Jointly published by Akadémiai Kiadó, Budapest and Springer, Dordrecht

Scientometrics, Vol. 76, No. 1 (2008) 135–158
DOI: 10.1007/s11192-007-1908-4

PageRank for bibliographic networks

DALIBOR FIALA,[a,b] FRANÇOIS ROUSSELOT,[b] KAREL JEŽEK[a]

[a] *University of West Bohemia in Pilsen, Pilsen (Czech Republic)*
[b] *INSA, Strasbourg (France)*

In this paper, 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 graph. We verify 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 E. F. Codd Innovations Award. Rankings based on both the citation and co-authorship information turn out to be "better" than the standard PageRank ranking.

Introduction

Notions of importance, significance, authority, prestige, quality and others 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 links between Web pages may serve as citations. Therefore, current Web search engines make use of various link-based

Received August 2, 2007

Address for correspondence:
DALIBOR FIALA
Department of Computer Science and Engineering, University of West Bohemia in Pilsen
Univerzitní 22, 30614 Plzeň, Czech Republic
E-mail: dalfia@kiv.zcu.cz

0138–9130/US $ 20.00

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 [BRIN & PAGE, 1998]. BIANCHINI & AL. [2005] and LANGVILLE & MEYER [2003] review the latest developments of PageRank thoroughly. 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 (e.g. [WAGNER & LEYDESDORHH, 2003; OTTE & ROUSSEAU, 2002; CUNNINGHAM & DILLON, 1997]). But their combination with citation graphs, which may lead to more "fair" rankings of authors, has been relatively little examined. In the following sections, 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.

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 $\exists (p_1,p_2) \in E^C \wedge \exists \{p_1,a_1\} \in E^P \wedge \exists \{p_2,a_2\} \in E^P \wedge A_{p1} \cap A_{p2} = \varnothing \wedge a_1 \neq a_2$.

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.

Rank computation

We associate a triple of weights ($w_{u,v}$, $c_{u,v}$, $b_{u,v}$) with each edge $(u, v) \in E$. $w_{u,v}$ and $c_{u,v}$ 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:

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

where

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

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 (1) becomes the weighted PageRank formula. (For instance, BOLLEN & AL. [2006] and XING & GHORBANI [2004] work with weighted PageRanks.) In addition to this, if all the weights $w_{u,v}$ are set to one, it is the standard PageRank [BRIN & PAGE, 1998]. The coefficients *c* and *b* are analogous to the co-authorship frequency and exclusivity in [LIU & AL., 2005] as noted on the related work.

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 1 shows edge weights for graph *G* in Figure 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 1 along its out-edges in the standard (PR) and weighted PageRank (w) and the variations a) – g) are given in Table.

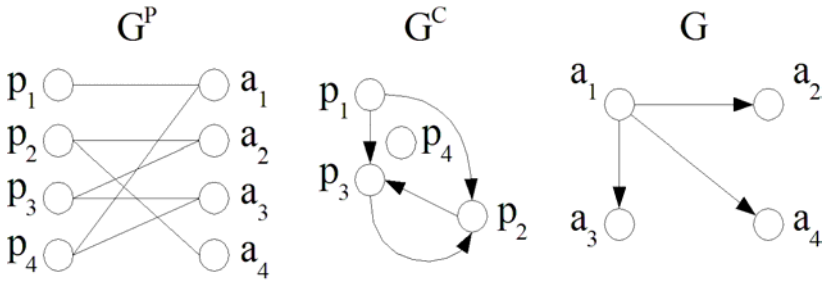

Figure 1. Examples of co-authorship, publication citation, and author citation graphs

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

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 2. Proportions of rank distributed by node a_1 in graph *G* in Figure 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
Σ	1	1	1	1	1	1	1	1	1

For example, to compute $\sigma_{a1,a2}$ for the variation *w*), we substitute in (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 (1) remains 2/4. Thus, one half of rank of author a_1 is transferred to author a_2 and so on.

