This is the preliminary JASIST submission

P-Rank: An indicator measuring prestige in heterogeneous scholarly networks

Erjia Yan[1], Ying Ding, Cassidy R. Sugimoto

School of Library and Information Science, Indiana University, Bloomington, USA

Abstract

Ranking scientific productivity and prestige are often limited to homogeneous networks. These networks are unable to account for the multiple factors that constitute the scholarly communication and reward system. This study proposes a new informetric indicator, P-Rank, for measuring prestige in heterogeneous scholarly networks containing articles, authors, and journals. P-Rank differentiates the weight of each citation based on its citing papers, citing journal, and citing authors. Articles from 16 representative library and information science journals are selected as the dataset. Principle Component Analysis is conducted to examine the relationship between P-Rank and other bibliometric indicators. We also compare the correlation and rank variances between citation counts and P-Rank scores. This work provides a new approach to examining prestige in scholarly communication networks in a more comprehensive and nuanced way.

1. Introduction

Citation analysis has long served as a formal instrument for quantitative scientific evaluation. Citations create a channel between scholarly communications (Cronin, 1984) and form the basis of the scientific reward system (Merton, 1968; Luukkonen, 1997). In this system, citations serve as "concept symbols" (Small, 1978), associating and crediting an author with a concept or contribution to the literature. The accumulation of these citations (either by a single author or aggregated to represent a journal, institution, etc.) represents the impact of that author (or aggregate) upon the domain. In accumulating citations, each citation is given equal weight, and thereby equal importance. In this way, the author with the largest number of citations has the greatest value within the system, regardless of the provenance of those citations.

[1] *Correspondence to*: Erjia Yan, School of Library and Information Science, Indiana University, 1320 E. 10th St., LI011, Bloomington, Indiana, 47405, USA. Email: eyan@indiana.edu

This paper argues that this equal weighting may be conflating the popularity of an article for prestige. As Pinski and Narin (1976) noted, "it seems more reasonable to give higher weight to a citation from a prestigious journal than to a citation from a peripheral one" (p. 298). Cronin (1984) also posited that the weight of citations should be differentiated to reflect the prestige of citing journals. Bollen, Rodriguez, and de Sompel (2006) argued that "popularity" and "prestige" are not identical measures of journal impact. Ding and Cronin (2010) defined author popularity as the number of times an author is cited and author prestige as the number of times an author is cited by highly cited papers. This paper adopts these notions of popularity and prestige and describes a model for evaluating scientific productivity by placing weights on 1) the citing articles, 2) the citing authors, and 3) the citing journals.

This paper extends Yan and Ding's (2010a) study of heterogeneous networks. In the present study, P-Rank is proposed as an indicator for identifying scholarly prestige based on the weight of citing papers, citing authors, and citing journals. Articles from 16 representation library and information science (LIS) journals are selected as the dataset. P-Rank is used to rank papers, authors, and journals within this domain.

This research is valuable to the scientometric community as it provides a more comprehensive and nuanced way to evaluate scholars and research aggregates. It may also be informative for administrators and policy makers looking to improve science indicators.

2. Related studies

In recent years, we have witnessed a trend of using scientific networks to evaluate scholars, institutions, countries, and other research aggregates, including coauthorship networks (Liu et al., 2005; Yin et al., 2006; Liu et al., 2007; Yan & Ding, 2009), paper citation networks (Chen et al., 2007; Ma, Guan, & Zhao, 2008), author citation networks (Radicchi et al., 2009), journal citation networks (Bollen et al., 2006; Leydesdorff, 2007, 2009), and author cocitation networks (Ding, Yan, Frazho, & Caverlee, 2009).

PageRank-like indicators denote a collection of algorithms based on Google's PageRank, such as AuthorRank (Liu et al., 2005), Y-factor (Bollen et al., 2006), CiteRank (Walker et al., 2007), FutureRank (Sayyadi & Getoor, 2009), Eigenfactor (Bergstrom & West, 2008), and SCImago Journal Rank (SCImago, 2007). Lopez-Illescas et al. (2008) found a high correlation between journal impact factor and SCImago journal rank for journals indexed in 2006. Fersht (2009) found there is a strong correlation between Eigenfactors and the total number of citations for journals. Leydesdorff (2009) compared PageRank with *h*-index, impact factor, centrality measures, and SCImago Journal Rank, and found that PageRank is mainly an indicator of size, but has important interactions with centrality measures. Bollen et al. (2009) conducted a Principal Component Analysis for

39 indicators on the basis of citation and usage data. They found that these indicators can be measured in two dimensions: prestige vs. popularity and rapid vs. delayed.

Most of these studies focus on one-mode networks a.k.a. homogenous networks. They aimed to differentiate the weight of citations based on citing paper, citing author, or citing journal separately. Besides homogenous scholarly networks, some studies combine different types of networks to form heterogeneous scholarly networks, for example, author-article networks (Zhou et al., 2007; Sayyadi & Getoor, 2009) and journal-article networks (Yan & Ding, 2010b). The co-ranking model (Zhou et al., 2007) coupled two networks: a coauthorship network and a paper citation network, and connected the two networks by a paper-author matrix. FutureRank (Sayyadi & Getoor, 2009) used coauthorship and citation networks to predict future citations. FutureRank has two main procedures: values for articles are first obtained by calculating PageRank for the article citation network, and values for authors are then obtained based on a paper-author matrix.

All of the previous studies focused on pairing two networks. However, the scholarly communication process involves more than two units. Therefore, the present study seeks to expand upon these studies by proposing a model that integrates papers, authors, and journals. The proposed heterogeneous scholarly network allows authors to interact with papers via paper-author adjacency matrix (authorship), journals to interact with papers via paper-journal adjacency matrix (journal-ship), and papers to interact with other papers via citations (Figure 1). This allows for a more comprehensive assessment of scholarly ranking than has been previously possible. The product of this model is called P-Rank.

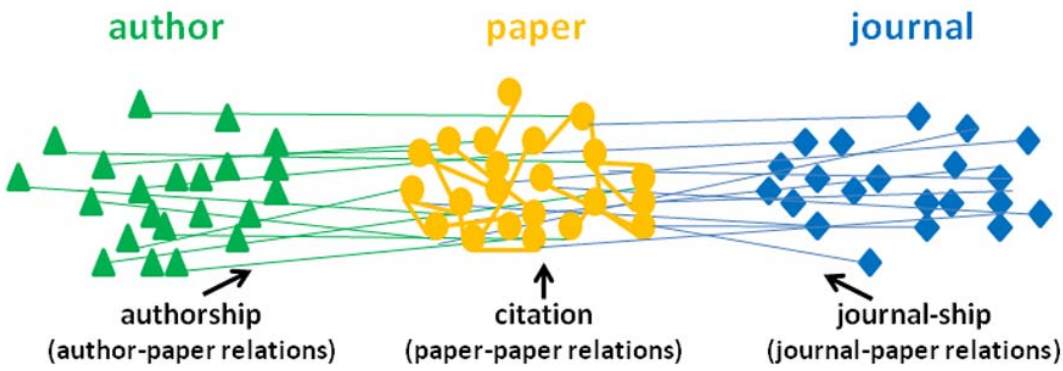

Figure 1. A heterogeneous scholarly network

3 Methods

3.1 Data collection

In order to examine the application of the P-Rank indicator for a given domain, 16 representative journals in LIS were selected: *Annual Review of Information Science and*

