lim_{t \downarrow 0} t^{-1} \mathbb{P}(\bar{F}_1(N) \leq tx \text{ or } \bar{F}_2(R) \leq ty) \tag{5.5}$$

and the function

$$r(x, y) := \lim_{t \downarrow 0} t^{-1} \mathbb{P}(\bar{F}_1(N) \leq tx, \bar{F}_2(R) \leq ty).$$

Provided that the limit in (5.5) exists, these are closely related since they satisfy $\ell(x, y) + r(x, y) = x + y$. The main result of this section gives the stable tail dependence function for N and R:

Theorem 5.1. *The function $r(x, y)$ for N and R is given by*

$$r(x, y) = \min\{x, y(\mathbb{E}A)^\alpha / K\}. \tag{5.6}$$

Consequently, $\ell(x, y) = \max\{y, x + y(1 - (\mathbb{E}A)^\alpha / K)\}$.

In the remainder of Section 5.2 we mainly work with $r(x, y)$ rather than $\ell(x, y)$, since its derivation is more appealing.

To prove Theorem 5.1 we need to use the following lemma.

Lemma 5.2. *As $u \to \infty$, the following asymptotic relation holds for any constant $C > 0$,*

$$\mathbb{P}(N > u, R > Cu) \sim \min\{f_1(u), (\mathbb{E}A/C)^\alpha f_1(u)\}.$$

We refer to Section 5.2.3 for the proof of this lemma, but the intuition behind it is clear. It follows from (5.1) and the strong law of large numbers that when N is large, we have $R \approx \mathbb{E}A \cdot N$. Therefore, when $\mathbb{E}A > C$, the event $\{R > Cu\}$ is already 'implied' by $\{N > u\}$, so the joint probability behaves just like $\mathbb{P}(N > u)$. When $\mathbb{E}A < C$, N needs to be larger for $R > Cu$ to hold, and the joint probability behaves like $\mathbb{P}(N > uC/\mathbb{E}A)$.

In order to understand Theorem 5.1 we fix $x, y > 0$ throughout this section and rewrite the joint probability in a form that enables application of Lemma 5.2. It will be convenient to use the functions

$$g_1(t) := f_1^{-1}(tx) \qquad \text{and} \qquad g_2(t) := f_2^{-1}(ty) = f_1^{-1}(ty/K) = g_1(ty/Kx), \tag{5.7}$$

which are well-defined for all $t > 0$, due to the monotonicity of f_1, and hence also f_2. The schematic derivation is as follows:

$$\begin{aligned}
&\mathbb{P}(\bar{F}_1(N) \le tx, \bar{F}_2(R) \le ty) \overset{1}{\sim} \mathbb{P}(f_1(N) \le tx, f_2(R) \le ty) \\
&= \mathbb{P}(N \ge g_1(t), R \ge g_2(t)) \overset{2}{=} \mathbb{P}\left(N \ge g_1(t), R \ge \left(\frac{y}{Kx}\frac{L(g_1(t))}{L(g_2(t))}\right)^{-1/\alpha} g_1(t)\right) \\
&\overset{3,1}{\sim} \mathbb{P}\left(N \ge g_1(t), R \ge \left(\frac{y}{Kx}\right)^{-1/\alpha} g_1(t)\right)
\end{aligned} \tag{5.8}$$

The statement of Theorem 5.1 now follows from Lemma 5.2 since $f_1(g_1(t)) = tx$, provided that each of the three steps indicated in (5.8) is justified. We resolve these issues as follows:

1. We deduce the asymptotic equivalence of the two probabilities from the asymptotic equivalence of the functions *inside* the probabilities. This step is intuitively clear but not mathematically rigorous. In the proof of Theorem 5.1 we will make the argument precise, see Section 5.2.3.

2. This step is fairly straightforward. Indeed, $v = f_1(u) = u^{-\alpha}L(u)$ implies $u = (v/L(u))^{-1/\alpha}$, so $f_1^{-1}(v) = (v/L(f_1^{-1}(v)))^{-1/\alpha}$. Hence, for $v = tx$, from (5.7) we obtain

   $$g_1(t) = \left(\frac{tx}{L(g_1(t))}\right)^{-1/\alpha} \quad \text{and also} \quad g_2(t) = \left(\frac{ty}{KL(g_2(t))}\right)^{-1/\alpha},$$

   so

   $$\frac{g_2(t)}{g_1(t)} = \left(\frac{y}{Kx}\frac{L(g_1(t))}{L(g_2(t))}\right)^{-1/\alpha}. \tag{5.9}$$

3. This is a consequence of the following statement (the proof of which can be found in Section 5.2.3), combined with issue 1.

   Lemma 5.3. *For all $x, y > 0$ we have $L(g_1(t)) \sim L(g_2(t))$ as $t \downarrow 0$.*

Now, in order to prove Theorem 5.1 we only need to resolve issue 1 twice in the derivation in (5.8). The formal proof is presented in Section 5.2.3.

5.2.2 Angular measure

In this section we find the angular measure that corresponds to the function $r(x, y)$ we found, but first we will give some preliminaries. In extreme value theory (see [11]), it has been shown that a unique (nonnegative) measure $H(\cdot)$ exists on the set $\Xi = \{\omega \in \mathbb{R}^2_+ \mid ||\omega|| = 1\}$, such that the stable tail dependence function ℓ can be expressed as

$$\ell(x, y) = \int_\Xi \max(\omega_1 x, \omega_2 y) H(d\omega). \tag{5.10}$$

Here $||\cdot||$ is a norm that may be chosen freely, but for (5.10) to hold, the measure has to be normalized in such a way that

$$\int_\Xi \omega_1 H(d\omega) = \int_\Xi \omega_2 H(d\omega) = 1,$$

so that we have $\ell(x, 0) = x$ and $\ell(0, y) = y$, as should. In this work we choose the $||\cdot||_1$ norm, for which $||\omega||_1 = |\omega_1| + |\omega_2|$, since that is easiest to work with. Then (5.10) can be rewritten as

$$\ell(x, y) = \int_0^1 \max\{wx, (1-w)y\} H(dw),$$

and the normalization becomes

$$\int_0^1 w H(dw) = \int_0^1 (1-w) H(dw) = 1. \tag{5.11}$$

Here we let $w = \omega_1$, and we identify the measures on Ξ and $[0, 1]$. By (5.11) it follows that the function $r(x, y)$ can be written as

$$\begin{aligned} r(x, y) &= \int_0^1 wx H(dw) + \int_0^1 (1-w)y H(dw) - \int_0^1 \max\{wx, (1-w)y\} H(dw) \\ &= \int_0^1 \min\{wx, (1-w)y\} H(dw). \end{aligned} \tag{5.12}$$

We will now derive the function $r(x, y)$ in case when the angular measure has masses in 0 and a only, as we suspect to be the case for in-degree and PageRank.

First of all, the normalization (5.11) boils down to $aH(a) = H(0) + (1-a)H(a) = 1$, which is easily solved to give

$$H(0) = 2 - 1/a \qquad \text{and} \qquad H(a) = 1/a. \tag{5.13}$$

Note that H has total measure 2 (as also follows for the general case by summing both integrals in (5.11)), and that $H(0) > 0$ implies that $a > 1/2$. Combining (5.12) and (5.13), the function $r(x, y)$ can now be written as

$$r(x, y) = \min\{x, y(1/a - 1)\}.$$

This is a very similar form as we found earlier in (5.6), and it is not difficult to see that the expressions are equal for $a = K/(K + (\mathbb{E}A)^\alpha)$. So by uniqueness, the angular measure of N and R is indeed a two-point measure, and after using (5.13) we arrive at

Theorem 5.4. *The angular measure with respect to the $||\cdot||_1$ norm of N and R is a two-point measure, with masses*

$$\begin{aligned} H(0) &= 1 - \frac{(\mathbb{E}A)^\alpha}{K} && \text{in } 0, \\ H(a) &= 1 + \frac{(\mathbb{E}A)^\alpha}{K} && \text{in } a = \frac{K}{K + (\mathbb{E}A)^\alpha}. \end{aligned}$$

Corollary 5.5. *If K is given by (5.3) with $\mathbb{E}N = d$ and $\mathbb{E}A = c/d$, then the angular measure of N and R is a two-point measure, with masses*

$$\begin{aligned} H(0) &= c^\alpha d^{(1-\alpha)} && \text{in } 0, \\ H(a) &= 2 - c^\alpha d^{(1-\alpha)} && \text{in } a = \left(2 - c^\alpha d^{(1-\alpha)}\right)^{-1}. \end{aligned} \tag{5.14}$$

5.2.3 Proofs

Proof of Lemma 5.2. The proof is based on the strong law of large numbers. Informally, we use the fact that if N is large, then (5.1) implies $R \approx \mathbb{E}A \cdot N$.

Assume first that $C < \mathbb{E}A$. Then we write

$$\mathbb{P}(N > u, R > Cu) = \mathbb{P}(N > u)\mathbb{P}(R > Cu | N > u), \tag{5.15}$$

and we further obtain

$$\begin{aligned} \mathbb{P}(R > Cu | N > u) &\geq \mathbb{P}\left(\sum_{i=1}^{\lfloor u \rfloor} A_i R_i + B > Cu\right) \geq \mathbb{P}\left(\sum_{i=1}^{\lfloor u \rfloor} A_i R_i > Cu\right) \\ &= \mathbb{P}\left(C^{-1}u^{-1}\sum_{i=1}^{\lfloor u \rfloor} A_i R_i > 1\right) \to 1 \text{ as } u \to \infty, \end{aligned}$$

where the convergence holds by the strong law of large numbers for any $C < \mathbb{E}A$. Hence when $C < \mathbb{E}A$ the result follows directly from (5.4) and (5.15).