Experiments

We tested our algorithms on the DBLP data available in XML (http://dblp.uni-trier.de/xml/). DBLP has established itself as a testbed for bibliographic studies in recent years (e.g. [ELMACIOGLU & AL., 2005; RAHM & THOR, 2005; BALMIN & AL., 2004; NASCIMENTO & AL., 2003]). We took advantage of the only time-stamped version of the collection from February 14, 2004 which may serve researchers as a testbed for experiments and comparisons. We extracted only *article* and *inproceedings* records exactly like SIDIROPOULOS & MANOLOPOULOS [2005, 2006].

DBLP testbed data

Table 3 summarizes some basic statistics of the DBLP data we work with. 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 3.

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

	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

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 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.

Publications

Let us return to Table 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 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 1). 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.

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 by SIDIROPOULOS & MANOLOPOULOS [2005]. See Figure 2 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+ (in-degree or citations of zero or more) with all the isolated authors included. It is not depicted in Figure 2. 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.

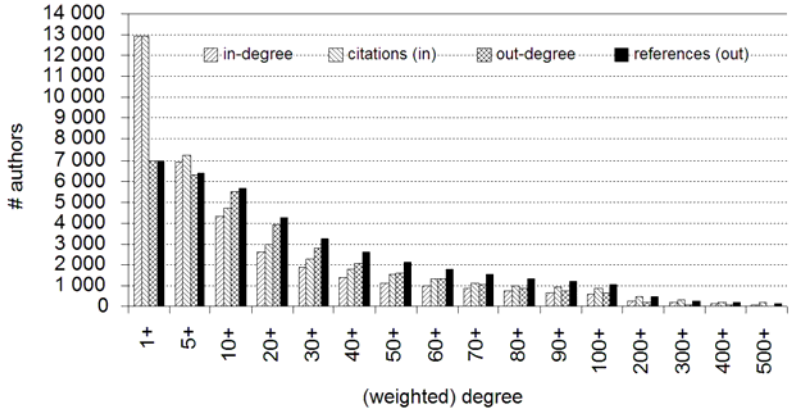

Figure 2. Cumulative histogram showing distribution of in- and out-degrees in *G*

Distribution of c and b coefficients

Figure 3 shows the cumulative distributions of various parameters defined 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 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 3 top right).

In Figure 3 bottom left, 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 3 bottom right 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 3 bottom right). 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.

