The effects of dangling nodes on citation networks

Erjia Yan and Ying Ding

{eyan, dingying}@indiana.edu
School of Library and Information Science, Indiana University, Bloomington, USA

Abstract
This study discusses the effects of dangling nodes on citation networks through the PageRank algorithm. The origins of dangling nodes for citation networks are introduced and three methods are applied to handle dangling nodes: retaining all dangling nodes, deleting dangling nodes, and clustering dangling nodes into one node. Correlation analyses are used to compare these three methods.

Introduction
In the language of network analysis, dangling nodes denote the nodes without outgoing links. With the advent of the Web, the concept of dangling nodes became a common topic. It is well understood that most web pages link to and are linked by other pages. But it is possible that some pages do not contain any valid hyperlinks, which may be broken pages (i.e., those that formerly contained hyperlinks but have now become "403/404 Error") or multimedia data types (i.e., PDF, JPG, PS, MOV). The problem of dangling nodes has become more evident with the proliferation of search engines. Search engines are reported to have low coverage of the entire Web (Lawrence & Giles, 1999; Bar-Ilan, 2002; Vaughan & Thelwall, 2004). Consequently, if a page's linked pages are not crawled by search engines, it would become a dangling node.
For citation networks, each node is a publication and each link is a citation tie. Dangling nodes represent publications cited by other publications, but do not cite others. Citing behaviors affect the generation of dangling nodes in citation networks, as papers can only cite papers published earlier. Disciplinarity and databases coverage can also result in dangling nodes in citation networks.
PageRank is chosen as the underlying algorithm to measure the impacts of dangling nodes on citation networks. PageRank is not new to citation analysis. More than 30 years ago, Pinski and Narin (1976) proposed the concept of "influence weights", which served as the archetype for PageRank. For citation networks, PageRank algorithm gives higher weight to highly cited articles or articles cited by other highly cited articles.
Previous studies on the Web manipulated dangling nodes for two major reasons. First, dangling nodes exits in large scale and thus it is computational intensive to calculate PageRank for large network. Second, dangling nodes receive PageRank scores but did not distribute them and thus skewed the scores of non-dangling nodes. In the past decade, several methods have been proposed to handle the negative effect of dangling nodes on the Web. Citation networks, however, usually are comparatively small in size (from thousands to millions), and required computing time and space is therefore less demanding. Specifically, using power method, the cost of computing PageRank is $O(n)$, and thus it is possible to store the citation matrix using a linear amount of memory and the vector-matrix multiplication has linear complexity (Franceschet, 2010). Based on this, dangling nodes on citation network need to be manipulated only if they negatively affect PageRank scores of non-dangling nodes. The null and research hypotheses are proposed as follows.

- H_0: dangling nodes do not affect PageRank scores of non-dangling nodes on citation networks;
- H_1: dangling nodes have effects on PageRank scores of non-dangling nodes on citation networks.

Two methods (deleting and lumping) are used to manipulate dangling nodes, and test the hypotheses through comparing the PageRank scores of non-dangling nodes on the manipulated network and original network.
The article is organized as follows: Section 2 conducts a literature review of relevant methods used to handle dangling nodes; Section 3 introduces the data set and methods; Section 4 applies PageRank algorithm to the data set and discusses the effects of dangling nodes on citation networks; and Section 5 draws the conclusion.

Related studies

In the original PageRank paper, Page et al. (1999) suggested removing dangling nodes from the graph, and calculating the PageRank on the remaining graph. Kamvar et al. (2003) suggested removing the dangling nodes and then re-inserting them "for the last iterations". This approach was also suggested by Brin et al. (1998). Eiron, McCurley, and Tomlin (2004), however, discussed the caveats of removing dangling nodes, which skews the results on the non-dangling nodes, since the outdegrees of non-dangling nodes are adjusted when dangling nodes are deleted. They also argued that the process of removing dangling nodes may itself produce new dangling nodes. Langville and Meyer (2006) also held that simply removing the dangling nodes biases the PageRank vector.
Another approach proposed by Lee, Golub, and Zenios (2003) clustered dangling nodes into one node. The algorithm they proposed exploits the "lumpability" of the Markov chain and proceeds in two stages. At the first stage, they computed the limiting distribution of a chain where the dangling nodes are combined into one super node; at the second stage, they computed the limiting distribution of a chain where only the non-dangling nodes are combined. Ipsen and Selee (2007) took the same approach, where all dangling nodes are lumped into a single node. They also showed that the PageRank of the non-dangling nodes can be computed separately, with the convergence rate as that of the power method applied to the full matrix. Other related methods include Eiron, McCurley, and Tomlin's (2004) notion of penalty pages. They proposed four methods to "penalize" the pages linking to dangling pages by reducing their PageRank scores.
For paper citation networks, Chen, Xie, Maslov, and Render (2007) applied PageRank to assess the relative importance of publications in the Physical Review family. They found that PageRank values and citations for each publication are positively correlated. Ma, Guan, and Zhao (2008) applied PageRank in evaluating research influence of countries in the fields of Biochemistry and Molecular Biology. They found that citation and PageRank are highly correlated, with correlation coefficient reaching to 0.9 at the 0.01 level. Another advance that utilizes the concept of PageRank is the SCImago Journal and Country Rank (SCImago, 2007). These studies applied PageRank to citation networks but did not consider the effects of dangling nodes. For this study, three methods are applied to handle the dangling nodes in citation networks: (1) keeping all dangling nodes; (2) deleting dangling nodes; and (3) clustering dangling nodes into one node.

