University Louis Pasteur Strasbourg I

University of West Bohemia in Pilsen

Doctoral Dissertation
under Joint Supervision

2007 **Dalibor FIALA**

University Louis Pasteur Strasbourg I
LGeCo, INSA Strasbourg

University of West Bohemia in Pilsen
Faculty of Applied Sciences

WEB MINING METHODS FOR THE DETECTION OF AUTHORITATIVE SOURCES

by

Dalibor FIALA

A dissertation under joint supervision submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in "Computer Science" and "Computer Science and Engineering"

Presented and defended publicly on November 30, 2007 before the board of examiners.

Pierre COLLET	**internal reviewer**	**University Louis Pasteur**
François JACQUENET	**external reviewer**	**University of Saint-Etienne**
Lubomír POPELÍNSKÝ	**external reviewer**	**Masaryk University Brno**
Bernard KEITH	**examiner**	**INSA Strasbourg**
Roland DE GUIO	**examiner**	**INSA Strasbourg**
Václav MATOUŠEK	**examiner**	**University of West Bohemia**
François ROUSSELOT	**supervisor**	**INSA Strasbourg**
Karel JEŽEK	**supervisor**	**University of West Bohemia**

Strasbourg / Pilsen 2007

Université Louis Pasteur Strasbourg I
LGeCo, INSA Strasbourg

Université de la Bohême de l'Ouest à Plzeň
Faculté des Sciences Appliquées

LES MÉTHODES DE LA FOUILLE DU WEB POUR LA DÉTECTION DES SOURCES FAISANT AUTORITÉ

par

Dalibor FIALA

Thèse en cotutelle présentée pour l'obtention du grade de Docteur de l'Université Louis Pasteur Strasbourg (spécialité Informatique) et de l'Université de la Bohême de l'Ouest (spécialité Informatique et ingénierie)

Soutenue publiquement le 30 novembre 2007 devant la commission d'examen.

Pierre COLLET	**rapporteur interne**	**Université Louis Pasteur**
François JACQUENET	**rapporteur externe**	**Université Saint-Etienne**
Lubomír POPELÍNSKÝ	**rapporteur externe**	**Université Masaryk Brno**
Bernard KEITH	**examinateur**	**INSA Strasbourg**
Roland DE GUIO	**examinateur**	**INSA Strasbourg**
Václav MATOUŠEK	**examinateur**	**Université de Plzeň**
François ROUSSELOT	**directeur de thèse**	**INSA Strasbourg**
Karel JEŽEK	**directeur de thèse**	**Université de Plzeň**

Strasbourg / Plzeň 2007

Université Louis Pasteur Strasbourg I
LGeCo, INSA Strasbourg

Západočeská univerzita v Plzni
Fakulta aplikovaných věd

METODY WEB MININGU PRO VYHLEDÁVÁNÍ AUTORITATIVNÍCH ZDROJŮ

Ing. Dalibor FIALA

Disertační práce pod dvojím vedením
k získání akademického titulu doktor
v oboru "Informatika" a "Informatika a výpočetní technika"

Předneseno a obhájeno veřejně před zkušební komisí dne 30. listopadu 2007.

Prof. Pierre COLLET		**Univ Louis Pasteur**
Prof. François JACQUENET		**Univ Saint-Etienne**
Doc. RNDr. Lubomír POPELÍNSKÝ, Ph.D.		**Masarykova univerzita**
Prof. Bernard KEITH		**INSA Strasbourg**
Prof. Roland DE GUIO		**INSA Strasbourg**
Prof. Ing. Václav MATOUŠEK, CSc.		**KIV ZČU v Plzni**
Dr. François ROUSSELOT		**INSA Strasbourg**
Doc. Ing. Karel JEŽEK, CSc.	**školitel**	**KIV ZČU v Plzni**

Strasbourg / Plzeň 2007

To Anna for her love and patience
and to all of my family for their support and encouragement

The work on this doctoral thesis was supported in part by the Ministry of Education of the Czech Republic under Grant 2C06009 COT-SEWing.

Declaration

I submit this dissertation for review and defence in partial fulfillment of the requirements for the degree of Doctor of Philosophy at the University of West Bohemia in Pilsen, Czech Republic and at the University Louis Pasteur Strasbourg, France.

I hereby declare that this doctoral thesis is completely my own work and that I used only the cited sources.

Pilsen, August 30, 2007 Dalibor Fiala

Abstract

The development of information society in recent decades has enabled collecting, filtering and storing huge amounts of data. These data must be further processed to gain valuable information and knowledge. The scientific field dealing with extracting information and knowledge from data has evolved rapidly to cope with the extent and growth of information sources the number of which has geometrically increased with the appearance of the World Wide Web. All traditional approaches in information retrieval, knowledge acquisition, and data mining must be adapted for the dynamic, heterogeneous, and unstructured data on the Web. Web mining has come into being as a fully-fledged research discipline.

The Web brings much specificity with it. The most salient feature is its link structure. The Web is a dynamic, linked network of nodes. Web pages contain links to other pages with similar contents, of a specific or more general interest, or otherwise related. Soon it was discovered that the link structure of Web is a vast resource of information and that it presents a wonderful field for applications from the social network domain as well as from the mathematical graph theory. Brin and Page have submitted the interlinkage of Web pages to an extensive research which resulted in the appearance of the now famous article "The anatomy of a large-scale hypertextual Web search engine" in 1998 introducing Google – a search engine for day-to-day usage by the whole Web community. The success of Google has been very much due to the underlying algorithm called PageRank, which makes use of the interconnection of billions of Web pages recursively so as to identify popular, prestigious, significant, or authoritative sources on the Web. The description of PageRank has been published and this results in a steady flow of new research papers on link-based methods that finally introduce a completely new group of algorithms – ranking algorithms. Each technique has its particular properties and is aimed at coping with specific problems. Although originally conceived for the Web, ranking algorithms are usable in every environment that can be modelled as a graph.

The innovative portion of this doctoral thesis deals with the definitions, explanations and testing of modifications of the standard PageRank formula adapted for bibliographic networks. The new versions of PageRank take into account not only the citation but also the co-authorship graph. We verify the viability of the new algorithms by applying them to the data from the DBLP digital library and by comparing the resulting ranks of the winners of the ACM SIGMOD E. F. Codd Innovations Award. The rankings based on both the citation and co-authorship information turn out to be better than the standard PageRank ranking. In another part of the disseration, we present a methodology and two case studies for finding authoritative researchers by analyzing academic Web sites. In the first case study, we concentrate on a set of Czech computer science departments' Web sites. We analyze the relations between them via hyperlinks and find the most important ones using several common ranking algorithms. We then examine the contents of the research papers present on these sites and determine the most authoritative Czech authors. In the second case study, we do exactly the same with French academic computer science Web sites to find the most significant French researchers in the field. We also discuss the weak points of our approach and propose some future improvements. To the best of our knowledge, it is the only attempt ever made at discovering authoritative researchers from the above countries by directly mining from Web data.

Keywords: Web mining, Web crawling, ranking algorithms, bibliographic networks, citations, co-authorships, authorities, bibliographic PageRank.

Résumé

Le récent développement de la société de l'information a permis de collecter, de filtrer et de stocker de grandes masses de données. Le problème est maintenant d'exploiter ces données pour obtenir des informations et des connaissances pertinentes. Les techniques d'extraction des informations et des connaissances à partir de données ont rapidement évolué à cause de la forte croissance des sources d'informations dont le nombre a augmenté de façon exponentielle après l'arrivée du Web. Il faut maintenant adapter toutes les approches traditionnelles de la recherche d'information, de l'acquisition des connaissances et de la fouille de données aux données dynamiques, hétérogènes et non structurées qui se trouvent sur le Web. La fouille du Web est devenue une discipline de recherche reconnue.

Le Web a beaucoup de spécificités. La propriété la plus caractéristique est sa structure de liens. Le Web est un réseau de noeuds liés et c'est aussi un réseau dynamique. Les pages Web contiennent des liens vers d'autres pages avec un contenu similaire, intéressant ou lié de façon quelconque. On a découvert assez tôt que la structure de liens du Web est une ressource énorme d'information et qu'elle représente un domaine typique d'application des réseaux sociaux aussi bien que de la théorie des graphes en mathématiques. Brin et Page ont largement étudié l'inter-connection des pages Web ce qui a résulté en la publication de leur célèbre article « The anatomy of a large-scale hypertextual Web search engine » en 1998. Dans leur article ils ont présenté Google – un nouveau moteur de recherche sur le Web qui est utilisé par des millions d'utilisateurs chaque jour jusqu'à présent. Le descriptif de PageRank a été publié et cela a eu pour effet la publication fréquente de nouveaux articles scientifiques sur les méthodes basées sur les liens. Les chercheurs ont finalement créé un nouvel ensemble d'algorithmes – des algorithmes de classement (ranking algorithms). Chaque méthode a ses qualités spécifiques et est réservée à la résolution de problèmes différents. Même si les algorithmes de classement ont été conçus pour le Web à l'origine, ils sont applicables à tout système modélisable sous forme de graphe.

La partie innovante de cette thèse porte sur les définitions, les explications et teste des modifications de la formule standard de PageRank adaptée aux réseaux bibliographiques. Les nouvelles versions de PageRank tiennent compte non seulement du graphe de citations mais aussi du graphe de collaboration. On vérifie l'applicabilité des nouveaux algorithmes en traitant des données issues de la bibliothèque numérique DBLP et en comparant les rangs des lauréats du prix « ACM SIGMOD E. F. Codd Innovations Award ». Les classements reposant sur les informations concernant à la fois les citations et les collaborations s'avèrent meilleurs que les classements générés par PageRank standard. Dans un autre chapitre de la thèse, on présente une méthodologie et deux études de cas concernant la recherche des chercheurs faisant autorité en analysant des sites Web académiques. Dans la première étude de cas, on se concentre sur une collection de sites Web des laboratoires d'informatique tchèques. On analyse les relations entre eux à l'aide de liens et on trouve les laboratoires les plus significatifs en utilisant plusieurs algorithmes d'évaluation courants. Ensuite, on examine le contenu des articles de recherche trouvés sur ces sites et on détermine les auteurs tchèques les plus importants. Dans la deuxième étude, on fait exactement la même chose avec des sites Web des laboratoires d'informatique français pour trouver les scientifiques français les plus éminents dans ce domaine. On discute également les difficultés de notre approche et on propose quelques améliorations envisageables dans le futur.

Mots-clés : fouille du Web, robots Web, algorithmes d'évaluation, réseaux bibliographiques, citations, co-auteurs, authorité, PageRank bibliographique.

Abstrakt

Rozvoj informační společnosti v posledních desetiletích umožňuje shromažďovat, filtrovat a ukládat obrovská množství dat. Abychom z nich získali cenné informace a znalosti, musejí se tato data dále zpracovávat. Vědecký obor zabývající se získáváním informací a znalostí z dat se překotně vyvíjí, aby zachytil vysoké tempo nárůstu zdrojů informací, jejichž počet se po vzniku celosvětové pavučiny (webu) zvyšuje geometrickou řadou. Všechny tradiční přístupy z oblasti získávání informací, dobývání znalostí a dolování z dat se musejí přizpůsobit dynamickým, heterogenním a nestrukturovaným datům z webu. Dolování z webu (web mining) se stal plnohodnotnou vědeckou disciplínou.

Web má mnoho speciálních vlastností. Tou nejvýznačnější je jeho struktura odkazů mezi stránkami. Web je dynamickou, propojenou sítí. Webové stránky obsahují odkazy na jiné stránky s podobným obsahem nebo na zajímavé či jinak spřízněné dokumenty. Velmi brzy se zjistilo, že webová struktura odkazů je ohromným zdrojem informací a že představuje rozsáhlé pole aplikací z oboru sociálních sítí a matematické teorie grafů. Brin a Page podrobili propojení webu intenzivnímu výzkumu a v roce 1998 vydali dnes už slavný článek „The anatomy of a large-scale hypertextual Web search engine“, v němž světu představili Google – webový vyhledávač pro každého. Úspěch Googlu spočívá především v algoritmu pro hodnocení webových stránek nazvaném PageRank. Ten využívá struktury webu k tomu, aby v něm rekurzivní metodou nalezl populární, důležité, významné a autoritativní zdroje. Technický popis PageRanku byl publikován a měl za následek doslova příval dalších odborných článků o metodách založených na propojení uzlů sítě, které nakonec daly vzniknout úplně nové skupině algoritmů – hodnoticím (ranking) algoritmům. Každá metoda má své zvláštnosti a umí se vypořádat s určitými problémy. Ačkoliv byly hodnoticí algoritmy původně vymyšleny pro web, jsou použitelné v každém prostředí, které lze modelovat grafem.

Inovativní část této doktorské práce se zabývá definicemi, vysvětlením a testováním modifikací standardního vzorce PageRanku uzpůsobeného pro bibliografické sítě. Takto vzniklé nové verze PageRanku berou v úvahu nejen citační graf, ale i graf spoluautorství. Použitelnost nových algoritmů ověřujeme jejich aplikací na data z digitální knihovny DBLP. Získané žebříčky významných autorů porovnáváme s držiteli ocenění ACM SIGMOD E. F. Codd Innovations Award. Ukazujeme, že hodnocení založené jak na citacích, tak na spolupracích dává lepší výsledky než standardní PageRank. V jiné části disertace představujeme metodologii a dvě případové studie vyhledávání autoritativních vědců analyzováním univerzitních webů. První studie se zaměřuje na množinu webových stránek českých kateder informatiky. Zkoumáme zde propojení mezi jednotlivými katedrami a několika běžnými hodnoticími metodami označujeme ty nejdůležitější. Poté analyzujeme obsah odborných publikací nalezených na daných stránkách a určujeme nejvýznačnější české autory. V druhé případové studii provádíme ten samý postup s francouzskými univerzitními weby pro nalezení nejvýznamnějších francouzských výzkumníků v oboru informatiky. Rovněž se zmiňujeme o slabých stránkách našeho přístupu a navrhujeme několik budoucích vylepšení. Na základě našich znalostí konstatujeme, že výše uvedené studie jsou jediným dosud publikovaným pokusem o vyhledávání autoritativních vědců z obou zemí přímým dolováním z webových dat.

Klíčová slova: dolování z webu, weboví pavouci, hodnoticí algoritmy, bibliografické sítě, citace, spoluautorství, autority, bibliografický PageRank.

Acknowledgements

I would first like to thank the Almighty to make me healthy and fit enough to come through all those university years and to be able to work on such interesting problems.

I am very grateful to my supervisor Karel Ježek that he enabled me to embark on doctoral studies in very special circumstances a couple of years ago and that he readily guided me towards this thesis. I am also indebted to my supervisor François Rousselot for taking part in the joint supervision and for making possible the French part of my doctoral studies.

I thank my colleagues and friends from the Text Mining Research Group for the beautiful ambiance we had together. Those years will belong to the best of my life! Thank you very much, all my friends from outside of the university! Thank you for the marvellous distractions, needed from time to time to escape from the world of science.

I must also thank all of my family for their support and encouragement. Without them it would have been much harder for me. Forgive me please my long absence! Finally, I would like to express my thanks and gratitude to my fiancé and future wife Anna for her love and patience without which I could not live, let alone writing a thesis!

Dalibor

Strasbourg, June 6, 2007

Table of Contents

List of Figures

List of Tables

Introduction

At the dawn of the World Wide Web in the early 1990s, nobody actually knew what kind of medium was emerging. The concept of hypertext coined by Tim Berners-Lee was not generally known to the public, and the underlying technological infrastructure, Internet, was not much spread beyond some university institutions. This was to change rapidly within the following decade in a breath-taking pace. Millions of Web servers began to host millions of documents of all kinds, and the Web's dimension doubled every six months. It became clear very soon that the new medium had a huge potential to exploit. Sergey Brin and Larry Page were among the first to recognize the amazing possibilities of what was now called the World Wide Web and to make practical attempts to turn it into something more manageable. From 1996 to 1998, they designed and implemented *Google*, a search engine for the Web. They were aware that the Web had one particularity that standard information retrieval (IR) systems of that time did not handle well. This feature was the presence of hyperlinks between Web documents. Brin and Page realized that links did not have just a navigational function, but that they were a kind of endorsement of a document by other documents. This analogy to bibliographic citations between publications made them invent and incorporate an algorithm called *PageRank* in their Web search engine.

Motivations

PageRank is a technique to order (rank) Web documents by importance, significance, authoritativeness, quality, prestige, influence, value, or whatever we may call it, but not by relevance. It is query independent, i.e. it is pre-computed and the ranks of Web pages are known long before they are used to sort the results for a given user query. PageRank is recursive – it assigns high ranks to pages that are linked to by documents that themselves have a high rank. With regard to the immense scope of the Web (billions of documents), PageRank must be calculated iteratively (i.e. approximately), and it is sometimes called the world's largest matrix computation. The exact synthesis of PageRank and other IR techniques in order to detect relevant and high quality Web pages is proprietary information and know-how of commercial Web search engines which, having seen the tremendous business success of Google, have all added some link-based evaluation of Web documents to their ranking schemes.

Google's PageRank was one of the first large-scale applications of Web structure mining, a subdomain of Web mining besides Web content and Web usage mining. I guess that it was in particular the commercial success of Google that triggered interest in Web mining and Web structure mining. Many researchers, including me, have since tried to explore and explain PageRank's properties, speed up its computation, propose its modifications, or adapt it for graphs different from the Web graph. The class of *ranking algorithms* has come into being, and Web mining has become a research discipline of its own. The seminal book on this topic by Soumen Chakrabarti from 2002 is being prepared for the second edition as disclosed in a personal communication with the author. The Web is the largest data repository mankind has ever had, and the information excess can be reduced only with filtering techniques that detect not only topic-relevant but also high quality information. Therefore, I reckon that the need for the detection of authoritative sources in the Web will still be growing.

Goals and results achieved

The main objectives of this doctoral dissertation can be divided into two groups. First, I wanted to modify the PageRank formula and embed in it some parameters from a co-

authorship graph so as to work better with citations between authors. In other words, I wanted the modified PageRank to produce a "fairer" ranking of authors by importance that is based on the citation as well as collaboration information. My assumption was that a citation between two frequently collaborating researchers was less valuable than that between two authors that had never published together. The standard PageRank does not enable such a distinction as it is based on the citation graph only. Related work on this topic includes publications by Liu at al. [Liu2005] and Sidiropoulos et al. [Sidiropoulos2005]. Second, I wanted to apply some ranking algorithms, not necessarily novel ones, to some real and raw data in order to find authoritative institutions and researchers in a domain close to mine. In particular, I was interested in influential computer science departments in the Czech Republic and in France and wondered what authors would appear as significant after analysis of research papers found on the Web sites of those departments. In my view, no such analysis had ever been published. One can encounter some similar work in the articles by Thelwall and his colleagues [Thelwall2001, Thelwall2002, Li2003], but they are interested in universities rather than departments, and they do not analyze documents on the Web sites.

Coherently to the goals above, I consider my main contributions to be:

- **Bibliographic PageRank**. I proposed and implemented several modifications of the standard PageRank formula so as to better suit the need for a fair ranking of authors. Unlike the standard PageRank, the new formula takes into account citations as well as collaborations of authors. I tested the new method on the data from the DBLP digital library and compared the new author rankings with a list of ACM award winners. I can conclude that the new methods generally outperform PageRank.

- **Mining the Czech and French academic Web**. I also mined Web sites of Czech and French computer science departments and determined authoritative institutions and researchers. Due to the noise in the data, I prefer to underline it as a unique case study, the first of its scope and domain, in which I combine Web mining and information extraction techniques. The methodology I use is quite general and is thus applicable to completely distinct fields as well.

Thesis outline and ommissions

In chapters I and II, I discuss state-of-the-art approaches to Web structure mining. – Web crawling, a prerequisite of mining, in Chapter I and ranking algorithms for Web pages (or sites), the main tool for the detection of authoritative sources on the Web, in Chapter II. Chapter III deals with social networks, a domain that strongly influences Web structure mining. In Chapter IV, I present a few systems available on the Web that may help, among others, identify influential researchers and thus may be used in comparisons of author rankings. I introduce DBLP here, further employed in Chapter V on the bibliographic PageRank, the main innovative part of the thesis. I describe experiments with mining the academic Web in Chapter VI, and I summarize the dissertation afterwards. Some results from chapters V and VI are shown in the appendix.

At the end of the thesis, I enclose a list of over a hundred article references and several dozens of Web references. Actually, there could be more of them – in an order of magnitude! Such is the scale of Web mining. Thus, I made a number of omissions in the state-of-the-art sections to keep a reasonable scope of the thesis. For instance, I do not cover vertical (focused) crawling, information extraction, PageRank energies, eigenvector theory, the work by Mike Thelwall and others.

I Web Graph and Its Crawling

The World Wide Web is a gigantic dynamic network currently containing tens of billions of nodes [e.g. Gulli2005]. Web pages appear and disappear, their contents are modified. Links between pages are added or removed and the Web of today is not what it was yesterday. How are the new Web documents created? Which nodes do they link to most frequently? What does the Web graph look like? Are there any regularities to observe? The behaviour of social networks, one of which is the Web graph, is far from being fully understood [Newman2003]. Nevertheless, much research has been devoted to the analysis of the Web as a graph with view of answering some of the questions asked. To be able to study the Web, we need to collect the Web data first. This process known as *crawling* is not trivial, and we present the state-of-the-art knowledge and current trends in the second part of this overview chapter.

I.1 Web as a Graph

The very early simple random graph model with the number of nodes n and the same uniform probability p of the appearance of each of $n(n-1)$ possible edges does not seem to be in accordance with the real Web graph [Chakrabarti2002, p. 243]. This model had to be improved and verified in practical experiments. Since the Web graph model is not the key element in this thesis, we just briefly mention two of its interesting properties, namely the power law degree distribution and the bow tie structure, and we refer to the most recent survey articles on this topic [Chakrabarti2006, Donato2007]. Moreover, the latter introduces a free software library for generating and measuring huge graphs. Among others, a deep understanding of Web graph models may have a great impact on the design and implementation of ranking algorithms for Web sites, the best-known of which we cover in detail in sections II.2 and II.3.

I.1.1 Power Law Degree Distribution

One of the first phenomena of the Web observed was the *power law degree distribution*. It answers the question with what frequency Web pages with a certain in- or out-degree occur in the Web graph. The power law resembles the Zipf's law, in which an object ranked on the k-

th position by the number of occurrences occurs approximately N/k-times where N is the total number of all objects' occurrences. In the power law, k is not a rank but a degree size, and there is an exponent over it. Thus, the probabilities $Pr(d_{in}(p) = k)$ and $Pr(d_{out}(p) = k)$ of Web page *p* having an in-degree k or out-degree k are the following:

$$\Pr(d_{in}(p) = k) \propto 1/k^{\alpha 1} \quad \text{(I.1)}$$

$$\Pr(d_{out}(p) = k) \propto 1/k^{\alpha 2} \quad \text{(I.2)}$$

where $\alpha 1$ and $\alpha 2$ are coefficients varying from 2.1 to 2.7 [Kumar1999, Barabási1999, Kleinberg1999a, Bröder2000]. See Figure I.1 for the plots of degree distributions. Note that the power law holds also for the Web graph when intra-site links have been removed (denoted as "remote only" in the figure).

Although the power law degree distribution has been determined empirically, it can be proven theoretically as well. Barabási and Albert [Barabási1999] proposed a Web-suited random graph model, in which new nodes are continuously added and preferentially attached to nodes that already have a large in-degree. This is sometimes called the "winner takes all" or "rich get richer" scenario (compare with PageRank in Section II.2). This model was later amended by Pennock et al. [Pennock2002] so as to give less popular nodes a greater chance to get in-links from newly added nodes. This refinement was found to better fit the power law function.

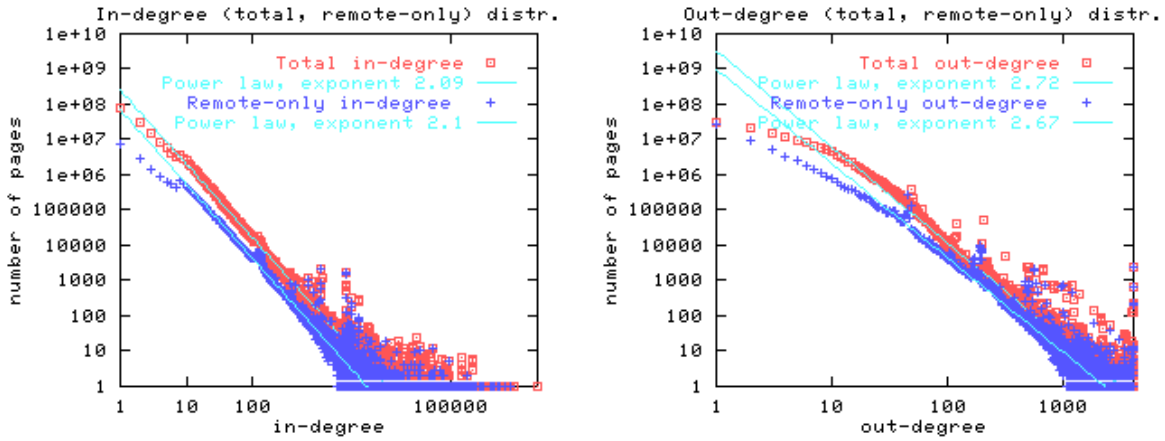


Figure I.1: Power law of in and out-degree distributions of Web pages [Bröder2000].

I.1.2 Bow Tie Structure

When analyzing the Web structure, Bröder and his colleagues [Bröder2000] discovered something unexpected. They examined two Web crawls of about 200 million pages and a billion and a half links in a half-year interval and repeatedly found out that the Web's connectivity was limited by a global structure shown in Figure I.2. They called it the "bow tie structure" of the Web. In their experiments, the weakly connected Web (i.e. connected when treated as an undirected graph) made up 90% of the whole Web crawled and consisted of four parts of about the same size. There was a *strongly connected core* (SCC), in which each node has a directed path to any other node, and three weakly connected components – IN with nodes having paths into SCC but not reachable from there, OUT with pages reachable from SCC but with no paths into SCC, and *tendrils*, which were weak components attached to IN and OUT. Some pages in OUT were reachable from IN via *tubes*, but not vice versa. Besides

the pages mentioned so far, there were also some 16 million Web pages forming separate disconnected components.

Bröder also made assumptions about the components' functionality. He suggested that the pages in IN were newly created pages not yet having been linked to from other pages. The pages in OUT might be corporate Web pages that never point to the "centre of events" in SCC. The authors of the experiment showed that the *diameter* of SCC was 28 at least. The diameter is the maximum of the shortest paths between any two nodes in that component. (See also Section III.3.) As for the graph as a whole, they determined that if there was a directed path between two nodes, its length was 16 on average. If there was an undirected path, its average length was six. Finally, the authors found out that the distributions of weakly and strongly connected components also followed the power law (see Section I.1.1). Although the experiments above are relatively old now, their conclusions were confirmed later on. Dill et al. [Dill2002] verified the existence of the bow tie structure even in subgraphs of the Web, e.g. given a top-level domain or a keyword occurrence, etc. Nevertheless, the Web is a dynamic organism, and it is unsure whether it will still adhere to the bow tie model in the future.

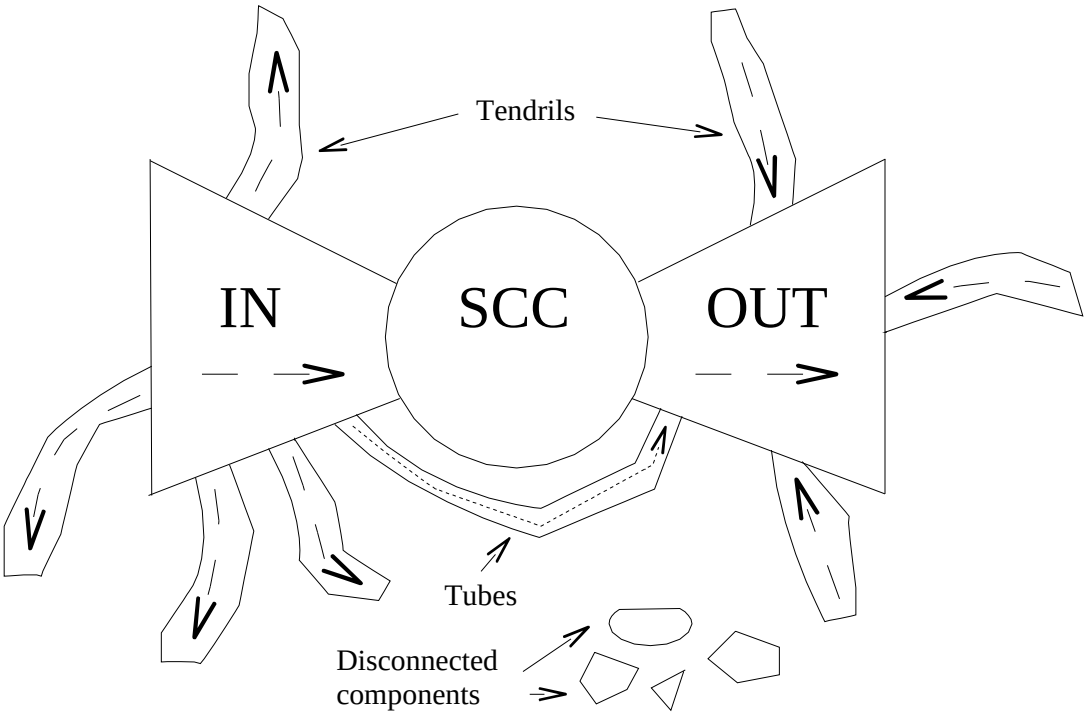


Figure I.2: The bow tie structure of the Web [Bröder2000].

I.2 Web Crawling

Web crawling or spidering is the process of collecting Web pages and other Web documents by recursively following the out-links from a set of starting (seed) pages. Its primary goal is to create a corpus of Web documents that could subsequently be indexed by a Web search engine in order to respond to users' requests. Every search engine relies on its indexed corpus and so the way of its creation is essential. The role currently played by Web search engines in the world is incontestable, and, therefore, it is somewhat surprising that crawling is still under-represented in the Web mining research. The experiments described in

Chapter VI could not have been conducted without Web crawling techniques, so we find useful to incorporate a section on this topic in this dissertation. Unless we indicate another source of information, the facts presented here come from our own experience, the most comprehensive overview of Web crawling strategies ever by Baeza-Yates et al. [Baeza-Yates2005] or from the Web mining book by Chakrabarti [Chakrabarti2002, ch. 2].

I.2.1 Architecture of a Web Crawler

Figure I.3 depicts the typical architecture of a large-scale Web crawler. By a large-scale crawler we mean a system capable of gathering billions of documents from the current World Wide Web. It is clear that with such a huge amount of data more sophisticated techniques must be applied than simply parsing HTML files and downloading documents from the URLs extracted from there. As we may see at the picture, much attention is paid to the problems of avoiding Web pages (URLs) already visited before, parallelizing crawling (fetching threads) and balancing the load of Web servers from which documents are obtained (server queues), and speeding up the access to Web servers (via DNS caching). We will give some notes on these issues further below.

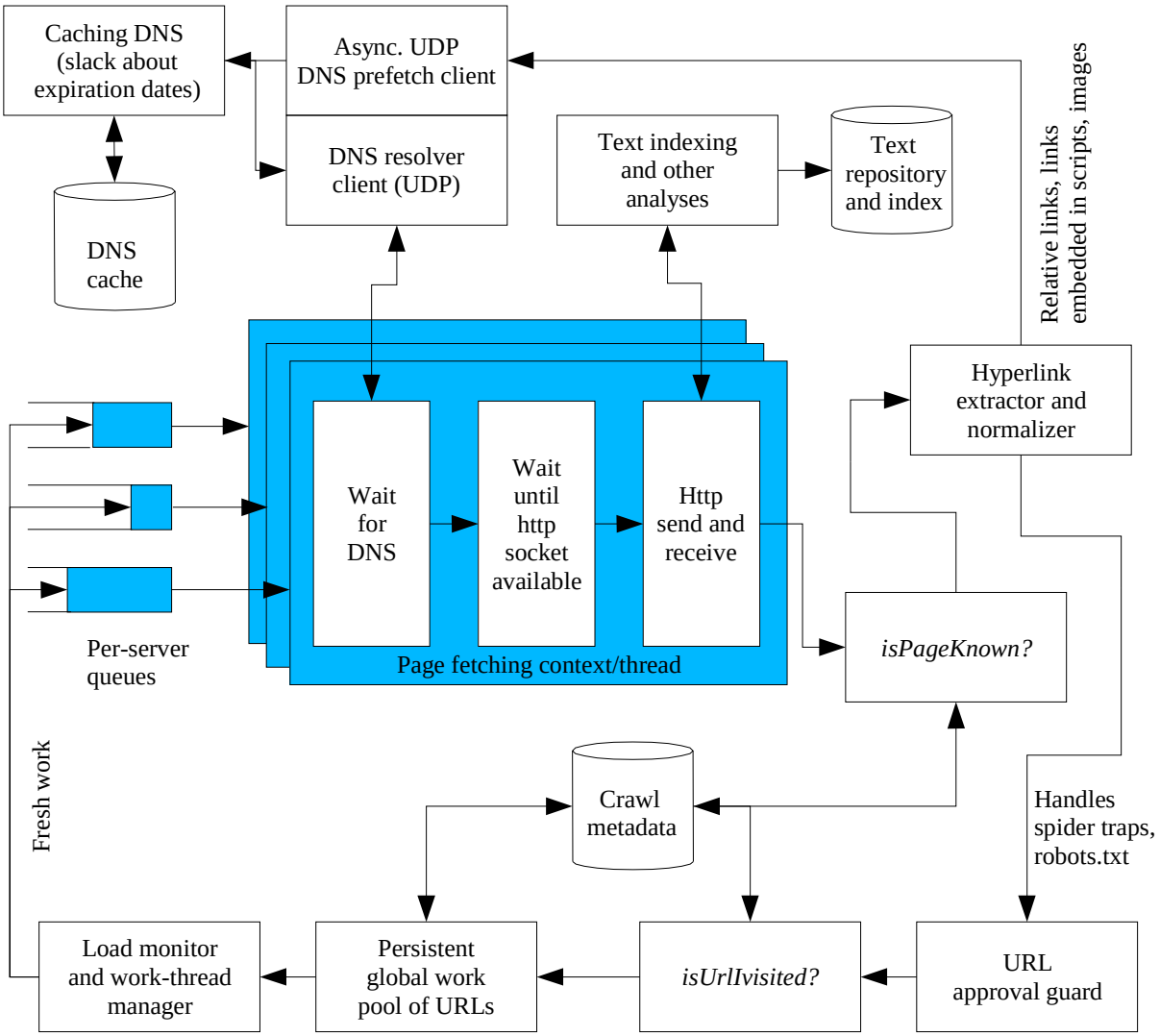


Figure I.3: Architecture of a typical Web crawler [Chakrabarti2002, ch. 2].

The role of Web crawling

Since Web crawling is at the heart of each Web search engine, rather general architectural descriptions of crawlers without important details have appeared so far. Commercial search engines treat their Web crawling techniques as business secrets and prefer not to give their rivals a chance to take advantage of their know-how. Another reason is to keep essential information on crawling away from *search engine spammers* who would abuse the information. (Search engine spammers deliberately create, remove or modify Web pages on the content as well as link level so as to promote in the result list to a given query those pages that would otherwise have been unnoticed. Search engines must defend themselves against such attempts and develop anti-spam techniques [Wu2005].) Some of the crawler architectures published are that of Alexa [Burner1997], which is still the Web robot of the Internet Archive [38], an early version of Googlebot [Brin1998], being the crawler of Google [39], Mercator [Heydon1999], which was the spider of AltaVista [40], Ubicrawler [Boldi2004b], and Dominos [Hafri2004]. Parallel crawling architectures are proposed in [Cho2002]. There have also been Web spiders released as free software under the GNU public licence [35], [36], [37].

In general, a Web crawler takes a URL from the queue of pending URLs, it downloads a new page from the URL, it stores the document to a repository and it parses its text to find hyperlinks to URLs, which it then enqueues in the queue of pending URLs in case they have not yet been downloaded ("fetched"). Ideally, crawling is stopped when the queue of pending URLs is empty. In practice, however, this will never happen as the universe of a large-scale Web crawler is almost infinite. The Web is steadily changing and will never be crawled as a whole. So a reasonable terminating condition must be set up for the crawler to stop. For example, a certain number of documents have been fetched, a specific number of terabytes of data has been downloaded, a particular time period has elapsed, or the crawler simply runs out of resources (main memory, storage capacities, etc.).

Internals

More specifically, a Web spider would like to do many activities in parallel in order to speed up the process of crawling. In fact, the DNS name resolving, i.e. getting IP address of an Internet host by contacting specific servers with name-to-IP mappings, and opening an HTTP connection to a Web server may take up to a second which is often more than receiving the response from a Web server (i.e. downloading a small or middle-sized document with a sufficiently fast connection). So the natural idea is to fetch many documents at a time. Current commercial large-scale Web robots fetch up to several thousands of documents in parallel and crawl the "whole" Web (billions of documents) within a couple of weeks. Interestingly, parallelization objects offered by operating systems such as processes and threads do not seem advantageous for multiple fetching of thousands of documents due to thread (process) synchronization overheads. Instead, a non-blocking fetching via asynchronous sockets is preferred. Indeed, present commercial search engines work with such huge amounts of data that they have to use technologies that are often beyond capabilities of traditional operating systems. Google, for example, has a file system of its own [Ghemawat2003].

Implementors of large-scale Web crawlers try to reduce the host name resolution time by means of DNS caching. The DNS server mapping host names to their IP addresses is customized and extended with a DNS *cache* and a *prefetching client*. The cache is preferably placed in the main memory for a very fast lookup in the table of names and IPs. In this way, server names that have already been put in the cache before can be found almost immediately.

New names, though, have still to be searched for on distant DNS servers. Therefore, the prefetching client sends requests to the DNS server right after URL extraction from a downloaded page and does not wait until the resolution terminates (non-blocking UDP datagrams are sent). Thus, the cache is filled up with corresponidng IPs long before they are actually needed. (DNS requests are kept completely away from a common Web surfer. It is the Web browser that gets all the work done.)

Avoiding redundancy

The biggest task of a crawler is to avoid redundancy by eliminating duplicate pages and links from the crawl. A crawler that does not respect this may easily end up in a *spider trap* – an infinite loop of links between the same pages. Such a trapped spider can "crawl" the Web for ages and collect petabytes of data, but it will be useless, because it gets stuck in just one place of the Web. There must be a module (*isUrlVisited?*) that checks whether or not a page has been already fetched before putting its URL to the working pool of pending documents (sometimes called *frontier*). The intuitive solution is to have a list of URLs already visited and to compare each newly extracted URL against this list. Unfortunately, many problems arise here:

- **Different forms of URLs**. URLs occur in various forms. They may be absolute or relative, they may or may not include port numbers, fragments, or queries, thay may contain special or even non-latin characters, they may be in lower case or upper case, etc. Before we can attempt to compare URLs, we have to normalize them and produce the so-called *canonical form*. In this form, every URL is absolute, with the host name in lower case, without non-latin characters and so on.

- **Too many URLs**. To crawl a significant portion of the Web, we would need to store somewhere a few billions URLs for further comparisons. Imagine that an average normalized URL is fifty characters long. Even for a one-billion-pages crawl, a storage capacity of 50 billion bytes (50 GB) would be required. Moreover, access to the list of URLs visited must be very fast as the check will be very frequent. How to resolve this difficulty? We can somewhat reduce the size of URLs by encoding them into MD5 fingerprints or CRC checksums. These fingerprints may be four to eight bytes long according as how many URLs we suppose to crawl. In addition, we can use each fingerprint as a hash and store the URLs in a hash table on disk. Disk seeks will still be slow, but we can improve this with a two-level hashing – host name hashing and path hashing will be done separately for each URL.

- **Duplicate pages with different URLs**. Even if we are careful enough and never crawl the same URL twice, we can still download pages with the same content if they have different URLs. In order to avoid adding links to the frontier that appear as new, because they are relative to the page with a different URL but with a duplicate content, but in reality have been added before, it is necessary for each newly fetched page to check whether it has been downloaded yet (*isPageKnown?* module in Figure I.3). Again, we can use the MD5 hash function here. We will maintain a list of fingerprints of fetched pages' contents and compare each new page against it. Unfortunately, only a very small difference between two pages that are otherwise considered as duplicate, such as a different time stamp at the bottom of the page, results in distinct fingerprints, and the duplicates recognition fails. Thus, the process must be enhanced by a technique called *shingling* [Bröder1997], which detects near duplicates.