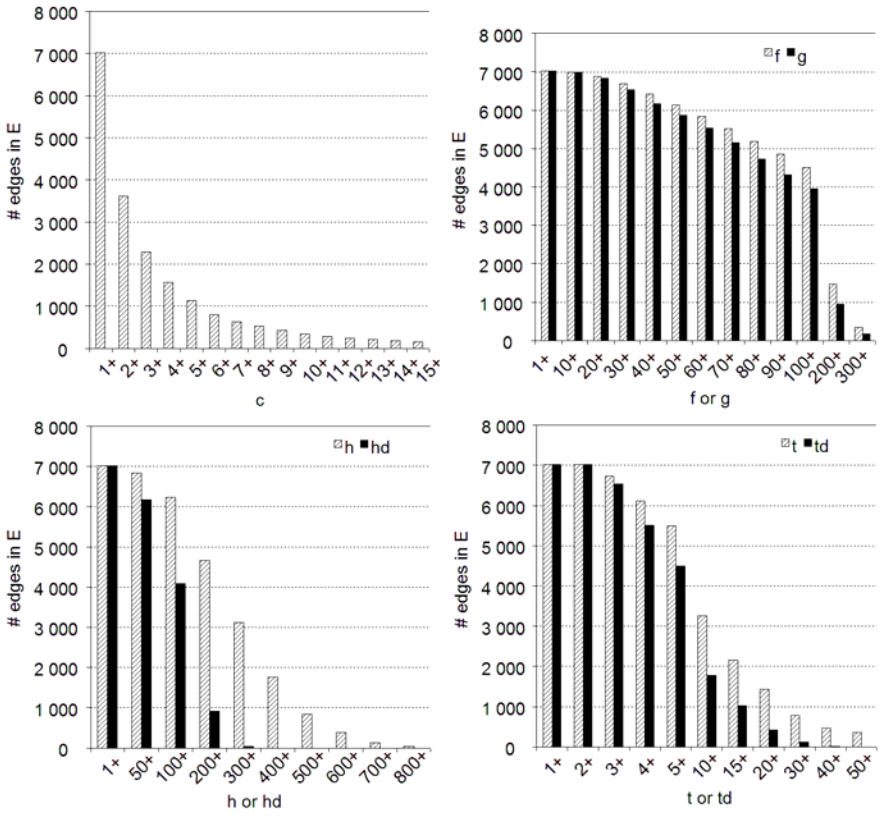

Figure 3. Cumulative distribution of values of parameters *c* (top left), *f*, *g* (top right), *h*, *hd* (bottom left) and *t*, *td* (bottom right) in graph *G*

Statistics of c and b coefficients

To terminate this subsection, Table 4 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 the definitions. 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 4 include: i) the maximum number of distinct co-authors in common publications by two specific authors is 67 (!), ii) 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 paper.

Table 4. Basic statistics of weight parameters for edges in *E* with non-zero *c*

	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

Computing ranks for authors

We exploited extensively the author citation graph *G* described in detail above. Altogether, twelve ranking methods were employed to evaluate the authors. In addition to the weighted (citation counting) and unweighted in-degree, HITS authorities, and the standard (unweighted) PageRank, we also applied the weighted and the bibliographic (seven variants) 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 available citation systems 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 [SIDIROPOULOS & MANOLOPOULOS, 2005] – 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 the related work section), the only award we can take advantage of is the ACM SIGMOD E. F. Codd Innovations Award (http://www.sigmod.org/sigmodinfo/awards/#innovations), 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 by LIU & AL. [2005]. 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 5 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 and "a" – "g"

correspond to the variations a) – g) mentioned at the very beginning of the section on rank computation. 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.

Table 5. E. F. Codd Innovations Award winners and their ranks in distinct methods

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

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 SIDIROPOULOS & MANOLOPOULOS [2005]. However, their HITS ranking was not obtained in the same way as ours.

Discussion of author ranks

The accompanying chart of Table 5 is in Figure 4. 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.)

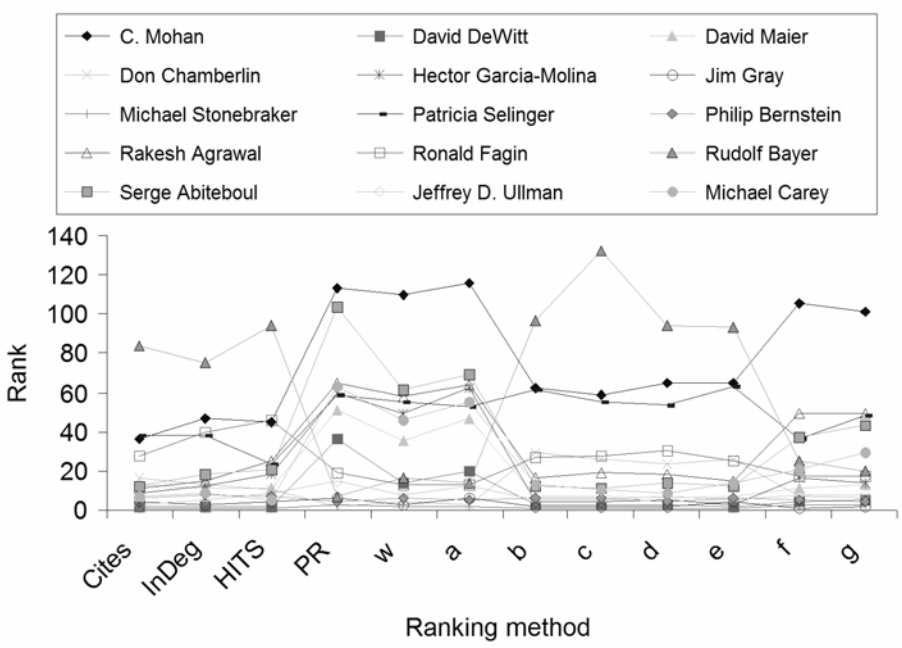

Figure 4. E. F. Codd Innovations Award winners and their ranks in distinct methods