Now assume that $C > \mathbb{E}A$. We would like to show that

$$\lim_{u\to\infty} \frac{\mathbb{P}(N > u, R > Cu)}{f_1([C/\mathbb{E}A]u)} \to 1. \tag{5.16}$$

Then the result of the lemma will follow since $L(u) \sim L([C/\mathbb{E}A]u)$ as $u \to \infty$. For the proof, we choose a sufficiently small δ so that we can break the joint probability into three terms:

$$\begin{aligned}\mathbb{P}(N > u, R > Cu) = \mathbb{P}(&N > [C/\mathbb{E}A + \delta]u, R > Cu)\\ &+ \mathbb{P}([C/\mathbb{E}A - \delta]u < N \le [C/\mathbb{E}A + \delta]u, R > Cu)\\ &+ \mathbb{P}(u < N \le [C/\mathbb{E}A - \delta]u, R > Cu).\end{aligned} \tag{5.17}$$

Exactly as in case $C < \mathbb{E}A$, using (5.4), we have

$$\lim_{u\to\infty} \frac{\mathbb{P}(N > [C/\mathbb{E}A + \delta]u, R > Cu)}{f_1([C/\mathbb{E}A]u)} = \lim_{u\to\infty} \frac{\mathbb{P}(N > [C/\mathbb{E}A + \delta]u)}{f_1([C/\mathbb{E}A]u)} = 1 + O(\delta). \tag{5.18}$$

Moreover, applying the argument as in the case when $C < \mathbb{E}A$, from the law of large numbers we obtain that

$$\mathbb{P}(R > Cu | u < N \le [C/\mathbb{E}A - \delta]u) \downarrow 0 \text{ as } u \to \infty,$$

and thus

$$\begin{aligned}0 &\le \lim_{u\to\infty} \frac{\mathbb{P}(u < N \le [C/\mathbb{E}A - \delta]u, R > Cu)}{f_1([C/\mathbb{E}A]u)}\\ &\le \lim_{u\to\infty} \frac{\mathbb{P}(N > u)\mathbb{P}(R > Cu | u < N \le [C/\mathbb{E}A - \delta]u)}{f_1([C/\mathbb{E}A]u)} = 0.\end{aligned} \tag{5.19}$$

Finally, we get

$$\begin{aligned}0 &\le \lim_{u\to\infty} \frac{\mathbb{P}([C/\mathbb{E}A - \delta]u < N \le [C/\mathbb{E}A + \delta]u, R > Cu)}{\mathbb{P}(N > [C/\mathbb{E}A]u)}\\ &\le \lim_{u\to\infty} \frac{\mathbb{P}([C/\mathbb{E}A - \delta]u < N \le [C/\mathbb{E}A + \delta]u)}{\mathbb{P}(N > [C/\mathbb{E}A]u)}\\ &= \lim_{u\to\infty} \frac{f_1([C/\mathbb{E}A - \delta]u) - f_1([C/\mathbb{E}A + \delta]u)}{f_1([C/\mathbb{E}A]u)} = O(\delta).\end{aligned} \tag{5.20}$$

The result (5.16) now follows from (5.17)–(5.20) by letting $\delta \downarrow 0$.

In the case $C = \mathbb{E}A$ the argument is similar, only we write

$$\begin{aligned}\mathbb{P}(N > u, R > \mathbb{E}Au) = \mathbb{P}(&N > [C/\mathbb{E}A + \delta]u, R > Cu)\\ &+ \mathbb{P}(u < N \le [C/\mathbb{E}A + \delta]u, R > Cu).\end{aligned}$$

This completes the proof of the lemma. □

Proof of Lemma 5.3. Applying the Potter bounds of Lemma 3.3 in Chapter 3, we obtain that for all $\vartheta > 1, \delta > 0$ one can choose t sufficiently small such that

$$\vartheta^{-1}\left[\max\left\{\frac{g_1(t)}{g_2(t)}, \frac{g_2(t)}{g_1(t)}\right\}\right]^{-\delta} \leq \frac{L(g_1(t))}{L(g_2(t))} \leq \vartheta\left[\max\left\{\frac{g_1(t)}{g_2(t)}, \frac{g_2(t)}{g_1(t)}\right\}\right]^{\delta}$$

which by (5.9) is the same as

$$\vartheta^{-1}\left[\max\left\{\frac{g_1(t)}{g_2(t)}, \frac{g_2(t)}{g_1(t)}\right\}\right]^{-\delta} \leq \frac{Kx}{y}\left(\frac{g_1(t)}{g_2(t)}\right)^{\alpha} \leq \vartheta\left[\max\left\{\frac{g_1(t)}{g_2(t)}, \frac{g_2(t)}{g_1(t)}\right\}\right]^{\delta}.$$

From the first inequality above we get

$$\liminf_{t\downarrow 0} \vartheta^{1/\alpha}\left[\max\left\{\frac{g_1(t)}{g_2(t)}, \frac{g_2(t)}{g_1(t)}\right\}\right]^{\delta/\alpha}\left(\frac{Kx}{y}\right)^{1/\alpha}\frac{g_1(t)}{g_2(t)} \geq 1$$

for all $\vartheta > 1$, $\delta > 0$. Taking $\vartheta \to 1$ and $\delta \downarrow 0$ we obtain that

$$\liminf_{t\downarrow 0}\left(\frac{Kx}{y}\right)^{1/\alpha}\frac{g_1(t)}{g_2(t)} \geq 1.$$

Analogously, we can show that

$$\limsup_{t\downarrow 0}\left(\frac{Kx}{y}\right)^{1/\alpha}\frac{g_1(t)}{g_2(t)} \leq 1.$$

so that the limit of the left-hand side is 1. This implies the result, again by (5.9). □

Proof of Theorem 5.1. Since $\bar{F}_i(u) \to 0$ and $|\bar{F}_i(u) - f_i(u)| = o(\bar{F}_i(u))$, $i = 1, 2$, as $u \to \infty$, then for any small $\varepsilon > 0$ we can choose t_1 small enough so that for any $t \leq t_1$ and $u > 0$ that satisfy $\bar{F}_1(u) \leq tx$ we also have $|\bar{F}_1(u) - f_1(u)| \leq \varepsilon|\bar{F}_1(u)|$, and hence $|\bar{F}_1(u) - f_1(u)| \leq \varepsilon tx$. Moreover, we can choose $t_2 \leq t_1$ small enough such that $\bar{F}_2(u) \leq ty$ implies $|\bar{F}_2(u) - f_2(u)| \leq \varepsilon ty$ for all $t \leq t_2$. Also, for any small $\delta > 0$ it follows from Lemma 5.3 that there exists a positive number $t_3 \leq t_2$ such that for all $t \leq t_3$,

$$1 - \delta \leq \frac{L(g_1((1+\varepsilon)t))}{L(g_2((1+\varepsilon)t))} \leq 1 + \delta.$$

If we now fix some small $\varepsilon > 0$ and $\delta > 0$, the above implies for any $t \leq t_3$ that

$$\begin{aligned}
&\mathbb{P}(\bar{F}_1(N) \leq tx, \bar{F}_2(R) \leq ty) \\
&\quad = \mathbb{P}(f_1(N) \leq (f_1(N) - \bar{F}_1(N)) + tx, f_2(R) \leq (f_2(R) - \bar{F}_2(R)) + ty) \\
&\quad \leq \mathbb{P}(f_1(N) \leq (1+\varepsilon)tx, f_2(R) \leq (1+\varepsilon)ty) \\
&\quad = \mathbb{P}\left(N \geq g_1((1+\varepsilon)t), R \geq g_2((1+\varepsilon)t)\right) \\
&\quad = \mathbb{P}\left(N \geq g_1((1+\varepsilon)t), R \geq \left(\frac{y}{Kx}\frac{L(g_1((1+\varepsilon)t))}{L(g_2((1+\varepsilon)t))}\right)^{-1/\alpha} g_1((1+\varepsilon)t)\right) \\
&\quad \leq \mathbb{P}\left(N \geq g_1((1+\varepsilon)t), R \geq \left(\frac{y}{Kx}(1+\delta)\right)^{-1/\alpha} g_1((1+\varepsilon)t)\right).
\end{aligned}$$

Note that the above closely follows the derivation in (5.8), with $\sim$ signs replaced by inequalities; in particular the 4th and 5th lines follow immediately from (5.7) and (5.9) upon replacing t by $(1+\varepsilon)t$. Now we can apply Lemma 5.2 to the above (note that $f_1(g_1((1+\varepsilon)t)) = (1+\varepsilon)tx$) and then let $t \to 0$, to obtain

$$\limsup_{t\to 0} t^{-1}\ \mathbb{P}(\bar{F}_1(N) \le tx, \bar{F}_2(R) \le ty) \ \le\ (1+\varepsilon)\min\{x, (1+\delta)y(\mathbb{E}A)^\alpha/K\}.$$

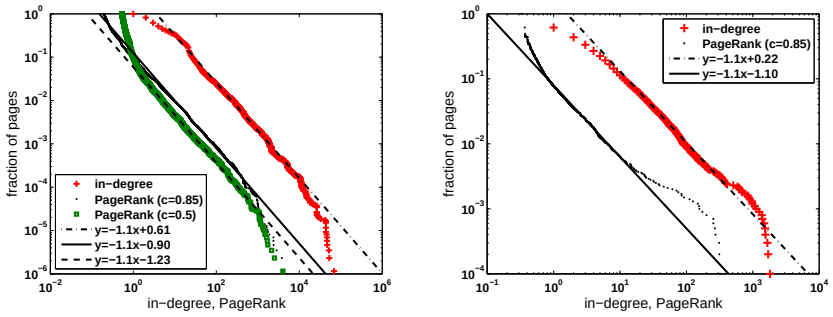
Similarly we can obtain

$$\liminf_{t\to 0} t^{-1}\ \mathbb{P}(\bar{F}_1(N) \le tx, \bar{F}_2(R) \le ty) \ \ge\ (1-\varepsilon)\min\{x, (1-\delta)y(\mathbb{E}A)^\alpha/K\},$$

so that the statement of the theorem follows by letting ε and δ go to 0. □

5.2.4 Examples and discussion

We compare the above results to the measurements on two different network structures: the EU-2005 data set and the Growing Network from Section 4.1. In Figure 5.1 we present log-log plots for in-degree and PageRanks with fitted straight lines.



(a) Web data set (b) Growing Network data set

Figure 5.1: Cumulative log-log plots for in-degree and PageRanks.

Following [11, p.328] we define an estimator of the angular measure. We are interested in the dependencies between two regularly varying characteristics of a node, namely the in-degree N and the PageRank R. Recall that w is the number of nodes in the graph. Let (N_j, R_j) be observations of (N, R) for the corresponding node j. Then we use the rank transformation of (N, R), leading to $\{(r_j^N, r_j^R),\ 1 \le j \le w\}$, where r_j^N is the descending rank of N_j in $(N_1, \ldots, N_n)$ and r_j^R is the descending rank of R_j in $(R_1, \ldots, R_w)$. Next we apply a coordinate transform $(r_j^N, r_j^R) \longrightarrow (r_j, \Theta_j)$,

given by

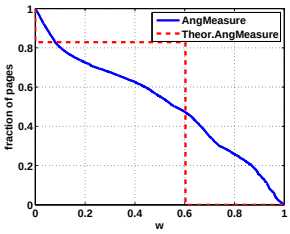
$$(r_j, \Theta_j) = \text{Trans}\left(\frac{1}{r_j^N}, \frac{1}{r_j^R}\right),$$

where we set $\text{Trans}(x, y) := (x+y, x/(x+y))$ since all results of this section are proven for the $||\cdot||_1$ norm. Alternatively, in Sections 5.3 and 5.4 we use the polar coordinate transformation: $\text{Trans}(x, y) := \left(\sqrt{x^2+y^2}, \arctan(y/x)\right)$. However, in this case we need to transform the angular measure in Theorem 5.4 to the corresponding measure w.r.t. the $||\cdot||_2$ norm using formula (8.38) in [11]. Now we need to consider k points $\{\Theta_j : r_j \geq r_{(k)}\}$, where $r_{(k)}$ is the kth largest in $(r_1, \ldots, r_w)$, and make a plot for the cumulative distribution function of Θ, which gives the estimation of the probability measure $H(\cdot)/2$. The question how to choose the right k can be solved by employing the Starica plot (see Section 5.3.1).