Care must be taken not to overload Web servers with requests. Not only does it prevent a denial of service by the Web servers, but it is also a measure of politeness to other Web users. Ideally, the *load monitor & manager* distributes requests evenly among servers, for each of which there is a queue pending URLs. It controls that the interval between two requests sent to the same server be no less than, say, a minute. Besides others, fetching pages uniformly from distinct servers reduces the risk of getting stuck in a spider trap.

Dynamic pages
We have seen that the greatest danger for a Web crawler consists in not recognizing that a Web page has already been fetched before. If this happens, the spider may easily crawl a very small part of the Web infinitely long. The main source of such difficulties are page duplication and site mirroring (i.e. duplication of whole Web sites), dynamically generated pages and Web host aliases. A computer with a certain IP address may be represented by one or more host names (virtual servers). On the other hand, a Web site may be hosted by several machines with distinct IPs. This many-to-many relation between host names and IPs along with aliases (synonymous names of a Web site) makes the recogniton of known URLs even more difficult. Besides shingling for duplicate pages, there exist techniques for the detection of mirrored Web sites [Bharat2000, Cho2000, Kumar1999] thay may help resolve this problem as well. But by far the biggest trouble is with dynamic pages such as CGI, PHP, or Java scripts.

Dynamic pages are dangerous in that they can generate whatever content (including what we are not at all interested in), and that their number may be virtually infinitely large. Dynamic pages often contain generated URLs that differ only in one parameter of their query part. Also, they are often results of a database query depending on what the Web user types in a Web form, etc. It is feasible to store nor fingerprints of their URLs neither of their contents because of their immense number. How can we overcome this problem? The most robust spider would just ignore dynamic pages. However, it would probably miss a lot of important data. There have even been attempts to crawl the hidden Web behind Web forms [Raghavan2001]. In practice, we must still observe crawling statistics and set bounds for various parameters such as the number of documents gathered on a site or the crawling depth (i.e. the number of links followed leading to the current page). Whenever a bound is exceeded, crawling as a whole or just on that particular site is stopped. For example, for the crawling in Section VI.1.1, we determined the maximum crawling depth to be eight. This is in accordance with Baeza-Yates [Baeza-Yates2005, Baeza-Yates2004]. He recommends five for static pages and fifteen for dynamic pages.

I.2.2 Crawling strategies

Assume for simplicity that we are to crawl a small part of the Web that is a tree. Because we are sure that this part of the Web is finite and that we are going to visit all of its pages, we can arbitrarily choose one of the two basic crawling methods – breadth-first or depth-first crawling. Let us recall that with breadth-first crawling, we first visit nodes with the same distance (number of links) from the root node. The data structure used here to store links extracted from pages is a queue. On the other hand, in depth-first crawling, we follow links as deep as we can. We put them on a stack. See Figure I.4 for a small example. Which of the two strategies is better? In this simple case, they are the same provided we are not interested in the order of visiting individual pages. At the end, we will have a set of documents which we can, for example, add to a corpus and build an index on it.

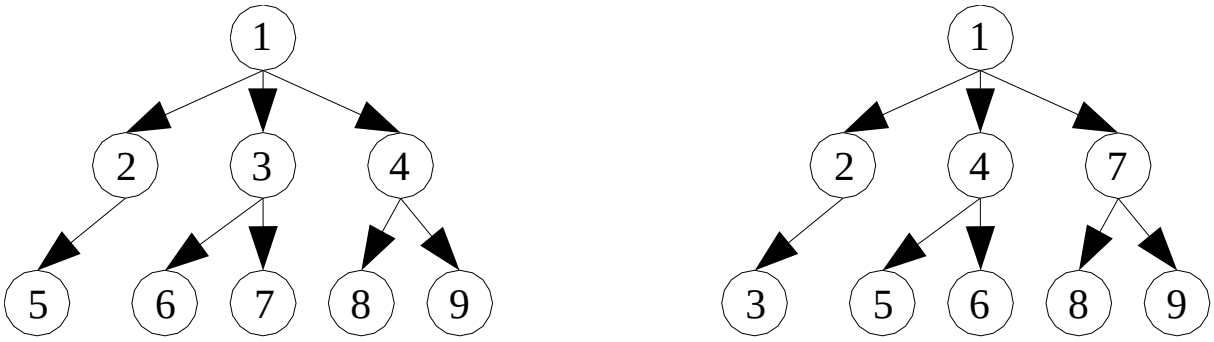


Figure I.4: Breadth-first (left) and dept-first (right) crawling of a simple Web tree.

In practice, however, neither is the Web graph a tree (and we must apply the techniques from Section I.2.1 to avoid crawling loops) nor can we collect all documents. Thus, if we know that we will not be able to crawl all pages, we would like to crawl the more important ones at least. Therefore, we expect a good crawling strategy to visit more important pages sooner during the crawl than a bad crawling strategy. We deal with the importance of Web pages in Chapter II and in Chapter III. Here, we only associate a value of significance with each Web page and set the total significance of all pages in the Web graph to be crawled to be one. Then, at any time point of the crawl, we can plot the importance value of all the pages crawled so far against the fraction of the total number of pages to crawl.

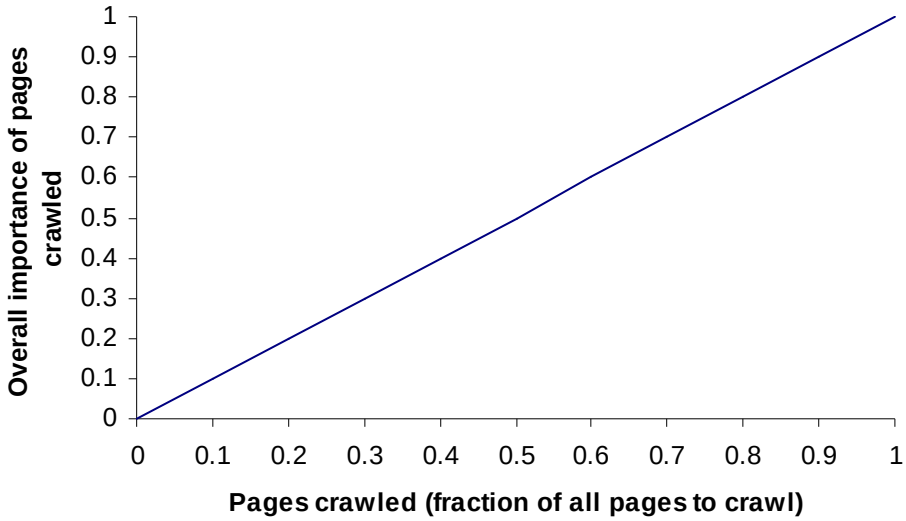


Figure I.5: Performance of a crawler sampling Web pages at random.

In a crawl, where pages would be picked up randomly (and uniformly) from the graph, the plot would be approximately diagonal like in Figure I.5. (In fact, a truly random sampling of Web pages from the real Web is quite a difficult task, which we do not cover in this thesis. See [Chakrabarti, pp.246-253] for more information on this.) The diagonal line may be considered as a baseline, and any crawler whose performance curve plotted on the chart is above the diagonal line is a more effective spider. Of course, normally we know neither the total number of pages on the Web nor their importance. Therefore, this measurement is possible for *synthetic* (artificial) *graphs*, when the number of pages and their importance are

known before, or for pre-crawled Web graphs with all the values required already computed. In both cases, we call these "artificial" spiders *crawling simulators*. Alternatively, we can measure crawling performance retroactively and compute all the values when the crawl has finished.

Baeza-Yates defines three groups of crawling strategies:

- **With no extra information**. When deciding which page to crawl next, the spider has no additional information available except knowing the structure of the Web crawled so far in the current crawl.

- **With historical information**. The crawler additionally knows the Web graph obtained in a recent "complete" crawl. This is common for search engine spiders that regularly crawl the Web in several-week intervals. Typically, the spider knows what pages existed a couple of weeks ago, what links they contained and what importance the pages had which was computed after the crawl. Although the Web changes very fast (about 25% new links are created every week [Ntoulas2004]), the historical data were too costly to acquire so that it could be entirely neglected. Thus, the selection of a next page to crawl will be based on the historical information.

- **With all information**. This is a theoretical strategy not usable in a real Web crawl. We will call it the **omniscient** method, which perfectly knows the whole Web graph that should be crawled including the values of importance of individual pages. This method always chooses the page with the highest importance from the frontier.

Crawling strategies with no extra information

- **Breadth-first**. We mentioned this technique earlier. It is reported to collect high quality (important) pages quite soon [Najork2001]. On the other hand, depth-first strategies are not much used in real Web crawling, also because the maximum crawling depth is worse controllable in them.

- **Backlink-count** [Cho1998]. Pages in the frontier with a higher number of in-links from pages already downloaded have a higher priority of crawl.

- **Batch-PageRank** [Cho1998]. We will talk about PageRank in Section II.2. Now, we can think of it as importance. This technique calculates PageRank values for the pages in the frontier after downloading every k pages. Of course, these PageRanks are based on the graph constituted of the pages downloaded so far, and they are only estimates of the real PageRanks derived from the whole Web graph. After each re-calculation, the frontier is prioritized according to the estimated PageRank and the top k pages will be downloaded next.

- **Partial-PageRank**. It is like Batch-PageRank but with temporary PageRanks assigned to new pages until a new re-calculation is done. These temporary PageRanks are computed non-iteratively unlike normal PageRanks as the sum of PageRanks of in-linking pages divided by the number of out-links of those pages (the so-called out-link normalization).

- **OPIC** [Abiteboul2003]. This technique may be considered as a weighted backlink-count strategy.

- **Larger-sites-first**. This method tries to cope best with the rule that Web sites must not be overloaded and choose preferentially pages from Web sites having a large number of pending pages. The goal is not to have at the end of the crawl a small number of large sites, because that would slower down crawling due to the delay required between two accesses to the same site.

Crawling strategies with historical information
Again, we would like to order the pages in the frontier by their PageRank and crawl the more important ones first. For the pages encountered in the current crawl that existed when the last crawl was run, we use their historical PageRank even though we are aware that their current PageRank may have changed. The pages that did not exist then have to be assigned some estimates. There are several methods how to deal with these new pages:

- **Historical-PageRank-Omniscient**. Again, it is a theoretical variant which knows the complete graph and assigns "true" PageRanks to the new pages.

- **Historical-PageRank-Random**. It assigns to the new pages random PageRanks chosen from those computed for the previous crawl.

- **Historical-PageRank-Zero**. New pages are all assigned a zero PageRank and are thus crawled after "old" pages.

- **Historical-PageRank-Parent**. Each new page is assigned an out-link-normalized PageRank of its parent page(s) linking to it. If a parent page is new as well (there is no historical PageRank associated with it) we obviously proceed to the grandparent and so forth.

Recommendations
Baeza-Yates and his colleagues conducted crawling simulations as well as real crawls of the Web graph of the whole Greek (.gr) and Chilean (.cl) national domains. The total number of Web pages crawled was in the order of millions of pages. Their experiments confirmed the following. The *omniscient* technique is the best as expected except the last crawl stages (see Figure I.6 with the performance chart of crawling the Greek Web in September 2004). It crawls important pages fast, perhaps too fast so that it needs to select pages more or less at random towards the end of the crawl in order not to overload Web sites. From the strategies with no extra information *backlink-count* and *partial-pagerank* are the worst. In some crawling stages they are even worse then the baseline random ("diagonal") method (not in figure). Breadth-first performs very well for the first 30% of total pages to crawl, then its efficiency slightly decreases. *Batch-pagerank*, *OPIC* and *larger-sites-first* are the most efficient crawling strategies. The importance of the first 25% of pages they collect is more than 50% of the overall significance spread over pages in the graph.

Figure I.7 shows the performance of the various methods using historical information compared with the *omniscient* variant and *OPIC*. The "historical" methods take advantage of a complete crawl from May 2004 when only 55% of pages in the current crawl existed. Surprisingly, *historical-page-rank-random* is doing quite well even though 45% of randomly evaluated pages may seem a lot. A possible explanation is that very important pages are

relatively stable, and that it is the "less significant" part of the Web that changes. *OPIC* is less efficient at the beginning, but it improves later on. After some further measurements, Baeza-Yates concludes that "historical" strategies are "marginally better" than *OPIC* and *larger-sites-first*, and he recommends to use the latter for practical reasons. *Larger-sites-first* is more suitable for distributed crawlers, for no communication is needed between crawlers to exchange information on weighted in-links to a given page like in *OPIC*.

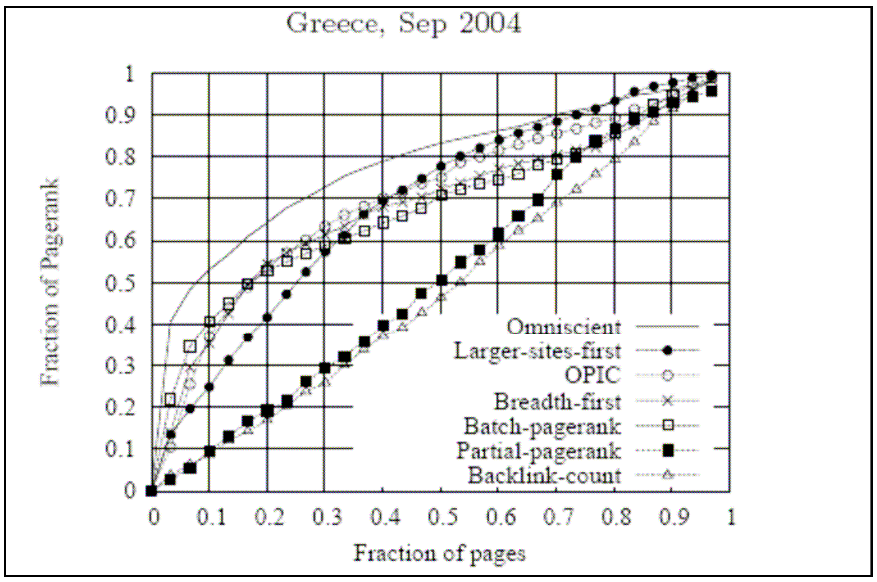


Figure I.6: Crawling methods performance with no extra information [Baeza-Yates2005].

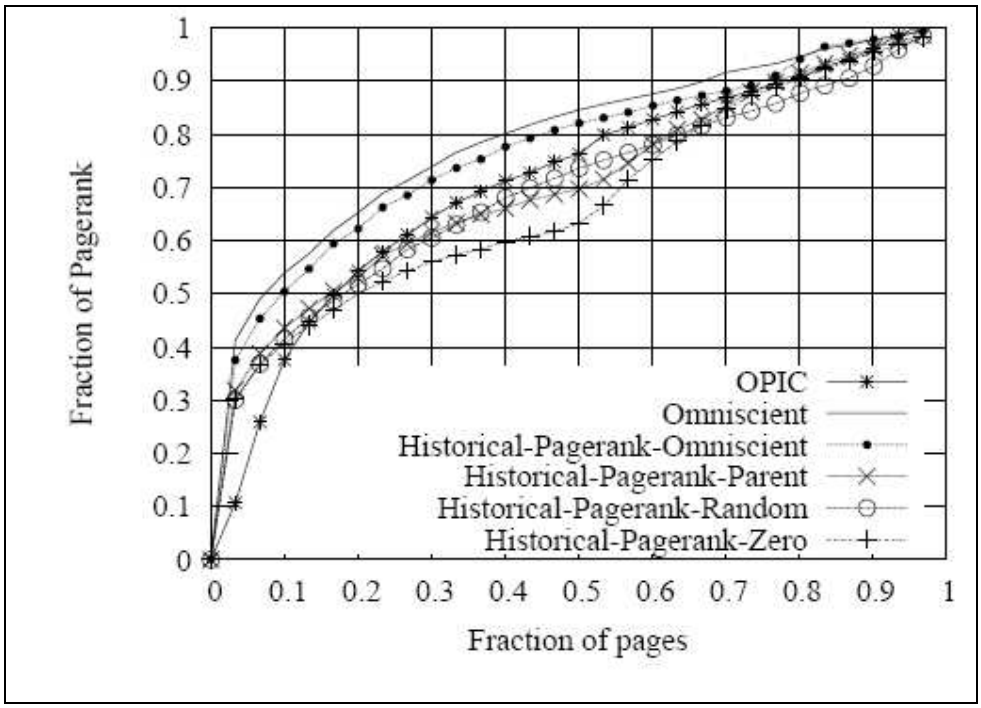


Figure I.7: Crawling methods performance with historical information [Baeza-Yates2005].

I.3 Summary

Studying the models of the Web helps us realize how vast, dynamic, and heterogenous it is. It also makes us wish to understand in detail its structure and behaviour. However, we cannot study anything before we crawl it. Crawling the Web is a prerequisite of Web mining. Before we can apply any Web mining techniques to some Web data, we first need to obtain the data somehow. Certainly, there exist archives of complete Web crawls available for researchers or free crawling software ready to use (both mentioned above). Nevertheless, they are difficult or even impossible to customize so as to meet very specific crawling needs. Furthermore, without realizing how difficult and complex Web crawling is and without understanding its internals to some extent at least, one could hardly consider oneself as a Web mining researcher. The code of a spider used in the present research can be found on the companion CD of this dissertation.

We did not cover *focused crawling* in this chapter. The aim of a focused (or also vertical) Web spider is to download topic-relevant documents and not all documents. Such crawlers need to classify pages in relevant (on a given topic) and irrelevant ones and follow links from the on-topic pages only. Performance of *vertical spiders* can be measured in the same way as we show in Section I.2.2, but instead of a general importance of pages crawled, we are interested in their topical importance. There are a number of methods that deal with this problem such as reinforcement learning [Rennie1999], context graphs [Diligenti2000], neural networks [Chau2003] or those proposed in [Chakrabarti1999]. There is an overview of focused crawling techniques in [Chakrabarti2000, pp. 268-283].

II Ranking Algorithms for Web Sites

We have understood the term *importance* used throughout Section I.2 rather intuitively until now. However, everyone recognizes that not all Web documents are important, valuable. The overwhelming majority of Web pages are useless for a particular Web user. They are not relevant to the topic the user is interested in. This can be figured out by filtering Web documents (classifying their textual content) and providing the user the relevant ones only. But do all the on-topic documents have the same quality? Everybody who has experience with searching for information on the Web will agree that they do not. Thus, in addition to relevance filters, some further criteria must decide which documents are worth our attention and which are not.

In 1998, two PhD students from the Stanford University in California, Sergey Brin and Larry Page, published a report on their project of a large-scale Web search engine called Google [Brin1998]. They described its architecture and also gave details of a new algorithm for ranking Web pages by importance – *PageRank*. (Curiously enough, page in its name can mean a Web page but also the surname of one its inventors.) What was then a University project has developed into a commercial multi-billion-dollar-revenue company operating a Web search engine serving hundreds of millions of users every day. Approximately at the same time, Jon Kleinberg proposed another algorithm for determining significant Web pages called *HITS* [Kleinberg1999b] but made to attempt to commercialize it. It remained as an academic foundation.

We can only guess that it was the commercial success of Google that raised an immense interest in the new group of algorithms for detecting significant Web pages that later earned the name of *ranking* or *topic-distillation algorithms*. New ranking methods and modifications appeared soon and the publication stream does not seem to fade out - SALSA, TruRank, BackRank, ObjectRank, AuthorRank, SCEAS Rank, etc. In this chapter, we will concentrate on PageRank and its modifications that has become extremely popular, and we will deal with HITS to much lesser extent.

II.1 First-Order Methods

A simple and intuitive procedure of ordering Web pages by importance is to count in-links of a Web page. To formalize, let $G = (V, E)$ be a directed, edge-weighted graph (Web graph), V a set of vertices (Web pages) and E a set of edges (hyperlinks among Web pages). Then, we can calculate the number of in-links of each node u in the graph like its in-degree:

$$D_{in}(u) = \sum_{(v,u)\in E} w(v,u) \tag{II.1}$$

where $w(v, u)$ is the weight of the edge pointing from node v to node u and we assume that all edge weights are set to one. We will refer to this ranking mechanism as In-Degree. The In-Degree ranking is called a *first-order* or *radius-1* method. The score obtained for a node depends only on its direct neighbours. In other words, nodes not sharing the same edge have no influence on each other. This is in contradiction with real life as objects in social networks (Web may be considered a social network – see Chapter III) often have an indirect impact on one another. PageRank and HITS are *higher-order* techniques and take this into account. If the values of weights w are allowed to be more than one, we call the in-degree a *weighted in-degree*. Although it is usually not much useful to determine *authoritativeness* of a Web page or Web site (according to the level we are interested in) by means of weighted in-degrees, because parallel edges between them are mostly ignored, it is appropriate to do so in other graphs such as bibliographic citation graphs. We can then call the weighted in-degree a *citation count* or simply citations (compare with Section III.2.1 and V.3.3).

II.2 PageRank

We will first introduce PageRank as presented in [Brin1998, Page1999] in a intuitive manner in Section II.2.1, and then we will enhance it within a linear system formulation [Bianchini2005] in Section II.2.2 and a probabilistic framework for ranking methods by [Diligenti2004] in Section II.2.3. In Section II.2.4, we discuss convergence issues, and we describe a PageRank modification that is most related to the innovative work in this thesis in Section II.2.5 – PageRank for publications by [Sidiropoulos2005]. Finally, we enumerate current research issues and trends on this topic [Langville2003] in Section II.2.6.

II.2.1 Primer

Using the Web graph $G = (V, E)$ from Section II.1, the PageRank score PR(u) for page u introduced by Brin and Page is defined as follows:

$$\mathrm{PR}(u) = \frac{1-d}{|V|} + d \sum_{(v,u)\in E} \frac{\mathrm{PR}(v)}{D_{out}(v)} \tag{II.2}$$

where d (*damping factor*) is an empirically determined constant usually set between 0.8 and 0.9 and $D_{out}(v)$ is the out-degree of node v computed analogically to (II.1). Note that PageRank of one node is dependant on PageRanks of other nodes, which can, in turn, be directly or indirectly (via other nodes) supplied with PageRank from the current node. So there is a recursion that allows influencing any other node to which there is a path from the current node.

Normalization of the rank obtained from in-linking nodes by their out-degree is a salient feature of PageRank. It penalizes nodes linking to many others. This is in accordance with real world situations: a citation by a researcher citing often is less valuable than that made by

someone who cites rarely. Figure II.1 shows the idea of such an *out-link normalization* in PageRank computation. *(1 - d)* is a randomizing factor representing the possibility to jump to any node in the graph regardless of the out-edges from the current vertex. On the contrary, *d* stands for the probability of following an out-link from the present page. Introducing the random term prevents loops of nodes (called *rank sinks*) from accumulating too much rank and not propagating it further. See Figure II.2 for a rank sink example. There is also a difficulty with nodes with no out-links (referred to as *dangling pages*) that would not distribute their PageRank either. In fact, zero-out-degree Web pages and rank sinks are the main obstacles in a straightforward computation of PageRank. Why are pages with no out-links and closed loops of pages so annoying and how are these problems resolved will be shown later on. On the other hand, nodes without in-links are not harmful, and their PageRank is always smaller than that of any nodes with some in-links as follows from (II.2).

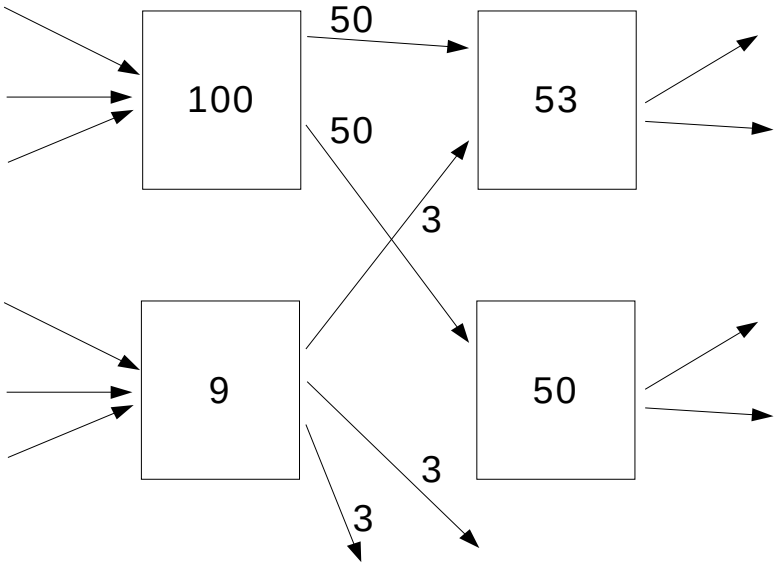


Figure II.1: Main idea of a PageRank calculation [Page1999].

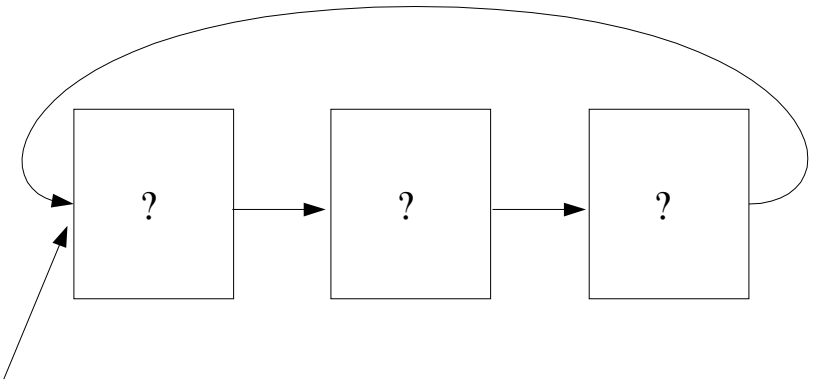


Figure II.2: Rank sink example.

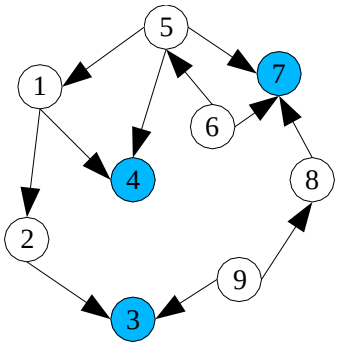
Iterative Calculation
In practice, we compute PageRank as follows.

1. We remove duplicate links and self-links from the graph.
2. We set initial PageRanks of all nodes in the graph uniformly so that the total rank in the system is one. This is the zeroth iteration.
3. We remove nodes having no out-links iteratively because removing one zero-out-degree node may cause another one to appear.
4. We compute the PageRank scores for all nodes in the residual graph according to (II.2) using the scores from the previous iteration. We perform an L_1 normalization so that the total rank in the system (including the vertices removed in step 3) is one again.
5. We repeat step 4 until convergence. Numerical convergence of the scores is usually not necessary. An ordering of nodes (by PageRank) that does not change (or changes relatively little) is satisfactory [Chakrabarti2002, p. 211].
6. We gradually add back the nodes removed in step 3, compute their rank score like in (II.2) and re-normalize the whole system.

Properties
The number of iterations needed depends on the number of nodes and edges in the graph. For a Web graph with over 320 million pages, roughly 50 iterations were required [Page1998]. The order of the nodes added back in as well as the frequency of normalization may affect the final rank scores, however, it should not have a large effect on the ranking itself. The property of the overall rank being one at each time step justifies the explanation of PageRank calculation in terms of a random walk. In fact, the PageRank score of a Web page is then a fraction of time spent on this page by a random Web surfer browsing on the Web for some infinitely long time. For a detailed analysis of the random walk framework, see [Diligenti2004] and an excerpt in Section II.2.3. For some more details including matrix notation of (II.2), see [Chakrabarti2002, pp. 210-211] and [Ding2001b]. There exist PageRank modifications. For instance, the one proposed by [Sidiropoulos2005] is meant particularly for bibliographic citation graphs.

II.2.2 Linear System Guise

Web matrices
Before we can explain how to compute PageRank scores for a Web graph by solving a system of linear equations, we first need to define two matrices. Let $G = (V, E)$ be a Web graph as before, $\boldsymbol{A}$ its *adjacency matrix*, and $\boldsymbol{T}$ its *transition matrix*. Let A_{ij} be one if i links to j, i.e. if there exists $(i, j) \in E$, and zero otherwise. Clearly, $\boldsymbol{A}$ is asymmetric and it imposes no restrictions upon the existence of self-links or parallel edges. By normalizing elements in $\boldsymbol{A}$ by out-degree and transposing $\boldsymbol{A}$, we obtain the transition matrix: $T_{ji} = A_{ij} / \sum_k A_{ik}$. An example of a Web graph and its corresponding matrices $\boldsymbol{A}$ and $\boldsymbol{T}$ is shown in Figure I.3.



$$\begin{bmatrix} 0&1&0&1&0&0&0&0&0\\ 0&0&1&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&0\\ 1&0&0&1&0&0&1&0&0\\ 0&0&0&0&1&0&1&0&0\\ 0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&1&0&0\\ 0&0&1&0&0&0&0&1&0 \end{bmatrix} \qquad \begin{bmatrix} 0&0&0&0&1/3&0&0&0&0\\ 1/2&0&0&0&0&0&0&0&0\\ 0&1&0&0&0&0&0&0&1/2\\ 1/2&0&0&0&1/3&0&0&0&0\\ 0&0&0&0&0&1/2&0&0&0\\ 0&0&0&0&0&0&0&0&0\\ 0&0&0&0&1/3&1/2&0&1&0\\ 0&0&0&0&0&0&0&0&1/2\\ 0&0&0&0&0&0&0&0&0 \end{bmatrix}$$

$G = (V, E)$ $\boldsymbol{A}$ $\boldsymbol{T}$

Figure II.3: Example of a Web graph and its adjacency (***A***) and transition (***T***) matrix.

Linear system

Now, we put all PageRanks in the Web graph PR(1), PR(2),…, PR(*N*) into a *PageRank vector* $\boldsymbol{x} = [x_1, x_2, \ldots, x_N]^T$ and apply (II.2) to the whole graph:

$$\boldsymbol{x} = (1 - d)\boldsymbol{e}_N + d\boldsymbol{T}\boldsymbol{x} \tag{II.3}$$

where $\boldsymbol{e}_N = [1,...,1]^T$ is a unity column vector of N ones. We will denote the equilibrium solution of this system as $\boldsymbol{x}^* = [x_1^*, x_2^*, ..., x_N^*]^T$. Note that unlike (II.2), the first term is not divided by the number of all pages N or $|V|$. We will come back to this interesting point a little later.

Dynamical system

A *linear system* can normally be solved algebraically using Gaussian elimination. However, this would require $O(N^3)$ floating-point operations, which is absolutely unfeasible with regard to the number of pages in the Web (billions). Therefore, the system must be transformed into a corresponding *dynamical system* and solved numerically:

$$\boldsymbol{x}(t) = (1 - d)\boldsymbol{e}_N + \mathrm{d}\boldsymbol{T}\boldsymbol{x}(t - 1) \tag{II.4}$$

where $\boldsymbol{x}(t)$ is the PageRank vector at time step t. It can be proven [Bianchini2005] that this dynamical system is stable (i.e. it converges), and that the sequence $\boldsymbol{x}(1)$, $\boldsymbol{x}(2)$, $\boldsymbol{x}(3)$… converges to the solution of the linear system in (II.3) independently of the non-zero initial vector $\boldsymbol{x}(0)$ if $d < 1$. This can be fixed very simply by adhering to the recommendations of PageRank's inventors and setting d on 0.85. We will discuss the impact of d on PageRank computation further below. But there is also a problem when we want to make (II.4) coherent with the Markov process model (see Section II.2.3), which was the original framework for PageRank. In this probabilistic model, the total sum of PageRanks over all nodes must be equal to one at any time. For the system (II.4) to converge, ***T*** must be *stochastic* (i.e. non-negative with all columns summing up to one) . What about this condition? Is it easy to figure this out?

Removing dangling pages

When we have a look at Figure II.3, we can see immediately that $\boldsymbol{T}$ is not stochastic. Using it directly in (II.4) and forcing $\|\boldsymbol{x}(t)\|_1 = 1$ for $t \geq 0$ would not yield a fixed-point distribution of PageRanks over the graph. The sequence $\{\boldsymbol{x}(t)\}$ would not converge. There are three columns in $\boldsymbol{T}$ not summing up to one the source of which are the *dangling pages* 3, 4, and 7 in the graph. Dangling pages (originally called *dangling links* by Brin and Page to refer to pages the links to which have been encountered by the Web crawler but that have not yet been crawled themselves) have no out-links and hence the columns in $\boldsymbol{T}$ that sum up to null. In fact, there are very many pages without out-links in the Web, and (II.4) could be directly applied only to some selected portions of the Web such as the SCC component in Figure I.2.

There are two main theoretic approaches (see [Berkhin2005] for others) how to deal with dangling pages. Both of them are depicted in Figure II.4. The first is to add a *dummy page* with a self-link to the graph and let all dangling pages point to it. Thus the dimensions of the adjacency and transition matrices $\boldsymbol{A}_1$ and $\boldsymbol{T}_1$ increase by one as well as the PageRank vector $\boldsymbol{x}_1 = [x_1, x_2, \ldots, x_N, x_{N+1}]^{\mathrm{T}}$. We must then replace $\boldsymbol{T}$, $\boldsymbol{x}$, and $\boldsymbol{e}_N$ in (II.4) with $\boldsymbol{T}_1$, $\boldsymbol{x}_1$, and $\boldsymbol{e}_{N+1}$:

$$\boldsymbol{x}_1(t) = (1 - d)\boldsymbol{e}_{N+1} + \mathrm{d}\boldsymbol{T}_1\boldsymbol{x}_1(t - 1). \tag{II.5}$$

In the second approach, we make dangling pages link to all pages in the graph including themselves like in Figure II.4 bottom. All dimensions remain intact, and the linear system (II.3) changes into:

$$\boldsymbol{x}_2 = \frac{1-d}{N}\boldsymbol{e}_N + d\boldsymbol{T}_2\boldsymbol{x}_2 . \tag{II.6}$$

Normalization

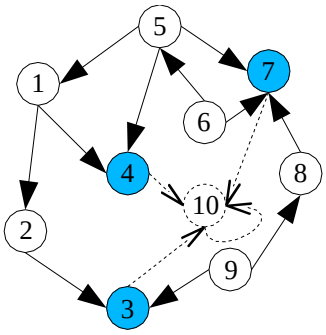
Bianchini et al. prove that equations (II.3) through (II.6) describe related systems, and that the ranking scheme provided by those four systems is the same. Moreover, if $\boldsymbol{x}_2^*$ is the stationary solution of (II.6), they show that it holds that this *normalized PageRank* can be obtained by normalizing the "*unnormalized*" PageRank $\boldsymbol{x}^*$:

$$\boldsymbol{x}_2^* = \boldsymbol{x}^* / \| \boldsymbol{x}^* \|_1 . \tag{II.7}$$

Thus, $\| \boldsymbol{x}_2^* \|_1 = \| \boldsymbol{x}_2(t) \|_1 = 1$ for $t \geq 0$ provided that $\|\boldsymbol{x}_2(0)\|_1 = 1$. On the other hand, the stationary PageRank vector $\boldsymbol{x}_1^*$ derived from (II.5) is not normalized, and its relation to $\boldsymbol{x}^*$ is the following:

$$\boldsymbol{x}_1^* = [\boldsymbol{x}^{*\mathrm{T}}, 1 + \frac{d}{1-d}\boldsymbol{R}\boldsymbol{x}^*]^T \tag{II.8}$$

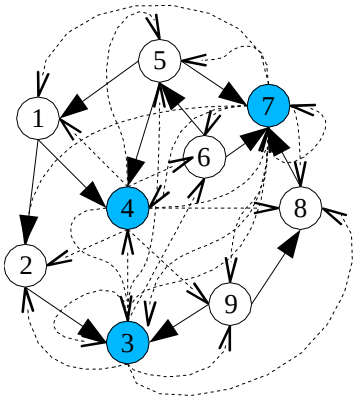
where $\boldsymbol{R} = [\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N]$, and $r_i = 1$ if i is a dangling page in the original graph G and $r_i = 0$ otherwise. In fact, $\boldsymbol{R}$ is the last row in $\boldsymbol{T}_1$ without the last element. For instance, $\boldsymbol{R}$ = [0,0,1,1,0,0,1,0,0] in Figure II.4.



$$\begin{bmatrix} 0&1&0&1&0&0&0&0&0&\mathbf{0}\\ 0&0&1&0&0&0&0&0&0&\mathbf{0}\\ 0&0&0&0&0&0&0&0&0&\mathbf{1}\\ 0&0&0&0&0&0&0&0&0&\mathbf{1}\\ 1&0&0&1&0&0&1&0&0&\mathbf{0}\\ 0&0&0&0&1&0&1&0&0&\mathbf{0}\\ 0&0&0&0&0&0&0&0&0&\mathbf{1}\\ 0&0&0&0&0&0&1&0&0&\mathbf{0}\\ 0&0&1&0&0&0&0&1&0&\mathbf{0}\\ \mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{1} \end{bmatrix}$$

$$\begin{bmatrix} 0&0&0&0&1/3&0&0&0&0&\mathbf{0}\\ 1/2&0&0&0&0&0&0&0&0&\mathbf{0}\\ 0&1&0&0&0&0&0&0&1/2&\mathbf{0}\\ 1/2&0&0&0&1/3&0&0&0&0&\mathbf{0}\\ 0&0&0&0&0&1/2&0&0&0&\mathbf{0}\\ 0&0&0&0&0&0&0&0&0&\mathbf{0}\\ 0&0&0&0&1/3&1/2&0&1&0&\mathbf{0}\\ 0&0&0&0&0&0&0&0&1/2&\mathbf{0}\\ 0&0&0&0&0&0&0&0&0&\mathbf{0}\\ \mathbf{0}&\mathbf{0}&\mathbf{1}&\mathbf{1}&\mathbf{0}&\mathbf{0}&\mathbf{1}&\mathbf{0}&\mathbf{0}&\mathbf{1} \end{bmatrix}$$

$G_1 = (V_1, E_1)$ $\quad$ $\boldsymbol{A}_1$ $\quad$ $\boldsymbol{T}_1$



$$\begin{bmatrix} 0&1&0&1&0&0&0&0&0\\ 0&0&1&0&0&0&0&0&0\\ 1&1&1&1&1&1&1&1&1\\ 1&1&1&1&1&1&1&1&1\\ 1&0&0&1&0&0&1&0&0\\ 0&0&0&0&1&0&1&0&0\\ 1&1&1&1&1&1&1&1&1\\ 0&0&0&0&0&0&1&0&0\\ 0&0&1&0&0&0&0&1&0 \end{bmatrix}$$

$$\begin{bmatrix} 0&0&1/9&1/9&1/3&0&0&0&0\\ 1/2&0&1/9&1/9&0&0&0&0&0\\ 0&1&1/9&1/9&0&0&0&0&1/2\\ 1/2&0&1/9&1/9&1/3&0&0&0&0\\ 0&0&1/9&1/9&0&1/2&0&0&0\\ 0&0&1/9&1/9&0&0&0&0&0\\ 0&0&1/9&1/9&1/3&1/2&0&1&0\\ 0&0&1/9&1/9&0&0&0&0&1/2\\ 0&0&1/9&1/9&0&0&0&0&0 \end{bmatrix}$$

$G_2 = (V_2, E_2)$ $\quad$ $\boldsymbol{A}_2$ $\quad$ $\boldsymbol{T}_2$

Figure II.4: Two ways of tackling dangling pages.

Now we can see that the division by N is not necessary in (II.3) provided we content ourselves with unnormalized PageRanks. Actually, the L_1-norm of the fixed point is always bounded by the number of all pages N:

$$\| \boldsymbol{x}^* \|_1 = N - \frac{d}{1-d} \sum_{i \in S} x_i^* \tag{II.9}$$

where S is a set of dangling pages. Therefore, the total unnormalized PageRank $\|\boldsymbol{x}^*\|_1$ in a graph without dangling pages is N, and any nodes with no out-links cause a loss of *energy*.

If we would like to have $\|\boldsymbol{x}^*\|_1 = \|\boldsymbol{x}(t)\|_1$ for $t \geq 0$ in order to have a stochastic system conforming to the random walk framework (see Section II.2.3), the transition matrix must be stochastic as well, i.e. it must be non-negative and have all columns summing up to one. Because $\boldsymbol{T}$ is not stochastic in general due to dangling pages and the system (II.4) would therefore not converge, we need to take advantage of $\boldsymbol{T}_1$ or $\boldsymbol{T}_2$ and the appropriate equations (II.5) or (II.6). Equation (II.5) resulting in a PageRank vector that should yet be normalized is preferred by [Bianchini2005]; (II.6) was originally used by PageRank's inventors [Page1999]. An alternative, practical approach is to preprocess the Web graph by iteratively removing

dangling pages, computing PageRank on the stochastic transition matrix, and then distributing the scores to previously removed nodes [Chakrabarti2002, p. 211]. This process is criticized by [Langville2003] for not being fair, but it is embedded in the iterative calculation outlined in Section II.2.1.