Technology; *Information Processing & Management*; *Scientometrics*; *Journal of the American Society for Information Science and Technology* (*Journal of the American Society for Information Science*); *Journal of Documentation*; *Journal of Information Science*; *Information Research-An International Electronic Journal*; *Library & Information Science Research*; *College & Research Libraries*; *Information Society*; *Online Information Review* (*Online and CD-ROM Review*; *On-Line Review*); *Library Resources & Technical Services*; *Library Quarterly*; *Journal of Academic Librarianship*; *Library Trends*; and *Reference & User Services Quarterly*. These journals were selected based on perception (Nisonger & Davis, 2005) and citation-based rankings (Journal Citation Reports). Additionally, only those journals indexed by Thomson Reuters' Web of Knowledge (WoK) were included. Using WoK, all articles published in the selected journals between 1988 and 2007 were identified. The results were refined by document type "article" and "review article". In total, 10,344 articles were identified. Records for all articles (including citation information) were downloaded from WoK for processing[2]. Table 1 describes the dataset.

Table 1. Summary statistics of the data

	Number
Number of citing articles	10,344
Number of cited references	205,283
Total times cited	298,830
Number of cited authors/group authors*	89,301
Number of cited journals/proceedings/books	87,610
Dimension of paper citation matrix	205,283*205,283
Dimension of paper-author adjacency matrix	205,283*89,301
Dimension of paper-journal adjacency matrix	205,283*87,610

*The cited references in WoK only contain the first author.

It should be noted that the issue of self citations is considered. Self-citations can be defined at the author level (Aksnes, 2003; Hyland, 2003; Glänzel & Thijs, 2004), at the journal level (Tsay, 2006; Krauss, 2007), or at the research group level (van Raan, 2008). Indices such as the Eigenfactor, a bibliometric indicator incorporated into the Journal Citation Report since 2007, excludes journal self-citations to avoid over-inflated journals that engage in the practice of opportunistic self-citations (Franceschet, 2009; West, Bergstrom, & Bergstrom, 2010). However, although it is recognized that manipulation can occur in self-citation practices, self-citations can also be a legitimate form of citing behavior. If an author consistently builds upon their past work, citing themselves can be fundamental for the arguments they propose. Therefore, the P-Rank indicator includes self-citations, but provides a lower weight for these citations: the value of 1 is assigned to a non-self-citation, 0.5 to a journal self-citation, and 0.25 to an author self-citation.

[2] The Pajek formatted data can be found at: http://ella.slis.indiana.edu/~eyan/papers/LIS.net; and the Matlab formatted data can be found at: http://ella.slis.indiana.edu/~eyan/papers/LIS.mat.

3.2 P-Rank

The P-Rank indicator introduces weighted citations, in order to more comprehensively evaluate scholars and sources. The indicator is predicated on the following assumptions:

1. Articles are more important if they are cited by other important articles (Chen et al., 2007; Ma et al., 2008; Maslov & Redner, 2008; Ding & Cronin, 2010);
2. Authors have a higher impact if their articles are cited by important articles, and articles are important if they are cited by prestigious authors (Zhou et al., 2007; Sayyadi & Getoor, 2009);
3. Journals have a higher impact if their articles are cited by important articles, and articles are important if they are cited by prestigious journals (Pinski & Narin, 1976; Cronin, 1984; Davis, 2008; Yan & Ding, 2010b); and

Note that citation context is not considered in the three assumptions. A paper can be cited for different purposes, such as background reading, crediting, validating, correcting, criticizing, etc. (Garfield, 1965). Therefore, references cited for the purpose of crediting may be more important to the citing paper; references cited for the purpose of background reading may be less important; and references cited for the purpose of criticizing and disputing may even have negative importance (Garfield, 1979). Due the complexity of identifying citation types, we do not distinguish citation contexts in the study.

The references listed for each assumption demonstrate that these issues have been studied by previous authors. However, few researchers have sought to incorporate all these assumptions into a single indicator. Therefore, the P-Rank indicator measures the prestige of an article by examining three factors: 1) the papers that cite the article, 2) the journals that cite the article, and 3) the authors who cite the article. As indicated by the factors, the unit measured is the journal article. However, rankings for an author can be determined by the status of all articles written by that author; similarly, the ranking of a journal can be determined by the status of all articles published within that journal (Figure 2).

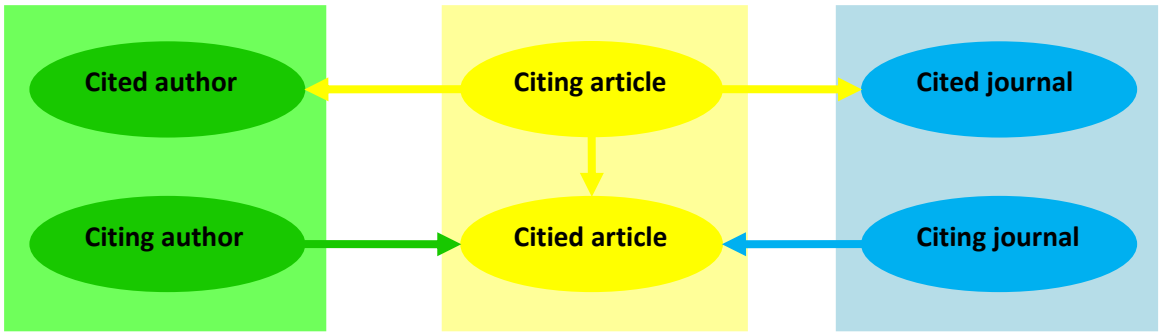

Figure 2. The P-Rank for heterogeneous scholarly network

The heterogeneous graph of authors, journals and articles can be represented as: $G=(V,E)=(V_{AU}\cup V_{AR}\cup V_J, E_{AR}\cup E_{AR-AU}\cup E_{AR-J})$, where V_{AR} represents the article set and E_{AR} represents the link set between articles and citations. Therefore, $G_{AR}=(V_{AR},E_{AR})$ is the unweighted direct graph (citation network) of articles and $G_{AR-AU}=(V_{AR}\cup V_{AU},E_{AR-AU})$ is the unweighted bipartite graph of authors and $G_{AR-J}=(V_{AR}\cup V_J,E_{AR-J})$ is the unweighted bipartite graph of journals. Edges in E_{AR-AU} connect each article with its authors and edges in E_{AR-J} connect each article with its journal.

The proposed scholarly network contains three walks: an intra-class walk within the paper citation network G_{AR} and two inter-class walks, between article and author G_{AR-AU} and between article and journal G_{AR-J}. PageRank is used as the underlying algorithm for the intra-class walk. Let M be the $n \times n$ matrix for the paper citation matrix, where n is the number of nodes in the network:

$$M_{i,j}=\begin{cases}1 & \text{if paper } i \text{ cites paper } j\\ 0 & \text{otherwise}\end{cases} \quad (1)$$

$\overline{M}$ is the fractioned citation matrix where $\overline{M}_{i,j}=\frac{M_{i,j}}{\sum_{i=1}^{n} M_{i,j}}$. Let e be the n-vector whose elements are all ones and v is an n-vector, also referred to as personalized vector (Haveliwala, Kamvar, & Jeh, 2003); and let $x(v)$ be the PageRank vector corresponding to the personalized vector v. Based on this, $x(v)$ can be computed by solving $x=\overline{\overline{M}}x$ (Haveliwala et al., 2003), where $\overline{\overline{M}}$ is the stochastic matrix and $\overline{\overline{M}}=d\overline{M}+(1-d)ve^T$. Therefore, x can be calculated as:

$$x=(1-d)(I-d\overline{M})^{-1}v \quad (2)$$

By letting $N=(1-d)(I-d\overline{M})^{-1}$, then $x=Nv$. According to Haveliwala et al. (2003), N comprises a complete basis for personalized PageRank vectors, since any personalized PageRank vector can be expressed as a convex combination of the columns of N. For any v, the corresponding personalized PageRank vector is given by Nv.