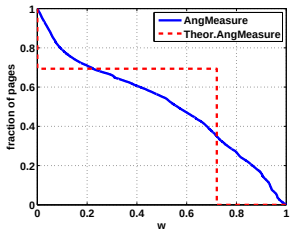
From (5.14) we can calculate the predicted angular measure concentrated in 0 and a. For the Web data sample with average in-degree $d = 22.2974$, taking $c = 0.5$ and $c = 0.85$, we obtain that $a_{0.5} = 0.6031$, $H(a_{0.5})/2 = 0.8290$, and $a_{0.85} = 0.7210$, $H(a_{0.85})/2 = 0.6934$, respectively. Recall that the values of $H(a)/2$ estimate the fraction of highly ranked pages whose large PageRank is explained by large in-degree. Observe that according to the model, this fraction becomes larger if c decreases.

In Figure 5.2(a) and 5.2(b) we plot the theoretical angular measures together with the empirical ones. The comparison between the graphs shows that there is only a very rough similarity to be seen, in the sense that the value of $H(0)/2$ is a reasonable estimate for the fraction of pages with high PageRank and small in-degree (corresponding to the 'turn' around 0.8). However, the 'point mass' at a seems to be spread out in an almost uniform manner. To understand this, we should realize that the theoretical two-point measure we found is only a formalization of the idea that each large PageRank value has to be either due to a large in-degree, or due to a large contributing PageRank. In the data (representing 'reality'), such a strict division is not reasonable; for instance there will surely be pages with high PageRank due to a high in-degree *and* a high contributing PageRank, or due to more than one high contributing PageRanks. Thus we see that although our model roughly captures the idea of different causes for large PageRank values, it is not subtle enough to properly represent the angular measure as found from a realistic data set. Future work could try to investigate how to improve the model in that respect, mainly by studying the dependencies amongst the R_i in (5.1), or between the R_i on the one hand and N on the other.

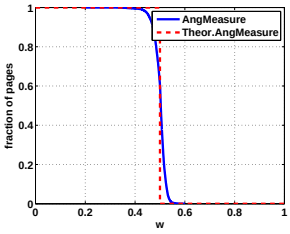
Finally, we perform experiments on the Growing Network. Clearly, in our model based on stochastic equation (5.1) we can not assume anymore that R is distributed as the R_i's since R_i's are the ranks of 'younger' nodes, and presumingly, the R_i will have lighter tails than R itself. Assuming that $P(R_i > u) = o(P(N > u))$ as $u \to \infty$, from Lemma 2.3(iv) we obtain that for this simple model the value of K is just $K = (c/d)^\alpha$. Substituting this into (5.14) gives us $a = 1/2,\ H(a) = 2,\ H(0) = 0$, i.e. the measure is concentrated in one point $a = 1/2$. In Figure 5.2(c) we again



(a) Web data set: c=0.5, k=100.000



(b) Web data set: c=0.85, k=100.000



(c) Growing Network data set: c=0.85, k=6.000

Figure 5.2: Angular measure and theoretically predicted angular measure.

plot the empirical and theoretical measures, which match perfectly. We see that in synthetic graphs constructed by the preferential attachment rule, large PageRank is always due to large in-degree, and this can be easily captured by our stochastic model.

In further research, it will be interesting to consider other graph models of the Web, for instance, a configuration model (see Section 1.3.3). The configuration model is not as centered as the preferential attachment network, and it is known to be close to the tree structure. Thus, one may expect that the stochastic equation provides an accurate description of the dependencies between in- degree and PageRank in such a model.

The derived two-point measure is only a first-order approximation of the complex angular measure observed on the data, since the realistic situation is way more complex than our simplified model. Further modifications of the model are needed in order to adequately describe the dependencies in real-life networks.

5.3 Evaluating statistical dependencies in Web graphs

In this section, we follow the book of Resnick [100]. We compute angular measures for in-degrees, out-degrees and PageRank scores in three large data sets studied in Section 4.1. Our experimental results reveal significant differences in dependence structures in these data sets.

5.3.1 Angular measure estimator

Suppose we are interested in analyzing the dependencies between two regular varying characteristics of a node, X and Y. As in Section 5.2.4 we define $\{(r_j^x, r_j^y),\ 1 \le j \le w\}$, such that r_j^x and r_j^y are the descending ranks of observations of X and Y for page j, respectively. Next we choose $k = 1, \ldots, w$ and apply the polar coordinate transform in (5.2.4):

$$\text{Trans}\left(\frac{k}{r_j^x}, \frac{k}{r_j^y}\right) = (R_{j,k}, \Theta_{j,k}). \tag{5.21}$$

Now, we obtain the following estimator for the angular measure:

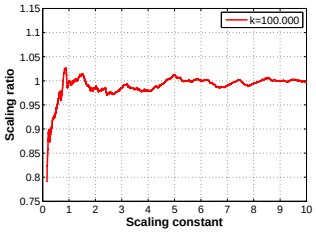
$$\frac{\sum_{i=1}^{w} \mathbf{1}[R_{i,k} > 1, \Theta_{i,k} \in A]}{\sum_{i=1}^{w} \mathbf{1}[R_{i,k} > 1]}, \tag{5.22}$$

where $\mathbf{1}[\cdot]$ is an indicator function. More details can be found in Chapter 9 of [100].

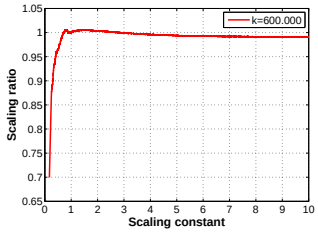
It was proved in [11, 100] that the empirical measure converges to a proper distribution on $[0, \pi/2]$ as $w, k \to \infty$, $k/w \to 0$. That is, ideally, we need to consider only a relatively small part of a large data set. In practice the problem remains: how to choose a suitable value of k? In the case of bivariate regular variation, this can be determined by making a *Starica* plot. This technique helps to determine where the bivariate power law behavior actually 'starts'. To make the Starica plot, we consider radii $R_{1,k}, \ldots, R_{w,k}$ from (5.21) and rank them in descending order $R_{(1)} \ge \ldots \ge R_{(w)}$ as before. To get Starica plot we graph

$$\left\{\left(\frac{R_{(j)}}{R_{(k)}}, \frac{R_{(j)}}{R_{(k)}} \cdot \frac{j}{k}\right),\ 1 \le j \le w\right\}, \text{ or } \left\{\left(R_{(j)}, \frac{R_{(j)} j}{\sum_{i=1}^{n} \mathbf{1}_{\{R_{i,k} \ge 1\}}}\right),\ 1 \le j \le w\right\}.$$

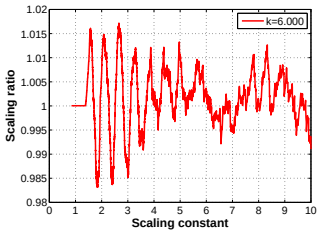
The idea is that for suitable k the ratio in the ordinate should be roughly a constant and equal 1 for the values of the abscissa in the neighborhood of 1. The plot looks different for the different parameters k and one can either find a suitable k by trial and error or use numerical algorithms to compute optimal k. A Starica plot for good k will have a region in the right neighborhood of $x = 1$ where the plot is hugging the $y = 1$ line. If the line is going steep up at $x = 1$ then the chosen k is too large. On the other hand, if the graph stabilizes around $y = 1$ for some $x < 1$ then it means that k is too small, and we miss some valuable tail data. We refer to Resnick [100] for more details and references.



(a) Eu-2005 data set: k=100.000



(b) Wikipedia data set: k=600.000



(c) Growing Network data set: k=6.000

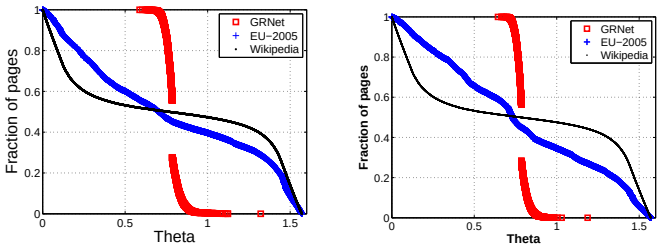
Figure 5.3: Starica plot for in-degree and PageRank ($c = 0.85$).

In Figure 5.3 we present Starica plots for the pair of in-degree and PageRank (c=0.85). The plots behave nicely in all three data sets, which makes our angular measure more reliable. The Growing Network exhibits an ideal Starica plot (Figure 5.3(c)). In [113] we provided the plots and the appropriate values of k for the other combinations: in-degree and PageRank ($c = 0.5$), in-degree and out-degree; and out-degree and PageRank ($c = 0.5$, $c = 0.85$).

5.3.2 Dependence measurements on the data

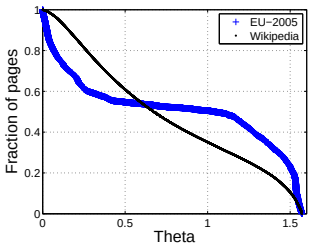
As in Section 4.1 we chose the EU-2005, the Wikipedia and the Growing Network data sets to represent different network structures. After defining a suitable k, we compute the pairwise angular measure. In Figure 5.4 we depict $\theta \in [0, \pi/2]$ against the estimated probability angular measure $[\theta, \pi/2]$, which, according to (5.22), equals to the fraction of pages i where the angle $\Theta_{i,k}$ is greater or equal to θ provided that $R_{i,k} > 1$.