Eigensystem
Theoretically, if the Web graph had no dangling pages, no rank sinks and was strongly connected (the transition matrix is then stochastic and said to be *primitive*), we could drop the factors d and $(1 - d)$ from (II.3) and could directly solve the system $\boldsymbol{x} = \boldsymbol{T}\boldsymbol{x}$ whose stationary solution is the *principal eigenvector* of $\boldsymbol{T}$ [Chakrabarti2002, p. 210]. See also Sections II.3.1 and III.2.2 for a similar concept. A survey of eigenvector ranking methods for the Web is available in [Langville2005].

II.2.3 Random Walk Guise

Diligenti [Diligenti2004] distinguishes *horizontal* and *vertical ranking* methods. Horizontal rankings are only based on the Web graph topology and do not take into account the contents of Web pages. PageRank and HITS (see Section II.3) are both horizontal. Vertical (focused) rankings classifying Web documents are useful for topic search. Diligenti's probabilistic framework is based on *random walks* in that the relevance (rank) x_p of a page p is computed as the probability of visiting that page in a random walk on the Web graph. The most popular pages (i.e. most often cited) are the most likely to be visited during a random walk.

Random walk
A random walk in the context of Web browsing is a mathematical model of actions taken by a generic Web surfer. At each step of the walk, the surfer can perform one of the following actions: jump to any Web page (action j), follow a link to another page (action l), follow a backlink from the current page (action b), stay where he or she is (action s). Thus, the set of atomic actions is $O = \{j, l, b, s\}$. At each step, the behaviour of the surfer depends on the current page. If he finds it interesting, he will probably click on a link there. If he finds it boring, he types another URL in the address bar of his Web browser. So the surfer's behaviour can be modelled by a set of conditional probabilities depending on the current page q:

- $x(p|q, l)$*: probability of following a link from page q to page p*
- $x(p|q, b)$*: probability of following a backlink from q to p*
- $x(p|q, j)$*: probability of jumping from q to p*
- $x(s|q)$*: probability of staying on q*

If $G = (V, E)$ is a Web graph defined as earlier and p and q are Web pages ($p,q \in V$) then the following constraints have to be satisfied for each page q:

$$\sum_{p \in V} x(p \mid q, j) = 1, \qquad \sum_{(q,p) \in E} x(p \mid q, l) = 1, \qquad \sum_{(p,q) \in E} x(p \mid q, b) = 1. \tag{II.10}$$

Evidently, the first constraint in (II.10) includes the case of remaining on the page because p can be the same as q. The probabilistic random walk model can be made use of to compute the probability $x_p(t)$ – that the surfer is located on page p in time t. The probability distribution on all pages is represented by the vector $\boldsymbol{x}(t) = [x_1(t),\ldots,x_N(t)]'$ where N is the total number of pages. The probabilities $x_p(t)$ are updated in each step of the random walk according to the following formula:

$$\begin{aligned} x_p(t+1) &= \sum_{q \in V} x(p \mid q) x_q(t) = \\ &= \sum_{q \in V} x(p \mid q, j) x(j \mid q) x_q(t) + \sum_{(q,p) \in E} x(p \mid q, l) x(l \mid q) x_q(t) + \\ &+ \sum_{(p,q) \in E} x(p \mid q, b) x(b \mid q) x_q(t) + x(s \mid p) x_p(t) \end{aligned} \tag{II.11}$$

where the probability *x(p|q)* of moving from page *q* to page *p* is expanded considering the user's actions. The probabilities *x(j|q)*, *x(l|q)*, and *x(b|q)* are general probabilities of jumping to a random page from page *q*, following a link from *q*, and following a backlink from *q*, respectively, without specifying a target page.

Now we will move to a matrix notation. The probabilities defining the random surfer model may be organized in a couple of $N \times N$ matrices:

- *forward matrix* $\boldsymbol{\Delta}$ whose element (p,q) is the probability x(p|q,l)
- *backward matrix* $\boldsymbol{\Gamma}$ of the probabilities x(p|q,b)
- *jump matrix* $\boldsymbol{\Sigma}$ gathering the probabilities x(p|q,j)

We can also define a set of *action* matrices that inform about the probabilities of the individual actions taken on each page *q*. These matrices are $N \times N$ diagonal matrices having *x(j|q)*, *x(l|q)*, *x(b|q)* and *x(s|q)* as their diagonal values *(q,q)*. We will denote those matrices $\boldsymbol{D}_j$, $\boldsymbol{D}_l$, $\boldsymbol{D}_b$ and $\boldsymbol{D}_s$, respectively. We can then restate (II.11) as

$$\boldsymbol{x}(t+1) = (\boldsymbol{\Sigma} \cdot \boldsymbol{D}_j)^{\mathrm{T}} \boldsymbol{x}(t) + (\boldsymbol{\Delta} \cdot \boldsymbol{D}_l)^{\mathrm{T}} \boldsymbol{x}(t) + (\boldsymbol{\Gamma} \cdot \boldsymbol{D}_b)^{\mathrm{T}} \boldsymbol{x}(t) + (\boldsymbol{D}_s)^{\mathrm{T}} \boldsymbol{x}(t). \tag{II.12}$$

By defining the transition matrix as

$$\boldsymbol{T} = (\boldsymbol{\Sigma} \cdot \boldsymbol{D}_j + \boldsymbol{\Delta} \cdot \boldsymbol{D}_l + \boldsymbol{\Gamma} \cdot \boldsymbol{D}_b + \boldsymbol{D}_s)^{\mathrm{T}}$$

we can write (II.12) in the following way:

$$\boldsymbol{x}(t+1) = \boldsymbol{T} \cdot \boldsymbol{x}(t). \tag{II.13}$$

From the initial distribution of probabilities $\boldsymbol{x}(0)$ we can compute a distribution in any time step *t*:

$$\boldsymbol{x}(t) = \boldsymbol{T}^{t} \cdot \boldsymbol{x}(0). \tag{II.14}$$

Equation (II.14) describes a Markov chain whose state transition matrix is $\boldsymbol{T}^{\mathrm{T}}$. The final rank of all pages in the graph is the stationary distribution $\boldsymbol{x}(\infty)$ of this chain. Diligenti further shows that on some conditions there must exist such a distribution and that it is independent of the initial vector $\boldsymbol{x}(0)$.

PageRank calculation based on a random walk

The single-surfer model may be extended to a *multisurfer walk* in which the things become slightly more complicated. In this model there are several surfers influencing one another. But we are more interested in how the PageRank calculation fits into the probabilistic single-surfer random walk framework.

PageRank is a special case of the single-surfer random walk in that it considers only two basic actions: jumping to a random page from the current page q with probability $x(j|q) = 1 - d$ and following a link from the page q with probability $x(l|q) = d$ (d may be chosen arbitrarily between 0 and 1 with the effect on convergence speed explained in II.2.4). The other two probabilities known from the general model ($x(b|q)$ and $x(s|q)$) are null. Obviously, all the probabilities are independent of the current page q. The target page p of a jump is selected uniformly from all the N pages in the graph G, thus $x(p \mid j) = 1/N, \forall p \in G$. The probability of following a link from page q to page p is $x(p|q,l) = \alpha_q$ where $\alpha_q = 1/h_q$ and h_q is the *hubness* of page q, i.e. the number of links pointing from q elsewhere (out-degree). Therefore, we can rewrite (II.11) as

$$x_p(t+1) = \frac{1-d}{N}\sum_{q \in V} x_q(t) + d\sum_{(q,p)\in E} \alpha_q x_q(t) = \frac{1-d}{N} + d\sum_{(q,p)\in E} \alpha_q x_q(t) \qquad \text{(II.15)}$$

where $\sum_{q \in E} x_q(t) = 1$.

The fact that $0 < d < 1$ implying $x(j|q) = 1 - d > 0$ guarantees that the PageRank vector converges to a distribution of scores independent of the initial distribution. Again, using a matrix notation, the computation of PageRank looks like this:

$$x_p(t+1) = \frac{1-d}{N}\boldsymbol{E} + d\boldsymbol{A}^{\mathrm{T}}\boldsymbol{Z}\boldsymbol{x}(t) \qquad \text{(II.16)}$$

where $\boldsymbol{E}$ is the $N \times N$ identity matrix, $\boldsymbol{A}$ is the adjacency matrix of the Web graph defined as before (i.e. an element $A_{pq} = 1$ if there is a link from p to q and it is zero otherwise), and $\boldsymbol{Z}$ is a diagonal matrix whose element $Z_{qq} = \alpha_q$.

There is a little problem with *sink pages* (we call them dangling pages in Section II.2.2) whose hubness is zero (i.e. $ch(q) = \varnothing$) and therefore we cannot compute the term $1/h_q$. Instead, it should be $x(l|q) = 0$ resulting in $x(j|q) = 1$ for any sink page q. So the PageRank equation must be modified in that $x(j|q) = 1 - d$ if $ch(q) \neq \varnothing$ and $x(j|q) = 1$ if $ch(q) = \varnothing$. Then the first term in (II.15) will not be constant but the probability $x(p \mid j,t) = \frac{1}{N}\sum_{q \in G} x(j \mid q)x_q(t)$ (jumping to p in time step t) needs to be computed at the beginning of each iteration.

Conclusions

Diligenti also presents HITS (another well-known ranking algorithm described in Section II.3) in terms of a multisurfer random walk notation (see Section II.3.1 for the eigenvector interpretation) and compares HITS with PageRank: Computation of PageRank is stable (see Section II.2.6) and it can be applied to large document collections because small communities are not overwhelmed by large ones. On the other hand, PageRank does not take into account complex relationships of Web page citations. HITS is not stable, only the largest community influences the ranking but HITS understands better relations among pages. As a result, [Diligenti2004] proposes a hybrid model called PageRank-HITS, which combines both of the algorithms.

Diligenti's probabilistic framework is also well suited for vertical ranking systems, which consider the contents of Web pages as well as the Web graph topology when assigning scores. Each page is represented by a set of keywords and it gets a relevance value by a classifier respecting the topic of interest. For instance, the page www.google.com is highly ranked by a general PageRank, but it would be little ranked by a PageRank focused on the *data mining* topic. From a couple of focused ranking algorithms "double focused PageRank" turned out to be the best.

II.2.4 Convergence Rate and the Effect of Factor *d*

The method represented by the dynamical system in (II.4) is the iterative Jacobi algorithm, which requires $O(m|E|)$ floating point operations. $|E|$ is the number of links in the Web graph, and m is the number of iterations. Bianchini and her colleagues show that m depends only on d and the relative error ε. In other words, the number of iterations in the system (II.4) needed to achieve a relatively stable vector of PageRanks depends neither on the size of the Web nor on its connectivity. They define the relative error at each time step as $\|\boldsymbol{e}_{\text{rel}}(t)\|_1 = \|\boldsymbol{x}^* - \boldsymbol{x}(t)\|_1 / \|\boldsymbol{x}^*\|_1$, and prove mathematically that in order to get the relative error under a certain threshold ε, it must be true that

$$t \geq \frac{\log((1-d)\varepsilon)}{\log d} \tag{II.17}$$

Number of steps

Thus, for $d = 0.85$ and $\varepsilon = 10^{-7}$, we get $t \geq 111$. That means, we need to iterate at least 111 times (i.e. m above will be 111) to get the error under the threshold. Changing d to 0.3 accelerates the computation to about 14 iterations whereas setting it on 0.99 yields 2 062 iterations! Now, when we make ε smaller, $\varepsilon = 10^{-15}$, t (and m) will have to be 3 895. Evidently a small ε slows down convergence, which is natural, but a small d speeds it up. The reported number of about fifty iterations by Brin and Page for their graph with 322 million nodes (see Section II.2.1) suggests that the authors contented themselves with $\varepsilon = 10^{-3}$. Let us recall that the number would have been the same if they had had a graph with just a million nodes. Apparently, the number of flops still relates to the number of edges in the graph, and, with regard to the immense size of the current Web, m greater than approximately one hundred would probably not be desirable even by top commercial search engines having enormous computing capacities.

So why not to set d very low if it speeds up convergence? In fact, $d = 0.85$, has been carefully chosen, and it is rarely set outside of the interval (0.8; 0.9). The fact is that the lower is d the less is respected the true link structure of the Web, because the term $(1 - d)$ represents the random jump of a Web surfer (see Section II.2.3). It is in accordance with intuition as well as with empirical observations that a random jump represents about one sixth of all transitions between Web pages. Thus, a low d overemphasizes random transitions between Web pages at the expense of existing links between them. What follows naturally is that different d's produce distinct rankings. Two orderings of Web pages may differ substantially. Also for this reason, the convergence criterion in practical applications is not the relative error of rank scores or a difference between two subsequent rank vectors, but the iterative process stops as soon as the ordering of pages does not change (much). Two subsequent orderings may be compared by means of some well-known metrics such as Kendall's tau or Spearman correlation coefficient.

Boundary values
What about the boundary values of *d*? If we set $d = 0$, all unnormalized PageRanks are equal to one. On the other hand, if $d = 1$ then (II.4) may not converge. Moreover Bianchini et al. show that when $d \to 1$, some nodes gain advantage over others, and the distribution of PageRank is biased towards them. They call these nodes *essential nodes* and we can recognize that a rank sink like that in Figure II.2 is always composed of one or more essential nodes. There is now way of escape from a rank sink except for a random jump outside of the loop of essential nodes. On the other hand, the stationary PageRank score x_p^* of any node *p* that is outside of rank sinks (an *inessential node*) is zero when $d \approx 1$. Thus, setting *d* higher or lower, we can also regulate how much rank we wish to confer to nodes in rank sinks. But *d* set to zero or one should be avoided, because the ranking scheme does not work anymore then.

II.2.5 PageRank for Publications

Although PageRank was originally conceived to help search engines rank Web pages by importance, it was clear that it could be applied to any graph-like structure, not only the Web graph. An evident application field is the network of bibliographic citations. We devote an entire Chapter III to the study of various social and information networks that invite the usage of PageRank-like methods. In this section, we will deal with the research carried out in [Sidiropoulos2005] that inspired our work described in Chapter V. The extent to which our work is different from the following is clarified in Section V.4.

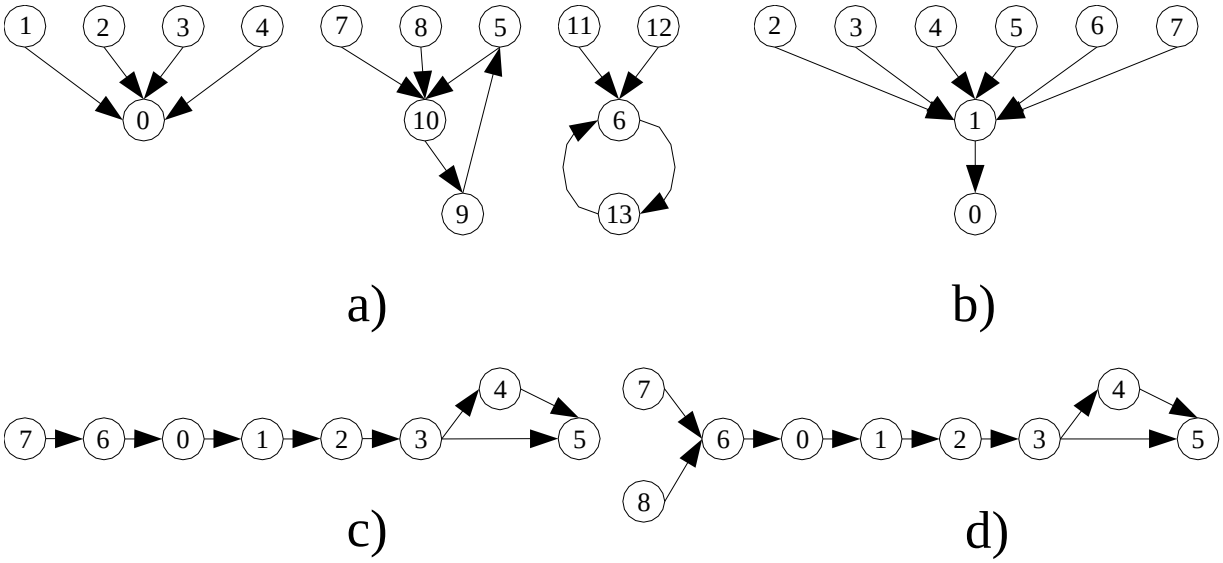


Figure II.5: Examples of graphs where standard PR does not work well [Sidiropoulos2005].

PageRank's drawbacks
Sidiropoulos and Manolopoulos are concerned with citation networks of scientific publications and find out that the standard PageRank metric is not appropriate for bibliographic measurements in some cases. More specifically, they show that in situations such as those depicted in Figure II.5 a modification of PageRank would be desirable. For example, in case a), nodes 10 and 6 are ranked higher than 0, because they are part of cycles. (In terms of the terminology we have learned in the previous sections, we would say that they are part of rank sinks or that they are essential.) However, in a graph where nodes are publications and edges are citations between them, a cycle is a kind of self-citation. Therefore,

we would rather node loops not to have much influence on rank distribution. Similarly, in case b), node 0 is ranked higher by PageRank than node 1 even though 1 has more "direct" citations which would be perceived as having more weight in bibliometrics. In the last examples c) and d), an additional node 8 points to 6, and this results in an increase of over 7% of PageRank of node 5. So even quite a distant change in the citation graph has some non-negligible influence on a particular node. Thus, the motivation for the following formula is to give more weight to direct citations and to make their impact smaller as the distance between the citing and the cited node gets larger:

$$R(u) = (1-d) + d \sum_{(v,u)\in E} \frac{R(v)+b}{D_{out}(v)} a^{-1} \qquad \text{(II.18)}$$

SCEAS Rank

In the above formula, R(u) is the rank score computed for node u (called SCEAS Rank) b is a factor enforcing direct citations, and a represents the speed with which an indirect citation impact converges to zero. Then the rank score of a cited node u is affected by a citing node v by the factor of a^{-k} when k other nodes lie between them. If b is zero and a is one, equation (II.18) is equivalent to the standard PageRank (II.2) except for the first term divided by the number of nodes which we explained earlier. If a ≥ 1 and b > 0 then one may even set d = 1 without convergence to zero unlike PageRank. Rankings produced by PageRank and SCEAS Rank for the graphs in Figure II.5 for d = 1, b = 1, and a = e (2.72) are shown in Table II.1. As e^{-7} is almost zero, a node citing from a distance of 7 and more has almost no impact on the cited node.

a)		**b)**		**c)**		**d)**	
PR	SCEAS	PR	SCEAS	PR	SCEAS	PR	SCEAS
6	0	0	1	5	5	5	5
13	6	1	0	3	3	3	6
10	10	2	2	2	2	2	0
9	13	3	3	1	1	1	1
5	9	4	4	4	0	0	2
0	5	5	5	0	6	4	3
1	1	6	6	6	4	6	4
2	2	7	7	7	7	7	7
3	3					8	8
4	4						
7	7						
8	8						
11	11						
12	12						

Table II.1: PageRank and SCEAS rankings for Figure II.5.

Authors of SCEAS experimentally prove that it converges twice as fast as PageRank. Moreover, they conduct a series of experiments with data from the DBLP digital library (see Section IV.1) and compare SCEAS rankings with several other ranking schemes including PageRank, HITS and a "baseline" ranking constituted of authors winning an ACM award. They show that their method is superior to the others. We adopted their comparison methodology to test our novel algorithm on real data in Chapter V.

II.2.6 Current Issues, Trends, and Areas of Future Research

Ranking algorithms have attracted much attention because of their evident practical usability, and the steady stream of new ideas, observations, modifications, and improvements does not seem to fade out. In spite of this, the theoretical properties of PageRank are still "only partially understood" [Bianchini2005]. The following topics appear to be the main research areas in this domain at present and will probably remain in the scientific mainstream in the next decade. We refer to some of the available literature only, please see [Langville2003] for further references.

- **Storage**. The vast dimensions of the Web transition matrix make its storage in main memory impossible. There are two principal approaches how to tackle this problem. First, we can compress the transition matrix, which is normally very sparse, store it in main memory, and then modify the iterative (also called *power*) method in (II.3) so that it could work with the compressed matrix. Second, we can store the transition matrix on disk in an efficient manner and then find out methods that allow for a timesaving access to this matrix. The aim is to minimize the time needed for I/O operations.

- **Convergence speed**. Although the number of steps required for the *power iterations method* to converge is not more than a hundred whatever the size of the Web graph (see Section II.2.4), the number of floating point operations involved in that computation may be enormous. Therefore, much research effort is devoted to finding out techniques aiming at speeding up the calculation. Basically, one can either try to reduce the number of iterations (mostly by tinkering with computation parameters or by relaxing the convergence criterion) or to reduce the number of operations in an iteration.

- **Stability and sensitivity**. There have been some contradictory research reports concerning the scale of change in the PageRank vector when the Web structure varies. Some authors [Ng2001a, Ng2001b] claim that PageRank ranking is stable and that it is not much affected by many poorly ranked pages that are modified. Nevertheless, some other researchers argue that it is mainly highly ranked pages which alter mostly and these modifications do have a great impact on the overall ranking. Finally, there are scientists who point out that rank stability should be observed rather than numerical stability.

- **Incremental computation**. The frequency of updating PageRank of Web pages should be high enough so as to reflect the dynamic nature of the Web. It is conforming to the time period between two consecutive crawls of the Web (see Section I.2), which amounts to weeks. The high computational costs of PageRank motivated endeavours to calculate it incrementally as the spider crawls the Web without needing to start over from scratch every time after a complete crawl [Desikan2005, Boldi2004a]. Alternatively, only the part of the Web that has changed since the previous crawl can be re-computed and then coherently incorporated in the overall ranking. Some techniques aim at "predicting" the Web structure [Yang2005].

- **Spamming**. There are estimates of millions of pages in the Web that have been created only for the purpose of *search engine spamming*. It means that they try to promote other pages or groups of pages in search engine rankings by linking to them. In other words, their goal is to unfairly increase the PageRank score of particular

pages. These unfair linkage patterns can be very complex and difficult to discover. Therefore, sophisticated algorithms [e.g. Wu2005] must be developed to help search engine ranking schemes (based on PageRank) evaluate Web pages fairly. An on-going clash of both parties – search engines vs. spammers – is to be expected in the years to come.

II.3 HITS

Web structure mining, one domain of Web mining, is concerned with exploring the topology of Web sites. The term topology is borrowed from graph theory and it means the structure of Web graph, in which the nodes are Web pages and the arcs are links pointing from one page to another. Obviously, it is a directed graph. Soon it was discovered that this structure could bear some information no less important than the actual contents of Web pages. For instance, researchers have noticed that Web pages can generally be divided into two categories: pages that link to many other pages and pages that are pointed to by many other pages. In fact, this behaviour resembles the human society when we think of Web pages as humans.

II.3.1 Authorities and Hubs

Gibson, Kleinberg and others [Gibson1998, Kleinberg1999b, Chakrabarti1998] explored the existence of *Web communities*. In doing so they introduced the notions of *authorities* and *hubs* and they developed a technique called HITS (Hyperlink-Induced Topic Search), which is based on them. The authors conducted a number of experiments with HITS or methods derived from HITS and they took a surprising conclusion that was in contrast to the common opinion that the World Wide Web was "becoming increasingly chaotic". A *hub* links to many pages, whereas an *authority* is linked to from many pages. Between these two entities there is a mutually reinforcing relationship – a good authority is linked to from many good hubs and a good hub links to many good authorities. A Web page can be a *hub* and *authority* at the same time. The following Figure II.6 shows an example of a Web community. The set S includes pages obtained with a query to a Web search engine, the extended set *T* contains, in addition, all the pages linking to the pages in the set *S* and all the pages that are linked to from the pages in the set *S*. The size of the set *S* is limited by choosing only a certain number of results from the search engine. We will denote such a graph as $G = (T, E)$ with E being the set of links as usual.

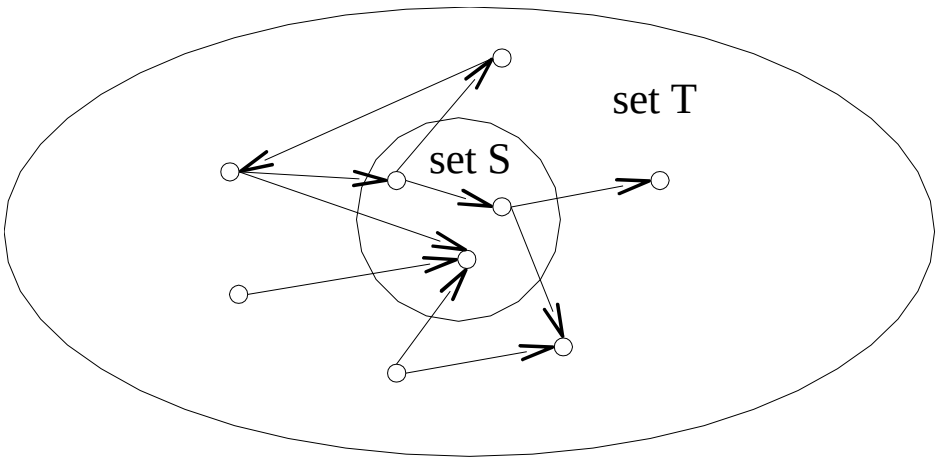


Figure II.6: Example of a Web community.

If we assign an *authority* weight $a(p)$ and a *hub* weight $h(p)$ to each page p, their values are computed as a sum of *hub* weights of the pages that link to it and as a sum of *authority* weights of the pages it links to, respectively. See equations (II.19) and (II.20).

$$a(p) = \sum_{(q,p)\in E} h(q), \tag{II.19}$$

$$h(p) = \sum_{(p,q)\in E} a(q). \tag{II.20}$$

At the beginning, we initialize the values of all $a(p)$ and $h(p)$ to 1. We update them according to the formulae (II.19) and (II.20) in each iteration. We proceed like this with all the pages and we normalize the weights after each iteration. The authors prove that this iterative process converges to stable sets of weights of *authorities* and *hubs*. Then say ten greatest authorities and ten greatest hubs can be denoted as the *core of a community*.

Previous thoughts can be expressed in the matrix notation. Again, let $\boldsymbol{A}$ be the adjacency matrix of a directed Web graph (similar to that in Figure II.6), where $A_{ij} = 1$ when the page i links to page j and 0 otherwise. Let $\boldsymbol{h}$ and $\boldsymbol{a}$ be vectors corresponding to the weights of hubs and authorities of all pages. We will repeatedly perform the following operations:

$$\boldsymbol{h} \leftarrow \boldsymbol{A}\,\boldsymbol{a}, \qquad \boldsymbol{a} \leftarrow \boldsymbol{A}^{\mathrm{T}}\,\boldsymbol{h}. \tag{II.21}$$

From the classical matrix theory it implies that with an appropriate renormalization $\boldsymbol{h}$ (respectively $\boldsymbol{a}$) converges to the principal eigenvector $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$ (respectively $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$). Kleinberg shows further that the *non-principal eigenvectors* of these matrices correspond to the authorities and hubs of the "non-principal" Web communities.

II.3.2 Extended Authorities and Hubs

We can also apply the idea of *hubs* and *authorities* to graphs with other kinds of nodes. In the previous case, each node is a Web page, but it could be a researcher, a research group or an institution as well. If a node is a researcher, the edges coming to this node are citations of this scientist (strictly said citations of his/her publications) made by other researchers (in their publications). On the other hand, the edges pointing from this node to the others are citations to other researchers. Of course, we might group researchers according to co-authorship, membership to institutions and so on. So the citations here are not meant to be references (links) from Web pages but directly from papers (publications). In the case of Web citations the cited entity can be easily recognized by its URL. It is more complicated with paper citations – it is necessary to find the references section in the paper, to retrieve the individual citations in it and possibly to determine the cited object. Here we must work with a certain ambiguity, because not all citations have the same format, not all authors are stated with their second given names and so on. See Chapter III for the concept of authorities in social networks and Chapter VI for case studies of finding authoritative institutions and researchers on the Web.

II.4 Summary

The idea of taking into account the link structure of the Web in order to rank pages by importance was a revolutionary step towards a unified view on methods seeking to detect authoritative sources in networked environments. An entirely new class of algorithms was born – ranking algorithms. Even though the first application area was the Web, these algorithms are suitable for any directed graphs. In this chapter, we concentrated mainly on the

best-known representative of ranking methods, PageRank, which is at the heart of the Google search engine. Although the most recent version of PageRank and its exact usage in combination with other techniques in Google is highly proprietary information, we attempted to summarize the state-of-the-art knowledge of this ranking method as understood in academia.

The contents of this chapter are based primarily on the original PageRank articles [Brin1998, Page1999], survey articles [Bianchini2005, Langville2003, Berkhin2005], the corresponding chapter in [Chakrabarti2002, pp. 209-216], and several research papers [Diligenti2004, Ding2001b, etc.]. We had to leave out those many modifications of PageRank due to space limitations such as SALSA [Lempel2000] – a stochastic method on the boundary between PageRank and HITS, TruRank [Vigna2005], which works also for $d \approx 1$, BackRank [Bouklit2005], which allows a random surfer to follow a backlink, or ObjectRank [Balmin2004], which is query dependant unlike PageRank. We just sketched out the features of SCEAS Rank [Sidiropoulos2005, Sidiropoulos2006] that was most relevant to our work, and for information on AuthorRank, another PageRank-based technique close to our research [Liu2005, Bollen2006], we refer to sections III.4.2 and V.4. As a final remark we would like to underline that some aspects related to ranking algorithms are also discussed in the context of social networks in Chapter III.

III Social Networks

The Web graph, about which we talk in Chapter I, is a network of pages connected via hyperlinks. Actually, every system that can be modeled as a graph is a network. The two expressions are synonymous, although mathematicians prefer speaking of graphs rather than networks, which is the terminology of social scientists. Wasserman and Faust [Wasserman1994] review social networks in detail and a very comprehensive overview of networked systems has been written by Newman [Newman2003]. Newman groups real world networks into social, information, technological, and biological networks. He considers citation networks and the World Wide Web (the structures we are interested in) as information networks, although the term "social" has been widely accepted and is often used in the context of citations or WWW, e.g. in [Chakrabarti2002, ch. 7] or [Liu2005].

The terms, algorithms, analyses, and models we discuss in the context of the Web are the results of mixing mathematical and social science approaches. If some decades ago social network models and theories were introduced which later had impact on the analysis of the Web (for instance bibliometric methods described in [Garfield1979] or [White1989]), now there are algorithms that have evolved in the Web environment and that, having been enriched with ingredients from the mathematical graph theory and numerical analysis, may be applied to original social networks again. Thus, webometrics influences bibliometrics. This is the case of ranking techniques covered in detail in Chapter II.

III.1 References and Citations

To avoid confusion between references and citations (which is sometimes the case even in the most accurate publications), we will strictly consider out-edges as references and in-edges as citations. Thus, articles (or authors) **refer to** other articles (authors) that are **cited by** them. A **citation by** *X* is an out-edge from *X*; a **reference to** *X* is an in-edge to *X*. A **citation** of *Y* **by** *X* is the same as a **reference to** *Y* from *X* meaning an edge from *X* to *Y*. So in the most strict sense, we should always use "refer" in the active form and "cite" in the passive form. Thus, the common phrase "author *X* cites author *Y*" would read only as "author *X* refers to author *Y*" or "author *Y* is cited by author *X*" in the most exact interpretation. However, the phrase "*X*

cites *Y*" in the sense of an edge from *X* to *Y* is so common in literaure that it is practically inevitable. Therefore, when appropriate we always indicate whether a relationship is an in-edge or out-edge throughout this thesis.

Definitions

We can gain valuable information from social networks when we have a look at *co-citations* and *co-references*. Let us recall what Ding et al. [Ding2001a, Ding2001b, Ding2002] say about them. Let $G = (V, E)$ be a directed graph of citations (set E) between publications (set V). Each edge (p_i, p_j) means that publication p_i cites publication p_j. Let $\boldsymbol{A}$ be the assymetric $N \times N$ adjacency matrix of such a citation network where N is the number of publications $|V|$, $A_{ij} = 1$ if p_i cites p_j and $A_{ij} = 0$ in other cases. The number of citations (or in-degree) of p_j is the sum of values in the j-th column of $\boldsymbol{A}$, i.e. $d_{in}(p_j) = \sum_{i=1}^{N} A_{ij}$ and the number of references (or out-degree) from p_i is the sum of values in the i-th row of $\boldsymbol{A}$, i.e. $d_{out}(p_i) = \sum_{j=1}^{N} A_{ij}$. Let $\boldsymbol{d}_{in} = [d_{in}(p_1), d_{in}(p_2),\ldots,d_{in}(p_n)]^{\mathrm{T}}$ be the vector of in-degrees of publications in V, $\boldsymbol{d}_{out} = [d_{out}(p_1), d_{out}(p_2),\ldots,d_{out}(p_n)]^{\mathrm{T}}$ be the vector of out-degrees, and $\boldsymbol{D}_{in} = diag(\boldsymbol{d}_{in})$ and $\boldsymbol{D}_{out} = \mathrm{diag}(\boldsymbol{d}_{out})$ be the corresponding diagonal matrices. The number of all interactions between publications is equal to the sum of all elements in $\boldsymbol{A}$: $|E| = \sum_{i=1}^{N}\sum_{j=1}^{N} A_{ij}$.

Co-citations

If publication p_1 references both p_2 and p_3 then p_2 and p_3 are co-cited by p_1. On the other hand, if both p_1 and p_2 reference p_3 then p_1 and p_2 co-reference p_3. See Figure III.1 for examples. Let us define the *co-citation index* C_{ij} of publications p_i and p_j as the number of other publications citing both p_i and p_j: $C_{ij} = \sum_{k=1}^{N} A_{ki}A_{kj} = (A^T A)_{ij}, i \neq j$ or by means of set notation $C_{ij} = |\{p_u \in V : (p_u, p_i) \in E, (p_u, p_j) \in E\}|$. The whole symmetric co-citation matrix will be denoted as $\boldsymbol{C}$. Although C_{ii} is not meaningful and is usually set to zero, $(A^T A)_{kk} = \sum_{j=1}^{N} A_{jk}A_{jk} = \sum_{j=1}^{N} A_{jk} = \boldsymbol{d}_{in}(p_k)$ is the in-degree of p_k. Thus, $diag(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}) = \boldsymbol{D}_{in}$. This results in an interesting relationship. The so-called authority matrix is a sum of the co-citation and in-degree matrices:

$$\boldsymbol{A}^T \boldsymbol{A} = \boldsymbol{D}_{in} + \boldsymbol{C} . \qquad \text{(III.1)}$$

The authority matrix $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ is the base for computing HITS authorities (compare with Section II.3.1) and it is surprising to see how close the authorities are to co-citations and in-degrees. Apparently, the higher the co-citation index, the more related are the co-cited publications. In fact, the co-citation index may be considered as a measure of similarity of two items [Small1973], and it can be utilized to cluster objects linking to each other such as publications [McCain1992], authors [Chen1999], or Web pages [Larson1996].

Co-references

It works similarly in the opposite direction of the citation. Two publications co-referencing some others (such as p_1 and p_2 on the right-hand side of Figure III.1) are likely to deal with the same topic. In bibliometrics, a co-reference is often referred to as bibliograhic coupling [Kessler1963]. The higher the number of co-referenced papers (*co-reference index*), the closer are the citing publications to one another. Let $\boldsymbol{R}$ be the co-reference matrix of papers in V. Then, the co-reference index R_{ij} of papers p_i and p_j is $R_{ij} = \sum_{k=1}^{N} A_{ik}A_{jk} = (AA^T)_{ij}, i \neq j$ or,

with set notation, $R_{ij} = |\{p_v \in V : (p_i, p_v) \in E, (p_j, p_v) \in E\}|$. Again, R_{ii} is set to zero, but $(AA^T)_{ii} = \sum_{k=1}^{N} A_{ik} A_{ik} = \sum_{k=1}^{N} A_{ik} = \boldsymbol{d}_{out}(p_i)$ is the out-degree of p_i. Thus, $diag(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}) = \boldsymbol{D}_{out}$, and we can express the hub matrix $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$ as the sum of co-reference and out-degree matrices:

$$\boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{D}_{out} + \boldsymbol{R}. \qquad \text{(III.2)}$$

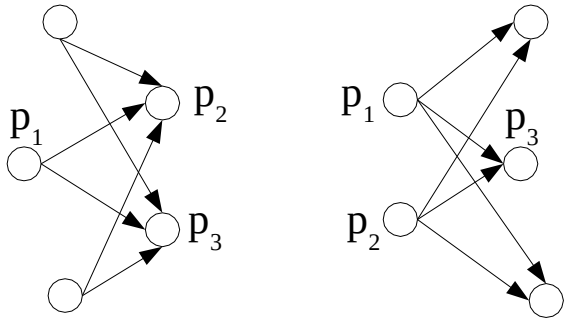


Figure III.1: p_2 and p_3 are co-cited by p_1 (left) and p_1 and p_2 co-reference p_3 (right).

Normalization

Ding [Ding2001a] further proves that the average co-citation index of p_i and p_j is $C_{ij} = \boldsymbol{d}_{in}(i)\boldsymbol{d}_{in}(j)/(N-1)$, and the average co-reference index is $R_{ij} = \boldsymbol{d}_{out}(i)\boldsymbol{d}_{out}(j)/(N-1)$ if the social network forms a fixed degree sequence random graph [Aiello2000]. (Compare with the Web graph model in Section I.1.) He also points out that both co-citation and co-reference indices should be normalized. We can explain this normalization looking at Figure III.2. Papers p_6 and p_7 (left) are co-cited by p_3, p_4, and p_6, but the co-citation by p_4 should weight less because p_1 and p_2 are also co-cited by p_4 in addition to p_6 and p_7. Actually, there are four out-links from p_4 and only two of them make the co-citation of p_6 and p_7. Thus, the co-citation by p_4 is 50% less valuable than that by p_3 or p_5 which have only two out-links each. This is called the normalization by out-links for co-citations, and it will be the normalization by in-links for co-references. Papers p_1 and p_2 (right) co-reference p_3, p_4, and p_5, but there are also p_6 and p_7 that co-reference p_4. Therefore, the contribution of p_4 to the co-reference index for p_1 and p_2 will be a half of the contribution made by p_3 or p_5.

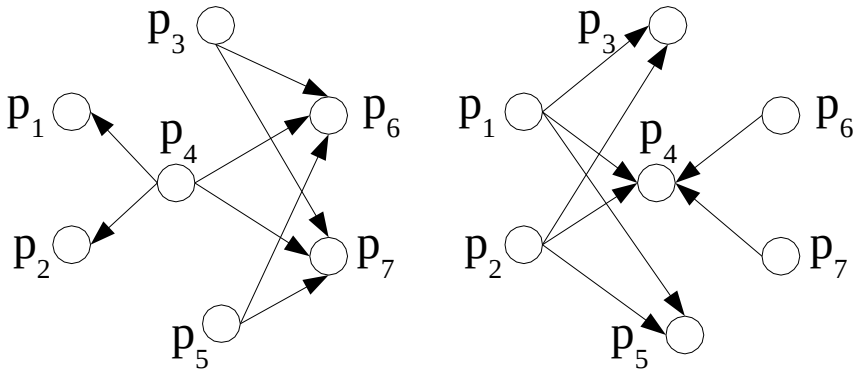


Figure III.2: Examples of co-citation and co-reference.