Let us have a look at some particularities in Figure 4. 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 [SIDIROPOULOS & MANOLOPOULOS, 2006] for some of them. We will briefly discuss the outcomes of three metrics – two numerical and one graphical. In Table 6 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"}.

The next comparison is based on the correlation between rankings. Table 7 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. With this metric, we consider the ranks of all authors that have some in-degree. (It is 12 934 as we mention above.) 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 6 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 6 has disappeared here. It seems that mismatches just accumulate in the upper part of rankings (which is more important than the lower one, though).

Table 6. Common elements in top 20 authors of different rankings

	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 7. Spearman correlation coefficients

	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

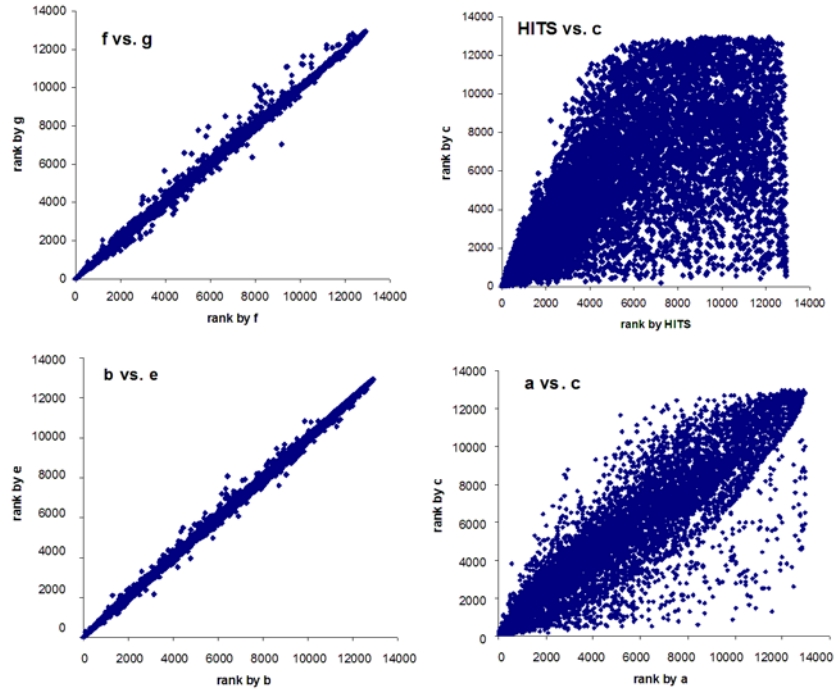

Figure 5. Some comparisons of rankings by means of q-q plots

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 5. 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").

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 6 where the damping factor (d in Equation 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. 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. 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 [SIDIROPOULOS & MANOLOPOULOS, 2005]. The newest citing paper is from 2001 as pointed out above.

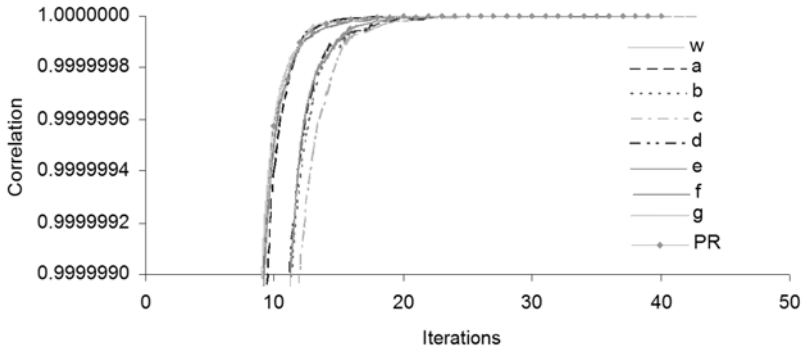

Figure 6. Convergence of standard (PR), weighted (w) and bibliographic (a–g) PR

Prediction

We show the top 40 authors for each ranking method in Table 8, Table 9, Table 10 and Table 11 in an 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.)