The results are striking. Let us first look at Figures 5.4(a),(b) that characterize

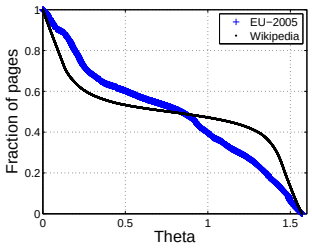


(a) in-degree and PageRank (c=0.85)

(b) in-degree and PageRank (c=0.5)



(c) in-degree and out-degree



(d) out-degree and PageRank (c=0.85)

Figure 5.4: Cumulative functions for Angular Measures.

the dependence between in-degree and PageRank. For the Wikipedia data set we observe that about half of the observations are concentrated around 0 whereas another half is close to $\pi/2$. This suggests an independence of the tails of in-degree and PageRank (c=0.85 and c=0.5). That is, in the Wikipedia data set an extremely high in-degree almost never implies an extremely high ranking. This may be due to the fact that many links in Wikipedia are created by automated processes rather than human processes. The picture is completely the opposite for Growing Networks, where the angular measure is entirely concentrated around $\pi/4$ indicating a complete dependence. Thus, in highly centralized preferential attachment graphs, most connected nodes are also most highly ranked.

Finally, the Web graph exhibits a subtle dependence structure that results in angular measure which is almost uniform on $[0, \pi/2]$. This suggest that PageRank popularity measure can not be replaced by in-degree without significant disturbance in the ranking (of course, in-degree can not be used as a popularity measure for

many other reasons, for instance, because it is easy to spam by creating link farms; we refer to [69] for further discussion of PageRank and other popularity measures).

The picture is different in Figure 5.4(c) where we depict the angular measure for in-degree and out-degree in the Web and in Wikipedia. In the Web, the in- and out-degree tend to be independent which justifies the distinction between hubs and authorities [63]. In Wikipedia the in- and out-degrees are dependent but this dependence is not absolute.

Finally, the dependence between out-degree and PageRank in the Web and Wikipedia in Figure 5.4(d) resembles the patterns observed for in-degree and PageRank.

Analysis of dependencies in real-life graph and synthetic data contributes towards a better understanding and modeling of complex graph structures. Clearly, for adequate modeling, it is not sufficient to maintain power laws. For instance, it was already argued in [38] that robustness of Internet power law router graph is in strong disagreement with a preferential attachment model. Likewise, our analysis clearly reveals a striking disagreement of the preferential attachment graph with dependence structure of the Web and the Wikipedia. Better models have to be sought and existing models have to be thoroughly analyzed before we can conclude that they adequately reflect important features of complex networks.

5.4 The Θ rank correlation measure

We start by noting again that the estimator of the angular measure described in Sections 5.2.4 and 5.3.1 is based on a rank transformation. This is clearly seen from formula (5.21) where only rank of the parameters X and Y plays a role. This observation naturally leads to a new measure for rank correlations.

In summary, our idea is as follows. As before, we define r_i^1 and r_i^2 as a ranking order of page i in scheme 1 and 2, respectively, where $i = 1 \dots n$. Now we suggest to represent the difference between the two rank positions of i by the angle

$$\Theta_i = \arctan(r_i^1 / r_i^2).$$

For example, in Figure 5.5, Θ_i is depicted for a node that has rank 3 in scheme 1 and rank 6 in scheme 2. Note that this is exactly the angle in $[0, \pi/2]$ computed in (5.21) in order to construct the angular measure estimator. The value Θ close to $\pi/4$ means a relatively small change in ranking. On the other hand, Θ around $\pi/2$ means that the node i is much better off with scheme 2, and the value close to 0 says that the node is ranked much higher by scheme 1. Thus, we actually measure the rank difference for node i in radians! Having computed Θ_i for every i (or for a certain set of highly ranked nodes i) we construct a corresponding empirical cumulative distribution function for Θ. As in the previous section, the resulting angular measure can be used to characterize the rank correlations.

We note that we characterize the rank correlation by a measure or a plot rather than a number. Compared to the common rank correlation measures such as Kendall's

τ [62] and Spearman's ρ [107], our proposed measure has an important advantage that it is able to reveal the slightest rank disturbance among highly rank nodes while neglecting even moderate perturbations among lowly ranked nodes. Indeed, the Kendall's τ and the Spearman's ρ are defined as

$$\tau = 1 - \frac{2d_\Delta}{n(n-1)}, \text{ and } \rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \tag{5.23}$$

where d_Δ is the number of pairs in the symmetric difference of $\{(r_i^1, r_j^1), 1 \le i < j \le n\}$ and $\{(r_i^2, r_j^2), 1 \le i < j \le n\}$; and $d_i = r_i^1 - r_i^2$ is the difference between two ranks of page i. Thus, the Θ rank correlation measure actually evaluates the rank disturbance visible for users. Certainly, the arctan$(\cdot)$ function makes our measure symmetric with respect to the schemes 1 and 2

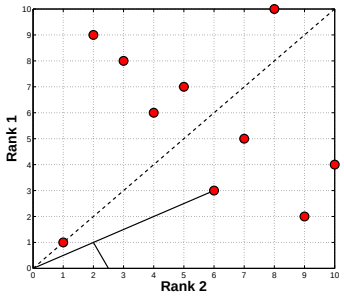
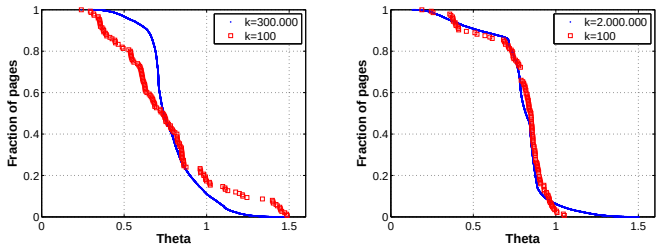


Figure 5.5: Rank Correlation.

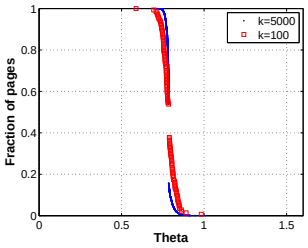
Naturally, in this framework, it is also possible to compute such angular measure only for the top ranked pages. This can be done along the same lines as in Section 5.3 as follows. Based on the polar transformation (5.21) we can separate top ranked pages by considering only points $\{\Theta_{i,k} : R_{i,k} > 1\}$. Here the question of choosing k does not arise anymore. Indeed, the technique involving Starica plot was needed to get an idea where the power law behavior 'starts' in order to measure statistical dependency for the heavy-tailed data as in [100]. On the other hand, if we are interested in rank correlations, we may simply pick the k that gives us the top proportion of pages we are interested in. Note that by increasing k we do not change the observed values of Θ, we merely increase their number. As a result, in the angular measure, each observation will simply have less weight. On contrary, decreasing k means 'zooming in' the rank perturbations on the top.

One more advantage of the proposed correlation measure is the fast and easy implementation since for each node i, only the fraction r_i^1/r_i^2 has to be computed.

Below we present the example of the proposed rank correlation measure in the Growing Network, the Web and the Wikipedia data sets from Section 4.1. For every data set we rank pages according to values of the PageRank with damping factor



(a) EU-2005 (b) Wikipedia



(c) Growing Network

Figure 5.6: Cumulative functions for Angular Measures for PageRank (c=0.5) and PageRank (c=0.85)

$c = 0.5$ (ranking scheme 1), and $c = 0.85$ (ranking scheme 2). In Figure 5.6 we plot cumulative functions for angular measures for $k = 100$ and the values of k's that have been chosen according to the Starica plots as described in Section 5.3 (see e.g. Figure 5.3).

For Growing Network data set we observe the strong correlation between ranking schemes. We can also conclude that in Wikipedia the change in the damping factor affects only about 20% of considered pages, in the top-hundred group as well as in the larger group. For the Web data, the correlation between ranking is not significant for approximately half of the pages. However, for the top pages, the difference in the damping factor mixes up the order of ranking.

The idea of a angular measure estimator is naturally extended to yield the Θ rank correlation measure. The main idea of this measure is that we characterize the rank correlations by a cumulative distribution of Θ_i's, where $i = 1, \ldots n$. This way, one can actually see how many pages change their ranks significantly. Such measure is substantially more informative than just one number, that represents the correlation

in the whole graph. For instance, Melucci [80] noticed that Kendall's τ tends to grow close to one for large data sets. The author provides an example where Kendall's τ for ranking orders of only a few hundred Web pages becomes almost 1, in spite of the large number of rank perturbations. We remark however that if for some reason having one number is necessary, one can always compute, e.g. the expected deviation of Θ from $\pi/4$.

As mentioned before, the proposed correlation measure is quite harsh with respect to lowly ranked nodes. Indeed, the node ranked 1000 must fall all the way to 2000 to make the same effect as number 1 becoming number 2. We would like to emphasize that such discrepancy is especially suitable for ranking order emerging from a heavy-tailed data, such as PageRank or in-degree. This is because in such data, there is a huge difference between the highest values of the realizations, cf. [43].

In the next chapter we apply the Θ rank correlation measure for various problems of rank aggregation.

CHAPTER 6

RANK AGGREGATION

In this chapter we report on work in progress, that was started during a research visit at `Yahoo!Research Barcelona` in November 2008. The goal of the project was to apply the extremal dependencies and angular measure to the problem of rank correlation.

Rank aggregation is an important and well-studied problem in Information Retrieval. The purpose of rank aggregation is to combine several ranking lists in one new list that obtains a ranking of better quality. Thus, rank aggregation gives a way to improve the quality of Web search results; in practice we aggregate results that are obtained according to different criteria, different parameters of retrieval algorithms, or by different search engines.

A significant component in the problem of rank aggregation is to evaluate how much better is the new ranking list compared with the input ranking lists. In the cases when an "ideal" ranking list is given, the evaluation is defined by the value of correlation between new ranking list and the ideal list. Thus, when the value of the correlation coefficient between the new list and the ideal list is close to 1, we consider that the rank aggregation algorithm produces a good ranking. There are several correlation measures, however, Kendall's τ [62] measure and Spearman's ρ [107] measure are the ones that are most commonly used. Here we recall the definition of these two measures. Given that $\sigma_1(k)$ and $\sigma_2(k)$ are the ranking orders of a page k, $k = 1 \dots n$, in two ranking schemes 1 and 2, respectively, we define Kendall's τ and Spearman's ρ as follows:

$$\tau = 1 - \frac{2d_\Delta}{n(n-1)}, \text{ and } \rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)},$$

where d_Δ is the number of pairs in the symmetric difference of $\{(\sigma_1(i), \sigma_1(j)), 1 \leq i < j \leq n\}$ and $\{(\sigma_2(i), \sigma_2(j)), 1 \leq i < j \leq n\}$; and $d_i = \sigma_1(i) - \sigma_2(i)$ is the difference

between two ranks of page i. From the above definitions it is clear that none of the two measures distinguishes rank disturbance in the high ranks or in the low ranks. In other words, they penalize errors equally at the top and at the bottom of the ranking lists. However, in many cases we are more interested in coincidences at the top than at the bottom of the lists. In particular, in Web search only 40% of users are viewing more than first 10 results returned for search query [57].