For the two inter-class walks, adjacency matrices are used to define the bipartite graphs. $A_{author\,i,j}$ is the $n \times m$ paper-author adjacency matrix, where n is the number of papers and m is the number of authors:

$$A_{author\,i,j} = \begin{cases} 1 & \text{if author } j \text{ writes paper } i \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

This matrix is used to link the citing authors to citing articles. Similarly, $A_{\text{journal}\,i,j}$ is the $n \times q$ paper-journal adjacency matrix, where *n* is the number of papers and *q* is the number of journals:

$$A_{\text{journal}\,i,j} = \begin{cases} 1 & \text{if paper } i \text{ is publised on journal } j \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

Hence, the P-Rank score of articles can be expressed as $x(v)_{article}$ in formula (2), where the personalized vector is

$$v = (\alpha((x(v)_{author} / np\_A)^T \times A^T{}_{author}) + \beta((x(v)_{journal} / np\_J)^T \times A^T{}_{journal}))^T ; \alpha + \beta = 1.$$

where *np_A* is a vector with the number of publications for each author, *np_J* is a vector with the number of publications for each journal. The intra-class and inter-class walks are coupled by α and β. α and β represents the mutual dependence of papers, authors, and journals (Zhou et al., 2007).

The P-Rank score of author can be expressed as:

$$x(v)_{author} = A_{author}{}^T \times x(v)_{article} \quad (5)$$

and the P-Rank score of journals can be expressed as:

$$x(v)_{journal} = A_{journal}{}^T \times x(v)_{article} \quad (6)$$

The damping factor for this study is set at 0.85 as default, and α and β are set at 0.5. Different damping factors and parameters may make a difference in the outcome, but we do not investigate it in the present study.

The following is the pseudocode of the algorithm applied. Each paper are allocated with the same score of 1/*n_P* where *n_P* is the number of papers; authors attain their scores via paper-author adjacency matrix A_{author} and journals attain their scores via paper-journal adjacency matrix $A_{journal}$; personalized vector *v* is then calculated; personalized PageRank is finally computed based on the personalized vector *v*. Above three steps are recursively implemented until convergence.

Algorithm: Intra- and inter- walks on the heterogeneous network

procedure $P - Rank(\overline{M}, A_{author}, A_{journal}, \alpha, \beta, \gamma, d)$

1 $x(v)_{article} = ones(n\_P,1)/n\_P$

2 **while** not converging

3 $x(v)_{author} = A_{author}^{T} \times x(v)_{article}$

4 $x(v)_{journal} = A_{journal}^{T} \times x(v)_{article}$

5 $v = (\alpha((x(v)_{author}/np\_A)^T \times A^T{}_{author}) + \beta((x(v)_{journal}/np\_J)^T \times A^T{}_{journal})$

6 $PageRank\ (\overline{M}, v)$

7 **end**

8 **return** $x(v)_{article}$, $x(v)_{author}$, and $x(v)_{journal}$

4 Results

4.1 Values for parameters

Two parameters can be manipulated in P-Rank: α and β. If $\alpha = 0, \beta = 0$, there is no coupling, which would be the situation of ranking the articles using a standard PageRank algorithm. However, if a new unit (namely, the journal) is introduced into the network, the parameters can be redefined as $\alpha = 0, \beta = 1$. This introduces one intra-walk (the citation network) and one inter-walk (the journal network) into the network, creating a heterogeneous network. The final manipulation involves adding authorship, which results in: $\alpha = \beta = 0.5$. The result is the combination of one intra-walk (citation network) and two inter-walks (journal and author networks) for the final heterogeneous network.

The values of the parameters depend upon the assumptions guiding the research. Four cases have been identified, using various combinations of the assumptions identified in the Methods section. Case 1 uses assumption 1; Case 2 uses assumption 1 and 3; Case 3 uses assumption 1 and 2; Case 4 uses assumption 1, 2, and 3. The cases and the associated parameters are labeled below:

- Case 1: Article citation network ($\alpha = 0, \beta = 0$)
- Case 2: Article-Journal citation network ($\alpha = 0, \beta = 1$)
- Case 3: Article-Author citation network ($\alpha = 1, \beta = 0$)
- Case 4: Article-Journal-Author citation network ($\alpha = \beta = 0.5$)

Using these case assumptions, journal rankings were calculated for four cases. The top 10 journals for each case are shown in Table 2.

Table 2. Top 10 journals (Cases 1-4)

Case 1	Case 2	Case 3	Case 4
J AM SOC INF SCI TEC	SCIENTOMETRICS	J AM SOC INF SCI TEC	J AM SOC INF SCI TEC

SCIENTOMETRICS	J AM SOC INF SCI TEC	SCIENTOMETRICS	SCIENTOMETRICS
COLL RES LIBR	COLL RES LIBR	COLL RES LIBR	COLL RES LIBR
J ACAD LIBR	J DOC LIBR TRENDS	J ACAD LIBR	J ACAD LIBR
LIBR J	J ACAD LIBR	LIBR J	LIBR J
LIBR TRENDS	RES POLICY	LIBR TRENDS	J DOC
J INFORM SCI	J INFORM SCI	J DOC	LIBR TRENDS
J DOC	LIBR J	J INFORM SCI	J INFORM SCI
INFORM PROCESS MANAG	INFORM PROCESS MANAG	INFORM PROCESS MANAGE	INFORM PROCESS MANAG
COMMUN ACM	SCIENCE	COMMUN ACM	COMMUN ACM

Journal rankings for the four cases are relatively stable: 14 journals occur at least once. Six journals appear top 10 for all four cases (*Journal of the American Society for Information Science and Technology*, *Scientometrics*, *College and Research Libraries*, *Journal of Academic Librarianship*, *Library Journal*, and *Journal of Information Science*). Four journals appear three times (*Communication of ACM*, *Information Processing and Management*, *Journal of Documentation*, and *Library Trend*) and four journals appear once.

Ranking were also generated for the top 10 publications, by case.