Related work

Sidiropoulos

SIDIROPOULOS & MANOLOPOULOS [2005] have proposed modifications of PageRank that would better meet needs for evaluating nodes in bibliographic networks. Their PageRank-based algorithm is called SCEAS. 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:

- 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 the time-stamped data instead, the input data would be the same and their results verifiable.
- 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.
- 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 [SIDIROPOULOS & MANOLOPOULOS, 2006] 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.

Liu and Bollen

LIU & AL. [2005] introduces 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 our definitions. 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 the definitions section is always set to one in their method. Interestingly, they do this for journal citation networks with a weighted PageRank algorithm [BOLLEN & AL., 2006]. 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.

Conclusions

Link-based ranking methods have become the standard way of determining authoritative Web pages. They may be easily applied in every environment that can be modelled as a graph and citation networks of authors or papers invite their usage. However, citation networks are only one part of bibliographic information. Collaboration networks are also a valuable source of information and their combination with citation graphs, which may lead to more "fair" rankings of authors, has been relatively little explored. Therefore, 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 graph. We verify 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 Edgar F. Codd Innovations Award. Rankings based on both the citation and co-authorship information tend to place the awarded authors higher than the standard PageRank ranking. In our future work, we would like to concentrate on the issue of incorporating the time factor in the bibliographic PageRank.

*

This work was supported in part by the Ministry of Education of the Czech Republic under Grant 2C06009.

References

BALMIN, A., HRISTIDIS, V., PAPAKONSTANTINOU, Y. (2004), ObjectRank: Authority-Based Keyword Search in Databases, Proc. 30th Int. Conf. Very Large Data Bases, Toronto, Canada, pp. 564–575.

BIANCHINI, M., GORI, M., SCARSELLI, F. (2005), Inside PageRank, *ACM Transactions on Internet Technology*, 5 (1) : 92–128.

BOLLEN, J., RODRIQUEZ, M. A., VAN DE SOMPEL, H. (2006), Journal status, *Scientometrics*, 69 (3) : 669–687.

BRIN, S., PAGE, L. (1998), *The Anatomy of a Large-Scale Hypertextual Web Search Engine*, Proc. 7th World Wide Web Conference, pp. 107–117.

CUNNINGHAM, S. J., DILLON, S. M. (1997), Authorship patterns in information systems, *Scientometrics*, 39 (1) : 19–27.

ELMACIOGLU, E., LEE, D. (2005), On six degrees of separation in DBLP-DB and more, *SIGMOD Record*, 34 (2) :33–40.

LANGVILLE, A. N., MEYER, C. D. (2003), Deeper inside PageRank, *Internet Mathematics*, 1 (3) : 335–380.

LIU, X., BOLLEN, J., NELSON, M. L., VAN DE SOMPEL, H. (2005), Co-authorship networks in the digital library research community, *Information Processing and Management*, 41 (6) : 1462–1480.

NASCIMENTO, M. A., SANDER, J., POUND, J. (2003), Analysis of SIGMOD's co-authorship graph, *SIGMOD Record*, 32 (3) : 8–10.

OTTE E., ROUSSEAU R. (2002), Social network analysis: a powerful strategy, also for the information sciences, *Journal of Information Science*, 28 (6) : 441–453.

RAHM, E., THOR, A. (2005), Citation analysis of database publications, *SIGMOD Record*, 34 (4) : 48–53.