Shien [106], and Pinto da Costa and Soares [96] suggest to use deferent weights to emphasize changes in the top ranks for Kendall's τ and Spearman's ρ, respectively. Recently, new approach based on average precision was introduced by Yilmaz et al. [116]. In our turn, we propose to use the angular measure as a correlation measure between different rankings. Indeed, the angular measure gives more weight to the errors at the high rankings. Moreover, since the angular measure is specially designed for measuring correlations between power law distributed parameters, we can assume that it provides a good measure of comparing Web-related rankings. In the next section, we define new distance measure between ranking lists based on the angular measure. Using this distance we perform experiments for two real data sets in Section 6.2. Further, in Section 6.3 we discuss future research.

6.1 Angular measure for rank correlation

In the previous chapter we defined the angular measure. Here we recall the definition. Let $\sigma_1(k)$ and $\sigma_2(k)$ be the ranking orders of a page k in two ranking schemes 1 and 2, respectively, where $k = 1 \dots n$. We note that we refer to the page with rank 1 as to the most "interesting" page. Now, we apply the transformation with respect to some norm

$$(R, \Theta_k) = \mathrm{Trans}(\sigma_1(k), \sigma_2(k)).$$

In the sequel we use $\mathrm{Trans}(x, y) = (x + y, x/(x + y))$.

Next, we suggest to represent the difference between the two rank positions of k by the angle $\Theta_k = \sigma_1(k)/(\sigma_1(k) + \sigma_2(k))$. Then we define the distance measure between two ranking lists as a sum of deviations from the case of $\sigma_1(k) = \sigma_2(k)$:

$$d(\sigma_1, \sigma_2) = \sum_{k=1}^{n} \left| \frac{1}{2} - \frac{\sigma_1(k)}{\sigma_1(k) + \sigma_2(k)} \right| = \frac{1}{2} \sum_{k=1}^{n} \frac{|\sigma_1(k) - \sigma_2(k)|}{\sigma_1(k) + \sigma_2(k)}. \tag{6.1}$$

Note that $d(\cdot, \cdot)$ is symmetrically defined, and it is a metric.

Remark 6.1. *If σ_1, σ_2 and σ_3 are full ranking lists, i.e., for every k and i rank $\sigma_i(k)$ is defined, then the triangle inequality holds:*

$$d(\sigma_1, \sigma_3) \le d(\sigma_1, \sigma_2) + d(\sigma_2, \sigma_3).$$

In the proof we justify that triangle inequality actually holds for every page k, $k = 1, \ldots, n$. For details we refer to Section 6.4. Then, similar to the analysis in Chapter 5 we can modify (6.1) for the case when we consider only pages that are in the top N either in one ranking list or in another. We denote $T(N) = \{k|\ \sigma_1(k) \leq N,\ \text{or}\ \sigma_2(k) \leq N\}$, and define the following distance:

$$d_N(\sigma_1, \sigma_2) = \sum_{k \in T(N)} \left| \frac{1}{2} - \frac{\sigma_1(k)}{\sigma_1(k) + \sigma_2(k)} \right| = \frac{1}{2} \sum_{k \in T(N)} \frac{|\sigma_1(k) - \sigma_2(k)|}{\sigma_1(k) + \sigma_2(k)}.$$

Compared to Kendall's τ and Spearman's ρ, the proposed measure is able to reveal the slightest rank disturbance among highly rank nodes while neglecting even moderate perturbations among lowly ranked nodes. Indeed, if we consider the nodes ranked $1, \ldots, n$, and swap the ranks 1 and 10, then we get $\tau = 1 - 2 * 18/n(n-1)$, $\rho = 1 - 6 * 162/n(n^2 - 1)$, and for our correlation measure at node 1 we obtain $\Theta_1 = 1/11 \approx 0.0910$ that is close to the x-axis, and is a visible deviation from $1/2$. On the other hand, swapping the numbers 1001 and 1010 yields the same values of τ and ρ, but $\Theta_{1001} = 1001/2011) \approx 0.4978$. Thus, the Θ rank correlation measure, and therefore the distance $d(\cdot, \cdot)$, actually evaluate the rank disturbance visible for users.

Now we formalize the rank aggregation problem. For full ranking lists $\sigma_1, \ldots, \sigma_m$, we seek to find a new ranking σ_* in the way that the sum of distances from σ_* to all other rankings σ_i, $i = 1, \ldots, m$, is minimized. In other words

$$\sigma_* = \arg\min \sum_{i=1}^{m} d(\sigma_*, \sigma_i). \tag{6.2}$$

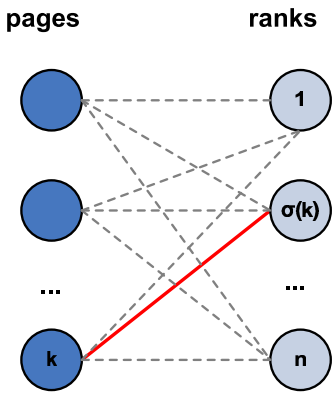


Figure 6.1: Bipartite graph for rank aggregation

In [39] Dwork et al. propose to use a minimum cost perfect matching in the bipartite graph in order to find the optimal ranking. Following the approach of

Dwork et al. we construct a complete bipartite graph, in which the one of part of nodes corresponds to pages, and the another to ranks. Every edge $(k, \sigma(k))$ has a weight that is defined by the sum $\sum_{i=1}^{m} d(\sigma(k), \sigma_i(k))$ (see Figure 6.1). Then, we need to find the perfect matching with the minimum cost. However, finding a perfect matching in the case of large data sets is computationally inefficient. One of the ways to approximate the optimal ranking is to use a specially constructed Markov Chain, as in [39]. Another approach is to obtain the optimal rank for every page independently.

In the latter case, for every k ($k = 1, \ldots, n$), and the corresponding ranks: $\sigma_1(k), \ldots, \sigma_m(k)$, we define a new rank as follows:

$$\sigma_*(k) = \arg\min \left(\sum_{i=1}^{m} \frac{|\sigma_i(k) - \sigma_*(k)|}{\sigma_i(k) + \sigma_*(k)} \right). \tag{6.3}$$

We note that this approach can be easily extended for the case of partial ranking lists, where for some pages ranks are not defined in some lists.

In the next sections we apply the new distance $d(\cdot, \cdot)$ to the problem of rank aggregation for real data sets. We compute rankings for a `Flickr` data set (Section 6.2.1) and a TREC data set (Section 6.2.2).

6.2 Numerical results

We start with $\sigma_1, \ldots, \sigma_m$ full ranking lists for pages $k = 1, \ldots, n$. Using the distance $d(\cdot, \cdot)$ we define new ranking list σ_{AM}. Moreover, we also define another list σ_B by applying *Borda's count* method. The Borda's count is a simple and very intuitive method, that is just the ranking of $\{\sum_{i=1}^{m} \sigma_i(1), \ldots, \sum_{i=1}^{m} \sigma_i(n)\}$ in increasing order.

For the case of the partial rankings we propose several extensions. For the `Flickr` data set in Section 6.2.1 we define *full lists* by the following procedures. We denote by M_i the largest score in the ith ranking list. Then, we assign rank $M_i + 1$ to all pages that are undefined in σ_i. We also use another method, when instead of assigning the same value we define scores for the pages in increasing order: $M_i + 1,\ M_i + 2, \ldots$ accordingly to some rules. Dwork et al. [39] argue that for any rule of assigning scores to unranked candidates, there are partial information cases in which undesirable outcomes for Borda's method may occur.

In Section 6.2.2 we use another way to deal with partial lists. In order to encourage pages that have been ranked in many lists, we first define new score of a page accordingly to (6.3), or as a sum of all ranks of this page, and then divide these new scores by the number of the lists that have ranked this pages. Finally, we rank these obtained scores in increasing order.

Since in our data we do not know the ideal ranking list, but we know which of the pages are relevant, then we use *precision* and *MAP (mean average precision)* for evaluation. Denote by Rlv the set of all relevant documents, and again by σ some

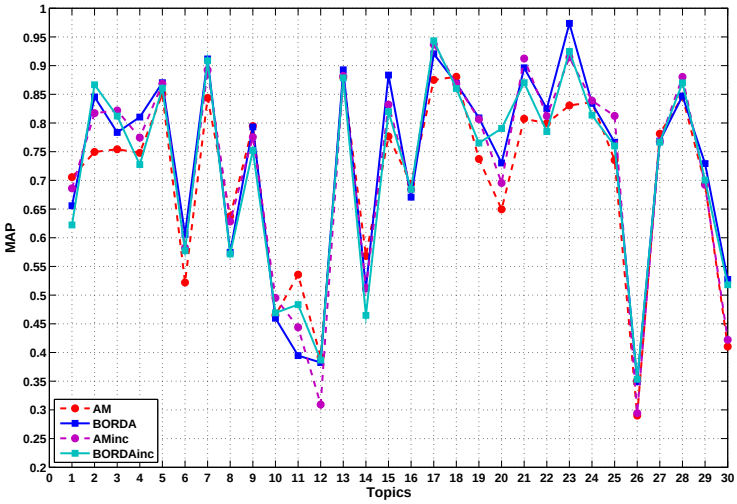


Figure 6.2: Flickr data: MAP's for various topics

ranking list. Then, precision at a ($P@a$) is defined as follows:

$$P@a = \frac{\#\{k|k \in Rlv, \text{ and } \sigma(k) \leq a\}}{a},$$

and mean average precision is defined as

$$\text{MAP} = \frac{\sum_{a \in Rlv} P@a}{\#Rlv}.$$

Here # states for the number of elements in the set. If $P@a$ and MAP are close to 1, then aggregated ranking is good.

Now, we present the experimental results.

6.2.1 Flickr data set

Here we use `Flickr` data set that was collected by Olivares et al. [91], and is a set of 30 topics, that were derived from `Flickr` search logs[1]. For each topic there are 10 lists of 25 ranked images. The only exception is for topic number 20, where we have only 9 lists. For more detail on this set we refer to [91].

In Figure 6.2 we present MAP values for various topic. For every topic i, $i = 1, \ldots, 30$ we aggregate four ranking that are defined by the following methods:

[1] On-line photo sharing service flickr.com; (Accessed in January 2009).