Table 3. Top 10 publications (Case 1-4)

Case 1	Case 2
Salton G, 1983, INTRO MODERN INFORMA	Lotka AJ, 1926, J WASHINGTON ACADEMY, V16, P317
Van Rijsbergen CJ, 1979, INFORMATION RETRIEVA	Bradford SC, 1934, ENGINEERING-LONDON, V137, P85
Garfield E, 1979, CITATION INDEXING	Salton G, 1983, INTRO MODERN INFORMA
Salton G, 1989, AUTOMATIC TEXT PROCE	Lotka AJ, 1926, J WASHINGTON ACADEMY, V16, P109
Lotka AJ, 1926, J WASHINGTON ACADEMY, V16, P317	Salton G, 1989, AUTOMATIC TEXT PROCE
Price DJD, 1963, LITTLE SCI BIG SCI	Garfield E, 1979, CITATION INDEXING
Price DJD, 1965, SCIENCE, V149, P510	Braun T, 1985, SCIENTOMETRIC INDICA
Garfield E, 1972, SCIENCE, V178, P471	Van Rijsbergen CJ, 1979, INFORMATION RETRIEVA
Lawrence S, 1999, NATURE, V400, P107	Price DJD, 1963, LITTLE SCI BIG SCI
Robertson SE, 1976, J AM SOC INF SCI TEC, V27, P129	Frame JD, 1977, INTERSCIENCIA, V2, P143
Case 3	**Case 4**
Salton G, 1983, INTRO MODERN INFORMA	Salton G, 1983, INTRO MODERN INFORMA
Lotka AJ, 1926, J WASHINGTON ACADEMY, V16, P317	Lotka AJ, 1926, J WASHINGTON ACADEMY, V16, P317
Garfield E, 1979, CITATION INDEXING	Garfield E, 1979, CITATION INDEXING
Vanrijsbergen CJ, 1979, INFORMATION RETRIEVA	Salton G, 1989, AUTOMATIC TEXT PROCE
Salton G, 1989, AUTOMATIC TEXT PROCE	Van Rijsbergen CJ, 1979, INFORMATION RETRIEVA
Schauder D, 1994, J AM SOC INF SCI TEC, V45, P73	Bradford SC, 1934, ENGINEERING-LONDON, V137, P85
Price DJD, 1963, LITTLE SCI BIG SCI	Price DJD, 1963, LITTLE SCI BIG SCI
Hirsch JE, 2005, P NATL ACAD SCI USA, V102, P16569	Braun T, 1985, SCIENTOMETRIC INDICA
Schubert A, 1989, SCIENTOMETRICS, V16, P3	Price DJD, 1965, SCIENCE, V149, P510
Braun T, 1985, SCIENTOMETRIC INDICA	Lawrence S, 1999, NATURE, V400, P107

In Case 1, the parameters $\alpha = 0, \beta = 0$ result in a pure citation network (the standard PageRank calculation). Case 2 adds the journal relation to the citation network. Two of Lotka's articles rank within the top 10, as they are cited by prestigious journals. Case 3 adds the author relation to the citation network. We find Hirsch's 2005 *h*-index article ranks 6th, for the reason that his article is cited more by renowned authors. As shown, for cases 1 to 4, more than half of the publications among the top 10 are monographs. This may be the result of an emphasis on dangling nodes (nodes cited that do not cite other nodes in the network) (Yan & Ding, 2010 submitted). However, when journal relations are added to the citation network (Case 2), the number of monographs in the top five decreases. This is likely the result of the different citing behaviors of communicative genres, as this calculation favors the journal-journal citation network (Sugimoto, 2010). Case 4 is the combination of all elements (citations, journals, and authors)—covering all three assumptions for the P-Rank indicator. In order to examine the relationship of Case 4 with the other cases, a correlation analysis was conducted (Figure 3).

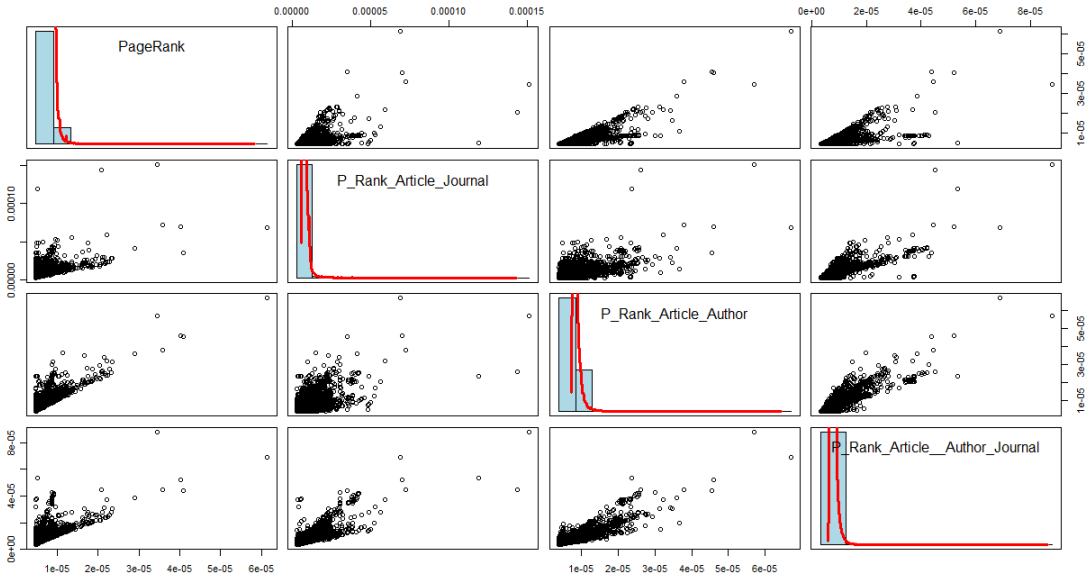

Figure 3. Comparison of different cases

Figure 3 illustrates the relationship between each case pair. Case 2 (P-Rank Article-Journal) and Case 3 (P-Rank Article-Author) have a strong relationship with Case 4 (P-Rank Article-Author-Journal), indicating that adding journal and author component respectively to the citation network partially changes P-Rank scores. If we compare the scores of Case 2 (P-Rank Article-Journal) and Case 3 (P-Rank Article-Author), however, we may find that they have weak relationship, which means that adding journal-ship or authorship can yield quite different results for P-Rank scores.

4.2 Top authors, journal, and publications

Informed by the case studies, the dataset was then analyzed according to the P-Rank indicator (Case 4). In addition, a citation count and corresponding rankings are provided for comparative purposes. Table 4 provides a listing of the top 20 authors, by P-Rank along with citation counts and rankings.

Table 4. Top 20 authors

Author	P-Rank		Citation		Author	P-Rank		Citation	
	Score	Rank	Count	Rank		Score	Rank	Count	Rank
Garfield E	3.33E-03	1	1348	2	Rousseau R	9.10E-04	11	451	23
Salton G	1.99E-03	2	1780	1	Narin F	9.05E-04	12	488	20
Egghe L	1.57E-03	3	906	3	Hernon P	9.01E-04	13	364	38
Spink A	1.26E-03	4	799	6	Borgman CL	8.83E-04	14	646	11
Cronin B	1.10E-03	5	712	9	Dervin B	8.76E-04	15	767	7
Tenopir C	1.07E-03	6	438	27	Braun T	8.34E-04	16	424	30
ALA	1.07E-03	7	245	59	Thelwall M	8.24E-04	17	601	12
Saracevic T	1.01E-03	8	906	4	Belkin NJ	8.02E-04	18	764	8
Lancaster FW	9.98E-04	9	508	19	Jacso P	7.58E-04	19	171	92
Leydesdorff L	9.59E-04	10	598	13	Bookstein A	7.43E-04	20	405	33

Since cited references in WoK only contain the first author, results in such case would favor first authors but not collaborative authors. Table 4 is thus used for illustrative purposes. In formula (3), the paper-author adjacency matrix links a paper with all its authors. P-Rank, therefore, is suitable for multi-authorship scholarly networks.