SIDIROPOULOS, A., MANOLOPOULOS, Y. (2005), A citation-based system to assist prize awarding, *SIGMOD Record*, 34 (4) : 54–60.

SIDIROPOULOS, A., MANOLOPOULOS, Y. (2006), A Generalized comparison of graph-based ranking algorithms for publications and authors, *Journal of Systems and Software,* 79 (12) : 1679–1700.

WAGNER, C., LEYDESDORFF, L. (2003), Mapping global science using international co-authorships: A comparison of 1990 and 2000, Proc. 9th Int. Conf. Scientometrics and Informetrics, Dalian, China, pp. 330–340.

XING, W., GHORBANI, A. (2004), Weighted PageRank algorithm, Proc. 2nd Conf. Communication Networks and Services Research, Fredericton, Canada, pp. 305–314.

Appendix

Table 8. Top 40 DBLP authors for each ranking (part 1)

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

	PR	w	a
1	E. F. Codd	E. F. Codd	E. F. Codd
2	**Donald D. Chamberlin**	**Michael Stonebraker**	**Michael Stonebraker**
3	**Michael Stonebraker**	**Jim Gray**	**Donald D. Chamberlin**
4	**Philip A. Bernstein**	**Donald D. Chamberlin**	Raymond A. Lorie
5	John Miles Smith	Raymond A. Lorie	**Philip A. Bernstein**
6	**Jim Gray**	**Philip A. Bernstein**	**Jim Gray**
7	**Rudolf Bayer**	John Miles Smith	John Miles Smith
8	Raymond A. Lorie	**Jeffrey D. Ullman**	Morton M. Astrahan
9	Morton M. Astrahan	Morton M. Astrahan	Irving L. Traiger
10	Kapali P. Eswaran	Irving L. Traiger	Eugene Wong
11	Eugene Wong	Eugene Wong	Kapali P. Eswaran
12	Irving L. Traiger	Kapali P. Eswaran	**Jeffrey D. Ullman**
13	Gerald Held	**Ronald Fagin**	**Ronald Fagin**
14	Hans Albrecht Schmid	**David J. Dewitt**	**Rudolf Bayer**
15	**Jeffrey D. Ullman**	Catriel Beeri	Catriel Beeri
16	Michael Hammer	**Rudolf Bayer**	William C. McGee
17	Mike W. Blasgen	William C. McGee	Gerald Held
18	Raymond F. Boyce	Gerald Held	Diane C. P. Smith
19	**Ronald Fagin**	Gianfranco R. Putzolu	Gianfranco R. Putzolu
20	Gianfranco R. Putzolu	Diane C. P. Smith	**David J. Dewitt**
21	Edward M. McCreight	Nathan Goodman	Nathan Goodman
22	Nathan Goodman	Michael Hammer	Michael Hammer
23	James W. Mehl	Mike W. Blasgen	Mike W. Blasgen
24	W. Frank King III	Stephen Todd	Hans Albrecht Schmid
25	Bradford W. Wade	Hans Albrecht Schmid	Stephen Todd
26	Paul R. McJones	Bradford W. Wade	Paul R. McJones
27	Robert C. Goldstein	James W. Mehl	Bradford W. Wade
28	Stephen Todd	Paul R. McJones	James W. Mehl
29	Patricia P. Griffiths	W. Frank King III	Patricia P. Griffiths
30	Diane C. P. Smith	Patricia P. Griffiths	W. Frank King III
31	Philip Yen-tang Chang	Alfred V. Aho	Alfred V. Aho
32	Peter Kreps	Peter Kreps	Peter Kreps
33	Vera Watson	Yehoshua Sagiv	Edward M. McCreight
34	Peter P. Chen	Edward M. McCreight	Robert C. Goldstein
35	Catriel Beeri	**David Maier**	Moshé M. Zloof
36	**David J. Dewitt**	Robert C. Goldstein	Philip Yen-tang Chang
37	Alfred V. Aho	Raymond F. Boyce	Raymond F. Boyce
38	John J. Donovan	Moshé M. Zloof	Vera Watson
39	Stuart G. Greenberg	Vera Watson	C. J. Date
40	Loius M. Gutentag	Umeshwar Dayal	Peter P. Chen
	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