Methodology

Data set

The field of informetrics is chosen, query recommended by Bar-Ilan (2008) is utilized and improved to search all relevant records in Web of Science (retrieval time: Jan 31st, 2009; time span: default all years): TS=(informetric* OR bibliometric* OR webometric* OR scientometric* OR citation analy* OR cocitation analy* OR co-citation analy* OR link analy* OR hyperlink analy* OR self citation* OR self-citation* OR impact factor* OR science polic* OR research polic* OR S&T indicator* OR citation map* OR citation visuali*

OR information visual* OR h-index OR h index OR Hirsch index OR patent analy* OR Zipf OR Bradford OR Lotka OR collaboration network* OR coauthorship network* OR co-authorship network*) OR SO=(Scientometrics OR Journal of Informetrics). Subject category INFORMATION SCIENCE & LIBRARY SCIENCE was used to narrow down the search results. The original data set covers 4,997 papers30 (articles and review articles) with 92,021 cited references.

Dangling nodes in citation networks

When constructing a citation network, relevant bibliographical data of certain field(s) are downloaded. This procedure resembles crawling on the Web, where more dangling nodes emerge with the expansion of crawled web pages. Unless building a citation network covering the entire body of literature ever accumulated, one can only construct citation networks for certain field(s), domain(s), time(s), etc. And when constructing such networks, some of the links will inevitably be excluded and thus produce dangling nodes.

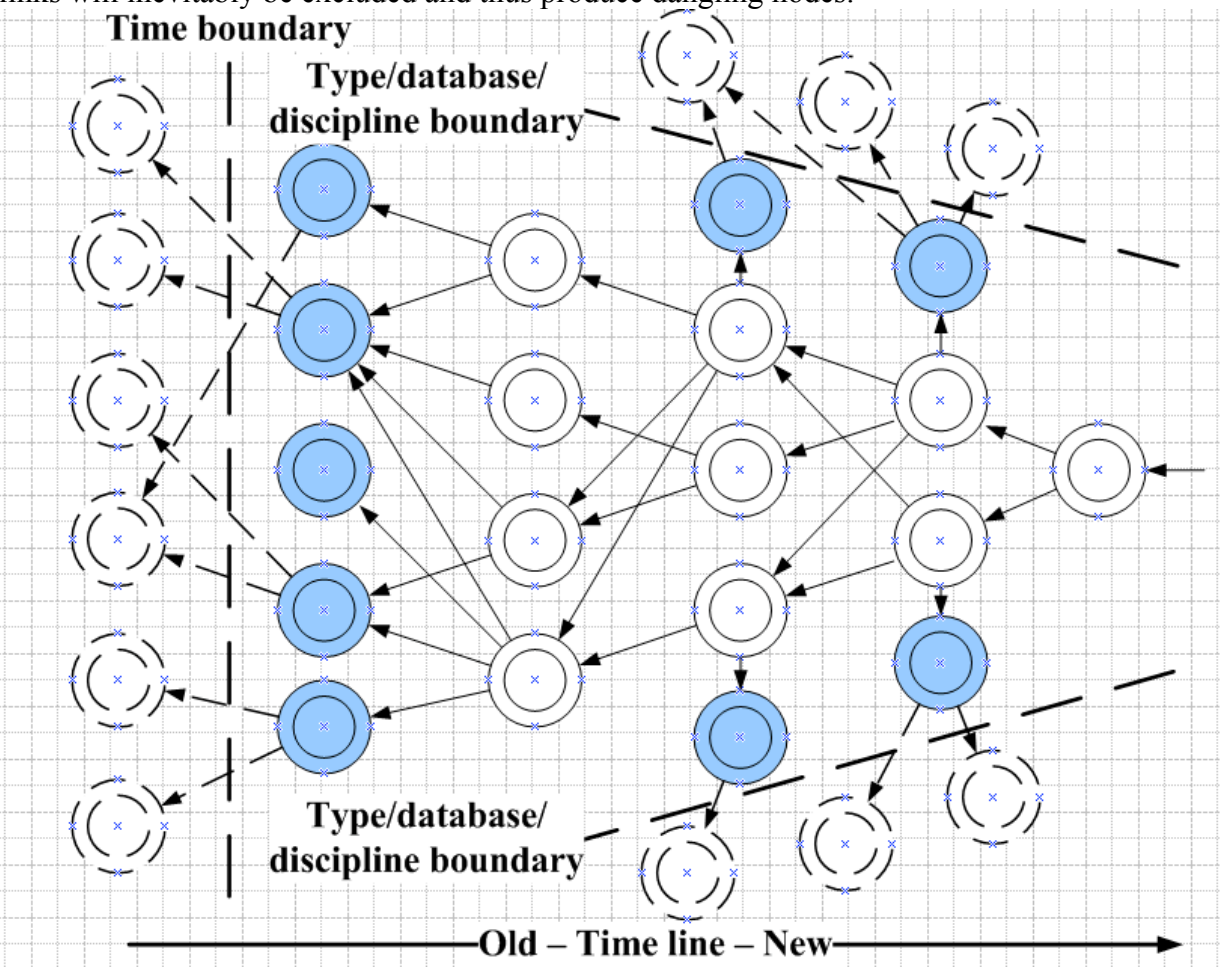

Figure 1. A small citation network with dangling nodes

Another feature of citation networks is that they strictly follow temporal order: only recent publications can cite previous ones. Figure 1 shows that publications at each time point can only cite those published before them, and thus the oldest publications in the data set would become dangling nodes. Also, types of literature, disciplinarity, and database coverage can all result in dangling nodes. Different types of literature, for example, such as citations from

[30] The record difference from "Discovering author impact: A PageRank perspective" is that 99 records that have no cited references are deleted in the current data set, which is resulted from the index issues of some records in the Web of Science database.

books, newspapers, or web pages, are usually not covered in academic databases, and therefore have no outbound links and will become dangling nodes. For this study, the focus is on informetrics research articles, where their cited articles may cover different disciplines, but these cited articles will not be covered in the original data set, and would end up as dangling nodes. Regarding different databases, some cited articles may not be collected by the database and will become dangling nodes accordingly.