The list of top twenty authors is indicative of some of the dominant areas of research within LIS—the authors could be divided into three main groups (in descending order of prominence within the list): scientometrics, information retrieval, and information seeking. We also find one group author in the top 20 list (American Library Association).

Table 5 provides a listing of the top 20 journals based on P-Rank score.

Table 5. Top 20 journals

Journal	P-Rank		Citation		5-year	Eigenfactor	
	Score	Rank	Count	Rank	Impact Factor	Eigenfactor	Article Influence
J AM SOC INF SCI TEC	1.58E-02	1	14747	1	2.18	0.010	0.67
SCIENTOMETRICS	1.32E-02	2	7357	2	2.30	0.006	0.50
COLL RES LIBR	8.08E-03	3	3846	5	1.16	0.002	0.58
J ACAD LIBR	7.15E-03	4	2184	9	0.68	0.002	0.26
LIBR J	6.45E-03	5	1847	17	0.28	0.002	0.12
LIBR TRENDS	6.30E-03	6	1860	11	0.61	0.001	0.19
J DOC	5.61E-03	7	4867	3	1.91	0.002	0.57
J INFORM SCI	5.45E-03	8	2387	7	1.35	0.002	0.34
INFORM PROCESS MANAG	4.69E-03	9	4564	4	2.02	0.005	0.54
COMMUN ACM	4.66E-03	10	2394	32	3.18	0.018	0.95
LIBR QUART	4.65E-03	11	2091	10	0.82	0.001	0.28

SCIENCE	4.27E-03	12	1451	14	30.27	1.583	16.29
LIBR INFORM SCI RES	3.80E-03	13	1853	12	1.30	0.001	0.44
ANNU REV INFORM SCI	3.72E-03	14	2237	8	2.95	0.001	0.96
RQ	3.26E-03	15	1215	18	-	-	-
CATALOGING CLASSIFIC	2.91E-03	16	843	22	-	-	-
ONLINE	2.73E-03	17	775	24	0.36	0.001	0.14
INFORMATION PROCESSI	2.72E-03	18	1175	19	-	-	-
RES POLICY	2.60E-03	19	1193	16	4.04	0.013	1.17
SERIALS LIBR	2.57E-03	20	667	51	-	-	-

P-Rank score and number of citations yield the same rank for top two journals: *Journal of the American Society for Information Science and Technology* and *Scientometrics*. As shown, there are also non-LIS journals within the top twenty: *Science*, *Communications of the ACM*, and *Research Policy*. This demonstrates that these works are extensively cited by LIS journals and have an impact upon the field. This may also demonstrate the high level of interdisciplinarity within the field. The difference between P-Rank rank and citation rank is not as noticeable for journals as with authors: there are only four journals that appear in the top twenty for P-Rank that do not occur in the top twenty for citation rank (*Communications of the ACM*, *Cataloging Classification*, *Online*, and *Serials Librarian*).

Table 6 displays the top 20 publications based on P-Rank score.

Table 6. Top 20 articles/books

Article	P-Rank		Citation	
	Score	Rank	Count	Rank
Salton G, 1983, INTRO MODERN INFORMA	6.55E-05	1	340	1
Lotka AJ, 1926, J WASHINGTON ACADEMY, V16, P317	6.42E-05	2	143	7
Garfield E, 1979, CITATION INDEXING	4.71E-05	3	160	5
Salton G, 1989, AUTOMATIC TEXT PROCE	4.22E-05	4	196	3
Van Rijsbergen CJ, 1979, INFORMATION RETRIEVA	4.21E-05	5	218	2
Bradford SC, 1934, ENGINEERING-LONDON, V137, P85	3.88E-05	6	86	31
Price DJD, 1963, LITTLE SCI BIG SCI	3.40E-05	7	99	21
Braun T, 1985, SCIENTOMETRIC INDICA	3.24E-05	8	44	129
Price DJD, 1965, SCIENCE, V149, P510	3.10E-05	9	110	16
Schubert A, 1989, SCIENTOMETRICS, V16, P3	2.70E-05	10	55	77
Garfield E, 1972, SCIENCE, V178, P471	2.54E-05	11	85	33
Moed HF, 1985, RES POLICY, V14, P131	2.51E-05	12	53	87
Lawrence S, 1999, NATURE, V400, P107	2.72E-05	13	94	25
Hirsch JE, 2005, P NATL ACAD SCI USA, V102, P16569	2.47E-05	14	24	429
Narin F, 1976, EVALUATIVE BIBLIOMET	2.47E-05	15	55	77
Robertson SE, 1976, J AM SOC INF SCI TEC, V27, P129	2.45E-05	16	117	12
Small H, 1973, J AM SOC INF SCI TEC, V24, P265	2.45E-05	17	111	14
Kuhlthau CC, 1991, J AM SOC INF SCI TEC, V42, P361	2.40E-05	18	158	6
Saracevic T, 1975, J AM SOC INF SCI TEC, V26, P321	2.40E-05	19	133	9

Schubert A, 1986, SCIENTOMETRICS, V9, P281	2.40E-05	20	52	89

The difference between P-Rank rank and citation rank is evident in the list of top twenty publications. This may have a direct relationship to the number of units within each component: there are likely more articles than authors, and more authors than journals. Therefore, as the lowest research aggregate, publications may have a less stable P-Rank than larger aggregates (such as authors and journals). In a paper citation network, senior nodes would always have higher probability to be cited as they have longer time for self-display. Therefore, it is not surprising to find that there are very few articles within a decade of the latest date of publication. Hirsch's (2005) *h*-index article is the most recent of the publications. A possible way to objectively evaluate these publications would be compare papers of the same publication year. It is noticeable, however, that authors of the top twenty publications are not the same as the top twenty authors—for example, Lotka, Hirsch, and Van Rijsbergen do not appear in the list of the top twenty authors by P-Rank, but they each have one of the top twenty publications. The same is true for some of the journals of the top cited publications. This may indicate that the main contribution of these units is from a single article.

5 Evaluation

5.1 Principal component analysis for journals

Another way to evaluate an indicator is to compare it with other indicators through principal component analysis (PCA). PCA is useful for reducing the dimensions and to study how different indicators relate with each other. Bollen et al. (2009) conducted a PCA for 39 journal measures on the basis of citation and usage data. Two components are extracted: rapid vs. delayed and popularity vs. prestige. Leydesdorff (2009) compared several journal indicators, including impact factors, *h*-index, centrality measures, and SCImago Journal Ranking, and found that two components, size and impact, are apparent. Here we compare P-Rank with other 12 indicators, including Journal Citation Reports impact factors (Impact_factor and 5_Year_IF), Eigenfactor measures (Eigenfactor and Article_Influence), and centrality measures (Closeness, degree, betweenness, and PageRank_normal) calculated by Leydesdorff (2009).