III.2 Popularity and Prestige

III.2.1 Popularity

Bollen et al. [Bollen2006] make a distinction between *popularity* and *prestige* when he discuss the significance of scientific journals. Before returning to this topic a little later, we incorporate these two terms into general social networks. In a social network like that in Section III.1, the popularity of a node *p* is clearly identified as the number of citations (in-links) or in-degree. Citations and in-degrees may sometimes not be the same quantitites (like in the author citation graph in Chapter V), and citations may then be considered as weights in an edge-weighted citation graph. To demonstrate the difference between in-degree and citations, we have to extend the graph $G = (V, E)$ from Section III.1 and associate weights with its edges. Let w_{ij} be the value assigned to edge $(i,j) \in E$. The popularity of node u is then

$$P(u) = \sum_{(v,u) \in E} w_{vu} \ . \tag{III.3}$$

Weighted and unweighted in-degrees

Obviously, if we set all weights *w* to one, the popularity of a node is its in-degree. If not all of the edges have a weight of one, the popularity is a *weighted in-degree*. Thus, counting citations for a particular author in a citation graph of authors, in which the weight w_{ij} means that author *j* is cited w-times by author *i*, is like calculating the weighed in-degree of that author. In Chapter V, we use "counting citations" and "calculating the weighted in-degree" synonymously. What is the relationship between the "normal" (unweighted) and weighted in-degree (citations)? Let us denote the weighted in-degree of node u as $P_w(u)$ and its unweighted in-degree as $P_u(u)$. If we suppose that w_{ij} is always greater or equal to one (as it should be in citation networks) then $P_w(u) \geq P_u(u)$. Consider the cases in Figure III.3.

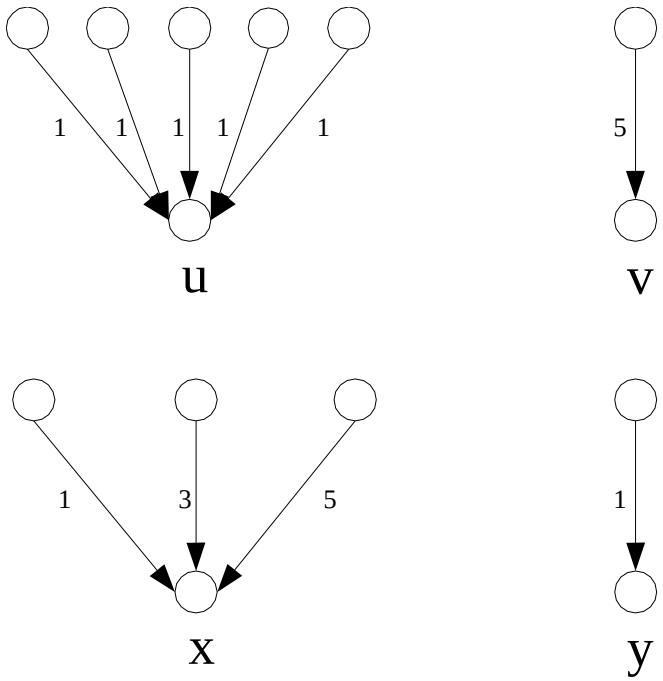


Figure III.3: Citations and in-degrees.

The weighted in-degrees of *u* and *v* are the same ($P_w(u) = P_w(v) = 5$) whereas the unweighted in-degrees are different ($P_u(u) = 5$, but $P_u(v) = 1$). We intuitively judge that *u* is more popular. On the other hand, nodes *v* and *y* have the same unweighted in-degree (1), but distinct weighted in-degrees (5 and 1). We will consider *v* as more popular, although we may be more reluctant here that in the previous case, because we associate the notion of poularity with the number of endorsing elements rather than with the strength of endorsment. The node *x* is more popular than both *v* and *y* in each aspect.

Until now, we could always decide upon popularity – the weighted and/or the unweighted in-degree was higher. But what if we are to compare more than two elements or if the in-degrees do not allow us to decide? For instance, the in-degrees of *u* and *x* are the following: $P_w(u) = 5$, $P_w(x) = 9$, $P_u(u) = 5$, and $P_u(x) = 3$. Can we normalize the in-degrees somehow? A short investigation shows us that we cannot. If we normalize citations by in-degree (i.e. P_w/P_u), we get 1 for *u*, 5 for *v*, 3 for *x*, and 1 for *y*. In case that we normalize in-degree by citations (i.e. P_u/P_w), we obtain 1 for *u*, 0.2 for *v*, 0.33 for *x*, and 1 for *y*. Among the many discrepancies we get, let us name just one. The "normalized" in-degree would be the same for *u* and *y* which is an obvious nonsense. Thus, the solution appears to be to have two separate rankings for the weighted in-degree (citations) as well as for the unweighted in-degree (like we do in Section V.3.3) and eventually to combine (average) the ranks from both rankings to produce a "universal" ranking like in [Sidiropoulos2006].

Balanced popularity

Finally, we will inspire ourselves by the genuine normalization of co-citations and co-references, and we will introduce the *balanced popularity* BP. Again, the balanced popularity presumes that the endorsment from a node linking to many other nodes is less significant than from a node having only few out-links. Sidiropoulos calls BP the balanced citation count and does not distinguish between weighted and unweighted citation counts, because he deals with publication citation graphs that normally do not have weights. We define the balanced popularity of node *u* as follows:

$$BP(u) = \sum_{(v,u)\in E} \frac{w_{vu}}{\sum_{(v,k)\in E} w_{vk}}. \qquad \text{(III.4)}$$

The denominator $\sum_{(v,k)\in E} w_{vk}$ is the out-degree of *v*. There may be an unweighted as well as a weighted variant of BP according as we set the weights *w*. The motivation behind the normalization by out-links is clear. For example, the citation of p_6 by p_3 should be twice as valuable as that by p_4, because p_4 has twice as many out-edges as p_3.

All the methods measuring popularity we have described so far are called *first-order* or *radius-1 methods*. The score obtained for a node depends only on its direct neighbours. In other words, nodes not sharing the same edge have no influence on each other. This is in contradiction with the real life as objects in social networks often have an indirect impact on one another. The next higher-order technique takes this into account.

III.2.2 Prestige

Prestige [Chakrabarti2002, p. 205] is defined recursively: the prestige score of a node depends on the prestige scores of nodes that point to the node, and their scores depend on nodes pointing to them and so forth. Thus, prestige Pr(u) of node *u* is computed as follows:

$$\Pr(u) = \sum_{(v,u)\in E} \Pr(v)\,. \tag{III.5}$$

Actually, prestige is authority in HITS (see Section II.3), and it can also be normalized by out-links which is, in turn, the base of PageRank described in detail in Section II.2. Prestige can be easily computed with *power iterations* like HITS with the omission of hubs. Let $\boldsymbol{p}(0) = [1,\ldots,1]^{\mathrm{T}}$ be the initial *prestige vector* of prestige scores of all nodes in the graph at time 0. We will then iteratively update the prestige vector, $\boldsymbol{p}(t+1) = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{p}(t)$, and L_1-normalize it (i.e. $\|\boldsymbol{p}(t)\|_1 = \sum_{i=1}^{N} p_i(t)$ should be one) after each iteration to avoid overflow. We will iterate until convergence of $\boldsymbol{p}$, which will give the final prestige scores of all nodes summing up to one. Of course, the greater the score, the higher the prestige. See the section on HITS for some notes on the eigenvectors of this system. Unfortunately, the prestige scores of nodes that are not in cycles and that are not even linked to by nodes in cycles converge to zero [Sidiropoulos2006].

This is not a problem in graphs with no cycles. (Theoretically, there should not be any loops in the graphs of citations between publications, although it is not impossible in practice.) But, of course, this fact is very annoying in the networks that are supposed to have many cycles such as citations between authors or the World Wide Web. The cycles in these types of networks are sometimes created deliberately so as to augment the prestige or any other recursive ranking score of theirs or of a particular node. See also the remarks on *spamming* in sections I.2.1 and II.2.6 and the paper on *link farm spams* by [Wu2005].

What about self-citations? We have not talked about them so far. There are no restrictions on the diagonal values of the node adjacency matrix $\boldsymbol{A}$. They can be 0 or 1 just like the other matrix elements. In fact, self-citations are very easily detectable small cycles. As we have just seen, loops may cause problems when applying recursive evaluation mechanisms. Therefore, we recommend to remove self-citing edges from the graph (like in Chapters V and VI) unless we are interested in some special graph properties.

III.3 Centrality

An alternative measure of importance of a node in a network is its *centrality*. We briefly mention three centrality metrics – *radius*, *closeness*, and *betweenness*. The radius $r(u)$ of node u is equal to $\max_v d(u,v)$ where $d(u,v)$ is the distance from u to (another) node v in graph G. Distance is the length of the shortest path from one node to another. Thus, radius is the distance to the most distant node in the graph. The node with the smallest radius is called the center of the graph. Evidently, nodes with a small radius have more influence in the network than nodes with a large radius. Closeness is somewhat similar to the radius, but it is represented by many distances to all other nodes instead of a single quantity. A central node should be close to any other node in the graph.

Betweenness is the number of times a particular node lies on the shortest path between any two nodes in the graph considering each possible pair of nodes. The node with a high betweenness centrality controls the information flow between other vertices. If we remove such a node, a large number of shortest paths get longer or the graph even breaks up into (more) components. A clear drawback of betweenness is its computational time complexity $O(n)$ with n as the number of nodes, which makes it impractical for large graphs. It is reasonable to calculate the radius and closeness only for the largest graph component. On the

other hand, we can compute the betweenness centrality in a graph with components, but it means examining many shortest paths "in vain". The high time complexity of betweenness would then suggest to use it only for the largest component, although this usually does not help much, because real-world networks often have a very big largest component either (see Section IV.1).

All centrality metrics can be adapted for directed, undirected, weighted as well as unweighted graphs. A single measure is not appropriate in all situations. Therefore, several techniques of importance evaluation should always be employed in parallel. Liu [Liu2005] uses prestige and centrality measures in the analysis of a special kind of social networks, which we will introduce in the next section.

III.4 Co-authorship networks

Co-authorship networks are a special case of social networks, in which the nodes are authors and edges mean collaboration between authors. Unlike the citation networks discussed in the previous sections, in which each edge is endorsement, recognition, acknowledgement, or express of debt, an edge between two authors in a co-authorship graph implies that those two authors have been or still are colleagues. They have co-authored one or more publications as a result of their collaboration and common research effort lasting for years or even decades. This is in contrast to citations, where the citing author often does not know the cited author in person, and they may be divided by a time span of up to centuries. In general, collaboration is much stronger tie than a citation: authors know each other personally. (Let alone the "fake" co-authorships that occur from time to time.)

Growing interest

In recent years, many research papers have appeared that deal with the analysis of co-authorship networks [e.g. Nascimento2003, Wagner2003, Smeaton2003, Farkas2002, Otte2002, He2002, Cunningham1997]. It is in relation to the emergence of a large number of electronic sources of bibliographic data. We cover some of them in Chapter IV. The analysis of co-authorship networks instead of citation networks is advantageous in that there is more reliable data to analyze. While citation indexing requires much manual labour and even if fully or partly automatized it is prone to errors, creating a co-authorship graph is by far not so demanding.

III.4.1 Network Types

Liu et al [Liu2005] enumerate three possible representations of co-authorship (collaboration) graphs – undirected unweighted (also called binary or, more generally, unit-weighted) graph, directed unweighted graph, and directed weighted graph. Let us have a paper p_1 co-published by authors a_1, a_2, and a_3 and a paper p_2 co-authored by a_1 and a_2. The three representations of this co-authorship network are visualized in Figure III.4 – as an undirected unweighted graph (left), directed unweighted graph (middle), and a directed edge-weighted graph (right) the weights of which will be explained bellow.

Can we measure prestige?

The undirected unweighted graph is the simplest form from which we can, however, gain all centrality information for each node. To be able to measure prestige and popularity, we must turn it into a directional network. So each original undirected edge is transformed into two inversely directed edges so that the relationship between the nodes sharing the original edge is symmetric. Let us recall that the endorsement (directed edge) is not a citation but a collaboration. Does it make sense to measure popularity or prestige on the basis of

collaborations and not on citations? Because of the lack of free availability of reliable citation data some researchers take advantage of directed collaboration networks and then try to identify authoritative authors as if it was a classical network of citations between authors. But is a researcher who has many collaborators more prominent than another scientist who has just a few colleagues? While citation indexing and analysis is an established means of determining significant sources, the analysis of co-authorship networks with view of finding authoritative researchers is not yet mature. A popular scientist in the collaboration network may be authoritative in some sense, but such authority will probably be different from the authority based on citations.

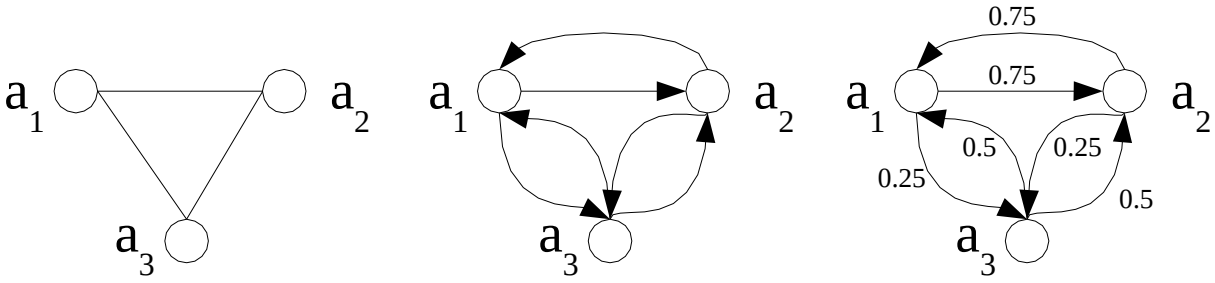


Figure III.4: Graph representations of a co-authorship network.

And, last but not least, authors who often publish their papers alone (without co-authors) are strongly handicapped if not entirely discarded in a collaboration authority ranking. The number of publications written by one author may be quite large. For example, in the DBLP data set we work with Section V.3, 149 031 papers of the total of 472 043 are single-author papers which accounts for 31.6 %. The publications examined by Liu et al. from the digital library research community were single-author articles in 19.6 %. Petříček [Petříček2005] reports about 10% single-author papers for DBLP and CiteSeer. The practical meaning of collaboration-based authorities is yet to be submitted to further research.

III.4.2 Weighted Networks

If we admit the existence of collaboration authorities, we might want the endorsements in a directed collaboration network not to weigh equally. In order to be able to assign weights to edges, we need some additional, explicit as well as implicit knowledge not included in the simple co-authorship graph. For example, co-authors of a paper jointly published by two authors are surely more connected than co-authors of a paper written by fifteen researchers. Also, collaboration between frequent collaborators is likely to be more intensive than between occasional co-authors and, therefore, the collaboration link should weigh more.

Frequency and exclusivity

Based on the ideas above, Liu defines two factors that will further determine weights in the directed edge-weighted collaboration graph – co-authorship *frequency* and *exclusivity*. The motivation behind introducing these two factors is to give more weight to collaboration links that connect authors who often co-publish together with a minimum number of other authors involved. Thus, the resulting graph is then $G = (V, E, W)$ where $V = \{a_1,\ldots,a_N\}$ is the set of authors as nodes, E is the set of co-authorship links between authors as edges, and W is the set of weights w_{ij} assigned to each edge (a_i, a_j). We also need a set of publications $P = \{p_1,\ldots,p_m\}$, because Liu does not consider a bipartite publication-author graph like we do in Section V.1. The weight w_{ij} is computed as follows:

$$w_{ij} = \frac{c_{ij}}{\sum_{k=1}^{N} c_{ik}} \tag{III.6}$$

where c_{ij} is the co-authorship frequency of authors a_i and a_j computed like this:

$$c_{ij} = \sum_{k=1}^{m} g_{i,j,k} \tag{III.7}$$

and $g_{i,j,k}$ is the co-authorship exclusivity of authors a_i and a_j ($a_i \neq a_j$) in publication p_k defined in the following way:

$$g_{i,j,k} = 1/(f(p_k) - 1) \tag{III.8}$$

where $f(p_k)$ is the number of authors of publication p_k.

The weights in (III.6) are normalized so that the sum of weights on edges emanating from a node is one. This is necessary, because a recursive PageRank-like algorithm will then analyze the graph and compute prestige, which requires this property for convergence. It also enables to interpret the weights as transition probabilities of a random walker on the graph that would not be allowed to randomly jump to an arbitrary node. An example of weight calculation is shown in Table III.1.

paper 1: {a_1, a_2, a_3}; paper 2: {a_1, a_2}		
Exclusivity		
paper	authors	result
p_1	a_1,a_2	0.5
p_1	a_1,a_3	0.5
p_1	a_2,a_3	0.5
p_2	a_1,a_2	1.0
Frequency		
authors	calculation	result
a_1,a_2	0.5 + 1	1.5
a_1,a_3	0.5	0.5
a_2,a_3	0.5	0.5
Weight		
edge	calculation	result
(a_1,a_2)	1.5 / (1.5 + 0.5)	0.75
(a_2,a_1)	1.5 / (1.5 + 0.5)	0.75
(a_1,a_3)	0.5 / (1.5 + 0.5)	0.25
(a_3,a_1)	0.5 / (0.5 + 0.5)	0.50
(a_2,a_3)	0.5 / (1.5 + 0.5)	0.25
(a_3,a_2)	0.5 / (0.5 + 0.5)	0.50

Table III.1: Weight calculation for graph in Figure III.4.

Context in our work
Liu's weights strengthen relationships between authors who often and exclusively co-publish. In Chapter V, we introduce similar factors into a citation graph. However, these factors affect the weights inversely, because a citation between two authors who frequently write articles together should weigh less than that between "foreign" authors. The advantage of considering these factors in a citation graph is that the detection of authoritative sources from a graph of citations is a fully recognized method whereas finding authorities in a collaboration network is still in experimental stages and has the drawbacks we discusses earlier. In addition, the directed weighted co-authorship graph representation of Liu et al. is not self-contained. To compute weights, we need additional information (such as a table of publications and their authors) that is nowhere to be found in the graph. On the other hand, our graph representation in Section V.1 is self-descriptive. We generalize and extend the notion of co-authorship frequency and exclusivity and employ them in a weighted citation network to identify "more fairly" authoritative researchers.

III.5 Scientometrics

The social networks we present in this chapter (co-authorship and citation networks) are often studied in bibliometrics – a scientific domain seeking to discover interesting publication patterns. Because Web pages can be regarded as publications, techniques and methodologies from bibliometrics have been widely adopted in webometrics to "measure" Web pages. Interestingly, some methods, such as prestige computation, have gone the opposite way – although originally developed for Web pages, they are now being applied to standard printed publications. If these publications are of scientific nature (journal articles, conference papers, technical reports, dissertations, technical books, patent documentations etc.), we can talk about scientometrics.
Scientometrics generally tries to measure research performance of individuals and groups of individuals (such as institutions or even countries) on the basis of the number of "generated" publications or patents and their *impact* on the research community. One can think of impact as a kind of popularity or prestige. The significance of scientometrics has been growing since research-funding bodies need some objective and quantifiable information to justify their funding policies. Their objective is to support high quality research and to limit aids to unproductive individuals or institutions. In this section, we will describe two metrics. One, quite well established although sometimes contested, for measuring impact of scientific journals and another one, rather new, for calculating research performance of individuals.

III.6 Impact Factor

The *journal impact factor* (IF) was first presented by Eugene Garfield in 1955 and reviewed many times in his subsequent publications [e.g. in Garfield1979, Garfield1999]. Nowadays, it plays a key role in the annual "Journals Citation Reports" issued annually by Thomson Scientific (see Section IV.3.3). In the light of the explanations above, we can regard it as a first-order popularity metric computed by normalizing weighted in-degrees. In fact, the impact factor of a journal in a given year is the average number of citations an article published in the journal in the two preceding years obtains from journal articles published in the current year.

We can group publications p from Section III.1 according to journals in which they appear and thus get a set of new nodes $V^{IF} = \{v_1,\ldots,v_N\}$ representing journals where $v_i = \{p_1,\ldots,p_k\}$. The original edges $(p_i,p_j)\in E$ for citations between publications will be grouped similarly so as to form new edges between journals $(v_p,v_q)\in E^{IF}$ if and only if there exists $(p_i,p_j)\in E$ such that $p_i\in v_p$ and $p_j\in v_q$. The weight $\theta_{pq}\in W^{IF}$ of $(v_p,v_q)\in E^{IF}$ will be the number of edges directed

from v_p to v_q, $\theta_{pq} = |\{(p_i,p_j) \in E: p_i \in v_p \wedge p_j \in v_q\}|$. Now that we have the journal citation graph $G^{IF} = (V^{IF}, E^{IF}, W^{IF})$, we further define $\mathrm{IF}(v_i,t)$ as the impact factor of journal v_i in year t, $c(v_j,v_i,t)$ as the number of citations to articles published in journal v_i in years $t-1$ and $t-2$ by articles published in journal v_j in year t, and $s(v_i)$ as the number of articles published in journal v_i in years $t-1$ and $t-2$. The impact factor of journal v_i in year t is then

$$\mathrm{IF}(v_i,t) = \frac{\sum_{j=1}^{N} c(v_j,v_i,t)}{s(v_i)}. \tag{III.9}$$

Criticism
Many objections to the calculation of IF may arise. First, only citations by articles in indexed journals are considered. What journals should be indexed? The selection of journals may immensely affect the IF computed. In addition, papers from conference proceedings are completely disregarded (i.e. citations made by them are never counted) which is an obvious problem in such a rapidly evolving field like computer science, in which some conferences are much more prestigious than journals. Second, the definition of IF admits self-citations which may result in a strong bias towards frequently self-cited journals. Third, why has the time delay of two years been chosen? Two years might be inconvenient for some research domains. And, last but not least, because of its popularity-based foundation, the impact factor measures quantity rather than quality. Should it not compute prestige instead?

This and other criticism has been expressed in numerous publications [Harter1997, Seglen1997, Nederhof2001, Bordons2002, Lewison2002, and Saha2003 among others]. Bollen [Bollen2006] proposes to replace IF with a weighted PageRank, in which weights are given as above in the journal citation graph, to determine prestige (*status* in his words) rather than popularity of journals. He conducts an interesting study on the data from "2003 Journal Citation Reports" and identifies journals for which the popularity and prestige ranks significantly differ.

III.7 Index H

The index *H* (also called *h*-index, *h*-score or Hirsch-Index) is a simple metric of research performance proposed by J. E. Hirsch in 2005 [Hirsch2005]. A researcher has a score *h* if *h* papers by him have at least *h* citations (in-links) each and the other papers have at most *h* citations each. For instance, a scientist with $h = 30$ has thirty publications each of which has been cited thirty times at least. The calculation of *h* is quite simple – we just sort publications of a particular researcher by citations descendingly and denote them with 1, 2, 3, etc. We then start from 1 and proceed until we found a publication number *g* that is larger than the citation count of that publication. The *h* index is *g* minus one. Clearly, there may be publications with the same citation count *h* that do not contribute to the *h*-index because they lie on the h+1st, h+2nd, etc. position.

Properties
Hirsch finds two interesting relations. The first one is a very rough estimate of the number of citations with regard to *h*:

$$T = ah^2 \tag{III.10}$$

where T is the total number of citations to publications of a scientist and a is a coefficient between 3 and 5 determined empirically. In the following equation, n is the number of years in service of a researcher usually measured from the year of the first publication and m is the slope of h versus n.

$$h \approx mn\,. \tag{III.11}$$

Obviously, if a researcher has $h = 20$ after twenty years of research or $h = 40$ after forty years of scholarly work, m is 1. On the other hand, both a researcher with $h = 20$ after ten years and a scholar with $h = 60$ after thirty years of service have the same $m = 2$ and their research output may be considered as comparable. The parameter m allows for evaluating researchers at different levels of seniority. Hirsch concludes that a scientist (in physics) with m close to one is successful and a scientist with m about three is an outstanding individual.

The index H has some significant advantages over traditional scientometric techniques. It is a single number (compared to citation counts of the most highly cited papers), it does not prefer quantity to quality (compared to the number of publications), and it acknowledges a systematic long-term work rather than a few frequently cited research results. On the other hand, it has some inconveniences. Similarly to citations, it cannot be compared across different scientific fields and subfields because of distinct citation patterns. For example, the top h-scores in physics and computer science are about 70, whereas their counterparts in medical and biological sciences can be twice as high. Also, the h-index of a scientist who stops publishing can never increase in spite of his publications being cited.

Bibliographic notes

The h-index is very new and is subject to some controversy and amendment proposals. Meanwhile, it has been suggested for journal evaluation [Braun2006] and compared with standard bibliometric measures and peer review judgements [Raan2006, Bornmann2005]. Bornmann and Daniel [Bornmann2007] summarize the state-of-the-art knowledge about the index h. However, its simplicity and availability makes it suitable for comparisons with other scientometric or bibliometric rankings such as those presented in Section V.3.3. A list of computer science researchers with the largest h-index is available at [24]. The index h can be retrieved automatically from Google Scholar (see Section IV.3.2) – a script for this purpose sorting an author's publications by citation counts may be found at [26] and a user interface at [25] or [34].

III.8 Summary

In this chapter, we have discussed social networks and we have shown how closely bibliometrics, webometrics, and scientometrics are related to each other via applications of social networks. Therefore, some parts of this chapter are on the boundary and could be placed in Chapter I on ranking algorithms for Web sites. We have presented several well-known metrics for evaluating nodes in a social network. In particular, we point out Section III.4.2 on weighted co-authorship networks. The bibliographic information used here to calculate edge weights in a collaboration graph is extended and newly adopted for citation graphs in Chapter V.

IV Web Systems for Researchers Support

In this chapter we will describe several on-line systems that may serve for the support of researchers by providing bibliographic information, citation indices, search services, or repositories of scientific publications. Basically, we can group these systems into free and charged applications, and into manually maintained and automated ones as is shown in Table IV.1. We will take a more datailed look at two prominent representatives – DBLP for the part of man-made services and CiteSeer from the group of computerized systems. The other services (Google Scholar,Rexa, ISI Web of Science, and ACM Portal) will be mentioned briefly. The importance of this chapter for the core of the thesis (detection of authoritative sources) is in that it introduces already existing services that allow for searching for authorities or provide data that might be used for the search itself or for the verification of search results in the field of (computer) science scholarly publications

IV.1 DBLP

When not stated otherwise, the information in this section comes from [Ley2006] and [6]. The DBLP digital library is a collection of bibliographic data from the field of computer science. It is manually maintained and freely available at [5]. As of April 2007, it contains over 870 000 bibliographic records. Its history reaches as far back as 1993, shortly after the appearance of the World Wide Web. Although more specific at the beginning, it gradually began to cover the whole domain of computer science and can be read as Digital Bibliography & Library Project now.

Features

DBLP is updated regularly and some 110 thousand new records are added each year. However, this is far below the number of new computer science papers that appear. It has been estimated that only about one fourth of new papers is input into the DBLP [Petříček2005]. The authors of DBLP, a small research group at the University of Trier in Germany, admit that the selection of papers for the digital library is more or less random. Nevertheless, the "all or nothing policy" is applied – whenever a journal issue or a conference proceedings book is selected, all of its papers are input. Some basic funding enables to hire

students to enter bibliographic data. The justify the huge manual effort needed with respect to fully autonomous systems such as CiteSeer, some benefits of this approach must prevail. Therefore, a great care is taken when entering author names. The objective is to unify the spelling of names and to disambiguate authors with the same names. Both manual and automated techniques are employed in this process and, actually, this takes most of the time of entering new data. As a result, the information on publications by an individual author in DBLP is quite reliable. Moreover, even diacritical symbols may be used by means of special character codes. A feature that often does not work well in other on-line bibliographic systems.

The most significant property is the availability of all the data in an XML file [8]. This allows for numerous bibliographic studies based on DBLP such as [Sidiropoulos2005], [Sidiropoulos2005b], [Sidiropoulos2006], [Bani-Ahmad2005], [Cai2005], [Liu2005], [Rahm2005], [Mohan2005], [Elmacioglu2005], [Nascimento2003], [Hassan2004] to name a few. DBLP has clearly established itself as a provider of high quality data for data mining methods. We have respected its outstanding role in this context, and we base our experiments in Section V.3 on DBLP. Besides the regularly updated data file, there is also a "preserved" file which enables different researchers to conduct experiments with the same data and to compare their results. Unfortunately, researchers often neglect this option.

Citations

Unfortunately, only a very small part of DBLP publications contains references to other papers. It is only a few percent. See [Sidiropoulos2005] for a list of conferences and journals whose papers include references. These are primarily papers that are part of the so-called "SIGMOD Anthology" [9]. In fact, about eighty percent of those 100 000 citation links in DBLP are citations made by SIGMOD Anthology publications. The anthology consists of articles the full-text versions of which are digitized and distributed on CDs (DVDs) for a fee. There are over 14 000 PDF files (as of January 2006). For some of the papers in the anthology the reverse "cited by" information was also added. Interestingly, this information is sometimes not disclosed by the DBLP Web interface and/or by the off-line browser [7] even if it evidently appears in the XML file. It is unclear why. For instance, for [Brin98] no citations (in-links) are shown on the Web site whereas the off-line browser finds eight citations in the September 2006 data file. It cannot be explained by different data files because the off-line file can never be more recent that the on-line data.

Also, it is not evident how to obtain the list of citing publications for a particular paper. Among others, to get the total citation count like in [10]. In fact, it is necessary to go from the anthology page or the corresponding BibTeX page over to the "citation" page via the <ee> element. For instance, from [11] or [12] to [13]. This "citation" page does not exist for some cited papers, however. The safest way to find all existing references in DBLP to a paper seems to look for that paper's key (ID) directly in the <cite> elements of the DBLP XML file. In general, citation analysis based on DBLP is still rather limited when compared to the extensive usage of its co-authorship graph. Therefore, some researchers even add directed edges into the co-authorship graph and consider it as a citation graph [Liu2005].

Statistics

The co-authorship graph has more 440 000 nodes in January 2006. There is one big component with over 330 000 authors whereas the second largest component has only 37 nodes! For more than 8 000 authors there exists a link to their personal homepage. We take advantage of this feature in Section VI.2.2 where we must decide on the "nationality" of a

researcher. For each publication record there is a BibTeX entry at least. For some publications a link to their on-line version is provided. Papers from the SIGMOD Anthology have their "local" links to a specific file on the CD/DVD included. Regarding the technology behind DBLP, surprisingly there is no underlying database [14]. Scripts and programs parsing data files and searching it in the main memory provide search results. This is true for both the on-line and off-line version. Therefore, before applying data mining methods to the DBLP data it is often needed to transfer it (or a portion of it) into a relational database.

We do not want to describe the internals of the XML data provided by DBLP, but we will rather terminate this section with a look at Figure IV.1 adopted from [6]. As we can see, the vast majority of publications in DBLP are either conference papers (inproceedings) or journal articles. The other publication types are negligible. That is why we analyze only papers and articles in Section V.3.

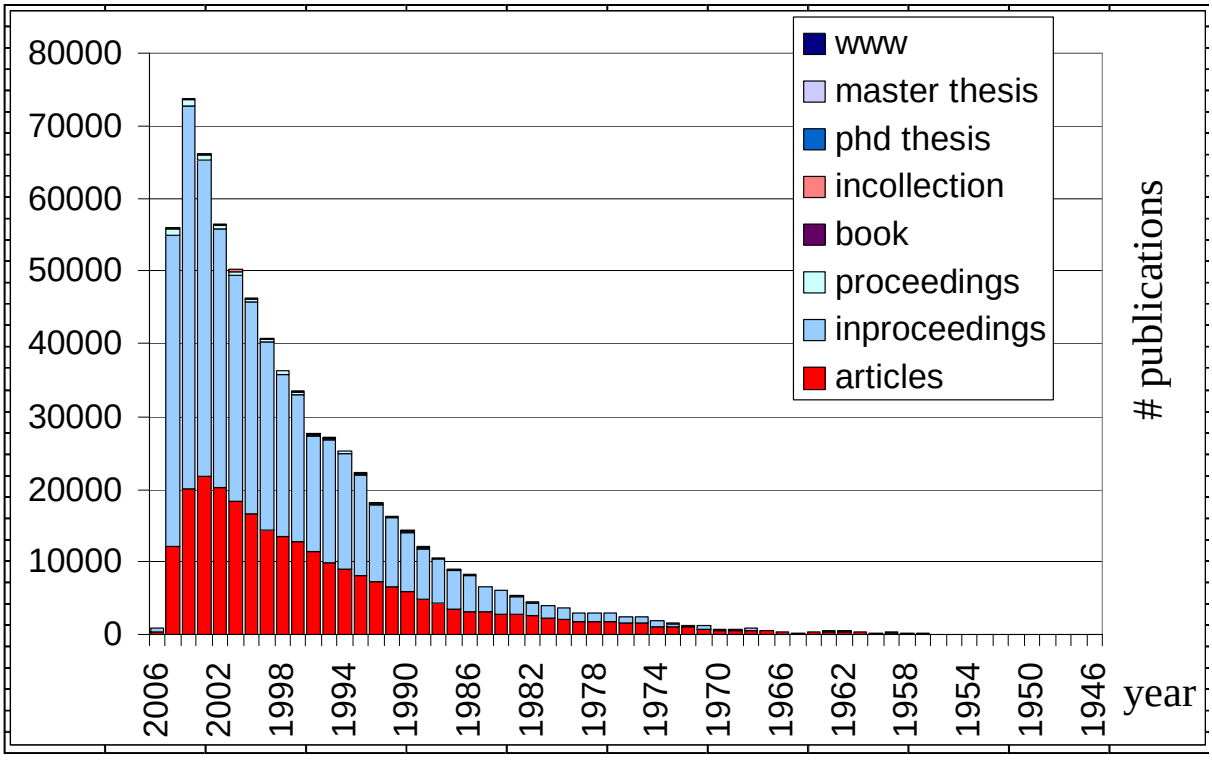


Figure IV.1: Distribution of various publication types and years in DBLP on 12 Jan 2006.

IV.2 CiteSeer

Services [15] as well as the full source code of CiteSeer (at the beginning called CiteSeer, then ResearchIndex, now CiteSeer again) are freely available. CiteSeer uses search engines (with queries "publications", "papers", "postscript", etc.) and crawling to efficiently locate papers on the Web. Start points for crawling may also be submitted by users who would like to have their documents indexed. It may take a few weeks after submitting to happen so. Its database is continuously updated 24 hours a day. Unlike DBLP or ISI Web of Science, the digital library and also its citation index (which is quite limited in DBLP as we mention in Section IV.1) are constructed in a fully automated way – no manual effort is needed. In April 2007, more than 760 000 documents are indexed.

Features
Operating completely autonomously, CiteSeer works by downloading papers from the Web and converting them to text. It then parses the papers to extract the citations and the context in which the citations are made in the body of the paper, storing this information in a database. CiteSeer includes full-text article and citation indexing, and allows the location of papers by keyword search or citation links. It can also locate papers related to a given article by using common citation information or word similarity. Given a particular paper, CiteSeer can also display the context of how subsequent publications cite that paper [Lawrence1999].

CiteSeer downloads Postscript or PDF files, which are then converted into text using PreScript from the New Zealand Digital Library project [16] explained in [Nevill-Manning1998]. It checks that the document is a research document by testing for the existence of a reference or bibliography section.

Once CiteSeer has a document in a usable form, it must locate the section containing the reference list, either by identifying the section header or the citation list itself. It then extracts individual citations, delineating individual citations by citation identifiers, vertical spacing, or indentation. CiteSeer parses each citation using heuristics to extract fields such as title, author, year of publication, page numbers, and the citation identifier. (Compare with the techniques of information extraction with hidden Markov models in [Seymore1999].) CiteSeer also uses databases of author names, journal names, and so forth to help identify citation subfields.

Internals
Citations to a given article may have widely varying formats. Much of the significance of CiteSeer derives from the ability to recognize that all of these citations might refer to the same article. Also, CiteSeer uses font and spacing information to identify the title and author of documents being indexed. Identifying the indexed documents allows analyzing the graph formed by citation links, which results in abundant citation statistics.

Several classes of methods for identifying and grouping citations to identical articles are applied [Lawrence1999]:

- String distance measurements, which consider distance as the difference between strings of symbols.
- Word frequency measurements, which are based on the statistics of words that are common to each string (TFIDF – Term Frequency vs. Inverse Document Frequency, common in information retrieval, see [Chakrabarti2002, p. 57]).
- Knowledge about subfields or the structure of the data.
- Probabilistic models, which use known bibliographic information to identify subfields of citations.

Furthermore, algorithms for finding related articles are used:

- Word vectors, a TFIDF scheme used to locate articles with similar words.
- Distance comparison of the article headers, used to find similar headers.
- Common Citation vs. Inverse Document Frequency (CCIDF), which finds articles with similar citations.

Data
Besides links to the original documents on the Web, which have been downloaded and processed, and document copies in its repository, CiteSeer provides links to corresponding document pages in DBLP and on the ACM Portal as well. Whether these DBLP and ACM references have been added manually is unclear. Thanks to its automated citation indexing, CiteSeer is able to publish lists of the most cited researchers on a regular basis [17]. However, the computer-generated citation rankings sustain the same problems we talk about in Section VI.2.1 – incorrectly recognized author names, ambigous names, difficulties with diacritics, etc. It is also possible to get all CiteSeer's bibliographic data in an XML-like format including references and author affiliations for some publications or just BibTeX records with basic information [18]. Suprisingly, CiteSeer's data are not widely used in scientometric research – an exception is [An2004].

IV.3 Other Systems

IV.3.1 Rexa

Rexa [19] is a service similar to CiteSeer that has evolved prom a project called Cora. It is newer and it has only about a half of the number of documents in its database (about 380 000 in April 2007). However, it enables to search bibliographic references to a total of seven million research papers, and its user interface appears to be more comfortable and intuitive. Documents are not added continuously to the digital library, though. The last extensive Web crawl was performed in 2005 and the next one is in preparation. Rexa is based on the work of Andrew McCallum and his colleagues, and the technology behind it is relatively well documented compared to CiteSeer [Seymore1999, McCallum1999a, McCallum1999b, McCallum2000 and others]. Emphasis is put on extracting information from scientific publications with machine learning techniques and on creating networks of linked objects – papers, authors, institutions (in development), etc. In particular, this enables a quick retrieval of the citation count of an individual author. For a more detailed overview of the information extraction methods employed, see [Seymore1999].

IV.3.2 Google Scholar

Google Scholar is a service provided by the Google Web search engine [20] in which it is also seamlessly integrated. It is a powerful tool for searching for bibliographic information operating entirely autonomously. Scientific documents are collected from all over the Web, indexed, and made available to the public via an effective user interface resembling that of its general search engine. It is not disclosed how many documents there are in the database, but we can learn on the Web site of Rexa that Rexa's size is approximately one fourth of the computer science papers indexed by Google Scholar. That would mean some 1.6 million documents let alone the other six scientific domains covered by Scholar and references to articles that have actually not been indexed, but whose bibliographic information is known (analogous to *dangling pages* from Section II.2.2). It is certainly the most comprehensive on-line repository of scientific bibliographic information at present.

On the other hand, Google Scholar may be considered less as a digital library that the other systems above. An open access to the cached versions of papers downloaded remains limited. Quite often only links to an abstract on a publisher's Web site are supplied, which then requires a subscription to get access to the full text of the article. Of course, articles on non-login-protected Web sites are still accessible like standard documents found by a search engine. Therefore, Scholar groups similar documents and usually offers free versions of a charged document retrieved somewhere else on the Web. (For instance, a preprint of a journal

article downloadable from the author's home page.) Moreover, there is a possibility to search library catalogues for print versions of papers. Scholar shows citation counts for individual publications, though not for individual researchers. The latter must be done manually or via scripts that communicate directly with Scholar's Web interface as is shown at [26] and mentioned in Section III.7 in the context of counting the *H* score.

IV.3.3 Web of Science

Thomson Scientific Web of Science (formerly ISI Web of Science [27] – ISI stands for Institute for Scientific Information) enables users to search a database consisting primarily of papers from about 8 700 research journals (5 900 journals with 10.8 million searchable articles for the "Science" domain which is still a superset of computer science). In addition to journals, specific Web sites are also included in the database. See [Testa1998], [28], and [29] for information on how the journals and Web sites are selected. The database covers 1978 to date, but only the 1991+ portion has English language abstracts. This amounts to approximately 70 % of the articles in the database. There are weekly updates, with items usually appearing 3 to 8 weeks after publication [30]. Its important feature is the *cited reference searching*. Citations mean later references citing an earlier article. Users can search for all references to specific papers, authors or even keywords. A related service provided by Thomson Scientific is "Journal Citation Reports" [31]. The complete statistics of citations between papers from the journals indexed are available there. This includes impact factors of individual journals – quantifiers that are discussed in Section III.6 and disputed by [Bollen2006]. Yet another derived Web site is [32] with a list of over 300 "highly cited" computer science researchers. The Web of Science is created and maintained manually. It is a commercial product.