Methods

This section explains the method used in this study. A five-paper graph example is referenced and presented it in a matrix (step 1); then three approaches are used to handle dangling nodes (step 2); and the transformed matrices are inputted to PageRank algorithm (step 3).

Step 1: Representing citation networks in a matrix

The papers and citations of citation networks can be presented in a directed graph. The nodes represent articles and the directed arcs represent citations. Figure 2 is an example of a five-paper citation network where paper 1 and 2 are dangling nodes.

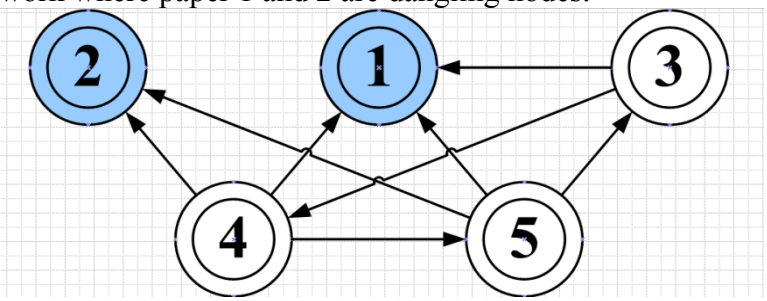

Figure 2. A five-page graph with dangling nodes

Let M be an adjacency matrix with the rows and columns corresponding to the directed graph of the network. For a weighted matrix, if there is a link from page j to page i, then the matrix entry m_{ij} has a value $1/N_j$, where N_j is the number of connections (right matrix in equation (1)). If there is no link from page j to page i, then the matrix entry m_{ij} is zero.

$$M = \begin{pmatrix} 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \Rightarrow \begin{pmatrix} 0 & 0 & 1/2 & 1/3 & 1/3 \\ 0 & 0 & 0 & 1/3 & 1/3 \\ 0 & 0 & 0 & 0 & 1/3 \\ 0 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 1/3 & 0 \end{pmatrix} \quad (1)$$

Step 2: Handling dangling nodes

A column where all entries are zero is a dangling node. The matrix M is irreducible only if there are no dangling nodes, i.e., all columns have norms of value one. One problem with the matrix M is that it is not stochastic (each column sums to one), and a Markov chain is defined only for stochastic matrices.

The first method is to retain all dangling nodes and replace each zero column (vector) with a dense column, thus transforming M into $\overline{M}_1$ (equation (2)). In equation (2), the dangling vector is replaced by a uniform vector e^T/n (e is the vector of all ones). For general application, a dangling vector is usually replaced by vector v^T, known as the personalization or teleportation vector (Langville & Meyer 2004). The corresponding network is called the whole/original network in following paragraphs.

$$\overline{M}_1 = \begin{pmatrix} 1/5 & 1/5 & 1/2 & 1/3 & 1/3 \\ 1/5 & 1/5 & 0 & 1/3 & 1/3 \\ 1/5 & 1/5 & 0 & 0 & 1/3 \\ 1/5 & 1/5 & 1/2 & 0 & 0 \\ 1/5 & 1/5 & 0 & 1/3 & 0 \end{pmatrix} \quad (2)$$

The second method is to delete all dangling nodes, as $\overline{M}_2$ in equation (3), and replace the non-stochastic column to a uniform vector e^T/n if necessary. The corresponding network is called the reduced network in following paragraphs.

$$\overline{M}_2 = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \quad (3)$$

The third method is to cluster all dangling nodes into one node, and then this node is replaced by a uniform vector e^T/n, as $\overline{M}_3$ in equation (4). The corresponding network is called the lumped network in following paragraphs.

$$\overline{M}_3 = \begin{pmatrix} 0 & 1 & 2 & 2 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \Rightarrow \begin{pmatrix} 1/4 & 1/2 & 2/3 & 2/3 \\ 1/4 & 0 & 0 & 1/3 \\ 1/4 & 1/2 & 0 & 0 \\ 1/4 & 0 & 1/3 & 0 \end{pmatrix} \quad (4)$$

Step 3: Calculating PageRank values for the transformed matrices

The last step is to input the transformed matrix $\overline{M}_1$, $\overline{M}_2$, and $\overline{M}_3$ to the PageRank algorithm: $\overline{\overline{M}} = \alpha\overline{M} + (1-\alpha)ee^T/n$, where $0 \le \alpha \le 1$ (0.85 for this study) and $E = \frac{1}{n}e^T$. $\overline{\overline{M}}$ is usually referred to as PageRank matrix. This combination of the stochastic matrix $\overline{M}$ and a stochastic perturbation matrix E ensures that $\overline{\overline{M}}$ is both stochastic and irreducible (no non-zero entries). The irreducibility adjustment also ensures that $\overline{\overline{M}}$ will converge to the stationary vector π^T (Langville & Meyer, 2004), called PageRank vector.