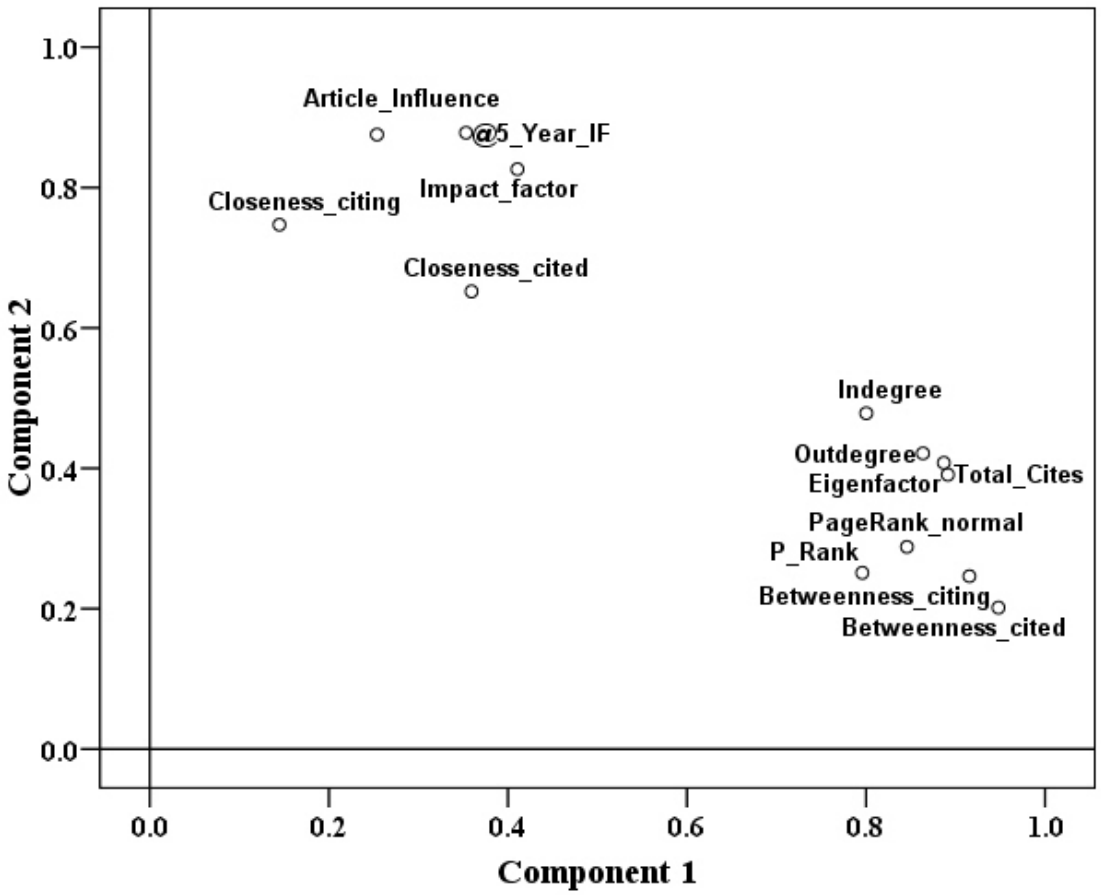

Figure 4. Principal component analysis for journals

Figure 4 shows the result of PCA (using varimax rotation). Two components account for 86% of the total variance. Two groups are evident in Figure 4: group 1 in the top left quadrant that contains impact factor, Article Influence, and closeness centrality, and group 2 in the bottom right quadrant that contains total citation counts, degree centrality, betweenness centrality, Eigenfactor, standard PageRank, and P-Rank. Indicators in group 1 focus on per article impact (impact factor, 5-year impact factor, and Article Influence) or the virtual distance between journals (closeness centrality). These indicators are size independent. Indicators in group 2 focus on the overall performance of a journal (degree centrality, betweenness centrality, total citations, Eigenfactor, PageRank, and P-Rank). These indicators are size dependent, in that a productive journal may have a higher value on degree, total citations, PageRank, or P-Rank. Furthermore, within group 2, there are two sub-groups: one includes Eigenfactor, degree centrality, and total citations, and the other includes betweenness centrality, standard PageRank, and P-Rank. The results are consistent with findings by Bollen et al. (2009) and Leydesdorff (2009) that PageRank and betweenness centrality are collocated and they are in different clusters with citations per article indicators.

5.2 Popularity (number of citations) vs. prestige (P-Rank score)

Social exchange theory considers prestige as an endorsement (Blau, 1964; Coleman, 1990; Henrich & Gil-White, 2001), and prestige is accumulated through each endorsement exchange. If each endorsement is treated as equal, then prestige is the same as popularity discussed in the scientometric community (Bollen et al., 2006; Franceschet, 2009; Yan & Ding, 2010b). If treated with different weights, the sociological version of prestige has the same interpretation as the scientometric one. From the scientometric perspective, prestige is therefore the weighted popularity. The PageRank-like algorithms simulate this prestige recognition procedure: at first, every actor has the same status. Each actor will then deliver its endorsement based on the number of endorsees, and after many rounds of exchanges, actors will have stable endorsements. The number of citations a paper, an author, or a journal receives can thus be considered as scholarly popularity, and the P-Rank score can be considered as prestige.

The units for comparison are citation per publication (CPP) and P-Rank score per publication (PPP). They are size independent, which can avoid the pitfall of using correlation coefficient to measure two variables which share a common size factor, i.e. number of publications, as pointed out by West, Bergstrom, and Bergstrom (2010).

Table 7 shows the Spearman's ranking correlation between CPP and PPP for papers, authors, and journals. We also list the correlation coefficients for larger and more representative research aggregates.

Table 7. Spearman's correlation (CPP vs. PPP)

		Size	Spearman's Correlation
Paper	All	205,283	0.3369
	No. of citations > 5	6,229	0.6131
Author	All	89,301	0.3235
	No. of publications > 5	7,584	0.6175
Journal	All	87,610	0.2747
	No. of publications > 5	4,547	0.5543

Spearman's correlations between CPP and PPP for all three research aggregates are correlated. We also filter out the publications at the "long tail" through number of citations and number of publications, and calculate the correlation between CPP and PPP for larger units, and find that larger research units have higher correlation between CPP and PPP.