-	average MAP	min MAP	max MAP
AM	0.7030	0.2902	0.8806
BORDA	0.7297	0.3492	0.9735
AMinc	0.7220	0.2943	0.9367
BORDAinc	0.7202	0.3534	0.9432

Table 6.1: Flickr: MAP statistics

- AM: We define full lists by adding the same ranks $M_i = 26$, and use Hungarian algorithm [66] to exactly define aggregated list for the angular measure distance;
- AMinc: We define full lists by adding increasing ranks $M_i = 26, 27, \ldots$ for every list, and again use Hungarian algorithm [66] to exactly define aggregated list according to the angular measure distance;
- BORDA: We again define full lists by adding the same ranks $M_i = 26$, and use Borda's method;
- BORDAinc: We define full lists by adding increasing ranks $M_i = 26, 27, \ldots$ for every list, and use Borda's method.

In Table 6.1 we present average, minimal and maximal values of MAP. As we can see that for this data set Borda's method performs better than angular measure based technique for the prevalent number of the topics. Thus, we can suppose that our measure is too rough for small data sets. Since Borda is much easier to calculate, we can suggest to use Borda's method for this case.

6.2.2 TREC Data

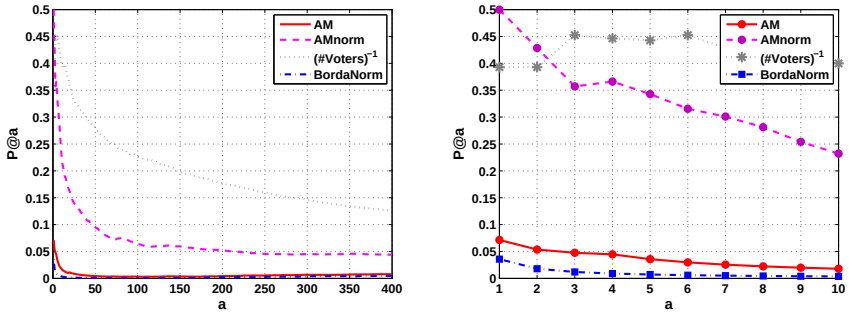
In order to perform experiment on the larger data set, we chose set of results from Web Track in the Text REtrieval Conference[1] 2000 (TREC-9). The set consists of 50 queries (topics), and for each of these topic 105 ranking results were assigned. Every ranking result is a list of top 1000 query relevant pages. Then, for every query we use the next rank aggregation methods:

- BordaNorm: For every page we consider only available ranks within a query, sum them and normalize by the number of the available ranks;
- AM: For every page we consider only available ranks within a query, and define the new rank as in (6.3), i.e. according to the angular measure distance;
- AMNorm: Similar to the second method, however we normalize AM-value by the number of the available ranks (to encourage ranks that come from the large number of voters);

[1] trec.nist.gov (Accessed in January 2009).

- $[\#\text{Voters}]^{-1}$: Simple count of the number of voters. (good page = many voters).

At the end of the analysis of each query, we rank all pages within a query accordingly to the obtained values in increasing order.



(a) a=400 (b) a=10

Figure 6.3: TREC: Average precision at a

In Figure 6.3 we plot the average precision at a. Here we clearly see that the Borda's method performs poorly on this data set. The number of voters, and normalized AM allow to achieve good precision. Moreover, we note that normalized AM wins over $[\#\text{Voters}]^{-1}$ at the first and second ranks. This confirms our assumption that the angular measure approach for the rank aggregation provides more relevant ranking for the top pages.

6.3 Discussion

In this work we made the first attempt for applying the angular measure to the rank aggregation problem. From numerical results we can conclude that methods that are defined by the angular measure can provide good precision for the top nodes in large data set, however they can fail in a small data set. It will be interesting to further specify the situations where the angular measure distance provide good results for rank aggregation.

Here we discuss some open problems for the future research. First of all the distance (6.1) is not normalized, thus we can not compare distances for rankings of different length. Second, we need to evaluate how good is (6.3) as an approximation of (6.2). We can also intend to formalize a Markov Chain approximation algorithm similar to the one proposed by Dwork et al. [39].

From the experimental point of view, we need to apply TREC methods for the

Flickr data, and possibly consider another relevance measures, like b-pref, and b-pref 10 (e.g., see [104]).

6.4 Appendix

Proof of Remark 6.1: We show that for every $1 \leq k \leq n$ the following inequality holds:

$$\frac{|\sigma_1(k) - \sigma_3(k)|}{\sigma_1(k) + \sigma_3(k)} \leq \frac{|\sigma_1(k) - \sigma_2(k)|}{\sigma_1(k) + \sigma_2(k)} + \frac{|\sigma_2(k) - \sigma_3(k)|}{\sigma_2(k) + \sigma_3(k)}.$$

Denote $x = \sigma_1(k)$, $y = \sigma_2(k)$ and $z = \sigma_3(k)$. Next, we consider the possible cases.

1: $x > y > z$

$$\begin{aligned}
&\frac{x-z}{x+z} \leq \frac{x-y}{x+y} + \frac{y-z}{y+z} = \frac{2y(x-z)}{(x+y)(y+z)} \\
&(x+y)(y+z) \leq 2y(x+z); \\
&(y-x)(y-z) \leq 0.
\end{aligned}$$

2: $x > z > y$

$$\begin{aligned}
&\frac{x-z}{x+z} \leq \frac{x-y}{x+y} + \frac{z-y}{y+z} = \frac{2(zx-y^2)}{(x+y)(y+z)} \\
&(x+y)(y+z)(x-z) \leq 2(zx-y^2)(x+z); \\
&(y-z)(3x(y+z) + x^2 + yz) \leq 0.
\end{aligned}$$

3: $y > x > z$

$$\begin{aligned}
&\frac{x-z}{x+z} \leq \frac{y-x}{x+y} + \frac{y-z}{y+z} = \frac{2(y^2-zx)}{(x+y)(y+z)} \\
&(x+y)(y+z)(x-z) \leq 2(y^2-zx)(x+z); \\
&(x-y)(3z(y+x) + z^2 + xy) \leq 0.
\end{aligned}$$

4: $y > z > x$

$$\begin{aligned}
&\frac{z-x}{x+z} \leq \frac{y-x}{x+y} + \frac{y-z}{y+z} = \frac{2(y^2-zx)}{(x+y)(y+z)} \\
&(x+y)(y+z)(z-x) \leq 2(y^2-zx)(x+z); \\
&(y-z)(3x(y+z) + x^2 + yz) \geq 0.
\end{aligned}$$

5: $z > x > y$

$$\frac{z-x}{x+z} \leq \frac{x-y}{x+y} + \frac{z-y}{y+z} = \frac{2(zx-y^2)}{(x+y)(y+z)}$$
$$(x+y)(y+z)(z-x) \leq 2(zx-y^2)(x+z);$$
$$(x-y)(3z(y+x)+z^2+xy) \geq 0.$$

6: $z > y > x$

$$\frac{z-x}{x+z} \leq \frac{y-x}{x+y} + \frac{z-y}{y+z} = \frac{2y(z-x)}{(x+y)(y+z)}$$
$$(x+y)(y+z) \leq 2y(x+z);$$
$$(y-x)(y-z) \leq 0.$$

□

BIBLIOGRAPHY

[1] W. Aiello, F. Chung, and L. Lu. Random evolution in massive graphs. In *42nd IEEE Symposium on Foundations of Computer Science (Las Vegas, NV, 2001)*, pages 510–519. IEEE Computer Soc., Los Alamitos, CA, 2001.

[2] R. Albert and A. L. Barabási. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[3] R. Albert and A. L. Barabási. Topology of evolving networks: Local events and universality. *Phys. Rev. Lett.*, 85(24):5234–5237, 2000.

[4] R. Albert, H. Jeong, and A.L. Barabási. The diameter of the World Wide Web. Technical Report 9907038, arXiv/cond-mat, 1999.

[5] R. Andersen, F. Chung, and K. Lang. Local graph partitioning using PageRank vectors. In *Proceedings of FOCS2006*, pages 475–486, 2006.

[6] K. Avrachenkov and D. Lebedev. PageRank of scale-free growing networks. *Internet Math.*, 3(2):207–231, 2006.

[7] K. Avrachenkov and N. Litvak. The effect of new links on Google PageRank. *Stoch. Models*, 22(2):319–331, 2006.

[8] K. Avrachenkov, N. Litvak, and K. S. Pham. Distribution of PageRank mass among principle components of the Web. In *Algorithms and Models for the Web-Graph, 5th International Workshop, WAW 2007, San Diego, CA, USA*, volume 4863 of *Lecture Notes in Computer Science*, pages 16–28. Springer, 2007.

[9] R. Baeza-Yates, C. Castillo, and E. N. Efthimiadis. Characterization of national Web domains. *ACM Trans. Interet Technol.*, 7(2):9, 2007.

[10] L. Becchetti and C. Castillo. The distribution of PageRank follows a power-law only for particular values of the damping factor. In *WWW'06: Proceedings of the 15th International Conference on World Wide Web*, pages 941–942. ACM Press, New York, NY, USA, 2006.

[11] J. Beirlant. *Statistics of Extremes: Theory and Applications*. Wiley, 2004.

[12] P. Berkhin. A survey on PageRank computing. *Internet Math.*, 2:73–120, 2005.

[13] K. Bharat and A. Broder. A technique for measuring the relative size and overlap of public Web search engines. *Comput. Networks*, 30(1-7):379–388, 1998.

[14] M. Bianchini, M. Gori, and F. Scarselli. Inside PageRank. *ACM Trans. Inter. Tech.*, 5(1):92–128, 2005.

[15] P. Billingsley. *Probability and measure*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, second edition, 1986.

[16] N. H. Bingham and R. A. Doney. Asymptotic properties of supercritical branching processes. I. The Galton-Watson process. *Advances in Appl. Probability*, 6:711–731, 1974.

[17] N. H. Bingham, C. M. Goldie, and J. L. Teugels. *Regular Variation*. Cambridge University Press, 1989.

[18] P. Boldi, B. Codenotti, M. Santini, and S. Vigna. Structural properties of the African Web. In *WWW '02: Poster proceedings of the 11th international conference on World Wide Web*, New York, NY, USA, 2002. ACM.

[19] P. Boldi, M. Santini, and S. Vigna. PageRank as a function of the damping factor. In *WWW '05: Proceedings of the 14th International Conference on World Wide Web*. ACM Press, 2005.

[20] P. Boldi and S. Vigna. The WebGraph framework I: Compression techniques. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 595–602, New York, NY, USA, 2004. ACM.