Results and analysis

Distribution and formation of dangling nodes

The citation network contains 95,340 nodes (4,997 original downloaded papers and their 92,021 cited references minus 1,678 overlapping records). In this network, 90,343 cited references are not covered in the original downloaded data set, and thus become dangling nodes in this citation network. This result is consistent with Langville and Meyer's (2004) finding that in some part of the Web, up to 80% of web pages are dangling nodes.

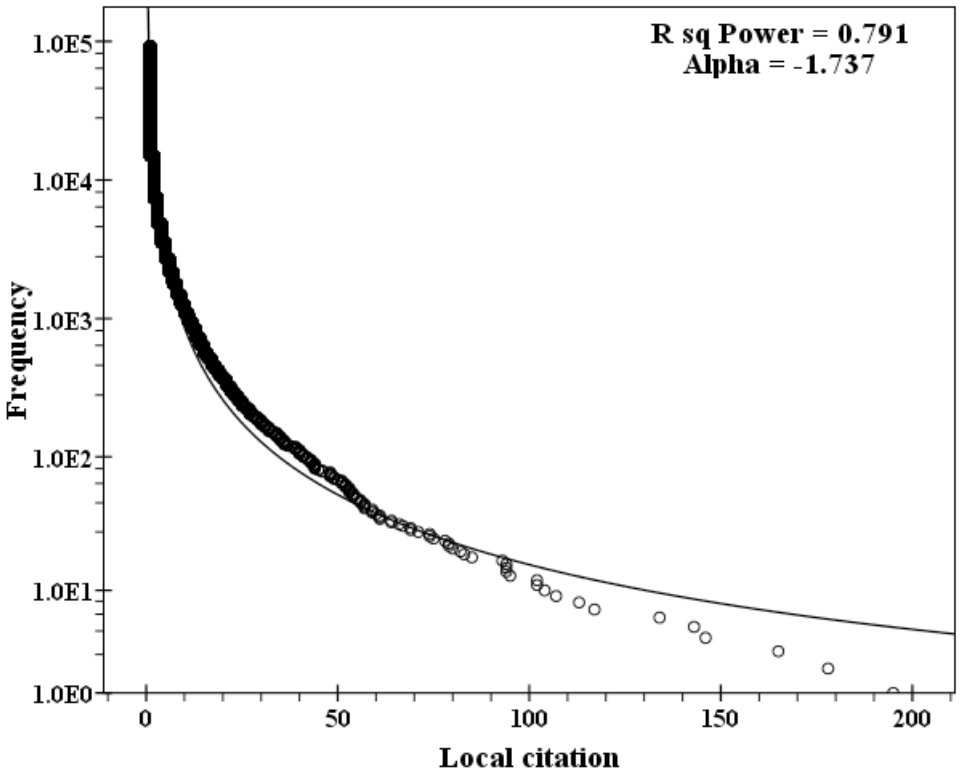

Figure 3. Distribution of number of citations for 90,343 dangling nodes

Local citation indicates the number of times an article is cited in the citation network, referred to as "internal citation" by some scholars (Ma et al., 2008). The distribution of dangling nodes shows the power-law feature: the alpha value $\alpha = -1.737$ is calculated through frequency curve estimation. A free toolkit can facilitate the calculation (Rousseau & Rousseau, 2000). Up to 90% of all references are only cited one or two times, indicating that these references are not echoed in other informetrics studies. On the other hand, 129 references are cited more than 30 times. These publications have high connections with informetrics research.

Table 1. Top 20 publications based on PageRank

PageRnk Rank	*First author*	*Title**	*Journal/Publisher***	*Year*	*Local Citation*	*Dangling Nodes*
1	Schubert A	Relative indicators and relational charts for comparative assessment of publication output and citation impact	Scientometrics	1986	74	FALSE
2	Braun T	Scientometric indicators	World Scientific	1985	55	TRUE
3	Lotka AJ	The frequency distribution of scientific productivity	Journal of the Washington Academy of Sciences	1926	195	TRUE
4	Garfield E	Citation Indexing	Wiley & Sons	1979	178	TRUE
5	Garfield E	Citation analysis as a tool in journal evaluation	Science	1972	146	TRUE
6	Schubert A	Scientometric data files	Scientometrics	1989	80	FALSE
7	Small H	Cocitation in scientific literature	JASIS	1973	165	FALSE
8	Price DJD	Networks of scientific papers	Science	1965	143	TRUE
9	Price DJD	Little science, big science	Columbia	1963	117	TRUE

			University Press			
10	Bradford SC	Sources of Information on Specific Subjects	Engineering (London)	1934	134	TRUE
11	Narin F	Evaluative bibliometrics	Computer Horizons	1976	94	TRUE
12	Hirsch JE	An index to quantify an individual's scientific research output	PNAS	2005	94	TRUE
13	Price DJD	General theory of bibliometric and other cumulative advantage processes	JASIS	1976	113	FALSE
14	Moed HF	The use of bibliometric data for the measurement of university-research performance	Research Policy	1985	69	TRUE
15	Small H	Structure of scientific literatures	Science Studies	1974	102	TRUE
16	Martin BR	Assessing basic research	Research Policy	1983	82	TRUE
17	Brookes BC	Bradford's law and bibliography of science	Nature	1969	71	TRUE
18	Egghe L	Introduction to informetrics	Elsevier	1990	79	TRUE
19	Bradford SC	Documentation	Crosby Lockwood	1948	61	TRUE
20	Beaver DD	Studies in scientific collaboration	Scientometrics	1978	57	FALSE

*Words after ":" and "-" are omitted;
**PNAS: Proceedings of the National Academy of Sciences; JASIS: Journal of the American Society for Information Science.