Table 10. Top 40 DBLP authors for each ranking (part 3)

	b	c	d
1	**Michael Stonebraker**	**Michael Stonebraker**	**Michael Stonebraker**
2	**Jim Gray**	**Jim Gray**	**Jim Gray**
3	**David J. Dewitt**	**David J. Dewitt**	**David J. Dewitt**
4	**Hector Garcia-Molina**	**Hector Garcia-Molina**	**Philip A. Bernstein**
5	**Jeffrey D. Ullman**	**Jeffrey D. Ullman**	**Hector Garcia-Molina**
6	**Philip A. Bernstein**	**Philip A. Bernstein**	**David Maier**
7	**David Maier**	**David Maier**	**Jeffrey D. Ullman**
8	Moshe Y. Vardi	Umeshwar Dayal	Umeshwar Dayal
9	E. F. Codd	Bruce G. Lindsay	**Michael J. Carey**
10	Catriel Beeri	**Michael J. Carey**	E. F. Codd
11	Umeshwar Dayal	**Serge Abiteboul**	Bruce G. Lindsay
12	**Serge Abiteboul**	Jeffrey F. Naughton	Catriel Beeri
13	**Michael J. Carey**	Catriel Beeri	Jeffrey F. Naughton
14	Yehoshua Sagiv	Hamid Pirahesh	**Serge Abiteboul**
15	Christos H. Papadimitriou	Moshe Y. Vardi	Hamid Pirahesh
16	**Rakesh Agrawal**	Hans-Jörg Schek	Goetz Graefe
17	Bruce G. Lindsay	E. F. Codd	Hans-Jörg Schek
18	Jeffrey F. Naughton	Yehoshua Sagiv	**Rakesh Agrawal**
19	Nick Roussopoulos	**Rakesh Agrawal**	Raymond A. Lorie
20	Hans-Jörg Schek	Raghu Ramakrishnan	Yehoshua Sagiv
21	Raghu Ramakrishnan	Goetz Graefe	Nick Roussopoulos
22	Hamid Pirahesh	Nick Roussopoulos	Gio Wiederhold
23	Goetz Graefe	Raymond A. Lorie	**Donald D. Chamberlin**
24	Raymond A. Lorie	Christos H. Papadimitriou	Moshe Y. Vardi
25	Alberto O. Mendelzon	Gio Wiederhold	Dina Bitton
26	Gio Wiederhold	**Donald D. Chamberlin**	Richard T. Snodgrass
27	**Ronald Fagin**	Richard T. Snodgrass	Christos H. Papadimitriou
28	Richard T. Snodgrass	**Ronald Fagin**	Raghu Ramakrishnan
29	**Donald D. Chamberlin**	Dina Bitton	Guy M. Lohman
30	François Bancilhon	Jennifer Widom	**Ronald Fagin**
31	Mihalis Yannakakis	Randy H. Katz	Randy H. Katz
32	Jennifer Widom	Alberto O. Mendelzon	François Bancilhon
33	Nathan Goodman	Guy M. Lohman	Alberto O. Mendelzon
34	Randy H. Katz	François Bancilhon	Jennifer Widom
35	H. V. Jagadish	H. V. Jagadish	Michael J. Franklin
36	Won Kim	Abraham Silberschatz	Irving L. Traiger
37	Irving L. Traiger	Irving L. Traiger	H. V. Jagadish
38	Abraham Silberschatz	Michael J. Franklin	Won Kim
39	Eugene Wong	Mihalis Yannakakis	Eugene Wong
40	Guy M. Lohman	Nathan Goodman	Nathan Goodman
	Missed: 61. Patricia Selinger, 62. C. Mohan, 97. Rudolf Bayer	Missed: 55. Patricia Selinger, 59. C. Mohan, 132. Rudolf Bayer	Missed: 54. Patricia Selinger, 65. C. Mohan, 94. Rudolf Bayer