IV.3.4 ACM Portal

The ACM Portal [33] is a Web interface of the digital library of the Association for Computing Machinery. The library is further divided into the Digital Library proper and into the Guide. There are some 200 000 articles published by ACM and partner societies in the Digital Library (April 2007). The full texts of ACM publications are available for subscribed members (the yearly rate is 198 USD). However, publications owned by third party publishers are still charged. In addition, the ACM Guide comprises the Digital Library plus more than 700 000 bibliographic records of articles cited by ACM papers. The ACM Portal is constructed manually, however, references in articles are extracted automatically using OCR techniques. References (out-links) and citations (in-links) are shown in principle only for ACM publications in the Digital Library. There is only an indirect way how to obtain the number of citations for an individual researcher – find all of his/her publications in the library and count their "citings" (ACM's expression for citations). Research conducted on the data from the ACM Digital Library includes e.g. [Kim2004].

IV.4 Summary

In this chapter, we discuss the topic of on-line systems assisting researchers in finding bibliographic information such as publication titles, dates, names of authors, references, citations or even providing them with access to the abstracts or full texts of the publications being searched for. We mention six such systems and omit others like INSPEC by IEE [23], Scirus run by Elsevier [22], or Academic Live Search by Microsoft [21]. The importance of this chapter for the thesis is in that it presents data usable in social network analysis (see Chapter III) and thus appropriate for testing and verification of methods introduced in chapters V and VI.

The summary Table IV.1 may be regarded as a feature matrix of the systems above. Some numbers are approximate only – such as those 1.6 million Google Scholar documents for computer science or 10.8 million Web of Science articles that cover all natural and technical sciences. By reference linking we mean whether one can navigate forward by following links to referenced papers and citations linking is the opposite – one can go back to citing publications. Let us remark in this context that a different terminology is used in Web of Science – navigating forward means forward in time and it is exactly the same that we call going back to citing papers. At the ACM Portal, reference and citation linking is possible only in the Digital Library, not in the Guide. In some systems we can find out the exact citation count for a particular scientist, in some others we have to count it indirectly by means of citations to the scientist's publications.

	DBLP	**ACM Portal**	**Google Scholar**	**Rexa**	**Web of Science**	**CiteSeer**
Free	yes	no	yes	yes	no	yes
Automated	no	no	yes	yes	no	yes
# documents	870 000	200 000	1 600 000	380 000	10 830 000	760 000
All bibl. data downloadable	yes	no	no	no	no	yes
Reference linking	partly	partly (DL)	no	yes	yes	yes
Citation linking	partly	partly (DL)	yes	yes	yes	yes
# citations for a publication	partly	partly (DL)	yes	yes	yes	yes
# citations for an author	partly indirectly	partly indirectly	indirectly	yes	yes	indirectly

Table IV.1: Feature matrix of systems as of April 2007.

We can conclude that DBLP appears to be the best repository for automated experiments with bibliographic data, for it is free, all of its data are easily downloadable and manageable (XML), and it is relatively free of errors (unlike CiteSeer) due to its manual creation.

V Bibliographic PageRank

Notions of importance, significance, authority, prestige, quality and other synonyms play a major role in social networks of all types. They denote an object that has a large impact on the other objects in the community. Perhaps the best example are bibliographic citations in the scientific literature. Counting citations of research publications is a relatively objective manner to determine quality research known since a long time ago. With the fast growth of the World Wide Web in the past ten years, this kind of analysis has become essential also in this domain in which citations are links between Web pages. Therefore, current Web search engines make use of various link-based quality ranking algorithms whose rankings they combine with the keyword search results to offer the user not only topic-relevant but also high quality Web pages. The best-known link-based ranking algorithm is PageRank (see Section II.2). This recursive algorithm is applicable to any directed graph – such as a graph of citations between authors or papers. However, bibliographic data usually offers more than just citations. Collaboration networks are also a valuable source of information and are often studied (see Section III.4). But their combination with citation graphs, which may lead to more fair rankings of authors, has been relatively little examined. In this chapter, we present several modifications of the classical PageRank formula adapted for bibliographic networks. Our versions of PageRank take into account not only the citation but also the co-authorship information.

V.1 Definitions

Let $G^P = (P \cup A, E^P)$ be an undirected, unweighted, bipartite graph (co-authorship graph), $P \cup A$ a set of vertices (P a set of publications, A a set of authors) and E^P a set of edges. Each edge $\{p, a\} \in E^P$, $p \in P$, $a \in A$ means that author a has (co-)authored publication p. Let $G^C = (P, E^C)$ be a directed unweighted graph (publication citation graph), P a set of vertices (the same set of publications), and E^C a set of edges (citations between publications). Now, based on the two graphs G^P and G^C, we will introduce yet another graph we will further work with. Let $G = (A, E)$ be a directed, edge-weighted graph (author citation graph), A a set of vertices (the same set of authors) and E a set of edges (citations between authors). For every $p \in P$ let $A_p = \{a \in A: \exists \{p,a\} \in E^P\}$ be the set of authors of publication p. For each (a_1,a_2), $a_1 \in A$,

$a_2 \in A$, $a_1 \neq a_2$ where there exists $(p_1,p_2) \in E^C$ such that $\{p_1,a_1\} \in E^P$ and $\{p_2,a_2\} \in E^P$ and $A_{p1} \cap A_{p2} = \varnothing$ (i.e. no common authors in citing and cited publications are allowed) there is an edge $(a_1,a_2) \in E$. Thus, $(a_1,a_2) \in E$ if and only if $\exists (p_1,p_2) \in E^C \wedge \exists \{p_1,a_1\} \in E^P \wedge \exists \{p_2,a_2\} \in E^P \wedge A_{p1} \cap A_{p2} = \varnothing \wedge a_1 \neq a_2$.

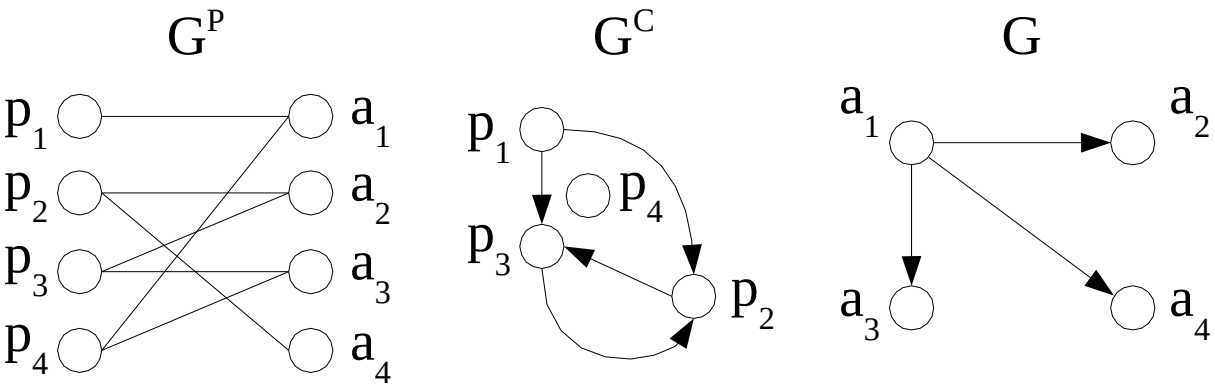


Figure V.1: Examples of co-authorship, publication citation, and author citation graphs.

Before assigning weights to edges in E, we further define:

- $w_{u,v} = |C|$ where $C = \{p_1 \in P: \exists \{p_1,u\} \in E^P \wedge \exists \{p_2,v\} \in E^P \wedge \exists \{p_1,p_2\} \in E^C \wedge p_1 \neq p_2\}$, as the number of citations from u to v,
- $f_{u,v} = |P_u| + |P_v|$ where $P_i = \{p \in P: \exists \{p,i\} \in E^P\}$, as the number of publications by u plus the number of publications by v,
- $c_{u,v} = |CP|$ where $CP = \{p \in P: \exists \{p,u\} \in E^P \wedge \exists \{p,v\} \in E^P\}$, as the number of common publications by u and v,
- $hd_{u,v} = |ADC_u| + |ADC_v|$ where $ADC_i = \{a \in A: \exists p \in P$ such that $\{p,a\} \in E^P \wedge \{p,i\} \in E^P\}$, as the number of all distinct co-authors of u plus the number of all distinct co-authors of v,
- $h_{u,v} = |ADC_u| + |ADC_v|$ where ADC_i is defined as above but it is a multiset, as the number of all co-authors of u plus the number of all co-authors of v,
- $td_{u,v} = |DCA|$ where $DCA = \{a \in A: \exists p \in P$ such that $\{p,a\} \in E^P \wedge \{p,u\} \in E^P \wedge \{p,v\} \in E^P\}$, as the number of distinct co-authors in common publications by u and v,
- $t_{u,v} = |DCA|$ where DCA is defined as above but it is a multiset, as the number of co-authors in common publications by u and v,
- $g_{u,v} = f_{u,v} - |SP_u| - |SP_v|$ where $SP_i = \{p \in P: \{p,i\} \in E^P \wedge d_{G^P}(p) = 1\}$, as the number of publications by u where u is not the only author plus the number of publications by v where v is not the only author.

Note that the current authors are considered as co-authors of themselves (variables hd, h, td, t). They should actually not be counted in but this would have no effect on the results.

V.2 Rank Calculation

We associate a triple of weights $(w_{u,v}, c_{u,v}, b_{u,v})$ with each edge $(u, v) \in E$ where $w_{u,v}$, $c_{u,v}$ are described above and $b_{u,v}$ can be equal to one of the seven following values according to the semantics of edge weights we want to stress: a) zero, b) $f_{u,v}$, c) $h_{u,v}$, d) $hd_{u,v}$, e) $g_{u,v}$, f) $t_{u,v}$, g) $td_{u,v}$. We then define the rank R(u) for author u as follows:

$$\mathrm{R}(u) = \frac{1-d}{|A|} + d \sum_{(v,u)\in E} \mathrm{R}(v) \frac{\sigma_{v,u}}{\sum_{(v,k)\in E} \sigma_{v,k}} \tag{V.1}$$

where

$$\sigma_{v,k} = \frac{w_{v,k}}{\frac{c_{v,k}+1}{b_{v,k}+1} \sum_{(v,j)\in E} w_{v,j}} \tag{V.2}$$

and d is the damping factor, an empirically determined constant usually set to about 0.9.

In all the variations above, we penalize the cited author for the frequency of collaboration with the citing author. We suppose that a citation obtained from a frequent co-author (colleague) is less valuable than that from a foreign researcher. Therefore, the contribution from citing authors is inversely proportional to the number of common publications with the cited author. This happens in case a). On the other hand, we mitigate this penalization under some circumstances. In cases c), d), f), and g) we recognize that the relationship between two authors is weaker if they have many co-authors in general – cases c) and d) – or in common publications – cases f) and g). We also distinguish between all co-authors – cases c) and f) - and distinct co-authors – cases d) and g). In case b) we claim that two authors are more closely related if they have relatively many common publications in relation to the total number of publications by both of them and less related in the opposite case. The same holds for case e) where the total number of publications by each author as the only author is counted. When all the coefficients *c* and *b* are equal to zero, equation (V.1) becomes the weighted PageRank formula. (For instance, [Bollen2006] and [Xing2004] work with weighted PageRanks.) In addition to this, if all the weights $w_{u,v}$ are set to one, it is the standard PageRank [Brin1998]. The coefficients *c* and *b* are analogous to the co-authorship frequency and exclusivity in [Liu2005] which is mentioned in Section V.4.

Zero c coefficients

Certainly, there will be many author pairs in *G* for which *c* is zero. Does it make sense to have a non-zero coefficient *b* if *c* is equal to zero? It surely does not when *b* is *t* or *td*. If there are no common publications, there are no co-authors in common publications either. Other parameters (*f*, *g*, *h*, *hd*) may (or even must) be greater than zero even if *c* is zero. But modifying the portion of rank distributed between authors only on the basis of all their publications (*f*), all their co-authors (*h*), etc. without the context of their common publications ($c = 0$) does not look meaningful. Why should author *x* obtain more rank than author *y* from a particular citing author only for the reason that he/she has written more publications? Briefly, we set *b* to zero whenever *c* is zero.

Example

Table V.1 shows edge weights for graph *G* in Figure V.1. The coefficients *f*, *g*, *h*, and *hd* are zero when *c* is zero as mentioned in the paragraph above, but their non-zero variants are also presented in parentheses for illustration. Edges (p_2,p_3) and (p_3,p_2) have no effect because they are considered as self-citations (author a_2 has co-authored both of them). The proportions of rank distributed by author a_1 in graph *G* in Figure V.1 along its out-edges in the standard (PR) and weighted PageRank (w) and the variations a) – g) are given in Table V.2.

Edge	*w*	*c*	*f*	*g*	*h*	*hd*	*t*	*td*
$\{a_1,a_2\}$	2	0	0 (4)	0 (1)	0 (7)	0 (4)	0	0
$\{a_1,a_3\}$	1	1	4	1	7	3	2	2
$\{a_1,a_4\}$	1	0	0 (3)	0 (1)	0 (5)	0 (4)	0	0

Table V.1: Edge weights for graph *G* in Figure V.1.

Edge	PR	w	a	b	c	d	e	f	g
$\{a_1,a_2\}$	1/3	2/4	4/7	4/11	2/7	2/5	2/4	4/9	4/9
$\{a_1,a_3\}$	1/3	1/4	1/7	5/11	4/7	2/5	1/4	3/9	3/9
$\{a_1,a_4\}$	1/3	1/4	2/7	2/11	1/7	1/5	1/4	2/9	2/9
$\sum$	1	1	1	1	1	1	1	1	1

Table V.2: Proportions of rank distributed by node a_1 in graph *G* in Figure V.1.

For example, to compute $\sigma_{a1,a2}$ for the variation w), we substitute in (V.2);

$$\sigma_{a1,a2} = \frac{2}{\frac{0+1}{0+1}(2+1+1)}$$

which is 2/4. Since $\sigma_{a1,a2} + \sigma_{a1,a3} + \sigma_{a1,a4} = 2/4 + 1/4 + 1/4 = 1$, the proportion $\frac{\sigma_{a1,a2}}{\sum_{(v,k)\in E}\sigma_{v,k}}$ from (V.1) remains 2/4. Thus, one half of rank of author a_1 is transferred to author a_2 and so on.

V.3 Experiments

We tested our algorithms on the DBLP data available in XML. We took advantage of the only time-stamped version of the collection from February 14, 2004 [2] which may serve researchers as a testbed for experiments and comparisons. We extracted only *article* and *inproceedings* records exactly like in [Sidiropoulos2005].

V.3.1 DBLP Testbed Data

Statistics

Table V.3 summarizes some basic statistics of the DBLP data we work with. (For more details on DBLP, see Section IV.1.) We spend some time discussing it here as a good understanding of it is vital for everyone wishing to reproduce our experiments. The data contained 173 630 *article* records (journal papers) and 298 413 *inproceedings* records (conference papers) that we imported into a relational database. These numbers are in cells B2 and C2, respectively. The total number of *article* and *inproceedings* records (i.e. their corresponding XML elements), which we will refer to as papers, is 472 043 (D2). The number of papers having some references is only 8 188 (D3) which is less than two percent of the total. In addition, a large part of all references from papers (D6) are references to undisclosed publications outside of the DBLP library. The references within DBLP (D7) can be further decomposed into references to papers (D8) and references to other kinds of publications such as books, theses, etc. The corresponding numbers of papers with references within DBLP publications and with references to papers are D4 and D5. Exactly 18 285 distinct papers are

cited (D11). Time spans are not shown in Table V.3. However, the most recent paper is from 2004, the oldest one is from 1936. The time period of citing papers is 1970 – 2001, that of cited papers is 1945 - 2001 We can also obtain other information from Table V.3, such as the number of references from journal papers to conference papers (B10), the number of conference-to-conference references (C10), the number of journal papers with references to papers (B5), etc.

	A	B	C	D
1		**articles**	**inproceedings**	**total**
2	#	173 630	298 413	472 043
3	**# with ref.**	1 818	6 370	8 188
4	**# with ref. within DBLP**	1 791	6 212	8 003
5	**# with ref. to papers**	1 771	6 177	7 948
6	**# references**	47 329	120 822	168 151
7	**# ref. within DBLP**	30 186	79 003	109 189
8	**# ref. to papers**	27 801	72 853	100 654
9	**# ref. to articles**	13 330	29 247	42 577
10	**# ref. to inproc.**	14 471	43 606	58 077
11	**# distinct cited**	7 391	10 894	18 285

Table V.3: Statistics of *article* and *inproceedings* records in DBLP 14 Feb 2004.

Problems with article and inproceedings elements

The number of papers with references in D3, D4, and D5 is decreasing as well as is the number of references themselves in D6, D7, and D8. This results from the fact that if M is a set of all publications in the world, Q is a set of publications in the DBLP digital library and P is a set of DBLP journal and conference papers then $P \subset Q \subset M$. The relationship $P \subset Q$ is completely disregarded in the statistics on DBLP presented in [Sidiropoulos2005]. For the reader who would like to verify our results we provide a small hint in Table V.4. It shows occurrences of *article* and *inproceedings* DBLP records along with their keys. Also we must be aware that some other DBLP XML elements use the “journals”, “conf”, “tr”, and “persons” keys. Thus, the key itself does not indicate whether or not a cited publication is a journal or conference paper.

tag	**key**	**#**	
article	journals	173 085	
article	persons	10	
article	tr	535	173 630
inproceedings	conf	298 322	
inproceedings	journals/jods	9	
inproceedings	journals/lncs	80	
inproceedings	persons/Codd74	1	
inproceedings	persons/JohnLM94	1	298 413
			472 043

Table V.4: Key and tag distribution in our DBLP data.

V.3.2 Co-Authorship and Citation Graphs

Publications

Let us return to Table V.3. The publication citation graph G^C based on the *articles* and *inproceedings* records will thus have 472 043 nodes ($|P|$ in D2) and 100 654 edges ($|E^C|$ in D8). So the references not pointing to papers or even pointing outside of DBLP have absolutely no effect. 7 948 nodes (D5) will have some out-edges and 18 285 nodes (D11) will have some in-edges. There will be 5 389 nodes with both in- and out-degree non-zero (not shown in Table V.3). The other graph constructed from the DBLP records is the co-authorship graph G^P. This graph has $|P| + |A|$ nodes (publications plus authors) which is 472 043 + 315 485 = 787 528 vertices in total. The number of edges $|E^P|$ is 1 070 643. This is actually the number of publication – author pairs (see G^P in Figure V.1). See Figure V.2 for a histogram of the number of co-authors in publications, i.e. of the degrees of publication nodes in G^P. The most frequent number of co-authors is two and a publication has 2.27 co-authors on average. Interestingly, there are also publications without any authors which is an obvious omission in DBLP.

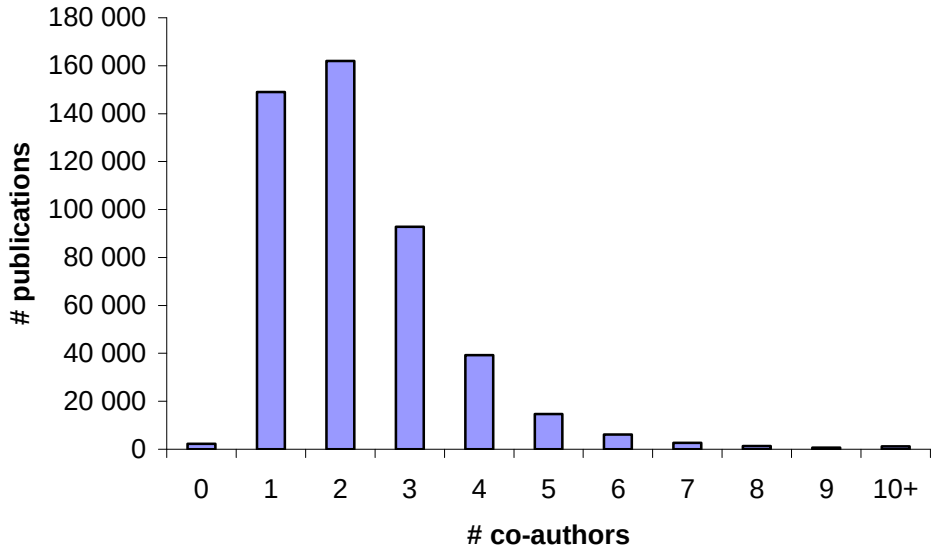


Figure V.2: Histogram of the number of co-authors in DBLP publications.

Author citation graph

The resulting citation graph of authors G had 295 531 edges (no self-citations are allowed and citations between publications that have at least one common author are considered as self-citations) which is $|E|$. Obviously, $|A|$ is still 315 485. 12 934 nodes had a non-zero in-degree, 6 992 nodes had a non-zero out-degree. 4 748 nodes had both a non-zero in-degree and a non-zero out-degree. Only 15 178 authors were not isolated. This low inter-linkage of nodes in G is a result of the nature of the DBLP data. Citations were systematically input only for a small number of journals and conferences, such as SIGMOD Record or VLDB Journal, as was already mentioned in [Sidiropoulos2005]. See Figure V.3 for a cumulative distribution of in- and out-degrees and their weighted variations (citations and references) in graph G. The maximum value for in-degree is 1 857, for out-degree 834, for citations (in) 5 346 and for references (out) 2 594. Apparently, the largest bin would be 0+ with all the isolated authors included. It is not depicted in Figure V.3. As we may see, the four series are quite well correlated. The number of authors with a specific degree decreases as the degree gets bigger.

There are no evident outliers. Perhaps the most interesting feature is the sudden drop in the number of authors for 1+ (having one or more) and 5+ (having five or more) in-degree and citations. This is not the case for out-degree or references. This means that 5 is quite a boundary for less and more cited authors. Also, the superiority of references over citations which begins with 10+ and terminates with 200+ indicates that the group of highly cited authors is greater than that of highly citing authors.

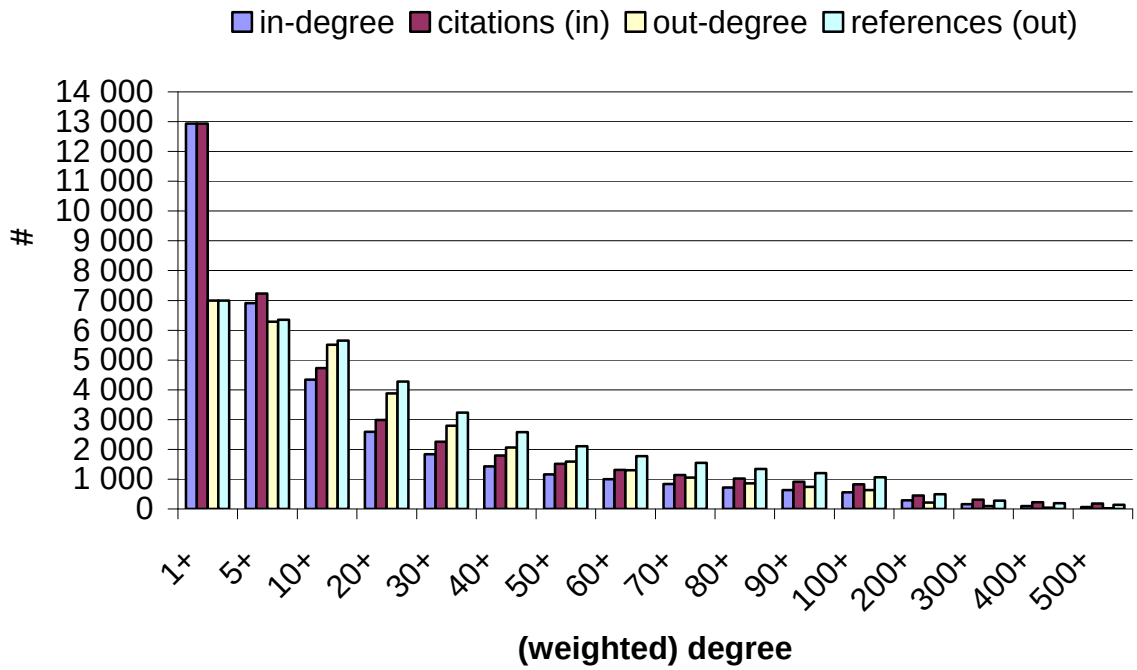


Figure V.3: Cumulative histogram showing distribution of in- and out-degrees in *G*.

Distribution of *c* and *b* coefficients

Figure V.4, Figure V.5, Figure V.6, and Figure V.7 show the cumulative distribution of various parameters defined in Section V.1 in the weights of edges in *E* of graph *G*. The size of the bin 0+ for each series of each graph would be 295 531, i.e. $|E|$. The number of edges in each 1+ bin is always 7 017 since this is the number of edges in *E* between authors that have some common publications. This number will never be exceeded by values of other parameters because in Section V.2 we have defined the parameters *f*, *g*, *h*, *hd*, *t*, *td* to be zero whenever *c* is zero. Now, let us make a few examples of interpretation of the data in the figures. For instance, the number of edges in *E* for which the parameter *c* is five or more is a little greater than one thousand. This means that there are some one thousand author pairs having five common publications at least that cite each other (not necessarily at the same time). The author pairs are ordered, so if the authors cite one another at the same time, i.e. there are two edges in *E* for this pair, the pair is counted twice. Another example: there are some 5 000 author pairs having some common publications whose sum of publications is 70 at least (see Figure V.5). In Figure V.6, we can observe that there are no collaborating authors that would have 400 or more distinct co-authors in total. The bins 1+ and 2+ in Figure V.7 are the same because each common publication of two authors has two (distinct) co-authors at least. The largest number of author pairs have between five and ten distinct co-authors in their common publications (see Figure V.7). If we subtract the citing and the cited author, it is between three and eight. In general, it holds that $f \geq g$, $h \geq hd$, $t \geq td$ as the second parameter in the couple is always more restrictive.

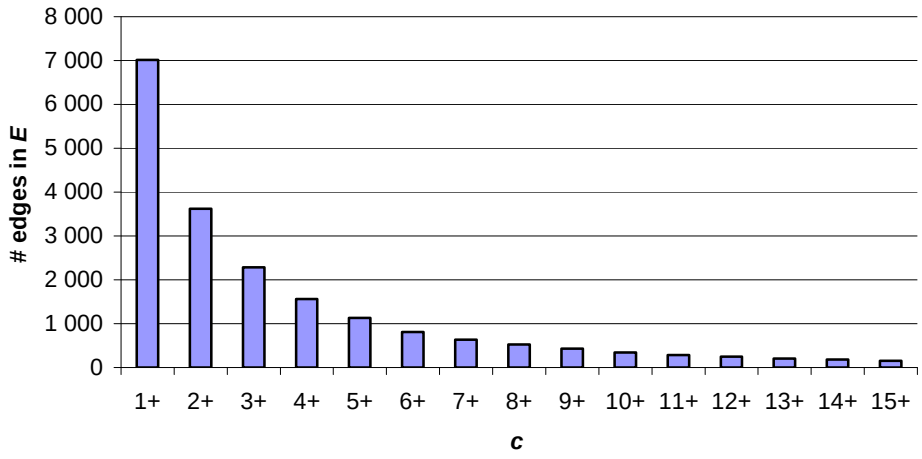


Figure V.4: Cumulative distribution of values of parameter c in graph *G*.

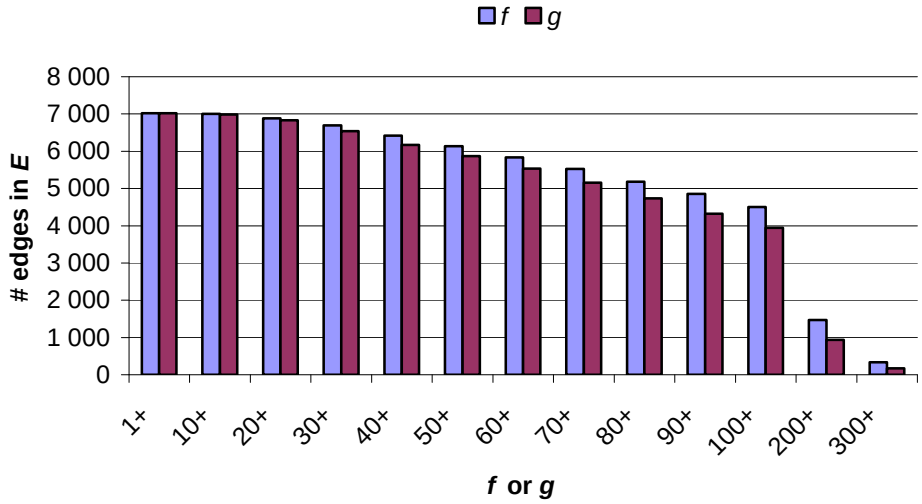


Figure V.5: Cumulative distribution of values of parameters *f* and *g* in *G*.

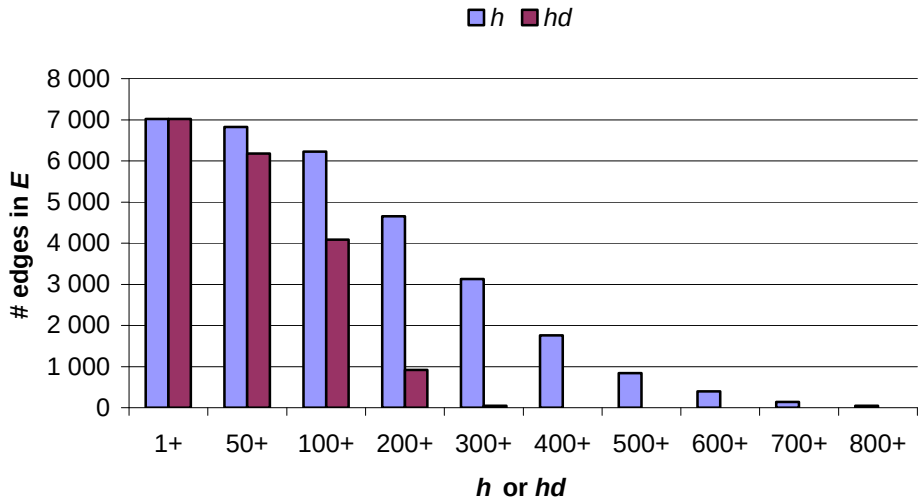


Figure V.6: Cumulative distribution of values of parameters *h* and *hd* in *G*.

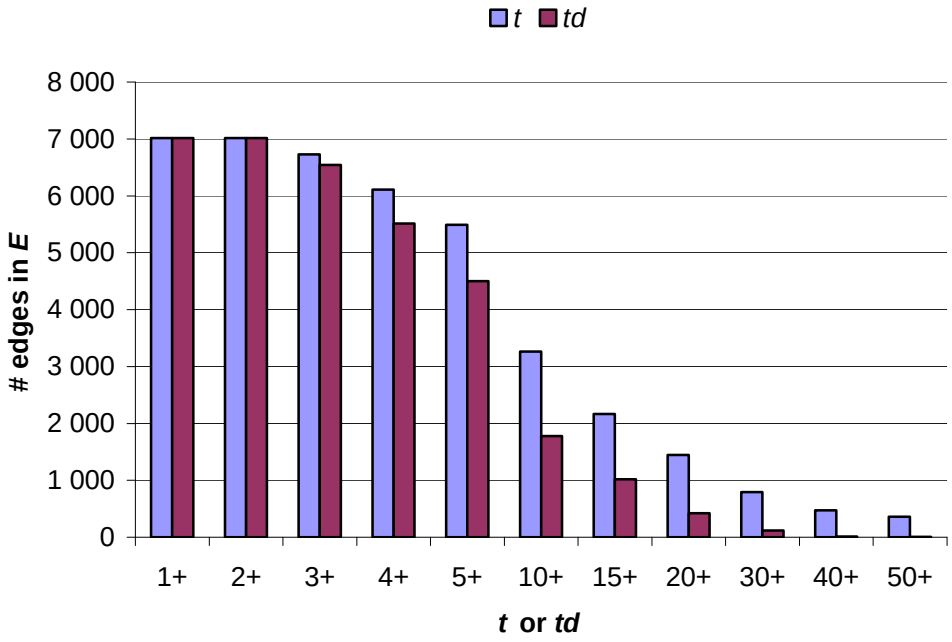


Figure V.7: Cumulative distribution of values of parameters *t* and *td* in *G*.

Statistics of *c* and *b* coefficients

To terminate this subsection, Table V.5 presents basic statistics of the *c* and *b* parameters in the weights of edges in graph *G*, which were commented on in the previous paragraphs. Parameter *b* is represented by the corresponding coefficients *f*, *g*, *h*, *hd*, *t*, and *td* as described in sections V.1 and V.2. Note that only those edges in *E* of *G* are considered for which *c* is non-zero, i.e. edges between authors who have some common publications. The number of these edges is 7 017 as mentioned above. Taking into account all of the edges in *E* would obviously decrease the mean values and set all medians and modes to zero. In total, we have found 10 902 author pairs having one common publication at least but not all of them have a citation edge in *E*, of course. Some interesting findings visible in Table V.5 include: i) the maximum number of distinct co-authors in common publications by two specific authors is 67 (!), iii) the most frequent number of the same is three (rather low), iii) the maximum total number of publications (counted separately) of two collaborating authors is 489, etc. Much more analysis (such as component analysis) of the co-authorship and citations graphs could be done but it is not the aim of this thesis.

	c	*f*	*g*	*h*	*hd*	*t*	*td*
min	1	4	2	2	2	2	2
max	56	489	443	977	355	210	67
avg	2.93	139.83	120.87	295.26	122.41	14.80	7.99
std. deviation	3.89	81.50	72.28	168.68	64.50	17.66	6.47
median	2	130	111	273	114	9	6
mode	1	153	134	188	59	3	3

Table V.5: Basic statistics of weight parameters for edges in *E* with non-zero *c*.

V.3.3 Computing Ranks for Authors

We exploited extensively the author citation graph *G* described in detail in Section V.3.2. Altogether, twelve ranking methods were employed to evaluate the authors. In addition to the weighted (citation counting) and unweighted in-degree, HITS authorities (see Section II.3), and the standard (unweighted) PageRank (see Section II.2), we also applied the weighted and the bibliographic (seven variants a) – g) from Section V.2) PageRank algorithms. In this way, we finally obtained twelve author rankings. The big problem that immediately arises is how to evaluate the quality of these rankings. The quality of a ranking is a highly subjective matter. A straightforward solution would be to compare the generated rankings with an official, "human-made" ranking. Unfortunately, this does not exist. Another possibility would be to make use of the various citation systems we talk about in Chapter IV and compare the new rankings with their citation-based rankings. The trouble here is that the citation data in DBLP is very incomplete and it is more or less concentrated on publications in a few particular journals and conferences. Thus, it would not be directly comparable.

Awards

It is remarkable in this context, that ACM SIGMOD Digital Review and ACM SIGMOD Record journals as well as the ACM SIGMOD Conference have their publications' citations included. This was perhaps what initially triggered the idea in [Sidiropoulos2005] – namely to compare author rankings with lists of ACM SIGMOD award winners. Quite logically, the authors expected that award winners should be placed higher in their rankings than other authors. In other words, the better a ranking, the higher ranks it associates with award winning authors. As our approach is somewhat different from theirs (more on this will be said in Section V.4), the only award we can take advantage of is the *ACM SIGMOD E. F. Codd Innovations Award* [1], which is awarded "for innovative and highly significant contributions of enduring value to the development, understanding, or use of database systems and databases."

Program committees

The only alternative approach to author ranking evaluation we are aware of is described in [Liu2005]. Here the newly derived rankings are compared to lists of program committee members (i.e. prestigious researchers) of conferences on digital libraries. A ranking with more authors being members of program committees is considered better than another one having only a few of them. This approach has two obvious drawbacks. First, it is domain specific. It is appropriate for rankings based on data from digital library conferences (as was the case). For other fields different program committees would have to be considered. But for general, non-specific data (more or less the case of DBLP) it is not reasonable. And second, actual ranks of authors are not taken account of. So two rankings with the same authors in a different order would be evaluated the same. (Although this can be improved easily by comparing a series of ranks rather than single total scores.)

Results

We thus compared the ranks achieved by fifteen winners of the ACM SIGMOD E. F. Codd Innovations Award from the years 1992 – 2006. We also expected that better rankings would place award winners higher. Let us have a look at Table V.6 with the actual ranks. The first three rankings (citations, in-degree and HITS authorities) are presented just for reference. The actual baseline ranking is "PR" (standard unweighted PageRank, in a darker column). In other words, the goal is to compare the new "bibliographic" PageRank rankings in columns "w" and "a" through "g" with the standard PageRank. The column "w" stands for the weighted PageRank from Section V.2 and "a" – "g" correspond to the variations a) – g) mentioned at

the very beginning of the same section. We can see that the weighted PageRank is much better than the classical one in terms of the sum of ranks (the smaller the better), the median rank and a little better as for the worst rank assigned to the award winners. The rankings "a" – "g" are always better than the standard PR regarding the sum of ranks and median rank and only "a" and "c" have a worse worst rank. The ranking "a" is also weaker than "w" in all metrics whereas "c" only with respect to the worst rank. The rankings "d" and "e" are the best in the sum of ranks and in the worst rank respectively. The median is better for "d" (9 versus 12). Let us recall that this ranking penalizes authors frequently cited by their co-authors but it weakens this handicap if the citing and cited authors have many distinct co-authors altogether. Moreover, the median rank 9 is the best of all in the table. Even the rankings not based on PageRank are worse in this respect.

As we may observe, simple citations counting and in-degree perform best. This is not astonishing since prestige, popularity, awards, and recognition generally still rely mostly on the number of an individual's citations. What is more surprising is the very good result of HITS which is in contradiction with the conclusions taken by [Sidiropoulos2005]. However, their HITS ranking was not obtained in the same way as ours (see Section V.4).

Year	Author	Cites	InDeg	HITS	PR	w	a	b	c	d	e	f	g
1992	Michael Stonebraker	1	1	1	3	2	2	1	1	1	1	3	3
1993	Jim Gray	4	3	4	6	3	6	2	2	2	4	1	2
1994	Philip Bernstein	6	8	7	4	6	5	6	6	4	6	5	4
1995	David DeWitt	2	2	2	36	14	20	3	3	3	2	4	5
1996	C. Mohan	36	47	45	113	110	116	62	59	65	65	105	101
1997	David Maier	13	11	11	51	35	47	7	7	6	7	11	13
1998	Serge Abiteboul	12	18	21	104	61	69	12	11	14	12	37	43
1999	Hector Garcia-Molina	9	12	18	60	49	62	4	4	5	3	16	14
2000	Rakesh Agrawal	11	15	25	65	58	64	16	19	18	15	49	49
2001	Rudolf Bayer	84	75	94	7	16	14	97	132	94	93	25	20
2002	Patricia Selinger	38	38	23	59	55	53	61	55	54	63	36	48
2003	Don Chamberlin	16	13	10	2	4	3	29	26	23	26	7	6
2004	Ronald Fagin	28	40	46	19	13	13	27	28	30	25	17	17
2005	Michael Carey	7	9	5	63	46	55	13	10	9	14	21	29
2006	Jeffrey D. Ullman	3	5	9	15	8	12	5	5	7	5	8	8
	Worst rank	**84**	**75**	**94**	**113**	**110**	**116**	**97**	**132**	**94**	**93**	**105**	**101**
	Sum of ranks	**270**	**297**	**321**	**720**	**480**	**541**	**345**	**368**	**335**	**341**	**345**	**362**
	Median rank	**11**	**12**	**11**	**36**	**16**	**20**	**12**	**10**	**9**	**12**	**16**	**14**

Table V.6: E. F. Codd Innovations Award winners and their ranks in distinct methods.

Discussion of author ranks

The accompanying chart of Table V.6 is in Figure V.8. We can easily capture the most significant trends there. The three lowest-ever ranked authors are Rudolf Bayer, C. Mohan, and Serge Abiteboul. At the same time, the positions of Rudolf Bayer and Serge Abiteboul are quite oscillating (both high and low ranks exist) whereas those achieved by C. Mohan remain more stable (rather low). There are two scientists who are always ranked in the top 10 – Michael Stonebraker and Jim Gray. Nevertheless, these two researchers were awarded first – in 1992 and 1993, respectively. Thus, there has been time enough for them to profit from the award and to collect citations. In this context, the high ranks of the most recently awarded researcher, Jeffrey D. Ullman, are very remarkable. (Of course, he may have won another one from the many awards before.)