Table 1 shows top 20 publications of the whole network (95,340 nodes) based on PageRank scores. Of the top 20 publications, 15 are dangling nodes. Seven dangling nodes are books that resemble PDF files on the Web, in that they usually have higher values but cannot cite or link to other resources. Some old journal articles are found to be dangling nodes. Examples as Lotka's article titled, "The frequency distribution of scientific productivity" published in 1926. Although it is a journal article which has references, they are not collected in the database due to its age. Some other dangling nodes (e.g., Citation analysis as a tool in journal evaluation) are resulted from the selection of data, since the data only cover records in the INFORMATION SCIENCE & LIBRARY SCIENCE subject category, and hence journals categorized as MULTIDISCIPLINE, MANAGEMENT and so on are unable to be included.

Citation in three networks vs. PageRank

In this section, three methods are applied to identify the effects of dangling nodes on citation networks by computing PageRank values under d=0.85 for: (1) the whole network; (2) the reduced network; and (3) the lumped network.

For the last two methods, the issue of defining dangling nodes is important, since the deletion of dangling nodes would result in new dangling nodes. For the whole network, if all 90,343 dangling nodes were deleted, 4,997 nodes would be left. For this remaining network, 2,314 new dangling nodes would emerge. Theoretically, this procedure can progress in this manner until no node is left; however, in practice, the real number is a little above zero (37 for this network), since some articles are cited in the preprint version by older publications.

Of the previous studies on deletion of dangling nodes, Page et al. (1999), Kamvar et al. (2003), and Brin et al. (1998) did not mention specifically how these nodes were removed. Page et al. (1999) and Brin et al. (1998) generalized this procedure as "remove links that point to any page with no outgoing links", and Kamvar et al. (2003) as "exclude dangling nodes from the Web graph until the final few iterations". But their experimental setups and results

imply that they only removed the dangling nodes in the original network, not the newly generated dangling nodes. Based on this, a similar approach is taken by only deleting the original dangling nodes.

In the whole network, PageRank values for all 95,340 are calculated, and those for the 4,997 non-dangling nodes are selected. For reduced network, PageRank scores for 4,997 nodes are computed after the deletion of 90,343 dangling nodes. In lumped network, 90,343 dangling nodes are first clustered into one super node, and then PageRank values are computed for this new network containing 4,998 nodes (4,997 non-dangling nodes plus the super node). The distribution between PageRank scores and local citation counts is illustrated in Figure 4.

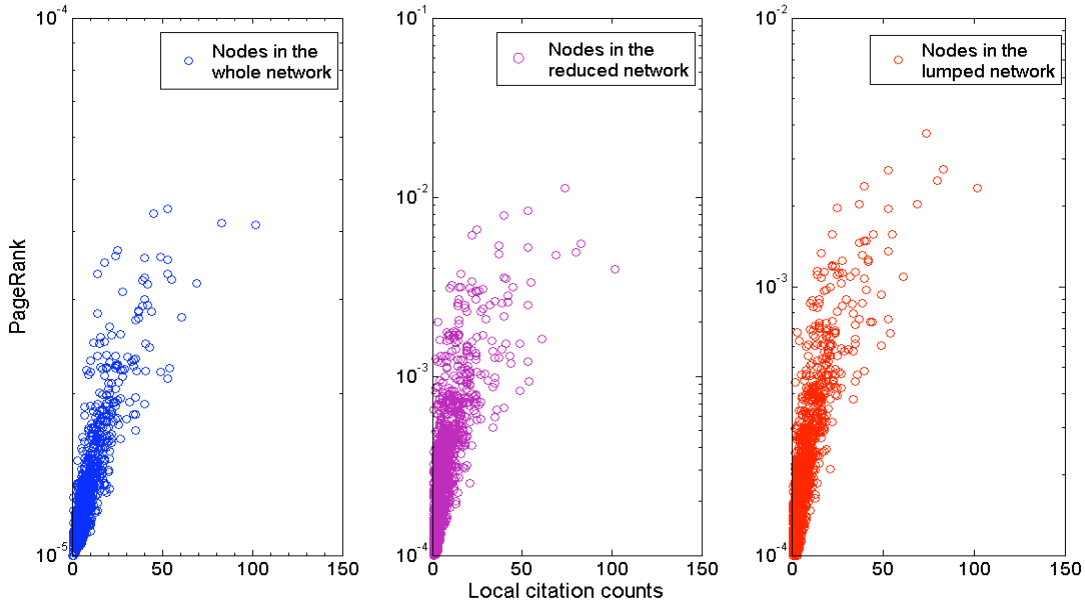

Figure 4. PageRank vs. Local citation counts for non-dangling nodes

As can be seen in Figure 4, local citation and PageRank are highly correlated, with Spearman's rank correlation coefficient above 0.9 (r_s= 0.9911, 0.9895, and 0.9931 respectively). Bollen et al. (2006) interpreted the number of citations as "popularity", measuring the bounds of impacts and PageRank as "prestige", which considers the quality of impacts. They also found a high correlation between citation and PageRank of ISI journals.