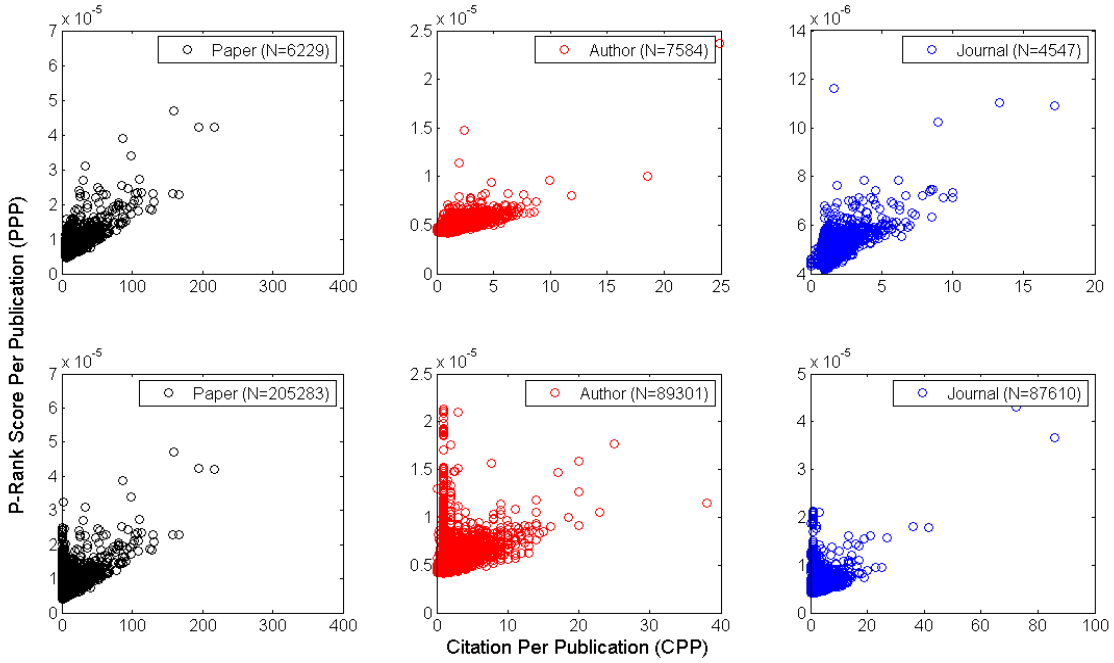

Figure 5. Scatter plots between popularity and prestige

In Figure 5, dots distributed near the virtual diagonal line have similar status on popularity and prestige. For dots above the virtual diagonal line, their prestige outweighs their popularity, and for dots below the virtual diagonal line, their popularity outweighs their prestige. The popularity and prestige for larger research units (figures in the first row) have stronger relationship. For all research units in the second row, papers, authors, or journals that have low CPP have unstable PPP: these PPP are vertically distributed instead of a diagonal distribution pattern.

Several studies have found high correlation between citation counts and scores of PageRank-like indicators for journals (Bollen et al., 2006; Davis, 2008; Lopez-Illescas et al., 2008; Fersht, 2009; Leydesdorff; 2009; Bollen et al., 2009; Franceschet; 2009) and for articles (Chen et al., 2007; Ma et al., 2008; Yan & Ding, 2010 submitted). In one collection, the majority of journals, authors, or articles may have similar status for popularity and prestige, while only a small portion of them have a different status, and hence it is not surprising to discover that discrepancies may occur at the local scale but cannot be reflected at the global level. Based on this outcome, the rank variances for papers, authors, and journals are compared.

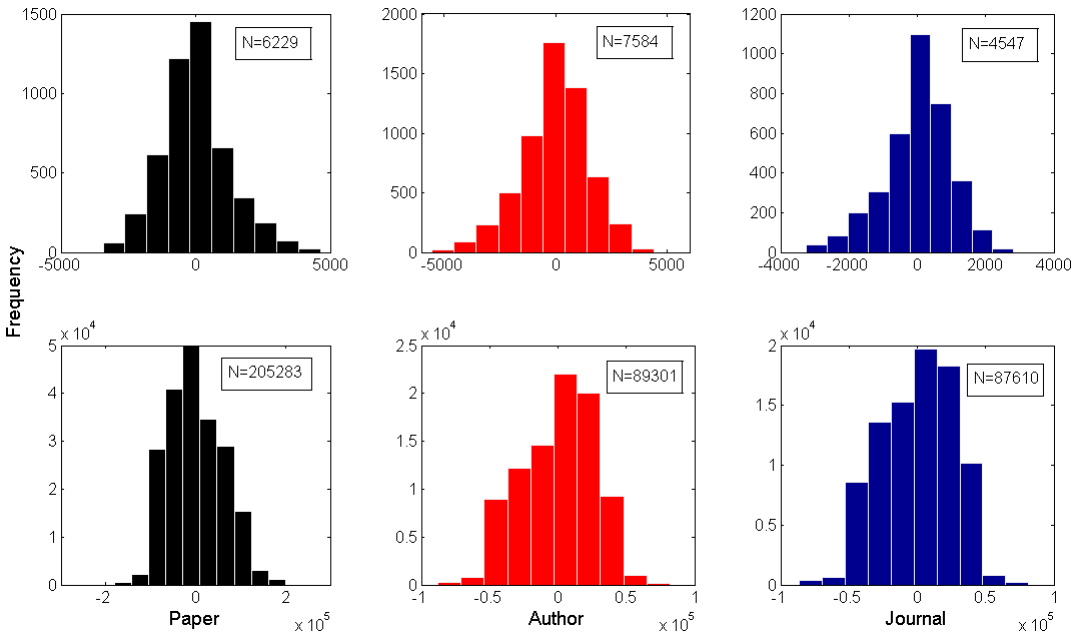

Figure 6. Rank Variances between popularity and prestige

As can be seen in Figure 6, the rank variances for papers, authors, and journals are normally distributed where the majority of them have similar popularity and prestige status, i.e., either low popularity-low prestige or high popularity-high prestige and only a small portion of papers, author, and journals have diverse status.

6 Conclusion

Citation analysis is an established tool for scientific evaluation. Yet although it is easy to comprehend and implement, this tool does not take into account the status of citing journals, authors, and articles. This study constructs a heterogeneous scholarly network and uses a new indicator called P-Rank to differentiate the weight of each citation. In this heterogeneous scholarly network, there are two inter-class walks and one intra-class walk. For the inter-class walks, authors interact with articles via the paper-author adjacency matrix, and journals interact with articles via the paper-journal adjacency matrix. For the intra-class walk, articles interact with other articles via citation links. P-Rank realizes the assumption that articles are more important if they are cited by other important articles, prestigious authors, and/or prestigious journals; authors have a higher impact if they are cited by important articles; and journals have a higher impact if they are cited by important articles.

Through PCA, we find that P-Rank is a size dependent indicator and is collocated with other size dependent indicators, such as normal PageRank and degree centrality. Citation counts of journals, authors, or articles can be considered as popularity, and P-Rank scores can be considered as an indicator of prestige since it considers the source of citation

endorsement. When conducting the correlation analysis for popularity and prestige, we find they are correlated. In order to understand how popularity and prestige are correlated, we calculate rank variances between citation counts and P-Rank scores for papers, authors, and journals. The majority of journals, authors, and articles are found to have an equivalent popularity and prestige status.

Citation time is a delicate issue in many scientific evaluation tasks. This study uses a 20-year dataset, and thus not surprisingly, many older, "classic" publications rank at the top. Therefore, the present research provides a description of the current and past LIS landscape. While it may be a useful starting point for anticipating future trends, it is unable to predict future developments with any certainty. Future research in this area should examine the time-dependency of P-Rank, by examining the ranking of papers, authors, and journals diachronically. This trend data may provide insight into predicting future directions in the field. Along these same lines, future work should seek to examine the topical element of these networks, in order to examine how knowledge diffuses in a heterogeneous network.

Acknowledgements

The authors would like to thank Ludo Waltman of Leiden University for his insightful comments on an early draft of this article.

References

Aksnes, D. W. (2003). A macro study of self-citation. Scientometrics, 56(2), 235-246.

Bergstrom, C. T., & West, J. D. (2008). Assessing citations with the Eigenfactor™ Metrics. Neurology, 71, 1850-1851.

Blau, P. M. (1964). Exchange and power in social life. New York: Wiley.

Bollen, J., Rodriguez, M. A., & Van De Sompel, H. (2006). Journal status. Scientometrics, 69(3), 669-687.

Bollen, J., Van de Sompel, H., Hagberg, A., & Chute, R. (2009) A principal component analysis of 39 scientific impact measures. PLoS ONE, 4(6), e6022. doi:10.1371/journal.pone.0006022.