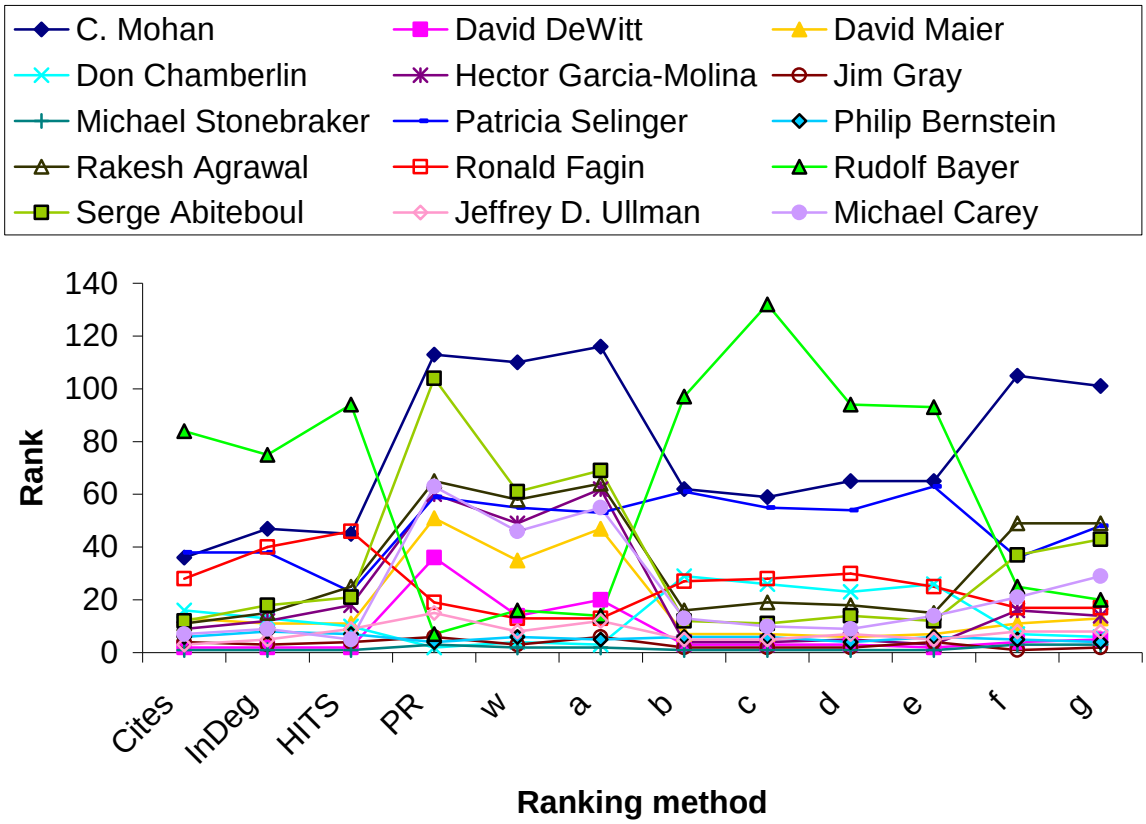


Figure V.8: E. F. Codd Innovations Award winners.

Let us have a look at some particularities in Figure V.8. For instance, Rudolf Bayer has relatively few citations and few distinct citing authors (citations and in-degree), but he is cited mostly by authoritative researchers ("PR" and "w") and not so much by his colleagues ("a"). Then he suddenly looses good positions which may indicate that his colleagues citing him have published rather little ("b" and "e") and that they usually have few co-authors in their publications ("c" and "d"). But the number of co-authors in the common publications with the researchers citing him is relatively high ("f" and "g"). Also, there is the biggest difference between "c" and "d" for Rudolf Bayer amongst all awarded authors. This may mean that there are less distinct co-authors in his publications (and/or in publications of his colleagues citing him) with respect to all co-authors than is the case with other award winners. It is somewhat inverse with Serge Abiteboul. He has many citations but is cited by less authoritative authors (a sudden drop with "PR"). However, if the frequency of endorsements is taken into account ("w"), Abiteboul's rank improves considerably (from over 100 to almost 60), etc. Certainly, all of the above explanations are not exclusive because there may be many other factors affecting the ranks that we are even not aware of. Also keep in mind that the results are based on the very incomplete data we work with. We do not present individual statistics over rankings for each author here since the objective is to compare rankings rather than authors.

Comparison of rankings

There are a number of metrics for comparison of rankings. See [Sidiropoulos2006] for some of them. We will briefly discuss the outcomes of three metrics – two numerical and one graphical. In Table V.7 we can see the number of common elements in the top twenty authors of two particular rankings. For instance, the ranking by citations has 16 authors in common with the ranking by in-degree in the Top 20. The number of common authors varies between

five and twenty. Of course, it does not reveal anything about the order of authors. It just says that 16 authors are the same. Theoretically, the ordering could be inverse. Two pairs of rankings have a complete match – “w” and “a”, and “b” and “e”. Also “f” and “g” have a rather great match (19 authors in common). On the other hand, the least observable match is produced by the standard PageRank – it shares just five authors with each “b”, “c”, and “e”. We can notice that there is a set of pairs of “twin” rankings that match quite well each other: {citations, in-degree}, {“PR”, “w”}, {“b”, “e”}, {“c”, “d”}, and {“f”, “g”}. The “twin” rankings are very close to each other in the definition of their coefficients, e.g. weighted or unweighted in-degree, co-authors or distinct co-authors, etc. This definition similarity results in the similarity of their top twenty authors. The only exception in this respect is the pair {“w”, “a”} that matches perfectly but whose definition is somewhat distinct. On the contrary, we may observe the smallest numbers between the rankings from {“b”, “c”, “d”, “e”}X{“PR”, “w”, “a”}.

	Cites	InDeg	HITS	PR	w	a	b	c	d	e	f	g
Cites	X	16	14	7	9	9	14	14	15	14	12	12
InDeg	16	X	16	9	10	10	12	12	13	12	13	13
HITS	14	16	X	11	12	12	11	12	13	11	16	15
PR	7	9	11	X	16	16	5	5	6	5	14	15
w	9	10	12	16	X	20	7	7	8	7	16	17
a	9	10	12	16	20	X	7	7	8	7	16	17
b	14	12	11	5	7	7	X	18	17	20	11	10
c	14	12	12	5	7	7	18	X	18	18	11	10
d	15	13	13	6	8	8	17	18	X	17	12	11
e	14	12	11	5	7	7	20	18	17	X	11	10
f	12	13	16	14	16	16	11	11	12	11	X	19
g	12	13	15	15	17	17	10	10	11	10	19	X

Table V.7: Common elements in top 20 authors.

	Cites	InDeg	HITS	PR	w	a	b	c	d	e	f	g
Cites	X	0.9904	0.8666	0.8119	0.8207	0.8188	0.8189	0.8079	0.8199	0.8203	0.8253	0.8237
InDeg	0.9904	X	0.8661	0.8178	0.8179	0.8163	0.8169	0.8072	0.8178	0.8180	0.8221	0.8207
HITS	0.8666	0.8661	X	0.7748	0.7496	0.7483	0.6786	0.6379	0.6831	0.6866	0.7473	0.7496
PR	0.8119	0.8178	0.7748	X	0.9806	0.9803	0.9168	0.8785	0.9213	0.9253	0.9751	0.9776
w	0.8207	0.8179	0.7496	0.9806	X	0.9993	0.9520	0.9197	0.9557	0.9586	0.9968	0.9981
a	0.8188	0.8163	0.7483	0.9803	0.9993	X	0.9452	0.9123	0.9491	0.9522	0.9938	0.9960
b	0.8189	0.8169	0.6786	0.9168	0.9520	0.9452	X	0.9935	0.9992	0.9995	0.9665	0.9620
c	0.8079	0.8072	0.6379	0.8785	0.9197	0.9123	0.9935	X	0.9921	0.9904	0.9376	0.9315
d	0.8199	0.8178	0.6831	0.9213	0.9557	0.9491	0.9992	0.9921	X	0.9993	0.9700	0.9657
e	0.8203	0.8180	0.6866	0.9253	0.9586	0.9522	0.9995	0.9904	0.9993	X	0.9722	0.9681
f	0.8253	0.8221	0.7473	0.9751	0.9968	0.9938	0.9665	0.9376	0.9700	0.9722	X	0.9994
g	0.8237	0.8207	0.7496	0.9776	0.9981	0.9960	0.9620	0.9315	0.9657	0.9681	0.9994	X

Table V.8: Spearman correlation coefficients.

The next comparison is based on the correlation between rankings. Table V.8 shows the Spearman correlation coefficients for each pair of rankings. They are all significant at the 0.01 level. An alternative metric would be Kendall’s tau (see Section II.2.4). With this metric, we consider the ranks of all authors that have some in-degree. (It is 12 934 as we mention in Section V.3.2.) Thus, few matches in the Top 20 may be easily compensated for with matches

of lower ranked researchers. All highly matching pairs of rankings from Table V.7 are represented by a large correlation coefficient. The highest correlation (0.9995) was measured between *b* and *e* where publications and "solo" publications are interchanged. On the other hand, the least correlation is reported between *c* and HITS (0.6379). However, the number of common top 20 authors is 12 which is by far not the worst. Evidently, there are many mismatches between lower-ranked scientists. The sector of small matches from Table V.7 has disappeared here. It seems that mismatches just accumulate in the upper part of rankings (which is more important than the lower one, though).

Finally, let us present a graphical representation called q-q plot. Ranks of authors generated by two different rankings are plotted against each other. Obviously, two perfectly matching rankings would produce a straight line. There are 68 ranking pairs, so it is impossible to show all charts. We have chosen four of them and show them in Figure V.10. The top-left and bottom-left charts are examples of highly matching "twin" rankings ("f" vs. "g" and "b" vs. "e", respectively). The top-right plot is for the least correlating pair (HITS vs. "c") and the bottom-right plot represents a "mediocre" ranking pair (namely "a" vs. "c").

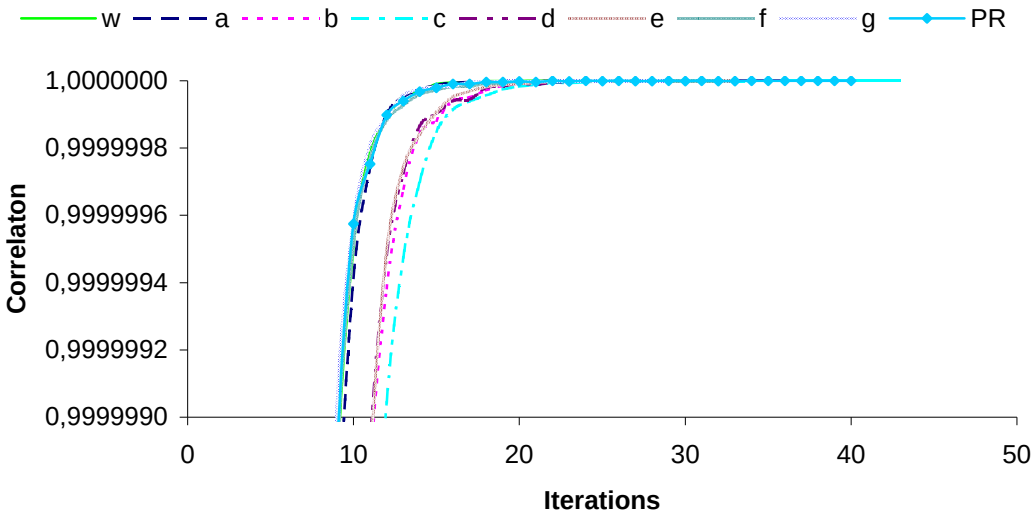


Figure V.9: Convergence rates of standard (PR), weighted (w) & bibliographic (a – g) PR.

Convergence

All in all, enhancing the citation graph with further bibliographic information proves to be very useful. The advantage over the standard PageRank is clear. Already assigning weights to the edges in the citation graph is very effective and adding data from the co-authorship network improves the results even more. The convergence rates of standard and bibliographic PageRanks are comparable. See Figure V.9 where the damping factor (*d* in equation (V.1)) is set to 0.9. The vertical axis in the figure represents the Spearman correlation coefficient between the rank vectors in the current and previous iteration. This simplified convergence criterion is often used instead of measuring the absolute error over rank scores (see Section II.2.4). In the single precision arithmetic (six or seven decimal digits), all algorithms converge in about ten iterations. Of course, the resulting rankings depend entirely on the structure of the citation and co-authorship graphs, i.e. on the DBLP data they are generated from. Remind that in our data collection, only 8 188 publications from the total 472 043 had references included. The rest could be used for the co-authorship graph only. Even though the DBLP collection dates from 2004, it still makes sense to take into account award winners from more recent

years because it usually takes a couple years for a publication to become cited and DBLP references to papers from years after 1997 are rather rare [Sidiropoulos2005]. The newest citing paper is from 2001 as pointed out in Section V.3.1.

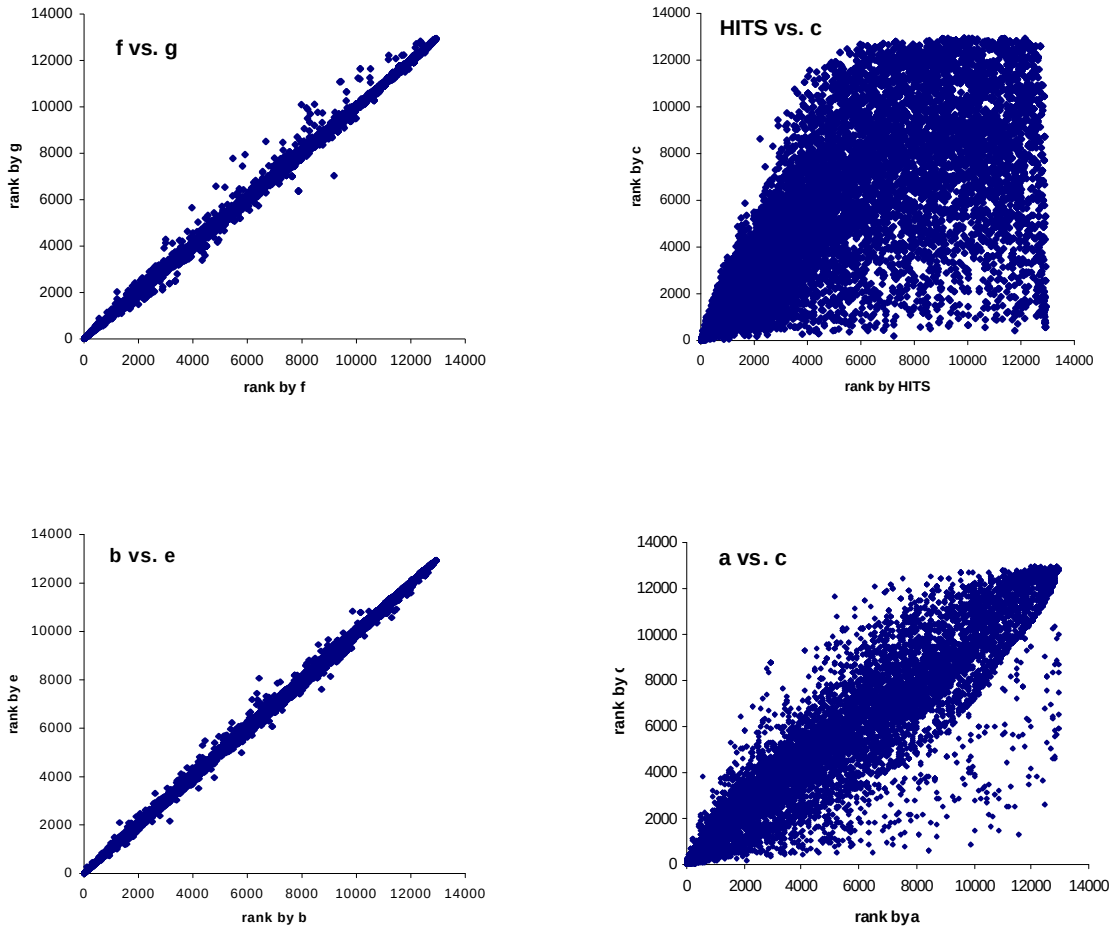


Figure V.10: Some comparisons of rankings by means of q-q plots.

Significance

To show some *statistical significance* of the improvement of the results of the baseline PR method by the new rankings (see Table V.6), we would need to reject the null hypothesis H_0: $\mu_{PR} - \mu_{NEW} = 0$, i.e. that the mean ranks of the baseline and of the new ranking are equal. However, to be able to perform such a test, the two rankings would have to be normal distributions, independent of each other, and their variances would have to be equal. At least the first two conditions are not satisfied. Therefore, we cannot say whether or not the improvements we have achieved are statistically significant. We can only demonstrate their *practical significance*.

Prediction

We show the top 40 authors for each ranking method in tables Table 1, Table 2, Table 3, and Table 4 in the appendix. E. F. Codd Award winners are in bold. Of course, the top ranked authors that have not yet been awarded have the greatest chance to win the award in future years. Raymond A. Lorie and Umeshwar Dayal appear among the best in each ranking. As the

awarding highly correlates with the ranking by citations, Won Kim is also a top candidate for the ACM SIGMOD E. F. Innovations Award in future years. (E. F. Codd himself died in 2003 and cannot be awarded.)

V.4 Related Work & Summary

Sidiropoulos

Sidiropoulos and Manolopoulos [Sidiropoulos2005] have proposed modifications of PageRank that would better meet needs for evaluating nodes in bibliographic networks. Their PageRank-based algorithm is called SCEAS Rank and is described in Section II.2.5. Although we adopted their testing methodology (DBLP and award winners) and tried our best for our results to be directly comparable, they are not. This has several reasons:

1. Different data. Unfortunately, authors use DBLP data from January 14, 2005. These data were probably up-to-date when they conducted their experiments but they are obsolete now and, in addition, they are not publicly available. Had they worked with [2] instead, the input data would be the same and their results verifiable.
2. No author citation graph. Only co-authorship graph G^P and publication citation graph G^C are constructed. All computations are performed upon G^C and rankings for authors are obtained by averaging ranks of their publications.
3. Not all publications considered. In addition, only the ranks of the 25 best-ranked publications of each awarded author are counted in for author ranks. The number 25 was selected because it appeared to be the global optimum of SCEAS Rank.

Evidently, the number of best publications selected can severely affect the ranking quality. If a global optimum for PageRank was chosen instead, one can assume that SCEAS Rank would come out much worse. Even for those 25 publications (optimal for SCEAS), PageRank has a smaller sum of ranks (200 against 207). The results of SCEAS would be comparable to ours if the ranks of all publications for each author were taken into account. The authors do not disclose these results. Working directly at the author level (and not at the publication level) avoids the problem of searching for the optimal number of best publications for authors (some authors may even not have the required number of publications) and, therefore, the resulting rankings are biased towards the method that the optimal number of top publications was selected for. Authors in [Sidiropoulos2006] try to amend the "number-of-publications" problem by aggregating the ranks of authors over several different numbers of top publications but still not all publications are considered which does not allow for an unbiased comparison of authors and methods. The inherent disadvantage of our author-level methodology is that it does not enable ranking publications.

Bollen

Liu, Bollen et al. [Liu2005] introduce co-authorship frequency and exclusivity computed from a co-authorship graph into PageRank (called AuthorRank) and rank authors from a few conferences on digital libraries. Co-authorship frequency and exclusivity are somewhat analogous to the *c* and *t* coefficients from Section V.1 and are explained in Section III.4.2. Their testing data originating from an undisclosed version of DBLP are rather small (759 publications) and domain-specific. They compare their rankings with relevant program committee members and conclude that "the results of PageRank and AuthorRank are highly correlated, but there is no conclusive evidence that one performs better than the other." However, they do not take advantage of distinct numbers of citations between authors, i.e. the parameter *w* from Section V.1 is always set to one in their method. Interestingly, they do this in [Bollen2006] for journal citation networks with a weighted PageRank algorithm. But no

co-authorship information was added to journals for obvious reasons. On the other hand, our "bibliographic" PageRank exploits both the co-authorship and citation information from bibliographic networks in a generalized manner.

Summary

In this chapter, we presented several modifications of the classical PageRank formula adapted for bibliographic networks. Our versions of PageRank take into account not only the citation but also the co-authorship graph. We verified the viability of our algorithms by applying them to the data from the DBLP digital library and by comparing the resulting ranks of the winners of the ACM SIGMOD E. F. Codd Innovations Award. Rankings based on both the citation and co-authorship information turned out to be better than the standard PageRank ranking. In the future work, we would like to concentrate on the issue of incorporating the time factor in the bibliographic PageRank. For instance, a citation between two authors made after their collaboration would be considered as less valuable than another one made before it, etc.

VI Mining the Academic Web

The succesful analysis of the well-structured DBLP data in the previous chapter invites us to try to discover authorities also in the world's biggest repository of unstructured data – on the Web. In this chapter, we present a methodology and two case studies for finding authoritative researchers by analyzing academic Web sites. In the first case study, we concentrate on a set of Czech computer science departments' Web sites. We analyze the relations between them via hyperlinks and find the most important ones using several common ranking algorithms. We then examine the contents of the research papers present on these sites and determine the most authoritative Czech authors. In the second case study, we do exactly the same with French academic computer science Web sites to find the most significant French researchers in the field. Unlike Chapter V, in which we work with quite non-noisy DBLP data, the results of the experiments we present in this chapter are inherently dependant on the structure and the content of the Web. Moreover, the Web data may be extremely noisy and biased. Thus, the outcomes should be considered as informative rather than conclusive. We also discuss the weak points of our approach and propose some future improvements. To the best of our knowledge, it is the only attempt ever made at discovering authoritative researchers from the above countries by directly mining from Web data.

This chapter comprises two sections. Section VI.1 deals with the analysis of Czech and French Web sites whereas Section VI.2 describes the process of examining the papers found thereon.

VI.1 Mining the Structure

The rapid growth of the Web has lead to fears of *information explosion*, *excess*, or *flooding*. There is too much information available, and we cannot handle all of it. The Web is a huge storehouse of data, information, and knowledge and in order to be able to get the maximum out of it, we must quickly recognize whether or not a source of information on the Web is valuable. Otherwise, we can easily waste our time studying Web documents that are irrelevant or of a poor quality. Like in the scientific literature where publications cite other publications, and we tend to refer to those highly cited ones, we prefer authoritative Web pages.

It is important for a Web surfer to search for information on quality pages to possibly gain advantage over others. It is no less relevant for a Web site creator to have his site perceived as valuable and thus attracting a larger number of visitors which may consequently imply a greater profit. Briefly, it is in the interest of the whole Web community to be capable of distinguishing between good and bad Web documents. In the Web domain, citations are links between Web pages or Web sites (when we talk about site level). Commercial Web search engines soon became aware of the potential of the Web link structure for the discovery of its authoritative resources, and a link-based quality judgement is a necessary complement of their content-based search techniques.

Algorithms for these judgements may be recursive, such as PageRank [Brin1998, Page1999, Chakrabarti2002, pp. 209-212] or HITS [Chakrabarti1998, Gibson1998, Kleinberg1999b] or simple like In-Degree which just counts in-links. Some studies [Ding2001b, Ding2002] have shown that the rankings produced by the three algorithms are highly positively correlated, but it has been contested by other researchers [Pandurangan2002]. Recursive methods have a strong probabilistic background [Diligenti2004] and there exist many modifications, e.g. PageRank for bibliographic citations [Sidiripoulos2005, Sidiripoulos2006]. We refer to Chapter II for in-depth information on ranking algorithms. Closest to our work is the research in [Thelwall2001, Li2003] not further described here due to space limitations, but in addition to the relations between Web sites we also studied the contents of the documents found on them. Other authors have tried to determine the importance of Web sites of Universities rather than departments as we have done [3].

VI.1.1 Czech University Computer Science Web Sites

Our first objective was to determine authoritative institutions among Czech computer science University departments. We have chosen this area because we know it well and we could expect that there would be enough data on the Web to analyze. At the same time, we supposed the data volume to be easily manageable. Even though we limited our experiments by topic and scope, the methodology we used was sufficiently general to be able of applying to a completely different scientific field.

Constraints

We have selected seventeen computer science Web sites from a Web directory of Czech academic institutions. Our selection had several constraints. First, we wanted to take account of their geographic location so as to include various regions of the Czech Republic. Second, each department had to have its home page on its own server. That means, we did not consider home pages being on a URL's path such as www.someuniversity.cz/somedepartment but only those like www.department.university.cz. Therefore, we had to eliminate departments whose home pages were located in their University domain, which was sometimes the case.

The reason for this is the fact that stand-alone servers can be manipulated more easily by a machine. A Web spider recognizes quickly whether or not a link on a department's Web page is internal (within department). No recognition of logical domains on Web sites is necessary, and we can get along without techniques similar to those in [Li2000], And third, we wanted the departments to correspond in the University hierarchy approximately to the level of our home department. This is somewhat tricky because not all of the Universities have the same structure of schools consisting of departments. For this reason, some institutions in our list are schools rather than departments.

Procedure
In December 2005, we let our Web spider crawl all of the seventeen servers. The spider stored information about hyperlinks between Web pages on the servers to a database and built a corpus of downloaded documents for further analysis (see Section VI.2). We repeated the same procedure two more times in a-few-days intervals and the results we obtained remained almost unchanged. We show those from the last experiment in Table VI.1.

We have to mention briefly a few Web crawling related issues which may have impact on the parameters we examined. We were interested only in links via the HTTP protocol and pointing to documents in certain formats. For instance, we did not consider video or audio documents, which is natural, but we also left out documents with extensions doc, rtf, txt, and ppt, which is more arguable. (However, taking account of these formats in one of the experiments caused only one change in the middle part of the chart in Table VI.1.) To prevent the spider from getting stuck in Web traps, we set the maximum depth of nesting in the Web graph to eight, which is empirically a good estimate for yielding reasonable results. (Documents in greater depths are usually duplicates with different names – URLs.)

Results
Our spider collected over 250 000 documents (in specific formats) and created a roughly 7 GB corpus. We found about 3.3 million links to those documents within the set of servers. We removed duplicate links and self-links (intra-site links). Duplicate links have the same source and target URL; self-links have a source and a target within the same server. After removal, there were 1 850 links left. The sites in Table VI.1 are ordered descendingly by the number of in-links (citations).

We can notice in Table VI.1 that the hosts are grouped into three clusters. At the top, there are three Web sites that are clearly ahead of the others. At the bottom, there are sites that have no or very few in-links. In between, there is the largest block of average departments. We show the number of the documents of our interest found on the individual servers as well. Of course, the number of in-links often depends on the number of documents on the target site. Their numbers vary greatly due to different sizes of hosting institutions (see also the constraints above), preference of various document formats and document generation (dynamic Web pages), etc. One way of tackling this problem is to normalize the number of citations somehow. For instance, it is possible to divide the number of citations by the number of documents on a particular site (the ratio in the last column of Table VI.1) or by the number of staff of the corresponding institution [Li2003]. In this context, it is interesting to note the very low total ratio. This means that in a closed set of Czech computer science institutions, the departments cite one another very rarely, which is somewhat astonishing.

Issues
There are some facts that may severely influence the ordering by in-links. One of them is the existence of server aliases. For instance, www.siteA.cz and www.siteB.zcu.cz is one machine with the same content. Thus, citations to both should be counted together. There may be a large number of aliases and ignoring them could lead to wrong results. It is not possible to replace host names with IP addresses either since more virtual servers can share one IP address.

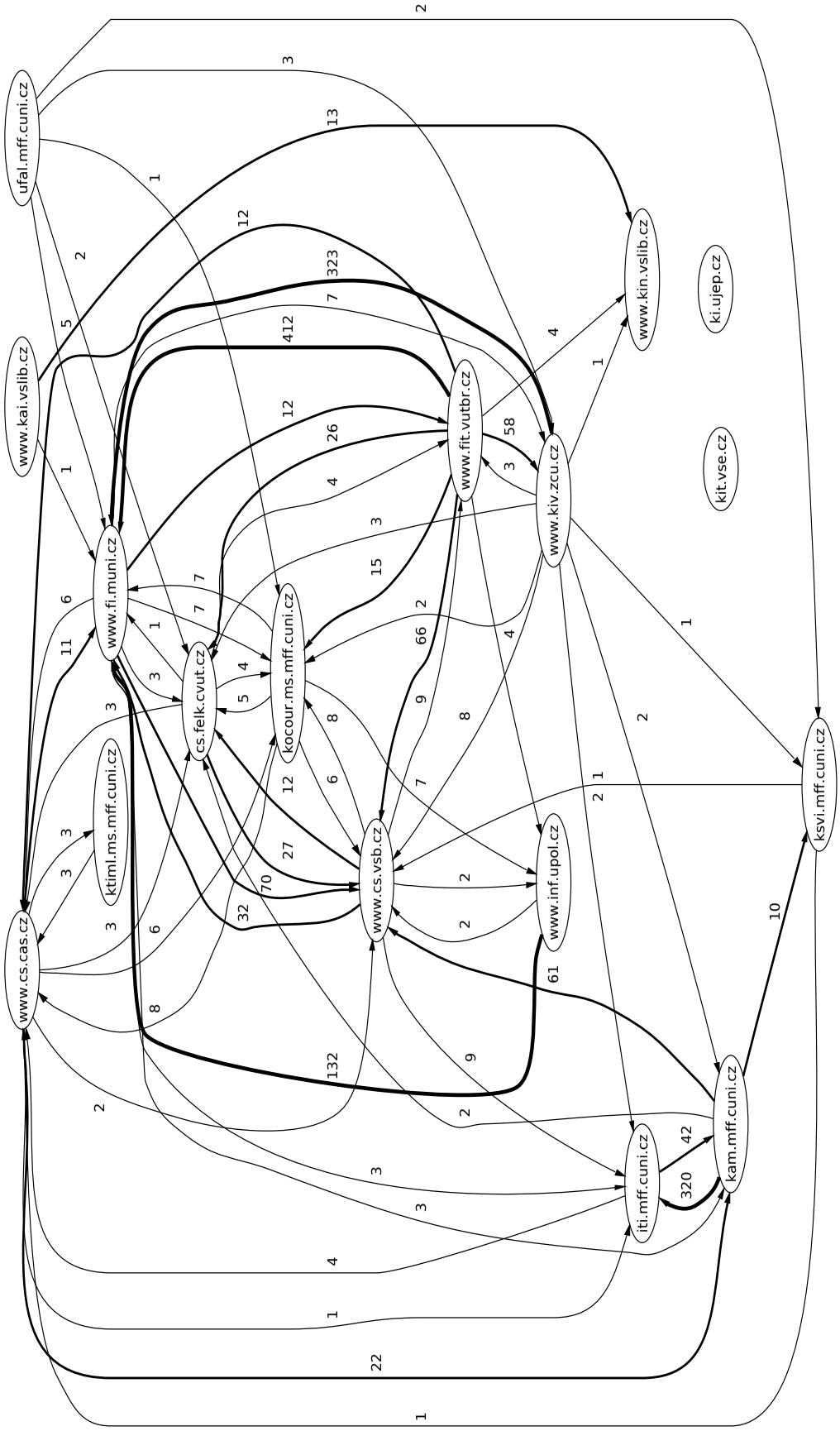


Figure VI.1: Citation graph of Czech Web sites.

Another problem is dynamically generated Web pages (see the Web site with a significantly higher number of documents). In such a case, two and more URLs (and two or more possible references) represent one document and citations should be counted only once then. This is very annoying, especially regarding the low inter-connectivity of the Web sites. Last but not least, there is a problem with document formats. If a server hosts documents in a format we ignore (e.g. rtf) to a greater extent than the other servers, it can automatically lose citations. All these issues (site mirrors, different site sizes, dynamic pages, etc.) must be taken into account when declaring the most authoritative institutions.

Server	# Docs	# In-Links	Ratio
www.fi.muni.cz	15 438	924	0.0599
iti.mff.cuni.cz	632	335	0.5301
www.cs.vsb.cz	18 325	243	0.0133
kam.mff.cuni.cz	10 952	69	0.0063
www.kiv.zcu.cz	12 309	68	0.0055
cs.felk.cvut.cz	16 422	56	0.0034
kocour.ms.mff.cuni.cz	11 860	43	0.0036
www.cs.cas.cz	3 226	37	0.0115
www.fit.vutbr.cz	148 682	28	0.0002
www.kin.vslib.cz	46	18	0.3913
www.inf.upol.cz	1 230	13	0.0106
ksvi.mff.cuni.cz	472	13	0.0275
ktiml.ms.mff.cuni.cz	847	3	0.0035
ki.ujep.cz	240	0	0
kit.vse.cz	273	0	0
ufal.mff.cuni.cz	8 316	0	0
www.kai.vslib.cz	2 423	0	0
Total	251 693	1 850	0.0074

Table VI.1: Czech Web sites analyzed.

Authoritative Institutions

The relations between the examined servers from Table VI.1 are depicted in Figure VI.1. The citation network is a directed graph with edge weights set to in-link numbers. To enhance visual perception we use three types of edges – normal width lines (less than ten citations), medium width lines, and thick lines (more than 99 citations). By simply looking at the network, we can immediately identify two major candidates for the most important hosts – www.fi.muni.cz and www.cs.vsb.cz. To verify it, we took advantage of the methods from Chapter II. First, we computed in-degrees of the nodes in the citation graph without respect to edge weights (i.e. each edge has a weight of one). Note that the in-links in Table VI.1 are actually in-degrees respecting edge weights. Then, we computed HITS authorities for the graph nodes and, finally, we generated PageRanks (HostRanks, in fact) for all of the nodes. Table VI.2 summarizes the rankings produced by all four algorithms.

We can see indeed that all four measures are strongly positively correlated. The hosts www.cs.vsb.cz and www.fi.muni.cz are in the top three servers whichever ranking method we applied; cs.felk.cvut.cz is highly ranked by In-Degree and HITS whereas www.cs.cas.cz is favoured by PageRank only. Number two by citations (in-links), iti.mff.cuni.cz, is handicapped by its strong support from more or less just one server as we may see in Figure VI.1. Naturally, the nodes (sites) with a zero in-degree end up at the bottom of each chart. Perhaps, we could prefer those with some out-links at least to those with a zero out-degree.

These nodes with no in-links and out-links are entirely isolated and do not participate in the community.

Site	In-Links	In-Deg	HITS	PageRank
cs.felk.cvut.cz	6	3	1	4
iti.mff.cuni.cz	2	6	5	6
kam.mff.cuni.cz	4	7 – 8	8	7
ki.ujep.cz	14 – 17	14 – 17	14 - 17	14 – 17
kit.vse.cz	14 – 17	14 – 17	14 – 17	14 – 17
kocour.ms.mff.cuni.cz	7	4 – 5	4	5
ksvi.mff.cuni.cz	11	9 – 12	12	11
ktiml.ms.mff.cuni.cz	13	13	13	12
ufal.mff.cuni.cz	14 – 17	14 – 17	14 – 17	14 – 17
www.cs.cas.cz	8	4 – 5	6	2
www.cs.vsb.cz	3	1 – 2	2	1
www.fi.muni.cz	1	1 – 2	3	3
www.fit.vutbr.cz	9	7 – 8	7	8
www.inf.upol.cz	11	9 – 12	10	9
www.kai.vslib.cz	14 – 17	14 – 17	14 – 17	14 – 17
www.kin.vslib.cz	10	9 – 12	11	13
www.kiv.zcu.cz	5	9 – 12	9	10

Table VI.2: Algorithms and rankings of Czech Web sites.

Correlation

Now that we have four different rankings: by in-links, in-degree (each edge has a weight of one), HITS (authority), and PageRank, we are interested in the correlations between these orderings. The Spearman correlation coefficients for each pair of rankings are presented in Table VI.3. They are all significant at the 0.02 level. The very high positive correlation between the four rankings was expected as it had already been reported before [Ding2001b, Ding2002].

	In-Links	In-Degree	HITS	PageRank
In-Links	X	0.89	0.89	0.86
In-Degree	0.89	X	0.96	0.96
HITS	0.89	0.96	X	0.95
PageRank	0.86	0.96	0.95	X

Table VI.3: Czech rankings correlation.

VI.1.2 French University Computer Science Web Sites

In this section, we will describe our experiment with the Web sites of French computer science departments. This data collection was also in the field of interest of this dissertation's author, but it was much larger than the Czech data set and, therefore, it required a different treatment. First, we had to draw up a list of laboratories. To do this, we looked up in Web directories and we also submitted queries to Web search engines. From these Web pages, we manually selected 80 final sites that constituted our set of departments. The selection was limited by the same constraints we discussed in the context of Czech Web sites. The first goal was to determine the most authoritative sites as of May 2006.

Procedure

To accelerate the process of creating the Web graph, we did not make use of a Web spider of our own, but we took advantage of a service provided by the search engine Yahoo! We submitted to it queries in this form:

site:www.loria.fr linkdomain:www.irisa.fr

which returns the number of documents on www.loria.fr containing at least one link to documents on www.irisa.fr. For us, it is a weight of the edge from www.loria.fr to www.irisa.fr. We had to construct 6 320 queries in this way. Of course, the construction and submission of queries, storing of results, and the graph creation were automated. (The complete figure of the Web graph with 393 edges is available on the accompanying CD and at [4]; its sketch without node labels and edge weights is in Figure VI.2.)

The drawbacks of relying solely upon search engines are discussed a great deal in [Thelwall2001, Li2003]. The problem consists primarily in "instability" of the results. This means that the results obtained one day differ from those of another one. Another disadvantage is that the results are not transparent. We do not know which document formats are taken into account, how duplicate documents are treated, etc.

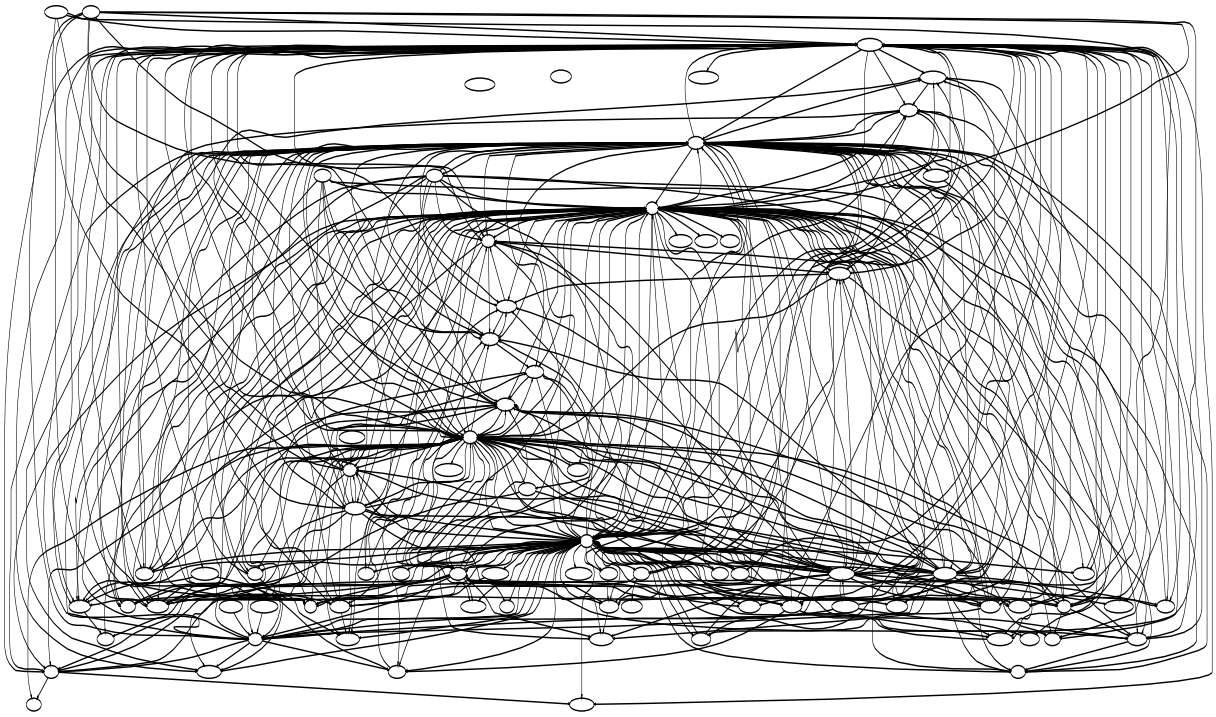


Figure VI.2: Citation graph of French Web sites.

Results and discussion

Again, we applied the four ranking methods to the Web graph of 80 sites of choice. We can see the results in Table VI.4 and Table VI.5. The sites are sorted by in-links (citations), i.e. by the total number of links to this site from other sites in the set (with some limitations imposed by the search engine). The first place belongs to www-futurs.inria.fr, whose positions achieved by the other methods, though, are much worse. We can suppose that the reason for this is a very strong support from a particular site. (After inspecting the Web graph, we can

see that it is www.lifl.fr.) The following sites always have high ranks - www-sop.inria.fr, www.loria.fr, www.lri.fr and www.lifl.fr. We can surely consider them as authoritative.