Table 11. Top 40 DBLP authors for each ranking (part 4)

	e	f	g
1	**Michael Stonebraker**	**Jim Gray**	E. F. Codd
2	**David J. Dewitt**	E. F. Codd	**Jim Gray**
3	**Hector Garcia-Molina**	**Michael Stonebraker**	**Michael Stonebraker**
4	**Jim Gray**	**David J. Dewitt**	**Philip A. Bernstein**
5	**Jeffrey D. Ullman**	**Philip A. Bernstein**	**David J. Dewitt**
6	**Philip A. Bernstein**	Raymond A. Lorie	**Donald D. Chamberlin**
7	**David Maier**	**Donald D. Chamberlin**	Raymond A. Lorie
8	Moshe Y. Vardi	**Jeffrey D. Ullman**	**Jeffrey D. Ullman**
9	Umeshwar Dayal	Irving L. Traiger	Irving L. Traiger
10	Catriel Beeri	Morton M. Astrahan	Morton M. Astrahan
11	E. F. Codd	**David Maier**	John Miles Smith
12	**Serge Abiteboul**	Eugene Wong	Eugene Wong
13	Yehoshua Sagiv	Catriel Beeri	**David Maier**
14	**Michael J. Carey**	John Miles Smith	**Hector Garcia-Molina**
15	**Rakesh Agrawal**	Bruce G. Lindsay	Catriel Beeri
16	Christos H. Papadimitriou	**Hector Garcia-Molina**	Kapali P. Eswaran
17	Bruce G. Lindsay	**Ronald Fagin**	**Ronald Fagin**
18	Jeffrey F. Naughton	Kapali P. Eswaran	Gerald Held
19	Nick Roussopoulos	Gerald Held	Umeshwar Dayal
20	Hans-Jörg Schek	Umeshwar Dayal	**Rudolf Bayer**
21	Raghu Ramakrishnan	**Michael J. Carey**	Michael Hammer
22	Hamid Pirahesh	Yehoshua Sagiv	Bruce G. Lindsay
23	Raymond A. Lorie	Gianfranco R. Putzolu	Nathan Goodman
24	Alberto O. Mendelzon	Nathan Goodman	Gianfranco R. Putzolu
25	**Ronald Fagin**	**Rudolf Bayer**	Stephen Todd
26	**Donald D. Chamberlin**	Mike W. Blasgen	Diane C. P. Smith
27	Gio Wiederhold	Michael Hammer	William C. McGee
28	Goetz Graefe	William C. McGee	Mike W. Blasgen
29	Nathan Goodman	Stephen Todd	**Michael J. Carey**
30	Mihalis Yannakakis	Diane C. P. Smith	Phyllis Reisner
31	François Bancilhon	Jeffrey F. Naughton	Paul R. McJones
32	Jennifer Widom	Thomas G. Price	Jeffrey F. Naughton
33	Randy H. Katz	Bradford W. Wade	Hamid Pirahesh
34	Richard T. Snodgrass	Hamid Pirahesh	Yehoshua Sagiv
35	Abraham Silberschatz	Phyllis Reisner	Bradford W. Wade
36	H. V. Jagadish	**Patricia G. Selinger**	Hans Albrecht Schmid
37	Guy M. Lohman	**Serge Abiteboul**	Nick Roussopoulos
38	Eugene Wong	W. Frank King III	Won Kim
39	Peter Buneman	François Bancilhon	James W. Mehl
40	Christos Faloutsos	James W. Mehl	W. Frank King III
	Missed: 63. Patricia Selinger, 65. C. Mohan, 93. Rudolf Bayer	Missed: 49. Rakesh Agrawal, 105. C. Mohan	Missed: 43. Serge Abiteboul, 48. Patricia Selinger, 49. Rakesh Agrawal, 101. C. Mohan