Chen, P., Xie, H., Maslov, S., & Redner, S. (2007). Finding scientific gems with Google's PageRank algorithm. Journal of Informetrics, 1(1), 8-15.

Coleman, J. S. (1990). Foundations of social theory. Cambridge, MA: Harvard University Press.

Cronin, B. (1984). The citation process: The role and significance of citations in scientific communication. London: Taylor Graham.

Davis, P. M. (2008). Eigenfactor: Does the principle of repeated improvement result in better estimates than raw citation counts? Journal of the American Society for Information Science and Technology, 59(13), 2186-2188.

Ding, Y., Yan, E., Frazho, A., & Caverlee, J. (2009). PageRank for ranking authors in co-citation networks. Journal of the American Society for Information Science and Technology, 60(11), 2229-2243.

Ding, Y., & Cronin, B. (2010). Popular and/or Prestigious? Measures of Scholarly Esteem. Information Processing and Management. Retrieved May 13, 2010 from DOI:10.1016/j.ipm.2010.01.002

Fersht, A. (2009). The most influential journals: Impact Factor and Eigenfactor. Proceedings of the National Academy of Science of the United States of America, 106(17), 6883-6884.

Franceschet, M. (2010). Ten good reasons to use the Eigenfactor™ metrics. Information Processing & Management, 46(5), 555-558.

Garfield, E. (1965). Can Citation Indexing Be Automated? Retrieved July 29, 2010 from http://www.garfield.library.upenn.edu/essays/V1p084y1962-73.pdf

Glänzel, W. & Thijs, B. (2004). The influence of author self-citations on bibliometric macro indicators. Scientometrics, 59(3), 281-310.

Haveliwala, T., Kamvar, S., & Jeh, G. (2003). An analytical comparison of approaches to personalizing PageRank. Stanford University Technical Report. Retrieved August 10, 2009 from http://infolab.stanford.edu/~taherh/papers/comparison.pdf

Henrich, J., & Gil-White, F.J. (2001). The evolution of prestige: Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. Evolution and Human Behavior, 22, 165-196.

Hyland, K. (2003). Self-citation and self-reference: credibility and promotion in academic publication. 54(3), 251-259.

Krauss, J. (2007). Journal self-citation rates in ecological sciences. Scientometrics, 73(1), 79-89.

Leydesdorff, L. (2007). Betweenness centrality as an indicator of the interdisciplinarity of scientific journals. Journal of the American Society for Information Science and Technology, 58(9), 1303-1319.

Leydesdorff, L. (2009). How are new citation-based journal indicators adding to the bibliometric toolbox? Journal of the American Society for Information Science and Technology, 60(7), 1327-1336.

Liu, L. G., Xuan, Z. G., Dang, Z. Y., Guo, Q., & Wang, Z. T. (2007). Weighted network properties of Chinese nature science basic research. Physica A-Statistical Mechanics and Its Applications, 377(1), 302-314.

Liu, X., Bollen, J. Nelson, M. L., & Sompel, H. V. (2005). Co-authorship networks in the digital library research community. Information Processing and Management, 41, 1462-1480.

Lopez-Illescas, C., de Moya-Anegn, F., & Moed, H. F. (2008). Coverage and citation impact of oncological journals in the Web of Science and Scopus. Journal of Informetrics, 2(4), 304-316.

Luukkonen, T. (1997). Why has Latour's theory of citations been ignored by the bibliometric community? discussion of sociological interpretations of citation analysis. Scientometrics, 38(1), 27-37.

Ma, N., Guan, J., & Zhao, Y. (2008). Bringing PageRank to the citation analysis. Information Processing and Management, 44, 800-810.

Maslov, S. & Redner, S. (2008). Promise and Pitfalls of Extending Google's PageRank Algorithm to Citation Networks. Journal of Neuroscience, 28(44), 11103-11105.

Merton, R. K. (1968). The Matthew Effect in Science: The reward and communication systems of science are considered. Science, 159(3810), 56-63.

Nisonger, T.E., & Davis, C.H. (2005). The perception of library and information science journals by LIS education deans and ARL library directors: A replication of the Kohl-Davis study. College & Research Libraries, 66, 341-77.

Pinski, G., & Narin, F. (1976). Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. Information Processing & Management, 12(5), 297-312.

Radicchi, F., Fortunato, S., Markines, B., Vespignani, A. (2009). Diffusion of scientific credits and the ranking of scientists. Physical Review E, 80, 056103.

Sayyadi, H., & Getoor, L. (2009). FutureRank: Ranking scientific articles by predicting their future PageRank. The Ninth SIAM International Conference on Data Mining. Retrieved August 31, 2009 from http://waimea.cs.umd.edu:8080/basilic/web/Publications/2009/sayyadi:sdm09/sayyadi_futureRank_sdm09.pdf

SCImago (2007). SJR: SCImago Journal & Country Rank. Retrieved August 31, 2009 from http://www.scimagojr.com

Small, H. (1978). Cited documents as concept symbols. Social Studies of Science, 8(3), 327-340.

Sugimoto, C.R. (2010). Looking across communicative genres: A call for inclusive indicators of interdisciplinarity. Scientometrics. doi: 10.1007/s11192-010-0275-8

Tsay, M. (2006). Journal self-citation study for semiconductor literature: Synchronous and diachronous approach. Information Processing & Management, 42(6), 1567-1577.

Van Raan, A.F.J. (2008). Self-citation as an impact-reinforcing mechanism in the science system. Journal of the American Society for Information Science and Technology, 59(10), 1631-1643.

Walker, D., Xie, H., Yan, K.K., & Maslov, S. (2007). Ranking scientific publications using a simple model of network traffic. Journal of Statistical Mechanics: Theory and Experiment, P06010, doi:10.1088/1742-5468/2007/06/P06010

West, J.D., Bergstrom, T.C., & Bergstrom, C.T. (2010). The Eigenfactor Metrics: A network approach to assessing scholarly journals. College and Research Libraries, 71(3), 236-244.

Yan, E. & Ding, Y. (2009). Applying centrality measures to impact analysis: A coauthorship network analysis. Journal of the American Society for Information Science and Technology, 60(10), 2107-2118.

Yan, E. & Ding, Y. (2010a). Measuring scholarly impact in heterogeneous networks. Proceedings of the ASIS&T 2010 Annual Meeting, October, 22-27, Pittsburgh.

Yan, E., & Ding, Y. (2010b). Weighted citation: An indicator of an article's prestige. Journal of the American Society for Information Science and Technology, 61(8), 1635-1643.

Yan, E. & Ding, Y. (2010 submitted). The effect of dangling nodes on citation networks.

Yin, L., Kretschmer, H., Hanneman, R. A., & Liu, Z. (2006). Connection and stratification in research collaboration: An analysis of the COLLNET network. Information Processing and Management, 42, 1599-1613.

Zhou, D., Orshanskiy, S. A., Zha, H., & Giles, C. L. (2007). Co-Ranking authors and documents in a heterogeneous network. 2007 Seventh IEEE International Conference on Data Mining. October 28-31, Omaha, Nebraska. pp.739-744.

Zhu, H., Wang, X., Zhu, J. Y. (2003). Effect of aging on network structure. Physical Review E, 68, 056121.