In-Links	Site	In-Degree	HITS	PageRank
1	www-futurs.inria.fr	45	41	53
2	www-sop.inria.fr	1	1	9
3	www.loria.fr	1	5	3
4	www.lri.fr	6	6	10
5	www-rocq.inria.fr	13	12	28
6	www.irisa.fr	4	3	18
7	www.lifl.fr	5	7	4
8	www.lix.polytechnique.fr	20	17	26
9	dpt-info.u-strasbg.fr	39	53	43
10	www.inrialpes.fr	6	8	2
11	www.irit.fr	9	4	8
12	www.liafa.jussieu.fr	13	15	39
13	www.lirmm.fr	1	11	1
14	www.labri.fr	13	13	30
15	www-leibniz.imag.fr	10	14	13
16	liris.cnrs.fr	13	16	11
17	www.prism.uvsq.fr	13	25	5
18	www.di.ens.fr	34	26	44
19	www.lip6.fr	20	21	40
20	www.laas.fr	6	2	27
21	dep-info.u-psud.fr	61	58	69
22	www-lil.univ-littoral.fr	25	34	35
23	www-verimag.imag.fr	25	37	16
24	www.i3s.unice.fr	25	31	7
25	eurise.univ-st-etienne.fr	25	23	32
26	www-lsr.imag.fr	34	26	37
27	www.info.unicaen.fr	13	10	14
28	www-timc.imag.fr	12	9	17
29	www-sic.univ-poitiers.fr	45	46	50
30	cedric.cnam.fr	25	22	38
31	www.dil.univ-mrs.fr	39	54	25
32	www-lmc.imag.fr	25	29	34
33	www.info.univ-angers.fr	34	44	24
34	lifc.univ-fcomte.fr	20	32	21
35	eric.univ-lyon2.fr	10	19	6
36	www-id.imag.fr	25	33	15
37	www-lipn.univ-paris13.fr	13	24	29
38	dept-info.labri.fr	25	18	36
39	www.isima.fr	39	43	48
40	sis.univ-tln.fr	20	28	12

Table VI.4: Ranking of French Web sites (1 – 40).

In-Links	Site	In-Degree	HITS	PageRank
41	www-clips.imag.fr	25	30	22
42	www.lisi.ensma.fr	39	40	33
43	www-info.iutv.univ-paris13.fr	61	69	72
44	www.lif.univ-mrs.fr	34	36	31
45	www.cril.univ-artois.fr	39	35	41
46	www.li.univ-tours.fr	34	42	45
47	citi.insa-lyon.fr	45	45	54
48	deptinfo.unice.fr	39	38	46
49	msi.unilim.fr	52	55	64
50	www.iut-info.univ-lille1.fr	61	62	65
51	www.lia.univ-avignon.fr	20	20	23
52	lil.univ-littoral.fr	52	48	57
53	lisi.insa-lyon.fr	45	39	47
54	www.isc.cnrs.fr	45	71	19
55	www.if.insa-lyon.fr	61	72	52
56	sirac.inrialpes.fr	61	62	62
57	phalanstere.univ-mlv.fr	45	65	20
58	www.lalic.paris4.sorbonne.fr	45	47	61
59	www.icp.inpg.fr	52	51	49
60	www-valoria.univ-ubs.fr	52	57	51
61	lihs.univ-tlse1.fr	52	48	60
62	www.epita.fr	52	67	42
63	llaic3.u-clermont1.fr	52	51	56
64	lsiit.u-strasbg.fr	52	48	57
65	liuppa.univ-pau.fr	52	56	66
66	wwwhds.utc.fr	61	66	55
67	www.depinfo.uhp-nancy.fr	61	68	59
68	lrlweb.univ-bpclermont.fr	61	62	62
69	www-lium.univ-lemans.fr	61	70	67
70	www.dptinfo.ens-cachan.fr	61	58	68
71	www.ai.univ-paris8.fr	61	58	69
72	www.lita.univ-metz.fr	61	58	69
73	dept-info.univ-brest.fr	73	73	73
74	lina.atlanstic.net	73	73	73
75	lis.snv.jussieu.fr	73	73	73
76	psiserver.insa-rouen.fr	73	73	73
77	www.listic.univ-savoie.fr	73	73	73
78	www-info.enst-bretagne.fr	73	73	73
79	www.info.iut.u-bordeaux1.fr	73	73	79
80	www.info.iut-tlse3.fr	73	73	79

Table VI.5: Ranking of French Web sites (41 – 80)

	In-Links	In-Degree	HITS	PageRank
In-Links	X	0.86	0.85	0.76
In-Degree	0.86	X	0.96	0.91
HITS	0.85	0.96	X	0.82
PageRank	0.76	0.91	0.82	X

Table VI.6: French rankings correlation.

The same difficulties as with the Czech sites persist – mirror sites, different logical Web sites (some departments may prefer separate sites for each of their projects), dynamic pages, etc. Moreover, some other errors introduced by the search engine may occur. The correlation between the individual rankings is rather high again (see Table VI.6).

	in-links	in-degree	out-links	out-degree
sum	5 160	393	5 160	393
min	0	0	0	0
max	917	15	1 476	54
avg	64.50	4.91	64.50	4.91
std. deviation	138.17	4.04	213.68	10.84
median	20,5	4	4	1
mode	0	1	0	0

Table VI.7: Statistics of the French Web graph.

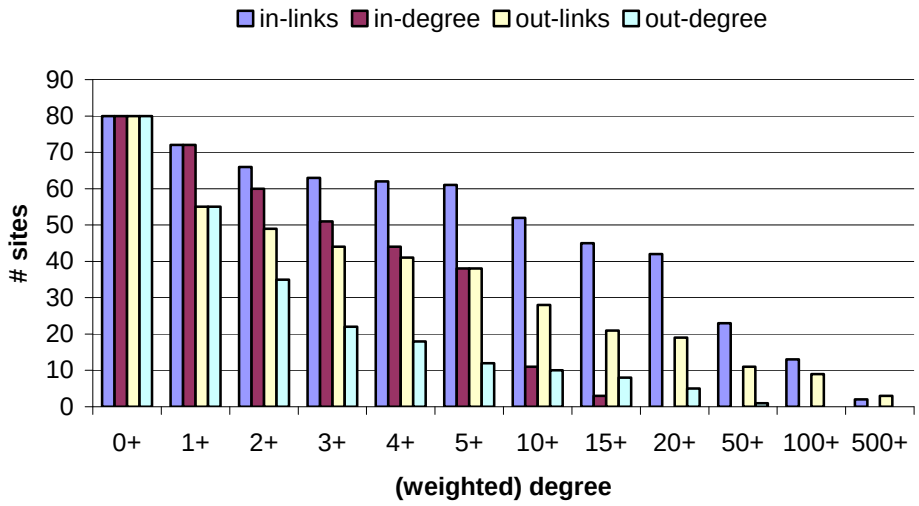


Figure VI.3: Cumulative distribution of degrees in the French Web graph.

Let us comment on some properties of the Web graph of French sites. Some statistics is shown in Table VI.7. Alphabetically sorted sites with the exact numbers of in-links, in-degree, out-links, and out-degree may be found in the appendix in Table 5 and Table 6. The Web graph has 80 nodes and 393 edges. 72 sites have some in-links, 55 sites have some out-links. 49 nodes have both a non-zero in-degree and a non-zero out-degree. Two Web sites are entirely isolated – www.info.iut.u-bordeaux1.fr and www.info.iut-tlse3.fr. They have no in-links as well as no out-links. The maximum in-degree (i.e. the maximum number of distinct sites citing a particular site) is 15 which is achieved by www.lirmm.fr, www.loria.fr, and www-sop.inria.fr. However, the maximum number of in-links to a site is much higher – 917 of www-futurs.inria.fr. www.lifl.fr and www.lri.fr are the top sites as for the out-links and out-degree (1 476 and 54, respectively).The median number of in-links is about twenty. Figure VI.3 shows the cumulative distribution of degrees in the French Web graph. If we have a more closer look at the sites that have much more in-links that out-links, these are www-futurs.inria.fr, www-rocq.inria.fr, www.lix.polytechnique.fr, dpt-info.u-strasbg.fr, www.inrialpes.fr, and www.liafa.jussieu.fr. The only Web sites that has strikingly more out-links than in-links is lsiit.u-strasbg.fr. (Though the two Strasbourg sites are more or less

complimentary.) The site with the most balanced relation between in-links and out-links is www.loria.fr (460 vs. 501).

VI.2 Mining the Content

The phase of finding significant institutions enables us to reduce the set of Web sites that we are going to analyze in the next stage in Section VI.2. For example, we might discard the last four sites in Table VI.1 or the last eight sites in Table VI.5, i.e. the least important sites. However, our case studies (Czech and French academic computer science Web sites) have still sufficiently small data sets so that no reduction is necessary. However, it might be inevitable with some very large data, such as American university Web sites. Measuring the quality of academic institutions with webometric tools is justified in [Li2003], where Web-based rankings correlated with official rankings.

VI.2.1 Czech Researchers

In addition to studying links in a collection of computer science Web sites, we were also interested in the documents themselves found on these Web sites. Thus, besides files containing hyperlinks (mainly HTML documents), we downloaded potential research papers as well. In practice, that meant collecting PDF and PostScript files because most research publications publicly accessible on the Web are in these two formats. First, we had to pre-process our download corpus. We unpacked archives and converted observed files to plain text via external utilities. So, at the beginning, we had a 12 thousand set of potential research papers. We discarded duplicates and examined the remaining documents. We used a simple rule to categorize the documents. In case they included some kind of references section they were considered as papers. In this way, we obtained some 3 600 papers in the end, i.e. over eight thousand documents did not look like research articles.

Information extraction

The next task is to extract information from the papers needed for citation analysis, i.e. names of authors, titles of papers, etc. We employ the same methodology with use of Hidden Markov Models (HMM) as that of McCallum and his colleagues [McCallum1999a, McCallum1999b, Seymore1999, McCallum2000]. A description of their approach is outside of the scope of this thesis. The difference is that we work with complete papers, not just with pre-processed headers and references. Moreover, the resulting text files analyzed by HMMs may often have been incorrectly converted to text before. Existence of diacritics in the Czech spelling also worsens the extraction. We did not measure the extraction accuracy due to lack of testing objects but, for the above reasons, we suppose it to be significantly lower than those 90 - 93% reported in [Seymore1999].

We stored the information to a database for a comfortable subsequent querying. The author citation graph G had over 28 000 non-isolated nodes and roughly 195 000 edges. Authors were represented by their surnames and their first name and, when applicable, middle name initials. Strictly said, words identified as surnames. Of course, many of these words were not surnames (they were incorrectly classified) or they were foreign surnames of international authors. From the citation graph with “surnames” as graph nodes we determined the most authoritative Czech authors using the three different ranking methods. (The recognition of a Czech surname was done manually for the top authors.) See Table VI.8 for details.

Rank	In-Degree	HITS	PageRank
1	Nešetřil J	Jančar P	Nešetřil J
2	Jančar P	Nešetřil J	Jančar P
3	Hajič J	Kučera A	Kučera A
4	Kučera A	Pala K	Pultr A
5	Matoušek J	Hajič J	Pala K
6	Panevová J	Oliva K	Smrž P
7	Pala K	Panevová J	Hajič J
8	Sgall P	Matoušek J	Panevová J
9	Kratochvíl J	Kratochvíl J	Matoušek J
10	Oliva K	Sgall P	Sedláček R

Table VI.8: Ten most authoritative Czech CS researchers.

Let us underline several facts. First, we did not disambiguate the names. Thus, a couple of authors may actually be represented by one name. Even adding first names does not resolve this problem. In addition, references in papers usually do not refer to full author names but to initials and surnames only. Thus, some mapping between these "short" names and full names is necessary. We contented ourselves with reducing even the full names in paper headers to short names and accepting some information loss. One solution of author disambiguation would be to cluster authors according to their co-authors or publication topics as it is done in [Han2005]. Authors report that this method works well with European (English) names but it achieves accuracy of only 60 – 70% with Chinese names. Second, duplicate citations are handled only in the sense that we remove duplicate documents before analysis. We do not examine whether two or more papers having perhaps only small differences are one publication in reality. Their references to another paper are counted separately.

Third, Czech names often contain diacritics. In international publications written in English, though, diacritics are left out sometimes. The spelling is not unified. Furthermore, conversion to plain text from PDF and PostScript files does not work well and produces more variants of one name. For instance, we found seven commonly used variations of the name "Hajič" in our database. In other words, names with no diacritics in their original spelling have a better chance to have their citations counted correctly. For all the surnames in Table VI.8, we tried to include their frequent versions in citations. The two-way name ambiguity (one author may be known under more names and one name may represent a couple of authors) is to be reflected in future improvements. For all these reasons, the actual citation numbers are less interesting than the ranking itself. Let us not forget that the ranking is a result of those 3 600 papers we got. The question is how it would change if more papers were analyzed.

Discussion

Again, no duplicate edges and self-citations were allowed in the citation graph of authors. The only two authors occurring among the top three researchers for each method are "Nešetřil J" and "Jančar P". Other highly ranked names include "Kučera A" or "Hajič J". Some of the names (such as "Kučera", "Matoušek", or "Sedláček") are very frequent Czech names and they might require further disambiguation even if we know the domain (computer science) and the first name initial.

Looking mostly just at the first page of results returned by a search engine we can make a guess about the probable affiliations of the authors. For example, for "Hajič" we got

ufal.mff.cuni.cz, for "Kučera" we obtained www.fi.muni.cz and kam.mff.cuni.cz, and for "Matoušek" we got kam.mff.cuni.cz and www.fit.vutbr.cz. When comparing the sites of these authoritative researchers to those in Table VI.8, we may observe that ufal.mff.cuni.cz, kam.mff.cuni.cz, and www.fit.vutbr.cz have no high positions there. Only www.fi.muni.cz is ranked high. Therefore, it is unclear what impact highly cited authors have on the importance of their institutions' Web sites. It would have to be submitted to an extensive research.

VI.2.2 French Researchers

We also gradually crawled all of the French sites and thus obtained a nearly 40 GB corpus of downloaded documents. So, at the beginning, we had about 45 thousand potential research papers. We treated them in the same way as the "Czech" articles and we obtained some 16 000 papers in the end. The final citation graph of authors *G* (without duplicate edges and self-citations) had almost 86 000 non-isolated nodes and about 477 000 edges. Unlike the Czech authors in Section VI.2.1, surnames alone did not turn out to be very discriminative. Thus, authors were represented by surnames and initials of their first and middle names. See Table VI.9 for details.

Rank	In-Degree	HITS	PageRank
1	Halbwachs N	Halbwachs N	Cahon S
2	Caspi P	Caspi P	Berry G
3	Sifakis J	Sifakis J	Filiol E
4	Berry G	Berry G	Halbwachs N
5	Benveniste A	Benveniste A	Zhang Z
6	Abiteboul S	Nicollin X	Benveniste A
7	Maler O	Cousot R	Lavallée S
8	Nicollin X	Raymond P	Dombre E
9	Cousot P	Cousot P	Boudet S
10	Cousot R	Abiteboul S	Dégoulange E
11	Raymond P	Maler O	Gourdon A
12	Bouajjani A	Asarin E	Abiteboul S
13	Asarin E	Comon H	Charpin P
14	Comon H	Bouajjani A	Carlet C
15	Zhang Z	Coupaye T	Cohen G
16	Berstel J	Berstel J	Troccaz J
17	Meyer B	David B	Abdalla M
18	Florescu D	Arnold A	Payan Y
19	Baccelli F	Pilaud D	Cousot R
20	Leroy X	Bruneton E	David R
21	Bruneton E	Maraninchi F	Cousot P
22	Flajolet P	Meyer B	Caspi P
23	Arnold A	Leroy X	Sifakis J
24	Graf S	Bensalem S	Deransart P
25	Cohen J	Graf S	Maler O
26	Coupaye T	Tripakis S	Bouajjani A
27	Pilaud D	Lakhnech Y	Dubois D
28	Lakhnech Y	Bozga M	Caron P
29	David R	Gautier T	Pierrot F
30	Faugeras O	Liu J	Raymond P

Table VI.9: Authoritative French CS researchers.

Results and discussion
The rankings produced by In-Degree and HITS are very similar (the top five researchers are exactly the same) whereas that by PageRank is rather different. The authors in In-Degree and HITS are more or less the same (only in various positions), but PageRank introduces some new names. However, there are two authors ("Halbwachs N" and "Berry G") occurring in top five of each ranking. We can certainly call these researchers authorities.

Deciding whether or not a researcher is French is inherently subjective. Our decision was based on searching with several general and specialized search engines. Ideally, we found the researcher's home page hosted by a French Web site or affiliation to a French institution given in an article. Of course, by French authors we also mean those who had lived and worked in France for a long time. We are aware that this feature is particularly fuzzy. Even with first name initials there are certainly more individuals with the same name. Again, the question is how the rankings would change if more than those 16 000 papers were analyzed.

VI.3 Summary & Future Work

Summary
Notions of popularity or authority, commonly used in social networks such as scientific publications, have also been adopted for the World Wide Web in recent years. The most popular ranking techniques are link-based methods like In-Degree, PageRank, and HITS. We present a methodology and two case studies of finding authoritative researchers on the Web. We applied these algorithms to a set of Czech and a set of French academic computer science Web sites and determined the most authoritative ones within each set. (We also tried to examine Slovak computer science departments, but the data set was too small.)

This step normally enables reducing the volume of data to be analyzed since we could continue finding researchers on the more important sites only. Further, we analyzed the research papers publicly available on the sites and we determined the most significant researchers by applying several ranking techniques to the citation graph. The method is a relatively objective means of presenting facts, but the interpretation is necessarily subjective. The results we achieved are not quite reliable due to the constraints and problems mentioned above, but we believe that our methodology is practical as we have shown in our experiments.

International authors
Unlike Section V.3.2, we do not provide exact information on the co-authorship and citation graphs (including statistics and histograms) in Section VI.2. Neither do we present the results of the PageRank-based methods introduced in Chapter V. We are aware that the Web-based bibliographic data are very incomplete and inaccurate. There is a great deal of noise. Therefore, it does not make much sense to attempt to be too accurate in this case. Even the rankings in Table VI.8 and in Table VI.9 should be considered as a hint rather than some precise measurements. However, all this information may be found on the accompanying CD including the complete graphs and rankings in the form of database tables.

To allow for some minimum comparison at least, Table 7 in the appendix shows top 40 international authors for three basic ranking methods applied to both the Czech and the French corpus. There are names of authors of all nationalities without diacritics and only with some evident inaccuracies removed. We summarize the numbers of common researchers in the Top 40 for each pair of rankings in Table VI.10. Apparently, rankings based on one corpus tend to be more similar than those from two corpora. The largest intersection is between HITS and in-degree rankings for each corpus (29 common scientists in the Czech

data and 32 in the French corpus). On the other hand, there is hardly any intersection between PageRank from one country and other methods from the other country. Nevertheless, there are a couple of authors who occur at the top in both countries – “Bouajjani A”, “Ullman J D”, and “Vardi M”. These scientists seem to be regarded as authoritative by both Czech and French computer science researchers. In addition, “Ullman J D” is one of the ACM SIGMOD E. F. Codd Innovations Award winners (see Section V.3). Another award winner is “Abiteboul S’ who appears among the top authors in the French corpus only.

	CZ InD	CZ HITS	CZ PR	FR InD	FR HITS	FR PR
CZ InD	X	29	16	5	4	1
CZ HITS	29	X	14	5	3	2
CZ PR	16	14	X	0	0	0
FR InD	5	5	0	X	32	11
FR HITS	4	3	0	32	X	9
FR PR	1	2	0	11	9	X

Table VI.10: Common authors in Top 40.

Future work

In the future, we would like to have yet another ranking for institutions based on citations in papers. This would mean enhancing assigning affiliations to each researcher. We will be interested in the difference between the top ranked sites determined via analysis of Web links on one hand and those based on paper citations on the other hand. We would like to discover any correlation between the link-based (Web) and citation-based (papers) ranking. The social networks formed by academic institutions and by their research publications are assumed to be different. They are each destined for a distinct audience. Nevertheless, in our future research we would like to concentrate on the issue of combining Web and paper authorities. The methodology we have developed is general, which will enable us to focus on other areas of the Web as well.

To the best of our knowledge, the two case studies presented above are the first attempt ever made at finding authoritative researchers in those two countries by directly mining from unstructured Web data.

Conclusions

Web mining is an exciting area of research. Although quite new (who has heard of it fifteen years ago?), it has been subject to study to such extent in recent years that the body of knowledge is growing constantly and so fast that survey articles and books do not catch up with covering this topic. It spans across many scientific disciplines including artificial intelligence, machine learning, data mining, knowledge acquisition, information retrieval, graph theory and others. It borrows concepts and techniques from these domains, and it enriches them with novel methods, algorithms, approaches, and empirical observations that turn out to be of a more general interest. Perhaps the most interesting finding so far of studying the Web is that it is developing into something more than we hoped. The patterns and regularities discovered in its scope, content, structure, usage, and behaviour disclose something amazing. It is no more just a network of documents. It is a kind of *living organism*. How will it evolve in the future? Is there something more we could know about it? With the arrival of Web 2.0 and the semantic Web even more space for research will be available, and I predict that, in the next decade at least, the study of Web mining techniques will be no less challenging than it has been until now.

Disclaimer

The eminent feature of the Web that excludes direct applications of classic information retrieval processes is its volatility and infinity. Web documents and links between them may change on a daily basis or even more often, and the Web sample we are analyzing is always "a picture of the past". It is never the true, real Web of a given moment, and it must be treated as such. We can never know precisely how much of the information on the Web we actually have at our disposal, how much is still hidden and yet to be discovered, and, therefore, we can never measure recall, a fundamental metric in information retrieval, but only make a guess about it. Another characteristic is its decentralized and "democratic" nature. It is a product of millions of humans and human-controled machines that can, more or less arbitrarily, modify its content and structure. There is no regulatory body, and it governs itself. As in each democracy, there is some self-control, but discrepancies are common. Therefore, all the knowledge mined form the Web is affected by the factors above, and we should avoid to draw too far-reaching conclusions from it.

Main contributions

In this doctoral dissertation, I concentrated on the issue of mining the Web structure in order to find authoritative sources. Besides surveying the current progress in related areas such as Web models, crawling techniques, ranking algorithms, and social networks, I made the following research contributions:

- **PageRank for bibliographic networks**. I proposed a modification of the well-known PageRank equation, this time suited for graphs of citations between publications and collaborations between authors. I extended and generalized the notions of collaboration frequency and co-authorship exclusivity by Liu, Bollen et al. by deriving them directly from the co-authorship graph and combining them with the information from the citation graph. Intuitively, this enables to rank authors "more fairly" by significance taking into account not only citations but also collaborations between them. In total, I proposed seven variants of the "bibliographic PageRank" formula. To test this new approach on real and non-noisy data, I applied the ranking algorithms to a data set from the DBLP digital library and used the methodology of Sidiropoulos

and Manolopoulos for ranking comparisons. I compared author rankings to a list of *ACM SIGMOD E. F. Codd Innovations Award* winners and found out that the new rankings reflected much better the prize awarding scheme than the baseline, "standard" PageRank ranking. It is not possible to compare directly my results with those of Sidiropoulos et al., because they utilized a slightly different data set and their method is primarily destined for publications, not for authors. This research contribution is described in Chapter V.

- **Mining the Czech and French academic Web**. I attempted to determine authoritative institutions from two collections of Czech and French computer science University Web sites by applying some well-known methods for exploiting the Web structure. Furthermore, I analyzed the contents of documents found on these Web sites, more specifically of research papers. Using existing techniques of information extraction, I found out the most significant Czech and French computer science researchers that can be retrieved from documents available on the Web. The approach I brought into play is not new but the application and synthesis of several data mining processes yes. The results are certainly influenced by the limitations I faced and the selections I made. Especially the data for author retrieval is quite noisy, and I even do not present all the results, although they are all available on the CD accompanying this thesis. Detection of authoritative sites and authors may be helpful to decision makers and funding agencies in their personal and financial policies. To the best of my knowledge, my experiments are the first attempts published at finding influential Czech and French computer science authors by directly mining from Web data. This research is close to the work of Mike Thelwall in some aspects and is explained in Chapter VI.

Future work

My research efforts are far from being accomplished. As their natural continuation I see in particular:

- **Stability and sensitivity analysis**. Analysis of stability and sesitivity of the bibliographic PageRank formula (5.2) to small perturbations in the citation and/or co-authorship graph would be desirable. Although the standard PageRank has been shown to be relatively stable (see Section II.2.6), the larger number of parameters involved in the calculation of (5.2) may negatively affect this property.

- **Inclusion of time**. The concept of a "fairer" ranking of researcher based not only on citations but also on collaborations invites the inclusion of the time factor. A citation between two scientists should certainly have a different meaning when it is made after their co-authorship of many articles or long before they get to know each other. This enhancement might add even more "justice" to the ranking.

- **Comparison of Web-based and paper-based authorities**. The ranking of institutions represented by their Web sites in Section VI.1 is based purely on Web links. It would be interesting to associate affiliations with authoritative researchers from Section VI.2 and to compare the two institutional rankings. I also see a great potential of the CiteSeer data (see Section IV.2) with affiliations already assigned, which may be useful for this purpose as well.

References

Printed

[Abiteboul2003] Abiteboul S., Preda M., Cobena G. *Adaptive on-line page importance computation*. Proceedings of the 12th international conference on World Wide Web (WWW'03), Budapest, Hungary, pp. 280-290, 2003.

[Aiello2000] Aiello W., Chung F., Lu L. *A random graph model for massive graphs*. Proceedings of the 32nd annual ACM symposium on Theory of computing, Portland, Oregon, USA, pp. 171-180, 2000.

[An2004] An Y., Janssen J., Milios E. E. *Characterizing and Mining the Citation Graph of the Computer Science Literature*. Knowledge and Information Systems, vol. 6, no. 6, pp. 664-678, 2004.

[Baeza-Yates2004] Baeza-Yates R., Castillo C. *Crawling the infinite Web: five levels are enough*. Proceedings of the third Workshop on Web Graphs (WAW), Rome, Italy, Lecture Notes in Computer Science, Springer, vol. 3243, pp. 156-167, 2004.

[Baeza-Yates2005] Baeza-Yates R., Castillo C., Marín M., Rodríguez A. *Crawling a country: better strategies than breadth-first for web page ordering*. Proceedings of the 14th international conference on World Wide Web (WWW 2005), Chiba, Japan, pp. 864-872, 2005.

[Balmin2004] Balmin A., Hristidis V., Papakonstantinou Y. *ObjectRank: Authority-Based Keyword Search in Databases*. Proceedings of the 30th International Conference on Very Large Data Bases (VLDB 2004), Toronto, Canada, pp. 564-575, 2004.

[Bani-Ahmad2005] Bani-Ahmad S., Cakmak A., Özsoyoglu G., Al-Hamdani A. *Evaluating Publication Similarity Measures*. IEEE Data Engineering Bulletin, vol. 28, no. 4, pp. 21-28, 2005.

[Barabási1999] Barabási A. L., Albert R. *Emergence of Scaling in Random Networks*, Science, vol. 286, no. 5439, pp. 509-512, 1999.

[Berkhin2005] Berkhin P. *A Survey on PageRank Computing*. Internet Mathematics, vol. 2, no. 1, pp. 73-120, 2005.

[Bharat2000] Bharat K., Bröder A., Dean J., Henzinger M. R. *A comparison of techniques to find mirrored hosts on the WWW*. Journal of the American Society for Information Science, vol. 51, no. 12, pp. 1114-1122, 2000.

[Bianchini2005] Bianchini M., Gori M., Scarselli F. *Inside PageRank*. ACM Transactions on Internet Technology, vol. 5, no. 1, pp. 92-128, 2005.

[Boldi2004a] Boldi P., Santini M., Vigna S. *Do your worst to make the best: Paradoxical effects in pagerank incremental computations*. Proceedings of the third Workshop on Web Graphs (WAW), Rome, Italy, Lecture Notes in Computer Science, Springer, vol. 3243, pp. 156-167, 2004.

[Boldi2004b] Boldi P., Codenotti B., Santini M., Vigna S. *UbiCrawler: a scalable fully distributed Web crawler*. Software Practice and Experience, vol. 34, no. 8, pp.711-726, 2004.

[Bollen2006] Bollen J., Rodriquez M. A., Van de Sompel H. *Journal status*. Scientometrics, vol. 69, no. 3, pp. 669-687, 2006.

[Bordons2002] Bordons M., Fernández M. T., Gómez I. *Advantages and limitations in the use of impact factor measures for the assessment of research performance in a peripheral country*. Scientometrics, vol. 53, no. 2, pp. 195-206, 2002.

[Bornmann2005] Bornmann L., Daniel H.-D. *Does the h-index for ranking of scientists really work?* Scientometrics, vol. 65, no. 3, pp. 391-392, 2005.

[Bornmann2007] Bornmann L., Daniel H.-D. *What do we Know About the h Index?* Journal of the American Society for Information Science and Technology, vol. 58, no. 9, pp. 1-5, 2007.

[Bouklit2005] Bouklit M., Mathieu F. *BackRank: an alternative for PageRank?* Proceedings of the 14th international conference on World Wide Web (WWW 2005), Chiba, Japan, pp. 1122-1123, 2005.

[Braun2006] Braun T., Glänzel W., Schubert A. *A Hirsch-type index for journals*. Scientometrics, vol. 69, no. 1, pp. 169-173, 2006.

[Brin1998] Brin S., Page L. *The Anatomy of a Large-Scale Hypertextual Web Search Engine*. Proceedings of the 7th World Wide Web Conference, pp. 107 – 117, 1998.

[Bröder1997] Bröder A., Glassman S. C., Manasse M. S., Zweig G. *Syntactic clustering of the Web*. Computer Networks and ISDN Systems, vol 29, no. 8-13, pp. 1157-1166, 1997.

[Bröder2000] Bröder A., Kumar R., Maghoul F., Raghavan P., Rajagopalan S., Stata R., Tomkins A., Wiener J. *Graph structure in the Web*. Computer Networks vol. 33, no. 1-6, pp. 309–320, 2000.

[Burner1997] Burner M. *Crawling towards Eternity*. Web Techniques, vol. 2, no. 5, pp. 37-40, 1997.

[Cai2005] Cai D., Shao Z., He X., Yan X., Han J. *Community Mining from Multi-relational Networks*. Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2005), Porto, Portugal, pp. 445-452, 2005.

[Chakrabarti1998] Chakrabarti S., Dom B. E., Gibson D., Kumar R., Raghavan P., Rajagopalan S., Tomkins A. *Spectral Filtering for Resource Discovery*. Proceedings of the ACM SIGIR Workshop on Hypertext Information Retrieval on the Web, Melbourne, Australia, pp. 13-21, 1998.

[Chakrabarti1999] Chakrabarti S., van den Berg M., Dom B. *Focused Crawling: A New Approach for Topic-Specific Resource Discovery*. Computer Networks, vol. 31, no. 11-16, pp. 1623-1640, 1999.

[Chakrabarti2002] Chakrabarti S. *Mining the Web: Analysis of Hypertext and Semi Structured Data*. Morgan Kaufmann Publishers, San Francisco, California, USA, 2002.

[Chakrabarti2006] Chakrabarti D., Faloutsos C. *Graph mining: Laws, generators, and algorithms*. ACM Computing Surveys, vol. 38, no. 1, 2006.

[Chau2003] Chau M., Chen H. *Comparison of three vertical search spiders*. Computer, vol. 36, no. 5, pp. 56-62, 2003.

[Chen1999] Chen C., Carr L. *Trailblazing the Literature of Hypertext: Author Co-Citation Analysis (1989-1998).* Proceedings of the 10th ACM Conference on Hypertext and Hypermedia: Returning to Our Diverse Roots (Hypertext '99), Darmstadt, Germany, pp. 51-60, 1999.

[Cho1998] Cho J., Garcia-Molina H., Page L. *Efficient Crawling Through URL Ordering*. Proceedings of the 7th international conference on the

World Wide Web (WWW7), Brisbane, Australia, pp. 161-172, 1998.

[Cho2000] Cho J, Shivakumar N., Garcia-Molina H. *Finding Replicated Web Collections*. Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, Dallas, Texas, USA, pp. 355-366, 2000.

[Cho2002] Cho J., Garcia-Molina H. *Parallel crawlers*. Proceedings of the 11th international conference on the World Wide Web (WWW'02), Honolulu, Hawaii, USA, pp. 124-135, 2002.

[Cunningham1997] Cunningham S. J., Dillon S. M. *Authorship patterns in information systems*. Scientometrics, vol. 39, no. 1, pp. 19-27, 1997.

[Desikan2005] Desikan P. K., Pathak N., Srivastava J., Kumar V. *Incremental page rank computation on evolving graphs*. Proceedings of the 14th international conference on World Wide Web (WWW 2005), Chiba, Japan, pp. 1094-1095, 2005.

[Diligenti2000] Diligenti M., Coetzee F., Lawrence S., Giles C. L., Gori M. *Focused crawling using context graphs*. Proceedings of the 26th International Conference on Very Large Data Bases (VLDB 2000), Cairo, Egypt, pp. 527-534, 2000.

[Diligenti2004] Diligenti M., Gori M., Maggini M. *A Unified Probabilistic Framework for Web Page Scoring Systems*. IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 1, pp. 4–16, 2004.

[Dill2002] Dill S., Kumar R., McCurley K. S., Rajagopalan S., Sivakumar D., Tomkins A. *Self-Similarity In the Web*. ACM Transactions on Internet Technology, vol. 2, no. 3, pp. 205-223, 2002.

[Ding2001a] Ding C., He X., Husbands P., Zha H., Simon H. *Link Analysis: Hubs and Authortities on the World Wide Web*. Lawrence Berkeley National Laboratory, University of California, Berkeley, California, USA, Technical Report 47847, May 2001.

[Ding2001b] Ding C., He X., Husbands P., Zha H., Simon H. *PageRank, HITS and a Unified Framework for Link Analysis*. Lawrence Berkeley National Laboratory, University of California, Berkeley, California, USA, Technical Report 49372, Nov. 2001.

[Ding2002] Ding C., He X., Husbands P., Zha H., Simon H. *PageRank, HITS and a Unified Framework for Link Analysis*. Proceedings of the 25th ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland, pp. 353–354, 2002.

[Donato2007] Donato D., Laura L., Leonardi S., Millozzi S. *The Web as a graph: How far we are*. ACM Transactions on Internet Technology, vol. 7, no. 1, 2007.

[Elmacioglu2005] Elmacioglu E., Lee D. *On six degrees of separation in DBLP-DB and more*. SIGMOD Record, vol. 34, no. 2, pp. 33-40, 2005.

[Farkas2002] Farkas I., Derényi I., Jeong H., Néda Z., Oltvai Z. N., Ravasz E., Schubert A., Barabási A. L., Vicsek T. *Networks in life: scaling properties and eigenvalue spectra*. Physica A: Statistical Mechanics and its Applications, vol. 314, no. 1-4, pp. 25-34, 2002.

[Garfield1979] Garfield E. *Citation Indexing: Its Theory and Application in Science, Technology, and Humanities*. John Wiley & Sons, New York, 1979.

[Garfield1999] Garfield E. *Journal impact factor: a brief review*. Canadian Medical Association Journal, vol. 161, no. 8, pp. 979-980, 1999.

[Ghemawat2003] Ghemawat S., Gobioff H., Leung S.-T. *The Google file system*.

Proceedings of the 19th ACM symposium on Operating systems principles, Bolton Landing, NY, USA, pp. 29-43, 2003.

[Gibson1998] Gibson D., Kleinberg J., Raghavan P. *Inferring Web Communities from Link Topology*. Proceedings of the 9th ACM Conference on Hypertext and Hypermedia, Pittsburgh, PA, pp. 225–234, 1998.

[Gulli2005] Gulli A., Signorini A. *The indexable web is more than 11.5 billion pages*. Proceedings of the 14th international conference on World Wide Web (WWW 2005), Chiba, Japan, pp. 902-903, 2005.

[Hafri2004] Hafri Y., Djeraba C. *High performance crawling system*. Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval, New York, NY, USA, pp. 299-306, 2004.

[Han2005] Han H., Zha H., Giles C. L. *Name Disambiguation in Author Citations Using a K-way Spectral Clustering Method*. Proceedings of the 5th ACM/IEEE-CS International Conference on Digital Libraries, Denver, CO, pp. 334-343, 2005.

[Harter1997] Harter S. P., Nisonger T. E. *ISI's impact factor as misnomer: A proposed new measure to assess journal impact*. Journal of the American Society for Information Science, vol. 48, no. 12, pp. 1146-1148., 1997.

[Hassan2004] Hassan H. E., Holt R. C. *The Small World of Software Reverse Engineering*. Proceedings of the 11th Working Conference on Reverse Engineering (WCRE 2004), Delft, Netherlands, pp. 278-283, 2004.

[He2002] He S., Spink A. *A comparison of foreign authorship distribution in JASIST and the journal of documentation*. Journal of the American Society for Information Science and Technology, vol. 53, no. 11, pp. 953-959, 2002.

[Heydon1999] Heydon A., Najork M. *Mercator: A scalable, extensible Web crawler*. World Wide Web, vol. 2, no. 4, pp. 219-229, 1999.

[Hirsch2005] Hirsch J. E. *An index to quantify an individual's scientific research output*. Proceedings of the National Academy of Sciences, vol. 102, no. 46, pp. 16569-16572, 2005.

[Kessler1963] Kessler M. M. *Bibliographic Coupling Between Scientific Papers*. American Documentation, vol. 14, no. 1, pp. 10-25, 1963.

[Kim2004] Kim S., Whitehead J. E. Jr. *Properties of academic paper references*. Proceedings of the 15th ACM Conference on Hypertext and Hypermedia (Hypertext 2004), Santa Cruz, California, USA, pp. 44-45, 2004.

[Kleinberg1999a] Kleinberg J., Kumar R., Raghavan P., Rajagopalan S., Tomkins A. *The web as a graph: Measurements, models and methods*. Proceedings of the 5th Annual International Conference on Combinatorics and Computing, Tokyo, Japan, Lecture Notes in Computer Science, Springer, vol. 1627, pp. 1-17, 1999.

[Kleinberg1999b] Kleinberg J. *Authoritative Sources in a Hyperlinked Environment*. Journal of the ACM, vol. 46, no. 5, pp. 604-632, 1999.

[Kumar1999] Kumar R., Raghavan P., Rajagopalan S., Tomkins A. *Trawling the Web for emerging cyber-communities*. Computer Networks, vol. 31, no. 11-16, pp. 1481-1493, 1999.

[Langville2003] Langville A. N., Meyer C. D. *Deeper Inside PageRank*. Internet Mathematics, vol. 1, no. 3, pp. 335-380, 2003.

[Langville2005] Langville A. N., Meyer C. D. *A Survey of Eigenvector Methods for*

Web Information Retrieval. SIAM Review, vol. 47, no. 1, pp. 135-161, 2005.

[Larson1996] Larson R. *Bibliometrics of the World Wide Web: An exploratory analysis of the intellectual structure of cyberspace*. Proceedings of the 59th Annual Meeting of the American Society for Information Science (ASIS'96), Baltimore, Maryland, USA, pp. 71-78, 1996.

[Lawrence1999] Lawrence S., Giles C. L., Bollacker K. *Digital Libraries and Autonomous Citation Indexing*. IEEE Computer, vol. 32, no. 6, pp. 67–71, 1999.

[Lewison2002] Lewison G. *Researchers' and users' perceptions of the relative standing of biomedical papers in different journals*. Scientometrics, vol. 53, no. 2, pp. 229-240, 2002.

[Ley2006] Ley M., Reuther P. *Maintaining an Online Bibliographical Database: The Problem of Data Quality*. Actes des sixièmes journées Extraction et Gestion des Connaissances (EGC'2006), Lille, France, Revue des Nouvelles Technologies de l'Information, vol. RNTI-E-6, pp. 5-10, 2006.

[Li2000] Li W.-S., Kolak O., Vu Q., Takano H. *Defining logical domains in a web site*. Proceedings of the 11th ACM Conference on Hypertext and Hypermedia (Hypertext 2000), San Antonio, Texas, USA, pp. 123-132, 2000.

[Li2003] Li X., Thelwall M., Musgrove P., Wilkinson D. *The Relationship Between the WIFs or Inlinks of Computer Science Departments in UK and Their RAE Ratings or Research Productivities in 2001*. Scientometrics, vol. 57, no. 2, pp. 239-255, 2003.