Table 2. Top 20 articles based on PageRank (non-dangling nodes)

Whole network	*Reduced network*	*Lumped network*
Schubert A, 1986, SCIENTOMETRICS, V9, P281	Schubert A, 1986, SCIENTOMETRICS, V9, P281	Schubert A, 1986, SCIENTOMETRICS, V9, P281
Schubert A, 1989, SCIENTOMETRICS, V16, P3	Brookes BC, 1968, J DOC, V24, P247	Almind TC, 1997, J DOC, V53, P404
Small H, 1973, J AM SOC INFORM SCI, V24, P265	Garfield E, 1979, SCIENTOMETRICS, V1, P359	Brookes BC, 1968, J DOC, V24, P247
Price DJD, 1976, J AM SOC INFORM SCI, V27, P292	Schubert A, 1983, SCIENTOMETRICS, V5, P59	Schubert A, 1989, SCIENTOMETRICS, V16, P3
Beaver DD, 1978, SCIENTOMETRICS, V1, P65	Small H, 1980, SCIENTOMETRICS, V2, P277	Garfield E, 1979, SCIENTOMETRICS, V1, P359
Moed HF, 1995, SCIENTOMETRICS, V33, P381	Almind TC, 1997, J DOC, V53, P404	Ingwersen P, 1998, J DOC, V54, P236
Schubert A, 1990, SCIENTOMETRICS, V19, P3	Yablonsky AI, 1980, SCIENTOMETRICS, V2, P3	Smith LC, 1981, LIBR TRENDS, V30, P83
Almind TC, 1997, J DOC, V53, P404	Small H, 1985, SCIENTOMETRICS, V7, P391	Yablonsky AI, 1980, SCIENTOMETRICS, V2, P3
Ingwersen P, 1998, J DOC, V54, P236	Schubert A, 1989, SCIENTOMETRICS, V16, P3	Schubert A, 1983, SCIENTOMETRICS, V5, P59

Schubert A, 1983, SCIENTOMETRICS, V5, P59	Beaver DD, 1979, SCIENTOMETRICS, V1, P231	Small H, 1985, SCIENTOMETRICS, V7, P391
Braun T, 1988, SCIENTOMETRICS, V14, P3	Smith LC, 1981, LIBR TRENDS, V30, P83	Small H, 1980, SCIENTOMETRICS, V2, P277
Vanraan AFJ, 1996, SCIENTOMETRICS, V36, P397	Ingwersen P, 1998, J DOC, V54, P236	Schubert A, 1990, SCIENTOMETRICS, V19, P3
Garfield E, 1979, SCIENTOMETRICS, V1, P359	Rabkin YM, 1979, SCIENTOMETRICS, V1, P261	Small H, 1985, SCIENTOMETRICS, V8, P321
Brookes BC, 1968, J DOC, V24, P247	Haitun SD, 1982, SCIENTOMETRICS, V4, P5	Haitun SD, 1982, SCIENTOMETRICS, V4, P5
Braun T, 1987, SCIENTOMETRICS, V11, P9	Brookes BC, 1977, J DOC, V33, P180	Brookes BC, 1977, J DOC, V33, P180
Braun T, 1987, SCIENTOMETRICS, V12, P3	Small H, 1980, J DOC, V36, P183	Beaver DD, 1979, SCIENTOMETRICS, V1, P231
Small H, 1985, SCIENTOMETRICS, V7, P391	Small H, 1985, SCIENTOMETRICS, V8, P321	Moed HF, 1995, SCIENTOMETRICS, V33, P381
Braun T, 1987, SCIENTOMETRICS, V11, P127	Bradley SJ, 1992, J INFORM SCI, V18, P225	Christensen FH, 1996, SCIENTOMETRICS, V37, P39
Egghe L, 1985, J DOC, V41, P173	Schubert A, 1990, SCIENTOMETRICS, V19, P3	Beaver DD, 1979, SCIENTOMETRICS, V1, P133
Small H, 1985, SCIENTOMETRICS, V8, P321	Christensen FH, 1996, SCIENTOMETRICS, V37, P39	Haitun SD, 1982, SCIENTOMETRICS, V4, P89

Table 2 lists top 20 articles based on PageRank for three networks. When compared to Table 1, much important literature is not included here, as the exclusion of dangling nodes has resulted in significant loss of information, with more than 90% of the records being excluded. Ten articles rank top 20 for all three networks, eight rank top 20 for two networks, and 14 rank top 20 for one network. As for authors, Small H has five articles ranked top 20 for one of the three networks, and other authors who have more than one articles ranked top 20 are: Schubert A (4 articles), Braun T (4 articles), Beaver DD (3 articles), Brookes BC (2 articles), and Haitun SD (2 articles). As for the year of publication, 1 article is published in 1960s, 8 in 1970s, 16 in 1980s, 7 in 1990s, and no article in 2000s. As for journals, Scientometrics is dominant, reaching 22 out of 32 unique articles in Table 2, followed by Journal of Documentation (6 articles) and Journal of American Society for Information Science (2 articles).

Comparing PageRank in three networks

Figure 5 shows the scatter plot for 4,997 nodes in three networks. The Spearman's rank correlation coefficient is 0.9872 for whole network vs. reduced network, and 0.9900 for whole network vs. lumped network, indicating that most non-dangling nodes have approximately same rank status for reduced network and lumped network, and dangling nodes do not have major impact on the non-dangling nodes.