[21] A. Bonato. A survey of models of the Web graph. In *Combinatorial and Algorithmic Aspects of Networking, First Workshop on Combinatorial and Algorithmic Aspects of Networking, CAAN 2004, Banff, Alberta, Canada, August 5-7, 2004, Revised Selected Papers*, volume 3405 of *LNCS*, pages 159–172. Springer, 2005.

[22] M. Bressan and E. Peserico. Choose the damping, choose the ranking? In *Algorithms and Models for the Web-Graph, 6th International Workshop, WAW 2009, Barcelona, Spain*, volume 5427 of *LNCS*, pages 76–89. Springer, 2009.

[23] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Comput. Networks*, 33:107–117, 1998.

[24] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Statac, A. Tomkins, and J. Wiener. Graph structure in the Web. *Comput. Networks*, 33:309–320, 2000.

[25] L. S. Buriol, C. Castillo, D. Donato, S. Leonardi, and S. Millozzi. Temporal analysis of the Wikigraph. In *Proceedings of International Conference on Web Intelligence (WI 2006)*, pages 45–51, 2006.

[26] A. Capocci, V. D. P. Servedio, F. Colaiori, L. S. Buriol, D. Donato, S. Leonardi, and G. Caldarelli. Preferential attachment in the growth of social networks: The internet encyclopedia Wikipedia. *Phys. Rev. E*, 74:036116, 2006.

[27] C. Castillo. *Effective Web Crawling*. PhD thesis, University of Chile, 2004. http://www.chato.cl/crawling/.

[28] D. Chakrabarti and C. Faloutsos. Graph mining: Laws, generators, and algorithms. *ACM Comput. Surv.*, 38(1):2, 2006.

[29] P. Chen, H. Xie, S. Maslov, and S. Redner. Finding scientific gems with Googles PageRank algorithm. *J.Informet.*, 1(1):8–15, 2007.

[30] P.G. Constantine and D.F. Gleich. Using polynomial chaos to compute the influence of multiple random surfers in the PageRank model. In *Algorithms and Models for the Web-Graph, 5th International Workshop, WAW 2007, San Diego, CA, USA*, volume 4863 of *LNCS*, pages 82–95, 2007.

[31] L. de Haan and J. de Ronde. Sea and wind: multivariate extremes at work. *Extremes*, 1(1):7–45, 1998.

[32] S. Dill, R. Kumar, K. S. McCurley, S. Rajagopalan, D. Sivakumar, and A. Tomkins. Self-similarity in the Web. *ACM Trans. Inter. Tech.*, 2(3):205–223, 2002.

[33] D. Donato, L. Laura, S. Leonardi, and S. Millozi. Large scale properties of the Webgraph. *Eur. Phys. J.*, 38:239–243, 2004.

[34] D. Donato, L. Laura, S. Leonardi, and S. Millozzi. The Web as a graph: How far we are. *ACM Trans. Inter. Tech.*, 7(1), February 2007.

[35] D. Donato, S. Leonardi, S. Millozzi, and P. Tsaparas. Mining the inner structure of the Web graph. *Journal of Physics A: Mathematical and Theoretical*, 41(22):224017, 2008.

[36] S. N. Dorogovtsev and J. F. F. Mendes. *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford University Press, 2003.

[37] S. N. Dorogovtsev, J. F. F Mendes, and A. N. Samukhin. Structure of Growing Networks with preferential linking. *Phys. Rev. Lett.*, 85(21):4633–4636, 2000.

[38] J. C. Doyle, D. L. Alderson, L. Li, S. Low, M. Roughan, S. Shalunov, R. Tanaka, and W. Willinger. The robust yet fragile nature of the Internet. *Proceedings of the National Academy of Sciences*, 102(41):14497–14502, 2005.

[39] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the Web. In *WWW '01: Proceedings of the 10th international conference on World Wide Web*, pages 613–622, New York, NY, USA, 2001. ACM.

[40] N. Eiron and K. S. McCurley. Locality, Hierarchy, and Bidirectionality in the Web. In *Workshop on Algorithms and Models for the Web Graph*, 2003.

[41] N. Eiron and K. S. McCurley. Links in hierarchical information networks. In *Algorithms and Models for the Web-Graph: Third International Workshop, WAW 2004, Rome, Italy*, volume 3243 of *LNCS*, pages 143–155. Springer, 2004.

[42] N. Eiron, K. S. McCurley, and J. A. Tomlin. Ranking the Web frontier. In *Proceedings of WWW2004*, pages 309–318, 2004.

[43] P. Embrechts, C. Kluppelberg, and T. Mikosch. *Modelling Extremal Events*. Springer, 1997.

[44] P. Embrechts, T. Mikosch, and C. Klüppelberg. *Modelling extremal events: for insurance and finance*. Springer-Verlag, London, UK, 1997.

[45] P. Erdös and A. Rényi. On random graphs. *Publ. Math. Debrecen*, 6(290), 1959.

[46] P. Erdös and A. Rényi. On the evolution of random graphs. *Bulletin of the Institute of International Statistics*, 38:343–347, 1961.

[47] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the Internet topology. In *Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication Table of Contents*, pages 251 – 262, 1999.

[48] A. Farahat, T. LoFaro, J. C. Miller, G. Rae, and L. A. Ward. Authority rankings from HITS, PageRank, and SALSA: Existence, uniqueness, and effect of initialization. *SISC*, 27(4):1181–1201, 2006.

[49] S. Fortunato, M. Boguñá, A. Flammini, and F. Menczer. Approximating PageRank from in-degree. In *Algorithms and Models for the Web-Graph, Fourth International Workshop, WAW 2006, Banff, Canada, Revised Papers*, volume 4936 of *LNCS*, pages 59–71, 2006.

[50] S. Fortunato and A. Flammini. Random walks on directed networks: the case of PageRank. Technical Report 0604203, arXiv/physics, 2006.

[51] A. Gulli and A. Signorini. The indexable Web is more than 11.5 billion pages. In *WWW '05: Special interest tracks and posters of the 14th international conference on World Wide Web*, pages 902–903, New York, NY, USA, 2005. ACM.

[52] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating Web spam with TrustRank. In *VLDB '04: Proceedings of the Thirtieth international conference on Very large data bases*, pages 576–587. VLDB Endowment, 2004.

[53] T. H. Haveliwala. Topic-sensitive PageRank: A context-sensitive ranking algorithm for Web search. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):784–796, 2003.

[54] T. H. Haveliwala, S. Kamvar, and G. Jeh. An analytical comparison of approaches to personalizing PageRank. Technical report, Stanford University, 2003.

[55] T. H. Haveliwala, S. Kamvar, D. Klein, C. Manning, and G. Golub. Computing PageRank using power extrapolation. Technical report, Stanford University, 2003.

[56] Y. Hirate, S. Kato, and H. Yamana. Web structure in 2005. In *Algorithms and Models for the Web-Graph, Fourth International Workshop, WAW 2006, Banff, Canada, Revised Papers*, volume 4936 of *LNCS*, pages 36–46. Springer, 2008.

[57] B. J. Jansen, A. Spink, and T. Saracevic. Real life, real users, and real needs: a study and analysis of user queries on the Web. *Information Processing and Management*, 36(2):207–227, 2000.

[58] G. Jeh and J. Widom. Scaling personalized Web search. In *WWW '03: Proceedings of the 12th international conference on World Wide Web*, pages 271–279, New York, NY, USA, 2003. ACM.

[59] A. H. Jessen and T. Mikosch. Regularly varying functions. *Publications de l'institut mathematique, Nouvelle série*, 79(93), 2006.

[60] S. D. Kamvar, T. H. Haveliwala, C. D. Manning, and G. H. Golub. Exploiting the block structure of the Web for computing. Technical report, Stanford University, 2003.

[61] S. D. Kamvar, T. H. Haveliwala, C. D. Manning, and G. H. Golub. Extrapolation methods for accelerating PageRank computations. In *WWW '03: Proceedings of the 12th international conference on World Wide Web*, pages 261–270, New York, NY, USA, 2003. ACM.

[62] M. G. Kendall. A new measure for rank correlation. *Biometrika*, 30(1-2):81–93, 1938.

[63] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *JACM*, 46(5):604–632, 1999.

[64] J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. The Web as a graph: Measurements, models and methods. In *Computing and Combinatorics, 5th Annual International Conference, COCOON '99, Tokyo, Japan, July 26-28, 1999*, volume 1627 of *LNCS*, pages 1–17. Springer, 1999.

[65] W. Kraaij, T. Westerveld, and D. Hiemstra. The importance of prior probabilities for entry page search. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 27–34, New York, NY, USA, 2002. ACM.

[66] H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistic Quarterly*, 2:83–97, 1955.

[67] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the Web graph. *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, page 57, 2000.

[68] A. N. Langville and C. D. Meyer. A survey of eigenvector methods for Web Information Retrieval. *SIAM Review*, 47(1):135–161, 2005.

[69] A. N. Langville and C. D. Meyer. *Google's PageRank and beyond: the science of search engine rankings*. Princeton University Press, Princeton, NJ, 2006.

[70] R. Lempel and S. Moran. The stochastic approach for link-structure analysis (SALSA) and the TKC effect. *Comput. Networks*, 33(1–6):387–401, 2000.

[71] J. Leskovec, J. M. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *KDD '05: Proceedings of the 11th ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 177–187, New York, NY, USA, 2005. ACM.

[72] L. Li, D. L. Alderson, J. C. Doyle, and W. Willinger. Towards a theory of scale-free graphs: definition, properties, and implications. *Internet Math.*, 2(4):431–523, 2005.

[73] F. Liljeros, C. R. Edling, L. A. N. Amaral, H. E. Stanley, and Y. Aaberg. The web of human sexual contacts. *Nature*, 411(6840):907–908, 2001.

[74] N. Litvak, W. R. W. Scheinhardt, and Y. Volkovich. In-degree and PageRank: Why do they follow similar power laws? *To appear in Internet Math.*

[75] N. Litvak, W. R. W. Scheinhardt, and Y. Volkovich. Probabilistic relation between in-degree and PageRank. In *Algorithms and Models for the Web-Graph, Fourth International Workshop, WAW 2006, Banff, Canada, Revised Papers*, volume 4936 of *LNCS*, pages 72–83. Springer, 2008.

[76] N. Litvak, W. R. W. Scheinhardt, Y. Volkovich, and B. Zwart. Characterization of tail dependence for in-degree and PageRank. In *Algorithms and Models for the Web-Graph, 6th International Workshop, WAW 2009, Barcelona, Spain*, volume 5427 of *LNCS*, pages 90–103. Springer, 2009.