[Liu2005] Liu X., Bollen J., Nelson M. L., Van de Sompel H. *Co-authorship Networks in the Digital Library Research Community*. Information Processing and Management, vol. 41, no. 6, pp. 1462-1480, 2005.

[McCain1992] McCain K. W. *Core journal networks and cocitation maps in the marine sciences: Tools for information management in interdisciplinary research*. Proceedings of the 55th Annual Meeting of the American Society for Information Science (ASIS'92), Medford, New Jersey, USA, pp. 3-7, 1992.

[McCallum1999a] McCallum A., Nigam K., Rennie J., Seymore K. *Building Domain-Specific Search Engines with Machine Learning Techniques*. Proceedings of the AAAI Spring Symposium on Intelligent Agents in Cyberspace, Stanford, California, USA, pp. 28-39, 1999.

[McCallum1999b] McCallum A., Nigam K., Rennie J., Seymore K. *A Machine Learning Approach to Building Domain-Specific Search Engines*. Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI99), Stockholm, Sweden, pp. 662-667, 1999.

[McCallum2000] McCallum A. K., Nigam K., Rennie J., Seymore K. *Automating the Construction of Internet Portals with Machine Learning*. Information Retrieval Journal, vol. 3, no. 2, pp. 127-163, 2000.

[Mohan2005] Mohan B. K. *Searching Association Networks for Nurturers*. IEEE Computer, vol. 38, no. 10, pp. 54-60, 2005.

[Najork2001] Najork M., Wiener J. L. *Breadth-first crawling yields high-quality pages*. Proceedings of the 10th international conference on the World Wide Web (WWW10), Hong Kong, pp. 114-118, 2001.

[Nascimento2003] Nascimento M. A., Sander J., Pound J. *Analysis of SIGMOD's co-*

authorship graph. SIGMOD Record, vol. 32, no. 3, pp. 8-10, 2003.

[Nederhof2001] Nederhof A. J., Luwel M., Moed H. F. *Assessing the quality of scholarly journals in Linguistics: An alternative to citation-based journal impact factors*. Scientometrics, vol. 51, no. 1, pp. 241-265, 2001.

[Nevill-Manning1998] Nevill-Manning C. G., Reed T., Witten I. H. *Extracting Text from PostScript*. Software Practice and Experience, vol. 28, no. 5, pp. 481–491, 1998.

[Newman2003] Newman M. E. J. *The Structure and Function of Complex Networks*. SIAM Review, vol. 45, no. 2, pp. 167-256, 2003.

[Ng2001a] Ng A. Y., Zheng A. X., Jordan. M. I. *Stable algorithms for link analysis*. Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 2001), New Orleans, Louisiana, USA, pp. 258-266, 2001.

[Ng2001b] Ng A. Y., Zheng A. X., Jordan. M. I. *Link Analysis, Eigenvectors and Stability*. Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI 2001), Seattle, Washington, USA, pp. 903-910, 2001.

[Ntoulas2004] Ntoulas A., Cho J., Olston C. *What's new on the web?: the evolution of the web from a search engine perspective*. Proceedings of the 13th international conference on the World Wide Web (WWW '04), New York, NY, USA, pp. 1-12, 2004.

[Otte2002] Otte E., Rousseau R. *Social network analysis: a powerful strategy, also for the information sciences*. Journal of Information Science, vol. 28, no. 6, pp. 441-453, 2002.

[Page1999] Page L., Brin S., Motwani R., Winograd T. *The PageRank Citation Ranking: Bringing Order to the Web*. Computer Science Department, Stanford University, California, USA, Technical Report 1999-66, Nov. 1999.

[Pandurangan2002] Pandurangan G., Raghavan P., Upfal E. *Using PageRank to Characterize Web Structure*. Proceedings of the 8th Annual International Conference on Computing and Combinatorics (COCOON 2002), Singapore, Lecture Notes in Computer Science, vol. 2387, pp. 330-339, 2002.

[Pennock2002] Pennock D. M., Flake G. W., Lawrence S., Glover E. J., Giles C. L. *Winners don't take all: Characterizing the competition for links on the web*. Proceedings of the National Academy of Sciences, vol. 99, no. 8, pp. 5207-5211, 2002.

[Petříček2005] Petříček V., Cox I. J., Han H., Councill I. G., Giles C. L. *A Comparison of On-Line Computer Science Citation Databases*. Research and Advanced Technology for Digital Libraries, Proceedings of the 9th European Conference, ECDL 2005, Vienna, Austria, Lecture Notes in Computer Science, Springer, vol. 3652, pp. 438-449, 2005.

[Raan2006] van Raan A. F. J. *Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups*. Scientometrics, vol. 67, no. 3, pp. 491-502, 2006.

[Raghavan2001] Raghavan S., Garcia-Molina H. *Crawling the Hidden Web*. Proceedings of the 27th International Conference on Very Large Data Bases (VLDB), Rome. Italy, pp.129-138, 2001

[Rahm2005] Rahm E., Thor A. *Citation analysis of database publications*.

SIGMOD Record, vol. 34, no. 4, pp. 48-53, 2005.

[Rennie1999] Rennie J., McCallum A. *Using reinforcement learning to spider the web efficiently*. Proceedings of the 16th International Conference on Machine Learning, Bled, Slovenia, pp. 335-343, 1999.

[Saha2003] Saha S., Saint S., Christakis D. A. *Impact factor: a valid measure of journal quality*? Journal of the Medical Library Association, vol. 91, no. 1, pp. 42-46, 2003.

[Seglen1997] Seglen P. O. *Why the impact factor of journals should not be used for evaluating research*. British Medical Journal, vol. 314, no. 7079, pp. 498-502, 1997.

[Seymore1999] Seymore K., McCallum A., Rosenfeld R. *Learning Hidden Markov Model Structure for Information Extraction*. Proceedings of the AAAI'99 Workshop on Machine Learning for Information Extraction, Orlando, FL, pp. 37–42, 1999.

[Sidiropoulos2005] Sidiropoulos A., Manolopoulos Y. *A Citation-Based System to Assist Prize Awarding*. SIGMOD Record, vol. 34, no. 4, pp. 54-60, 2005.

[Sidiropoulos2005b] Sidiropoulos A., Manolopoulos Y. *A new perspective to automatically rank scientific conferences using digital libraries*. Information Processing and Management, vol. 41, no. 2, pp. 289-312, 2005.

[Sidiropoulos2006] Sidiropoulos A., Manolopoulos Y. *Generalized comparison of graph-based ranking algorithms for publications and authors*. Journal of Systems and Software, vol. 79, no. 12, pp. 1679-1700, 2006.

[Small1973] Small H. *Co-citation in the scientific literature: A new measure of the relationship between two documents*. Journal of the American Society for Information Science, vol. 24, no. 4, pp. 265-269, 1973.

[Smeaton2003] Smeaton A. F., Keogh G., Gurrin C., McDonald K., Sødring T. *Analysis of papers from twenty-five years of SIGIR conferences: what have we been doing for the last quarter of a century?* SIGIR Forum, vol. 37, no. 1, pp. 49-53, 2003.

[Testa1998] Testa J. *The ISI Database: the journal selection process*. Ciência da Informação, vol. 27, no. 2, pp. 233-235, 1998.

[Thelwall2001] Thelwall M. *Extracting Macroscopic Information from Web Links*. Journal of the American Society for Information Science and Technology, vol. 52, no. 13, pp.1157-1168, 2001.

[Thelwall2002] Thelwall M. *Conceptualizing documentation on the Web: an evaluation of different heuristic-based models for counting links between university Web sites*. Journal of the American Society for Information Science and Technology, vol. 53, no. 12, pp. 995-1005, 2002.

[Vigna2005] Vigna S. *TruRank: taking PageRank to the limit*. Proceedings of the 14th international conference on World Wide Web (WWW 2005), Chiba, Japan, pp. 976-977, 2005.

[Wagner2003] Wagner C., Leydesdorff L. *Mapping global science using international co-authorships: A comparison of 1990 and 2000*. Proceedings of the 9th International Conference on Scientometrics and Informetrics, Dalian, China, pp. 330-340, 2003.

[Wasserman1994] Wasserman S., Faust K. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[White1989] White H. D., McCain K. W. *Bibliometrics*. Annual review of information science and technology, vol. 24, pp. 119-186, 1989.

[Wu2005] Wu B., Davison B. D. *Identifying link farm spam pages*. Proceedings of the 14th international conference on World Wide Web (WWW 2005), Chiba, Japan, pp. 820-829, 2005.

[Xing2004] Xing W., Ghorbani A. *Weighted PageRank algorithm*. Proceedings. of the 2nd Annual Conference on Communication Networks and Services Research, Fredericton, Canada, pp. 305-314, 2004.

[Yang2005] Yang H., King I., Lyu M. R. *Predictive ranking: a novel page ranking approach by estimating the web structure*. Proceedings of the 14th international conference on World Wide Web (WWW 2005), Chiba, Japan, pp. 944-945, 2005.

Web

[1] ACM SIGMOD Online | SIGMOD Awards: http://www.sigmod.org/sigmodinfo/awards/#innovations
[2] http://dblp.uni-trier.de/xml/dblp20040213.xml.gz
[3] World Universities' ranking on the Web: Home: http://www.webometrics.info/
[4] http://home.zcu.cz/~dalfia/papers/France.svg
[5] DBLP Computer Science Bibliography: http://dblp.uni-trier.de/
[6] http://dblp.uni-trier.de/xml/egc2006.ppt
[7] DBIS-Homepage - DBL-Browser: http://dbis.uni-trier.de/DBL-Browser/
[8] http://dblp.uni-trier.de/xml/
[9] ACM SIGMOD Anthology: http://dblp.uni-trier.de/db/anthology.html
[10] http://www.informatik.uni-trier.de/~ley/db/about/top.html
[11] http://www.informatik.uni-trier.de/~ley/db/journals/tods/Chen76.html
[12] http://dblp.uni-trier.de/rec/bibtex/journals/tods/Chen76
[13] http://dblp.uni-trier.de/db/journals/tods/Chen76.html
[14] DBLP FAQ: Software: http://www.informatik.uni-trier.de/~ley/db/about/faqsoft.html
[15] CiteSeer Publications ResearchIndex:: http://citeseer.ist.psu.edu/
[16] NZDL: PreScript: http://www.nzdl.org/html/prescript.html
[17] Most cited authors in Computer Science [CiteSeer.Continuity; Steve Lawrence, Kurt Bollacker, Lee Giles]: http://citeseer.ist.psu.edu/mostcited.html
[18] CiteSeer.PSU OAI: http://citeseer.ist.psu.edu/oai.html
[19] Rexa Search Engine: http://rexa.info/
[20] Google Scholar: http://scholar.google.com/
[21] Live Search Academic: http://academic.live.com/
[22] Scirus – for scientific information: http://www.scirus.com/
[23] Inspec: http://www.iee.org/publish/inspec/
[24] The h Index for Computer Science: http://www.cs.ucla.edu/~palsberg/h-number.html
[25] H-number (or H-index): http://www.brics.dk/~mis/hnumber.html
[26] http://www.cs.ucla.edu/~palsberg/hnum
[27] Web of Science - Thomson Scientific: http://scientific.thomson.com/products/wos/
[28] The Thomson Scientific Journal Selection Process - Thomson Scientific: http://scientific.thomson.com/free/essays/selectionofmaterial/journalselection/
[29] Current Web Contents Web Site Selection Criteria - Thomson Scientific: http://www.scientific.thomson.com/free/essays/selectionofmaterial/cwc-criteria/
[30] Science Citation Index help, Web version: Princeton Unviersity: http://biolib.princeton.edu/instruct/OLDSCI.html
[31] Journal Citation Reports - Thomson Scientific: http://scientific.thomson.com/products/jcr/
[32] [ISI Highly Cited Researchers Version 1.1]: http://www.isihighlycited.com/
[33] The ACM Portal: http://portal.acm.org/portal.cfm
[34] Harzing.com - Research in International and Cross-cultural Management: http://www.harzing.com/resources.htm#/pop.htm
[35] Larbin: http://larbin.sourceforge.net/index.html
[36] ht://Dig: http://www.htdig.org/
[37] webbase: http://www.nongnu.org/webbase/
[38] Internet Archive: http://www.archive.org/index.php
[39] Google: http://www.google.com/
[40] AltaVista: http://www.altavista.com/

Author's Publications

Journals (reviewed)

Fiala D., Rousselot F., Ježek K. *PageRank for Bibliographic Networks,* Scientometrics, vol. 76, no. 1, 2008. (to appear)

International conferences (reviewed)

Fiala D., Rousselot F., Ježek K. *Ranking Algorithms For Web Sites: Finding Authoritative Academic Web Sites and Researchers*. Proceedings of the 3rd International Conference on Web Information Systems and Technologies WEBIST'07, Barcelona, Spain, pp. 372-375, 2007.

Fiala D., Ježek K., Rousselot F. *Finding Authoritative Researchers on Academic Web Sites*. Proceedings of the 17th International Conference on Computer, Information, and Systems Science, and Engineering CISE'06, Cairo, Egypt, Enformatika Transactions on Engineering, Computing and Technology, vol. 17, pp. 74-79, 2006.

Fiala D., Tesař R., Ježek K., Rousselot F. *Extracting Information from Web Content and Structure*. Proceedings of the 9th International Conference on Information Systems Implementation and Modelling ISIM'06, Přerov, Czech Republic, pp. 133-140, 2006.

Tesař R., Fiala D., Rousselot F., Ježek K. A *comparison of two algorithms for discovering repeated word sequences*. Proceedings of the 6th International Conference on Data Mining, Text Mining and their Business Applications DATA MINING 2005, Skiathos, Greece, WIT Transactions on Information and Communication Technologies, vol. 35, pp. 121-131, 2005.

Fiala D., Ježek K. *Retrieving citations on the Web*. Proceedings of International Conference on Knowledge Engineering and Decision Support ICKEDS'04, Porto, Portugal, pp. 481 – 488, 2004.

National conferences (reviewed)

Fiala D., Ježek K., Rousselot F. *Využití struktury webu pro vyhledávání autoritativních institucí a osob* (Using the Web structure for the Search for Authoritative Institutions and Individuals). Proceedings of the 6th Annual Conference ZNALOSTI 2007, Ostrava, Czech Republic, pp. 300-303, 2007.

International workshops (unreviewed)

Belaïd A., Alusse A., Rangoni Y., Cecotti H., Farah F., Gagean N., Fiala D., Rousselot F., Vigne H. *Document retro-conversion for personalized electronic re-edition*. Proceedings of the International Workshop on Document Analysis IWDA'05, Calcutta, India, 2005.

Other publications

Fiala D., Tesař R., Ježek K. *Získávání informací z obsahu a topologie webu* (Extracting Information from Web Content and Topology). Final report on the project FRVŠ 1347/2005/G1, Dec. 2005.

Fiala D. *Web Mining and Its Applications to Researchers Support*. Department of Computer Science and Engineering, University of West Bohemia in Pilsen, Czech Republic, Technical Report DCSE/TR-2005-06, April 2005.

Fiala D. A *System for Citations Retrieval on the Web*. MSc. thesis, University of West Bohemia in Pilsen, 2003.

List of Abbreviations

ACM	Association for Computing Machinery
BP	Balanced Popularity
CCIDF	Common Citation vs. Inverse Document Frequency
CD	Compact Disk
CGI	Common Gateway Interface
CRC	Cyclic Redundancy Check
CS	Computer Science
DBLP	Digital Bibliography & Library Project
DL	Digital Library
DNS	Domain Name System
DVD	Digital Versatile (Video) Disk
GB	Gigabyte
HITS	Hyperlink-Induced Topic Search
HMM	Hidden Markov Model
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
ID	Identification
IF	Impact Factor
IP	Internet Protocol
IR	Information Retrieval
ISI	Institute for Scientific Information
MD5	Message Digest 5
OCR	Optical Character Recognition
OPIC	On-line Page Importance Computation
PDF	Portable Document Format
PHP	Hypertext PreProcessor
PR	PageRank
SALSA	Stochastic Approach for Link-Structure Analysis
SCC	Strongly Connected Core (Component)
SCEAS	Scientific Collection Evaluator with Advanced Scoring
SIGMOD	Special Interest Group Management of Data
TFIDF	Term Frequency vs. Inverse Document Frequency
UDP	User Datagram Protocol
URL	Uniform Resource Locator
VLDB	Very Large Databases
WWW	World Wide Web
XML	Extended Markup Language

Appendix

	Citations		In-degree		HITS
1	**Michael Stonebraker**	5 346	**Michael Stonebraker**	1 857	**Michael Stonebraker**
2	**David J. Dewitt**	4 865	**David J. Dewitt**	1 432	**David J. Dewitt**
3	**Jeffrey D. Ullman**	3 926	**Jim Gray**	1 347	Raymond A. Lorie
4	**Jim Gray**	3 702	Raymond A. Lorie	1 250	**Jim Gray**
5	Raymond A. Lorie	3 317	**Jeffrey D. Ullman**	1 156	**Michael J. Carey**
6	**Philip A. Bernstein**	2 893	Won Kim	1 113	Won Kim
7	**Michael J. Carey**	2 773	E. F. Codd	1 110	**Philip A. Bernstein**
8	E. F. Codd	2 732	**Philip A. Bernstein**	1 109	Umeshwar Dayal
9	**Hector Garcia-Molina**	2 696	**Michael J. Carey**	1 042	**Jeffrey D. Ullman**
10	Won Kim	2 670	Umeshwar Dayal	1 035	**Donald D. Chamberlin**
11	**Rakesh Agrawal**	2 640	**David Maier**	983	**David Maier**
12	**Serge Abiteboul**	2 601	**Hector Garcia-Molina**	974	Morton M. Astrahan
13	**David Maier**	2 448	**Donald D. Chamberlin**	940	François Bancilhon
14	Umeshwar Dayal	2 301	Peter P. Chen	896	Bruce G. Lindsay
15	Yehoshua Sagiv	2 160	**Rakesh Agrawal**	855	Kapali P. Eswaran
16	**Donald D. Chamberlin**	2 099	Morton M. Astrahan	829	Hamid Pirahesh
17	Catriel Beeri	2 089	Kapali P. Eswaran	820	E. F. Codd
18	François Bancilhon	2 059	**Serge Abiteboul**	809	**Hector Garcia-Molina**
19	Christos Faloutsos	1 970	Nathan Goodman	804	Eugene Wong
20	Jennifer Widom	1 937	François Bancilhon	802	Irving L. Traiger
21	Nathan Goodman	1 928	Hamid Pirahesh	765	**Serge Abiteboul**
22	Morton M. Astrahan	1 847	Bruce G. Lindsay	761	Nathan Goodman
23	Raghu Ramakrishnan	1 825	Irving L. Traiger	760	**Patricia G. Selinger**
24	Irving L. Traiger	1 708	Eugene Wong	742	Thomas G. Price
25	Jeffrey F. Naughton	1 704	Catriel Beeri	709	**Rakesh Agrawal**
26	Eugene Wong	1 600	Jennifer Widom	696	Catriel Beeri
27	Hamid Pirahesh	1 600	Randy H. Katz	676	Patrick Valduriez
28	**Ronald Fagin**	1 599	Jeffrey F. Naughton	675	Stanley B. Zdonik
29	Kapali P. Eswaran	1 595	Nick Roussopoulos	674	Yehoshua Sagiv
30	Bruce G. Lindsay	1 548	Stanley B. Zdonik	670	Lawrence A. Rowe
31	Peter P. Chen	1 511	Raghu Ramakrishnan	667	Jeffrey F. Naughton
32	Richard Hull	1 488	Yehoshua Sagiv	661	Randy H. Katz
33	Nick Roussopoulos	1 383	Shamkant B. Navathe	650	Jennifer Widom
34	Randy H. Katz	1 381	John Miles Smith	645	Raghu Ramakrishnan
35	Patrick Valduriez	1 373	H. V. Jagadish	640	Nick Roussopoulos
36	**C. Mohan**	1 350	Patrick Valduriez	621	Carlo Zaniolo
37	H. V. Jagadish	1 343	Henry F. Korth	619	Henry F. Korth
38	**Patricia G. Selinger**	1 341	**Patricia G. Selinger**	619	Mike W. Blasgen
39	Stanley B. Zdonik	1 336	Thomas G. Price	616	Goetz Graefe
40	Goetz Graefe	1 327	**Ronald Fagin**	613	Gianfranco R. Putzolu
	Missed: 84. Rudolf Bayer (845)		Missed: 47. C. Mohan (578), 75. Rudolf Bayer (466)		Missed: 45. C. Mohan, 46. Ronald Fagin, 94. Rudolf Bayer

Table 1: Top 40 DBLP authors for each ranking (part 1).

	PR	w	a	b
1	E. F. Codd	E. F. Codd	E. F. Codd	**Michael Stonebraker**
2	**Donald D. Chamberlin**	**Michael Stonebraker**	**Michael Stonebraker**	**Jim Gray**
3	**Michael Stonebraker**	**Jim Gray**	**Donald D. Chamberlin**	**David J. Dewitt**
4	**Philip A. Bernstein**	**Donald D. Chamberlin**	Raymond A. Lorie	**Hector Garcia-Molina**
5	John Miles Smith	Raymond A. Lorie	**Philip A. Bernstein**	**Jeffrey D. Ullman**
6	**Jim Gray**	**Philip A. Bernstein**	**Jim Gray**	**Philip A. Bernstein**
7	**Rudolf Bayer**	John Miles Smith	John Miles Smith	**David Maier**
8	Raymond A. Lorie	**Jeffrey D. Ullman**	Morton M. Astrahan	Moshe Y. Vardi
9	Morton M. Astrahan	Morton M. Astrahan	Irving L. Traiger	E. F. Codd
10	Kapali P. Eswaran	Irving L. Traiger	Eugene Wong	Catriel Beeri
11	Eugene Wong	Eugene Wong	Kapali P. Eswaran	Umeshwar Dayal
12	Irving L. Traiger	Kapali P. Eswaran	**Jeffrey D. Ullman**	**Serge Abiteboul**
13	Gerald Held	**Ronald Fagin**	**Ronald Fagin**	**Michael J. Carey**
14	Hans Albrecht Schmid	**David J. Dewitt**	**Rudolf Bayer**	Yehoshua Sagiv
15	**Jeffrey D. Ullman**	Catriel Beeri	Catriel Beeri	Christos H. Papadimitriou
16	Michael Hammer	**Rudolf Bayer**	William C. McGee	**Rakesh Agrawal**
17	Mike W. Blasgen	William C. McGee	Gerald Held	Bruce G. Lindsay
18	Raymond F. Boyce	Gerald Held	Diane C. P. Smith	Jeffrey F. Naughton
19	**Ronald Fagin**	Gianfranco R. Putzolu	Gianfranco R. Putzolu	Nick Roussopoulos
20	Gianfranco R. Putzolu	Diane C. P. Smith	**David J. Dewitt**	Hans-Jörg Schek
21	Edward M. McCreight	Nathan Goodman	Nathan Goodman	Raghu Ramakrishnan
22	Nathan Goodman	Michael Hammer	Michael Hammer	Hamid Pirahesh
23	James W. Mehl	Mike W. Blasgen	Mike W. Blasgen	Goetz Graefe
24	W. Frank King III	Stephen Todd	Hans Albrecht Schmid	Raymond A. Lorie
25	Bradford W. Wade	Hans Albrecht Schmid	Stephen Todd	Alberto O. Mendelzon
26	Paul R. McJones	Bradford W. Wade	Paul R. McJones	Gio Wiederhold
27	Robert C. Goldstein	James W. Mehl	Bradford W. Wade	**Ronald Fagin**
28	Stephen Todd	Paul R. McJones	James W. Mehl	Richard T. Snodgrass
29	Patricia P. Griffiths	W. Frank King III	Patricia P. Griffiths	**Donald D. Chamberlin**
30	Diane C. P. Smith	Patricia P. Griffiths	W. Frank King III	François Bancilhon
31	Philip Yen-tang Chang	Alfred V. Aho	Alfred V. Aho	Mihalis Yannakakis
32	Peter Kreps	Peter Kreps	Peter Kreps	Jennifer Widom
33	Vera Watson	Yehoshua Sagiv	Edward M. McCreight	Nathan Goodman
34	Peter P. Chen	Edward M. McCreight	Robert C. Goldstein	Randy H. Katz
35	Catriel Beeri	**David Maier**	Moshé M. Zloof	H. V. Jagadish
36	**David J. Dewitt**	Robert C. Goldstein	Philip Yen-tang Chang	Won Kim
37	Alfred V. Aho	Raymond F. Boyce	Raymond F. Boyce	Irving L. Traiger
38	John J. Donovan	Moshé M. Zloof	Vera Watson	Abraham Silberschatz
39	Stuart G. Greenberg	Vera Watson	C. J. Date	Eugene Wong
40	Loius M. Gutentag	Umeshwar Dayal	Peter P. Chen	Guy M. Lohman
	Missed: 51. David Maier, 59. Patricia Selinger, 60. Hector Garcia-Molina, 63. Michael Carey, 65. Rakesh Agrawal, 104. Serge Abiteboul, 113. C. Mohan	Missed: 46. Michael Carey, 49. Hector Garcia-Molina, 55. Patricia Selinger, 58. Rakesh Agrawal, 61. Serge Abiteboul, 110. C. Mohan	Missed: 47. David Maier, 53. Patricia Selinger, 55. Michael Carey, 62. Hector Garcia-Molina, 64. Rakesh Agrawal, 69. Serge Abiteboul, 116. C. Mohan	Missed: 61. Patricia Selinger, 62. C. Mohan, 97. Rudolf Bayer

Table 2: Top 40 DBLP authors for each ranking (part 2).

	c	d	e
1	**Michael Stonebraker**	**Michael Stonebraker**	**Michael Stonebraker**
2	**Jim Gray**	**Jim Gray**	**David J. Dewitt**
3	**David J. Dewitt**	**David J. Dewitt**	**Hector Garcia-Molina**
4	**Hector Garcia-Molina**	**Philip A. Bernstein**	**Jim Gray**
5	**Jeffrey D. Ullman**	**Hector Garcia-Molina**	**Jeffrey D. Ullman**
6	**Philip A. Bernstein**	**David Maier**	**Philip A. Bernstein**
7	**David Maier**	**Jeffrey D. Ullman**	**David Maier**
8	Umeshwar Dayal	Umeshwar Dayal	Moshe Y. Vardi
9	Bruce G. Lindsay	**Michael J. Carey**	Umeshwar Dayal
10	**Michael J. Carey**	E. F. Codd	Catriel Beeri
11	**Serge Abiteboul**	Bruce G. Lindsay	E. F. Codd
12	Jeffrey F. Naughton	Catriel Beeri	**Serge Abiteboul**
13	Catriel Beeri	Jeffrey F. Naughton	Yehoshua Sagiv
14	Hamid Pirahesh	**Serge Abiteboul**	**Michael J. Carey**
15	Moshe Y. Vardi	Hamid Pirahesh	**Rakesh Agrawal**
16	Hans-Jörg Schek	Goetz Graefe	Christos H. Papadimitriou
17	E. F. Codd	Hans-Jörg Schek	Bruce G. Lindsay
18	Yehoshua Sagiv	**Rakesh Agrawal**	Jeffrey F. Naughton
19	**Rakesh Agrawal**	Raymond A. Lorie	Nick Roussopoulos
20	Raghu Ramakrishnan	Yehoshua Sagiv	Hans-Jörg Schek
21	Goetz Graefe	Nick Roussopoulos	Raghu Ramakrishnan
22	Nick Roussopoulos	Gio Wiederhold	Hamid Pirahesh
23	Raymond A. Lorie	**Donald D. Chamberlin**	Raymond A. Lorie
24	Christos H. Papadimitriou	Moshe Y. Vardi	Alberto O. Mendelzon
25	Gio Wiederhold	Dina Bitton	**Ronald Fagin**
26	**Donald D. Chamberlin**	Richard T. Snodgrass	**Donald D. Chamberlin**
27	Richard T. Snodgrass	Christos H. Papadimitriou	Gio Wiederhold
28	**Ronald Fagin**	Raghu Ramakrishnan	Goetz Graefe
29	Dina Bitton	Guy M. Lohman	Nathan Goodman
30	Jennifer Widom	**Ronald Fagin**	Mihalis Yannakakis
31	Randy H. Katz	Randy H. Katz	François Bancilhon
32	Alberto O. Mendelzon	François Bancilhon	Jennifer Widom
33	Guy M. Lohman	Alberto O. Mendelzon	Randy H. Katz
34	François Bancilhon	Jennifer Widom	Richard T. Snodgrass
35	H. V. Jagadish	Michael J. Franklin	Abraham Silberschatz
36	Abraham Silberschatz	Irving L. Traiger	H. V. Jagadish
37	Irving L. Traiger	H. V. Jagadish	Guy M. Lohman
38	Michael J. Franklin	Won Kim	Eugene Wong
39	Mihalis Yannakakis	Eugene Wong	Peter Buneman
40	Nathan Goodman	Nathan Goodman	Christos Faloutsos
	Missed: 55. Patricia Selinger, 59. C. Mohan, 132. Rudolf Bayer	Missed: 54. Patricia Selinger, 65. C. Mohan, 94. Rudolf Bayer	Missed: 63. Patricia Selinger, 65. C. Mohan, 93. Rudolf Bayer

Table 3: Top 40 DBLP authors for each ranking (part 3).

	f	g
1	**Jim Gray**	E. F. Codd
2	E. F. Codd	**Jim Gray**
3	**Michael Stonebraker**	**Michael Stonebraker**
4	**David J. Dewitt**	**Philip A. Bernstein**
5	**Philip A. Bernstein**	**David J. Dewitt**
6	Raymond A. Lorie	**Donald D. Chamberlin**
7	**Donald D. Chamberlin**	Raymond A. Lorie
8	**Jeffrey D. Ullman**	**Jeffrey D. Ullman**
9	Irving L. Traiger	Irving L. Traiger
10	Morton M. Astrahan	Morton M. Astrahan
11	**David Maier**	John Miles Smith
12	Eugene Wong	Eugene Wong
13	Catriel Beeri	**David Maier**
14	John Miles Smith	**Hector Garcia-Molina**
15	Bruce G. Lindsay	Catriel Beeri
16	**Hector Garcia-Molina**	Kapali P. Eswaran
17	**Ronald Fagin**	**Ronald Fagin**
18	Kapali P. Eswaran	Gerald Held
19	Gerald Held	Umeshwar Dayal
20	Umeshwar Dayal	**Rudolf Bayer**
21	**Michael J. Carey**	Michael Hammer
22	Yehoshua Sagiv	Bruce G. Lindsay
23	Gianfranco R. Putzolu	Nathan Goodman
24	Nathan Goodman	Gianfranco R. Putzolu
25	**Rudolf Bayer**	Stephen Todd
26	Mike W. Blasgen	Diane C. P. Smith
27	Michael Hammer	William C. McGee
28	William C. McGee	Mike W. Blasgen
29	Stephen Todd	**Michael J. Carey**
30	Diane C. P. Smith	Phyllis Reisner
31	Jeffrey F. Naughton	Paul R. McJones
32	Thomas G. Price	Jeffrey F. Naughton
33	Bradford W. Wade	Hamid Pirahesh
34	Hamid Pirahesh	Yehoshua Sagiv
35	Phyllis Reisner	Bradford W. Wade
36	**Patricia G. Selinger**	Hans Albrecht Schmid
37	**Serge Abiteboul**	Nick Roussopoulos
38	W. Frank King III	Won Kim
39	François Bancilhon	James W. Mehl
40	James W. Mehl	W. Frank King III
	Missed: 49. Rakesh Agrawal, 105. C. Mohan	Missed: 43. Serge Abiteboul, 48. Patricia Selinger, 49. Rakesh Agrawal, 101. C. Mohan

Table 4: Top 40 DBLP authors for each ranking (part 4).

	Site	In-Links	In-Degree	Out-Links	Out-Degree
1	cedric.cnam.fr	32	6	0	0
2	citi.insa-lyon.fr	13	3	47	1
3	dep-info.u-psud.fr	59	1	0	0
4	dept-info.labri.fr	22	6	3	1
5	deptinfo.unice.fr	11	4	2	1
6	dept-info.univ-brest.fr	0	0	1	1
7	dpt-info.u-strasbg.fr	127	4	0	0
8	eric.univ-lyon2.fr	27	10	5	1
9	eurise.univ-st-etienne.fr	41	6	0	0
10	lifc.univ-fcomte.fr	28	7	9	2
11	lihs.univ-tlse1.fr	5	2	0	0
12	lil.univ-littoral.fr	10	2	2	1
13	lina.atlanstic.net	0	0	20	8
14	liris.cnrs.fr	80	8	17	2
15	lis.snv.jussieu.fr	0	0	13	7
16	lisi.insa-lyon.fr	7	3	36	2
17	liuppa.univ-pau.fr	2	2	0	0
18	llaic3.u-clermont1.fr	3	2	0	0
19	lrlweb.univ-bpclermont.fr	1	1	1	1
20	lsiit.u-strasbg.fr	2	2	371	18
21	msi.unilim.fr	11	2	0	0
22	phalanstere.univ-mlv.fr	5	3	0	0
23	psiserver.insa-rouen.fr	0	0	10	4
24	sirac.inrialpes.fr	6	1	7	1
25	sis.univ-tln.fr	21	7	1	1
26	www.ai.univ-paris8.fr	1	1	0	0
27	www.cril.univ-artois.fr	16	4	120	18
28	www.depinfo.uhp-nancy.fr	1	1	0	0
29	www.di.ens.fr	72	5	18	1
30	www.dil.univ-mrs.fr	32	4	0	0
31	www.dptinfo.ens-cachan.fr	1	1	13	2
32	www.epita.fr	4	2	0	0
33	www.i3s.unice.fr	44	6	82	4
34	www.icp.inpg.fr	5	2	6	2
35	www.if.insa-lyon.fr	6	1	0	0
36	www.info.iut.u-bordeaux1.fr	0	0	0	0
37	www.info.iut-tlse3.fr	0	0	0	0
38	www.info.unicaen.fr	40	8	5	3
39	www.info.univ-angers.fr	29	5	3	1
40	www.inrialpes.fr	125	12	5	2

Table 5: French sites and their graph properties (alphabetical order, 1 - 40).

	Site	In-Links	In-Degree	Out-Links	Out-Degree
41	www.irisa.fr	213	14	115	4
42	www.irit.fr	123	11	42	4
43	www.isc.cnrs.fr	6	3	0	0
44	www.isima.fr	22	4	4	2
45	www.iut-info.univ-lille1.fr	11	1	0	0
46	www.laas.fr	62	12	22	3
47	www.labri.fr	95	8	4	2
48	www.lalic.paris4.sorbonne.fr	5	3	0	0
49	www.li.univ-tours.fr	13	5	1	1
50	www.lia.univ-avignon.fr	10	7	35	19
51	www.liafa.jussieu.fr	122	8	2	1
52	www.lif.univ-mrs.fr	17	5	79	13
53	www.lifl.fr	209	13	1476	49
54	www.lip6.fr	64	7	6	4
55	www.lirmm.fr	117	15	480	40
56	www.lisi.ensma.fr	20	4	2	1
57	www.listic.univ-savoie.fr	0	0	1	1
58	www.lita.univ-metz.fr	1	1	0	0
59	www.lix.polytechnique.fr	177	7	3	2
60	www.loria.fr	460	15	501	37
61	www.lri.fr	225	12	992	54
62	www.prism.uvsq.fr	79	8	13	2
63	www-clips.imag.fr	20	6	10	3
64	www-futurs.inria.fr	917	3	4	1
65	wwwhds.utc.fr	2	1	23	12
66	www-id.imag.fr	24	6	38	1
67	www-info.enst-bretagne.fr	0	0	8	1
68	www-info.iutv.univ-paris13.fr	18	1	0	0
69	www-leibniz.imag.fr	93	10	0	0
70	www-lil.univ-littoral.fr	55	6	0	0
71	www-lipn.univ-paris13.fr	22	8	217	36
72	www-lium.univ-lemans.fr	1	1	6	4
73	www-lmc.imag.fr	31	6	9	2
74	www-lsr.imag.fr	41	5	10	1
75	www-rocq.inria.fr	223	8	1	1
76	www-sic.univ-poitiers.fr	33	3	0	0
77	www-sop.inria.fr	648	15	243	3
78	www-timc.imag.fr	37	9	2	2
79	www-valoria.univ-ubs.fr	5	2	14	2
80	www-verimag.imag.fr	50	6	0	0

Table 6: French sites and their graph properties (alphabetical order, 41 - 80).

	CZ In-Degree	CZ HITS	CZ PageRank	FR In-Degree	FR HITS	FR PageRank
1	Alon N	Jancar P	Moller F	Alur R	Halbwachs N	Cahon S
2	Nesetril J	Alon N	Nesetril J	Halbwachs N	Alur R	Berry G
3	Jancar P	Nesetril J	Jancar P	Zhang L	Abadi M	Milner R
4	Hell P	Christensen S	Kucera A	Abadi M	Zhang L	Shamir A
5	Milner R	Milner R	Hoppe H	Foster I	Caspi P	Filiol E
6	Caucal D	Hell P	Curless B	Caspi P	Sifakis J	Dubois M
7	Christensen S	**Bouajjani A**	Pultr A	Sifakis J	Berry G	Ullman J
8	Hajic J	Vardi M Y	Pala K	Berry G	Zhang H	Alur R
9	Burkart O	Caucal D	Lorensen W E	Pnueli A	Courcoubetis C	Halbwachs N
10	**Bouajjani A**	Thomas R	Alon N	Zhang H	Pnueli A	Karp R
11	Moller F	Moller F	Smrz P	Courcoubetis C	**Ullman J D**	Bellare M
12	Kucera A	Burkart O	Banaschewski B	**Ullman J D**	Benveniste A	Zhang Z
13	Hirshfeld Y	Kucera A	Hajic J	Benveniste A	Manna Z	Benveniste A
14	Matousek J	**Vardi M**	Caucal D	Abiteboul S	Nicollin X	Reps T
15	Panevova J	**Ullman J D**	Hirshfeld Y	Ullman J	Cousot R	Lavalle S
16	Robertson N	Rivest R L	Herrlich G	Gupta A	Raymond P	Dombre E
17	Golub G H	Hopcroft J E	Panevova J	Agrawal R	Cousot P	Boudet S
18	Esparza J	Pala K	Jerrum M	Manna Z	Foster I	Dgoulange E
19	Pala K	Robertson N	Sanguineti M	Maler O	Abiteboul S	Gourdon A
20	**Ullman J D**	Hajic J	Savick P	Nicollin X	Maler O	Abiteboul S
21	Johnson D S	Esparza J	Christensen S	Cousot P	Ullman J	Charpin P
22	Sgall P	Donald E	Esparza J	Thomas W	Asarin E	Carlet C
23	Graham R L	Oliva K	Jacobson N	Cousot R	Harel D	Abadi M
24	Rivest R L	Hirshfeld Y	Galluccio A	Raymond P	Olivero A	Gupta A
25	Oracles S	Panevova J	Winkler F	**Vardi M**	David A	Cohen G
26	Greenbaum A	Matousek J	Mcaloon K	Kesselman C	Henzinger T	Courcoubetis C
27	Agrawal R	Thomas W	Labahn G	Johnson D	**Vardi M**	Zhu X
28	Kratochvil J	Kratochvil J	Matousek J	Milner R	Comon H	Coppersmith D
29	Ganter B	Johnson D S	Johnson D S	Srikant R	Clarke E	Maier D
30	Thomas R	Richard J	Benzi M	**Bouajjani A**	Johnson D	Troccaz J
31	Clarke E M	Clarke E M	Welzl E	Dongarra J	Henzinger T A	Goldwasser S
32	Grumberg O	Sgall P	Sedlcek R	Olivero A	**Bouajjani A**	Williams M
33	**Vardi M**	Leiserson C E	Praehofer H	Asarin E	Johnson R	Taylor R
34	Sterling L	Grumberg O	Zeigler B	Johnson R	Gupta A	Zheng Y
35	Garey M R	Raspaud A	Kelton D	Harel D	Baader F	Dongarra J
36	Oliva K	Matthew L	Kim T	David A	Zwaenepoel W	Buhrman H
37	Imielinski T	Zhu X	Psutka J	Rivest R	Wolper P	Papadimitriou C H
38	Cormen T H	Seymour P	Sharma S	Helm R	Li K	Erdos P
39	Jerrum M	Graham R L	Jain P	Shenker S	Agrawal R	Abdalla M
40	Pach J	Hendler J	Pach J	Wolper P	Coupaye T	Payan Y

Table 7: Top 40 international authors in Czech and French corpora.