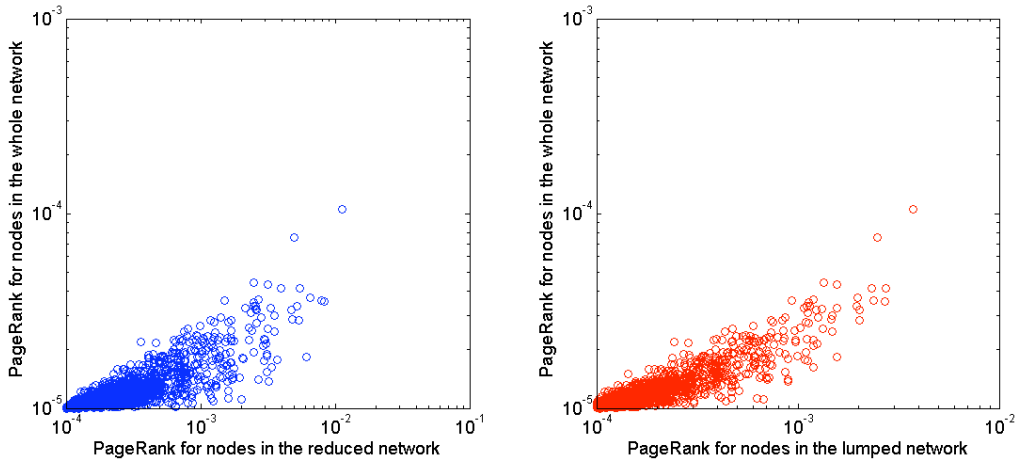

Figure 5. Correlation for three networks

Table 3. Number of dangling nodes for each level

Level	*Number of dangling nodes*	*Accumulated number of dangling nodes*	*Percentile*	*Accumulated percentile*
1--10	7	7	70.00%	70.00%
11--50	28	35	70.00%	70.00%
51--100	33	68	66.00%	68.00%
101-500	275	343	68.75%	68.60%
501--1000	390	733	78.00%	73.30%
1001-5000	3495	4228	87.38%	84.56%
5001--10000	4761	8989	95.22%	89.89%
10001--50000	39526	48515	98.82%	97.03%
50001--95340	41828	90343	92.25%	94.76%

Table 3 lists the number of dangling nodes for each PageRank ranking level. For each level, dangling nodes take a high percentile, ranging from 66% to 99%. Notably, for top 100 publications, 68% of them are dangling nodes. The results differ from the dangling nodes on the Web, where most dangling nodes are in the periphery of the Web and have low ranks (Langville & Meyer, 2004). The deleting or lumping dangling nodes on the Web, therefore, is appropriate for that environment. On the contrary, dangling nodes on citation networks are pervasive at each level. Deleting or lumping dangling nodes on citation networks will thus result in significant loss of data, especially for the top ranked publications (see Table 2). In addition, as can be seen in Figure 6, the rank variance between original and reduced network and rank variance between original and lumped network for most articles is zero, which means that the non-dangling articles in the network have almost identical rank status. The deleting and lumping dangling nodes in fact only have a minor impact on non-dangling nodes in the citation network, and thus do not change their overall ranking. Based on this, we argue that the methods of deleting and lumping dangling nodes are not appropriate for citation networks, as PageRank can yield comparatively similar results using the original network.