[77] G. Liu, Y. Yu, J. Han, and G. Xue. China Web graph measurements and evolution. In *Proccedings of APWeb2005*, pages 668–679. Springer, 2005.

[78] Q. Liu. Fixed points of a generalized smoothing transformation and applications to the branching random walk. *Adv. in Appl. Probab*, 30:85–112, 1998.

[79] Q. Liu. Asymptotic properties and absolute continuity of laws stable by random weighted mean. *Stochastic Processes and their Applications*, 95(1):83–107, September 2001.

[80] M. Melucci. On rank correlation in information retrieval evaluation. *SIGIR Forum*, 41(1):18–33, 2007.

[81] A. De Meyer and J. L. Teugels. On the asymptotic behaviour of the distributions of the busy period and service time in M/G/1. *J. App. Probab.*, 17:802–813, 1980.

[82] A. Micarelli, F. Gasparetti, F. Sciarrone, and S. Gauch. Personalized search on the World Wide Web. In *The Adaptive Web*, volume 4321 of *LNCS*, pages 195–230, 2007.

[83] T. Mikosch. Modelling dependence and tails in financial time series. In *Symposium in Honour of Ole E. Barndorff-Nielsen (Aarhus, 2000)*, volume 16 of *Memoirs*, pages 61–73. Univ. Aarhus, Aarhus, 2000.

[84] S. Milgram. The small world problem. *Psychology Today*, 2(1):60–67, 1967.

[85] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *IMC '07: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42, New York, NY, USA, 2007. ACM.

[86] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Math.*, 1(2):226–251, 2004.

[87] M. Mitzenmacher. Editorial: The future of power law research. *Internet Math.*, 2(4):525–534, 2005.

[88] M. E. J. Newman. The structure and function of complex networks. *SIAM Rev.*, 45(2):167–256, 2003.

[89] M. E. J. Newman. Power laws, Pareto distributions and Zipf's law. *Contemp. Phys.*, 46:323–351, 2005.

[90] M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E*, 64(2):026118, Jul 2001.

[91] X. Olivares, M. Ciaramita, and R. van Zwol. Boosting image retrieval through aggregating search results based on visual annotations. In *MM '08: Proceeding of the 16th ACM international conference on Multimedia*, pages 189–198, New York, NY, USA, 2008. ACM.

[92] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998.

[93] G. Pandurangan, P. Raghavan, and E. Upfal. Using PageRank to characterize Web structure. In *COCOON '02: Proceedings of the 8th Annual International Conference on Computing and Combinatorics*, pages 330–339, London, UK, 2002. Springer-Verlag.

[94] K. Park and W. Willinger. *Self-similar network traffic and performance evaluation*. Wiley, New York, 2000.

[95] D. M. Pennock, G. W. Flake, S. Lawrence, E. J. Glover, and C. L. Giles. Winners don't take all: Characterizing the competition for links on the Web. *Proceedings of the National Academy of Sciences*, 99(8):5207, 2002.

[96] J. Pinto da Costa and C. Soares. A weighted rank measure of correlation. *Australian & New Zealand Journal of Statistics*, 47(4):515–529, 2005.

[97] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 275–281, New York, NY, USA, 1998. ACM.

[98] S. Redner. How popular is your paper? An empirical study of the citation distribution. *The European Physical Journal B-Condensed Matter and Complex Systems*, 4(2):131–134, 1998.

[99] H. Reittu and I. Norros. On the power-law random graph model of massive data networks. *Perform. Eval.*, 55(1-2):3–23, 2004.

[100] S. I. Resnick. *Heavy-tail Phenomena*. Springer, New York, 2007.

[101] M. Richardson and P. Domingos. The intelligent surfer: Probabilistic combination of link and content information in PageRank. *Adv. NIPS*, 14:1441–1448, 2002.

[102] S. M. Ross. *Stochastic processes*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc., New York, second edition, 1996.

[103] S. M. Ross. The inspection paradox. *Probability in the Engineering and Informational Sciences*, 17:47–51, 2003.

[104] T. Sakai. Alternatives to bpref. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 71–78, New York, NY, USA, 2007. ACM.

[105] E. Seneta. *Regularly varying functions*, volume 508 of *Lecture Notes in Mathematics*. 1976.

[106] G. S. Shieh. A weighted Kendall's tau statistic. *Statistics & probability letters*, 39(1):17–24, 1998.

[107] C. Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 15(1):72–101, 1904.

[108] H. van den Esker, R. van der Hofstad, G. Hooghiemstra, and D. Znamenski. Distances in random graphs with infinite mean degrees. *Extremes*, 8(3):111–141, 2005.

[109] R. van der Hofstad, G. Hooghiemstra, and P. Van Mieghem. Distances in random graphs with finite variance degrees. *Rand. Str. and Alg.*, 27(1):76, 2005.

[110] R. van der Hofstad, G. Hooghiemstra, and D. Znamenski. Distances in random graphs with finite mean and infinite variance degrees. *Electron. J. Probab.*, 12:no. 25, 703–766 (electronic), 2007.

[111] Y. Volkovich and N. Litvak. Asymptotic analysis for personalized Web search. Memorandum 1884, Department of Applied Mathematics, University of Twente, Enschede, October 2008.

[112] Y. Volkovich, N. Litvak, and D. Donato. Determining factors behind the PageRank log-log plot. In *Algorithms and Models for the Web-Graph, 5th International Workshop, WAW 2007, San Diego, CA, USA*, volume 4863 of *LNCS*, pages 108–123, 2007.

[113] Y. Volkovich, N. Litvak, and B. Zwart. A framework for evaluating statistical dependencies and rank correlations in power law graphs. Memorandum 1868, University of Twente, Enschede, 2008.

[114] Y. Volkovich, N. Litvak, and B. Zwart. Measuring extremal dependencies in Web graphs. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 1113–1114, New York, NY, USA, 2008. ACM.

[115] Y. Volkovich, N. Litvak, and B. Zwart. Extremal dependencies and rank correlations in power law networks. In *Proceedings of the 1st International Conference on Complex Sciences: Theory and Applications (Complex2009)*, 2009.

[116] E. Yilmaz, J. A. Aslam, and S. Robertson. A new rank correlation coefficient for information retrieval. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 587–594, New York, NY, USA, 2008. ACM.

[117] A. P. Zwart. *Queueing Systems with Heavy Tails*. PhD thesis, Eindhoven University of Technology, 2001.

SUMMARY

Today, the study of the World Wide Web is one of the most challenging subjects. In this work we consider the Web from a probabilistic point of view. We analyze the relations between various characteristics of the Web. In particular, we are interested in the Web properties that affect the Web page ranking, which is a measure of popularity and importance of a page in the Web. Mainly we restrict our attention on two widely-used algorithms for ranking: the number of references on a page (in-degree), and `Google`'s PageRank. For the majority of self-organizing networks, such as the Web and the Wikipedia, the in-degree and the PageRank are observed to follow power laws. In this thesis we present a new methodology for analyzing the probabilistic behavior of the PageRank distribution and the dependence between various power law parameters of the Web. Our approach is based on the techniques from the theory of regular variations and the extreme value theory.

We start Chapter 2 with models for distributions of the number of incoming (in-degree) and outgoing (out-degree) links of a page. Next, we define the PageRank as a solution of a stochastic equation $R \stackrel{d}{=} \sum_{i=1}^{N} A_i R_i + B$, where R_i's are distributed as R. This equation is inspired by the original definition of the PageRank. In particular, N models in-degree of a page, and B stays for the user preference. We use a probabilistic approach to show that the equation has a unique non-trivial solution with fixed finite mean. Our analysis based on a recurrent stochastic model for the power iteration algorithm commonly used in PageRank computations. Further, we obtain that the PageRank asymptotics after each iteration are determined by the asymptotics of the random variable with the heaviest tail among N and B. If the tails of N and B are equally heavy, then in fact we get the sum of two asymptotic expressions. We predict the tail behavior of the limiting distribution of the PageRank as a convergence of the results for iterations. To prove the predicted behavior we use another techniques in Chapter 3.

In Chapter 3 we define the tail behavior for the models of the in-degree and the PageRank distribution using Laplace-Stieltjes transforms and the Tauberian theorem. We derive the equation for the Laplace-Stieltjes transforms, that corresponds to the general stochastic equation, and obtain our main result that establishes the

tail behavior of the solution of the stochastic equation.

In Chapter 4 we perform a number of experiments on the Web and the Wikipedia data sets, and on preferential attachment graphs in order to justify the results obtained in Chapters 2 and 3. The numerical results show a good agreement with our stochastic model for the PageRank distribution. Moreover, in Section 4.1 we also address the problem of evaluating power laws in the real data sets. We define several state of the art techniques from the statistical analysis of heavy tails, and we provide empirical evidence on the asymptotic similarity between in-degree and PageRank. Inspired by the minor effect of the out-degree distribution on the asymptotics of the PageRank, in Section 4.4 we introduce a new ranking scheme, called PAR, which combines features of HITS and PageRank ranking schemes.

In Chapter 5 we examine the dependence structure in the power law graphs. First, we analytically define the tail dependencies between in-degree and PageRank of a one particular page by using the stochastic equation of the PageRank. We formally establish the relative importance of the two main factors for high ranking: large in-degree and a high rank of one of the ancestors. Second, we compute the angular measures for in-degrees, out-degrees and PageRank scores in three large data sets. The analysis of extremal dependence leads us to propose a new rank correlation measure which is particularly plausible for power law data.

Finally, in Chapter 6 we apply the new rank correlation measure from Chapter 5 to various problems of rank aggregation. From numerical results we conclude that methods that are defined by the angular measure can provide good precision for the top nodes in large data sets, however they can fail in a small data sets.

ABOUT THE AUTHOR

Yana Volkovich was born on January 17, 1982, in Kohtla-Järve (Estonia). She attended public school 173, and graduated from Physical Mathematical Lyceum 239 at Saint Petersburg (Russia). In 1999 Yana started to study at the department of Mathematics and Mechanics at Saint Petersburg State University. In June 2004 she obtained her Specialist Degree (with honors) in Mathematics and Computer Science. Her thesis is entitled "Adaptive optimization for M/G/1 queue", supervised by Oleg Granichin.

In February 2005 she became a PhD student under the supervision of Nelly Litvak and Richard Boucherie at the Stochastic Operation Research group at the department of Applied Mathematics and Computer Science at the University of Twente. In 2008 she became the Google Europe Anita Borg finalist. In November 2008 she was visiting Yahoo! Research Barcelona. Yana defends her PhD thesis at the University of Twente on April 24, 2009.