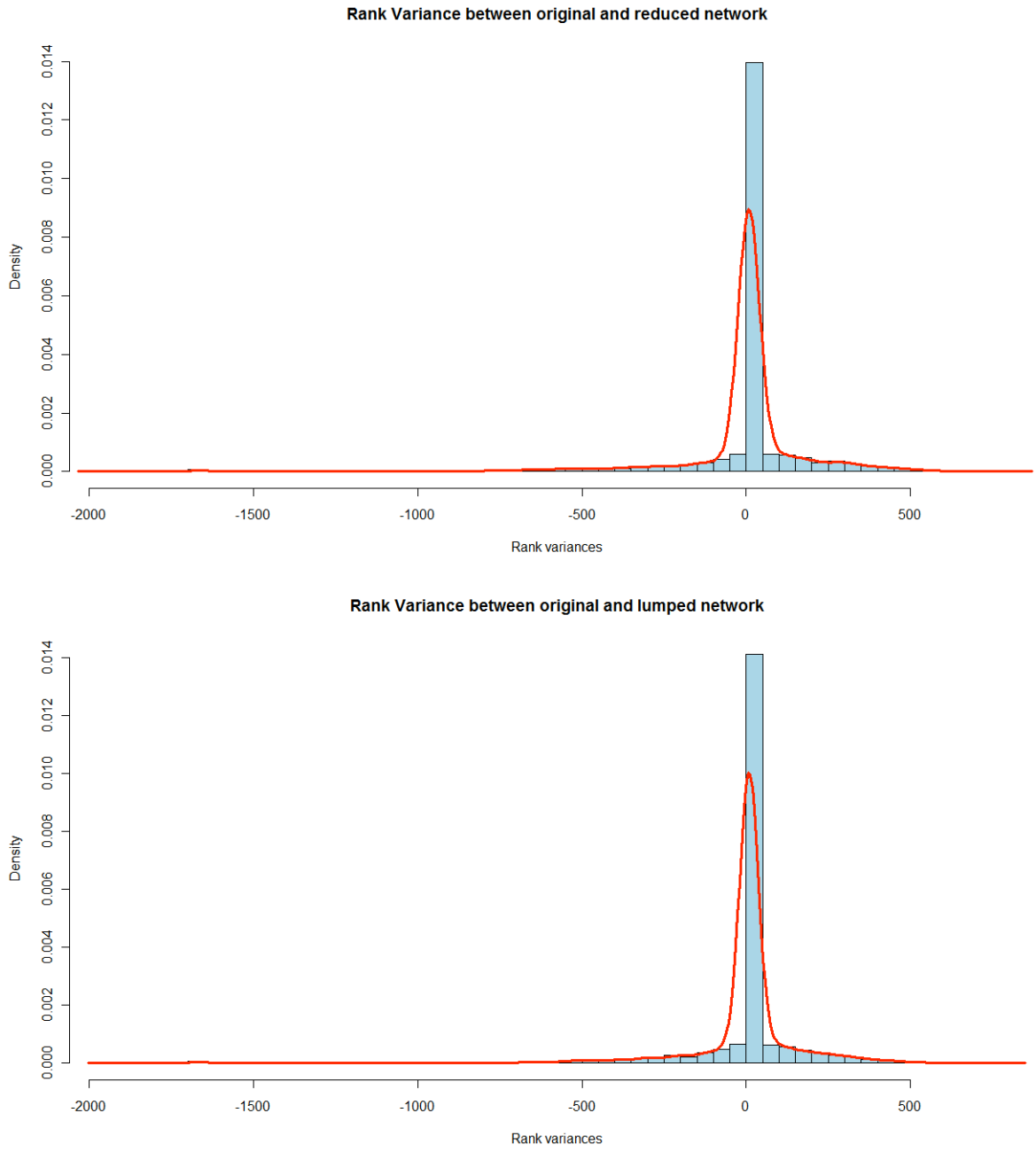

Figure 6. Rank variances

Conclusion

This article evaluates the effects of dangling nodes in citation networks through the PageRank algorithm. Hyperlinks on the Web can be updated frequently, whereas the publication aging problem is crucial for the formation of dangling nodes in citation networks: articles can only cite those published before them. Consequently, the oldest publications in a network unavoidably become dangling nodes. Moreover, since citations are linked, the selection of certain citations but not others will produce dangling nodes. Another major type of dangling node is caused by the coverage of databases.

Three methods are applied to handle dangling nodes in the citation network: retaining dangling nodes, deleting dangling nodes, and clustering dangling nodes into one node. Citation counts and PageRank values are correlated, with Spearman's rank correlation coefficients above 0.9. Through comparing the three methods, deleting and lumping methods

do not radically change the PageRank scores of non-dangling nodes, suggesting that dangling nodes on citation networks only have local impact on non-dangling nodes. Moreover, this study also compares the ranking variances between the original and reduced networks and rank variance between the original and lumped networks, and finds similar results: most non-dangling articles have identical rank for the original network and manipulated networks. Different from dangling nodes in the Web, highly cited dangling nodes in citation networks are important references, and therefore deleting or clustering them would result in loss of information and consequently prevent us from gaining an overview of the field. Based on these findings, the null hypothesis is thus retained: the non-manipulated network is preferable for handling dangling nodes, since PageRank can produce similar result without losing any information.

As the different rankings in Table 2 demonstrate, the rankings based on PageRank are sensitive to the data set selection, the network construction procedure, and the analyzing methods. Hence, studies that utilize the PageRank concept need to verify the sensitivity of the algorithm.

The limitation of this study is that only one data set is used, and thus the conclusion cannot be generalized to other scholarly data set. As the strict temporal order of citations affects citation networks, the older publications will have an advantage in accumulating citations. In future studies, it may be necessary to add temporal dimensions to citations and evaluate them based on different publication year.

Reference

Barabási, A. L. (2003). Linked: how everything is connected to everything else and what it means for business, science, and everyday life. New York: Plume.

Bar-Ilan, J. (2002). Methods for assessing search engine performance over time. Journal of the American Society for Information Science and Technology, 53(4), 308-319.

Bar-Ilan, J. (2008). Informetrics at the beginning of the 21st century: A review. Journal of Informetrics, 2, 1-52.

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

Brin, S., Motwani, R., Page, L., & Winograd, T. (1998). What can you do with a web in your pocket? Data Engineering Bulletin, 21, 37-47.

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

Eiron, N., McCurley, K. S., & Tomlin, J. A. (2004). Ranking the web frontier. In Proceedings of the Thirteenth International Conference on World Wide Web. New York: ACM Press.

Franceschet, M. (2010). PageRank: Stand on the shoulders of giants. Retrieved May 15, 2010 from http://arxiv.org/PS_cache/arxiv/pdf/1002/1002.2858v1.pdf

Ipsen, I. C. F., & Selee, T. M. (2007). PageRank computation, with special attention to dangling nodes. SIAM Journal of Matrix Analysis and Application, 29(4), 1281-1296.

Kamvar, S. D., Haveliwala, T. H., Manning, C.D., & Golub, G. H. (2003). Exploiting the block structure of the web for computing PageRank. Retrieved March 6, 2009 from http://citeseer.ist.psu.edu/kamvar03exploiting.html

Langville, A. N., & Meyer, C. D. (2004). Deeper inside PageRank. Internet Mathematics, 1(3), 335-380.

Lawrence, S., & Giles, C.L. (1999). Accessibility of information on the web. Nature, 400, 107-109.

Lee, C. P., Golub, G. H., & Zenios, S. A. (2003). A fast two-stage algorithm for computing PageRank and its extensions. Retrieved March 6, 2009 from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.133.5557

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

Page, L., Brin, S., Motwani, R., & Winograd T. (1999). The PageRank citation ranking: Bringing order to the web. Retrieved February 6, 2009 from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.38.5427

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

Rousseau, B., & Rousseau, R. (2000). LOTKA: A program to fit a power law distribution to observed frequency data. Cybermetrics, 4, http://www.cindoc.csic.es/cybermetrics/articles/v4i1p4.html

SCImago. (2007). SJR - SCImago Journal & Country Rank. Retrieved March 07, 2011, from http://www.scimagojr.com

Vaughan, L., & Thelwall, M. (2004). Search engine coverage bias: Evidence and possible causes. Information Processing and Management, 40, 693-707.