Scientometrics (2016) 107:1005–1020
DOI 10.1007/s11192-016-1891-8

Identifying key papers within a journal via network centrality measures

Saikou Y. Diallo[1] · **Christopher J. Lynch**[1] · **Ross Gore**[1] · **Jose J. Padilla**[1]

Received: 15 June 2015 / Published online: 15 February 2016

Abstract This article examines the extent to which existing network centrality measures can be used (1) as filters to identify a set of papers to start reading within a journal and (2) as article-level metrics to identify the relative importance of a paper within a journal. We represent a dataset of published papers in the Public Library of Science (PLOS) via a co-citation network and compute three established centrality metrics for each paper in the network: closeness, betweenness, and eigenvector. Our results show that the network of papers in a journal is scale-free and that eigenvector centrality (1) is an effective filter and article-level metric and (2) correlates well with citation counts within a given journal. However, closeness centrality is a poor filter because articles fit within a small range of citations. We also show that betweenness centrality is a poor filter for journals with a narrow focus and a good filter for multidisciplinary journals where communities of papers can be identified.

Keywords Paper filtering · Article-level metrics · Network centrality analysis

Mathematics Subject Classification 65F15

Introduction

Imagine researchers investigating a topic. With current scientometric knowledge, researchers can identify key journals, key authors, and several related journal level metrics that would allow them to assess the relative relevance of a journal on the topic of interest. However, even if the researchers can narrow their search to a reduced set of journals, they are left with the problem of filtering the thousands of articles in each journal to a

✉ Christopher J. Lynch
cjlynch@odu.edu

[1] Virginia Modeling Analysis and Simulation Center, Old Dominion University, 1030 University Boulevard, Suffolk, VA 23435, USA

meaningful initial set. The current practice for determining the importance of a paper is to track the number of times that the paper is cited (Abt 2000; Sikorav 1991; Zhu et al. 2004). However, citation counts are influenced by numerous variables and require years to accumulate even in the most highly regarded journals (Garfield 2006; Neylon and Wu 2009; Stringer et al. 2008). In addition, the reliance on citation counts only, could cause the researcher to overlook newly published papers as they have not yet had a chance to accumulate citations. Another approach is to rely on journal publisher and read the recommended articles. However these recommendations usually center on new articles at the exclusion of older ones. Ideally, researchers would have the ability to immediately identify important papers in a given journal such that they are aware of the existing body of knowledge that has accumulated in the journal over the years while being able to leverage new contributions as quickly as they are published.

In this article we propose the adoption of established network centrality measures that compute the importance of journals to determine the extent to which an individual paper is important or *key* within a given journal. Historically, the number of citations that a paper accrues serves as the currency through which we measure the importance of a paper and a myriad of journal importance measures have been proposed. For the most part, three measures related to network centrality serve as effective guidelines for: (1) authors making choices about where to disseminate their research; and (2) libraries making purchasing decisions (Gross and Gross 1927). These three network centrality measures are: betweenness centrality, closeness centrality, and eigenvector centrality. Each of these centrality metrics utilize the network of citations formed between publications to determine (1) which journals are important (Barnett et al. 2011; Griffin et al. 2015), (2) which authors are important across journals (Liu et al. 2005; Yan and Ding 2009), or (3) to identify important sentences or keywords within documents (Beliga et al. 2015; Erkan and Radev 2004; Khan and Wood 2015). *However, to our knowledge, the application of these metrics to filter out and determine the importance of papers within a single journal has not been explored*. As a result, we apply these centrality metrics to the papers within PLOS in order to determine which of these metrics can be used to effectively filter out important papers in a journal. Specifically, we ask the following questions:

1. Are important papers the papers that allow researchers to transition from one group of papers to another within a journal?
2. Are important papers the papers that a researcher is more likely to find when starting from any randomly selected paper?
3. Are important papers the papers that a researcher can reach from other important papers?

It is important to note that the types of papers enumerated in our research questions are not mutually exclusive and a paper can be important in more than one aspect. We restrict our analysis to creating networks for each paper within PLOS based on papers contained within PLOS. In order to verify our results, we investigate the correlation between a paper's citation count and its centrality value. The assumption here is that citation count is a good metric that indicates the historical value of a paper and therefore should correlate well with our recommended article-level metric.

The balance of the paper is organized as follows: In the "State of the art in network centrality metrics" section we conduct a literature review on network centrality metrics; In the "What makes a paper key" we describe the betweenness, closeness, and eigenvector centrality metrics in more detail and provide examples on how they identify the importance of a paper; In the "Methodology" section, we present our methodology; In the "Finding

key papers in PLOS" section, we apply our methodology to identify (1) key papers in PLOS and (2) the most appropriate centrality metric for describing the *keyness* of a paper within its journal; In the "Discussion" section, we present our findings and in the "Conclusions" section, we present conclusions and future research.

State of the art in network centrality metrics

Network centrality metrics identify the importance of publications within the scientific community by examining networks of papers based on their co-citation networks (links formed between papers based on their citations) or their co-author networks (links formed between papers based on their authors) for social network analysis based on the flow of information between publications (Leydesdorff 2007, 2011). Historically, point centrality seeks the identification of the person within a network that is structurally more central to the network than any other person in the network; however, a main difficult arises in determining if the central position is structurally unique (Freeman 1979). Therefore, the central node (be it a person or any other object within the network) represents the key position in the network. Various metrics have been introduced for determining whether a node is key based on (1) the node that is closest to all other papers (Freeman 1979), (2) the node positioned on the shortest path between the greatest number of other nodes (Freeman 1977; Leydesdorff 2007), and (3) the node that is key due to its connection to highly valued nodes (Bonacich 1987). These metrics are known as closeness centrality, betweenness centrality, and eigenvector centrality, respectively. They take into account the manner in which the citations: (1) accumulate over time for an individual journal; and (2) form a relationship among different journals.

Betweenness centrality focuses on identifying nodes which are frequently found on the shortest path between two other nodes (Freeman 1977); thus, betweenness centrality produces a relational value based on the local positions of the node with respect to the position of the nodes that it sits between (Leydesdorff 2007). Nodes that sit on the path between two other nodes serve to control the flow of information between the other nodes ranging from complete control (when only one path exists between the two other nodes) to limited control (when multiple paths exist between nodes) (Freeman 1979). Deleting nodes with high Betweenness centrality values from a network serves to break the network into coherent clusters (Leydesdorff 2007). These nodes also control the flow of knowledge between other nodes (Li-chun et al. 2006). Leydesdorff (2011) shows the application of betweenness centrality, among other metrics, to analyze the fields of nanotechnology and communications within a journal–journal citation network. Betweenness centrality computes the frequency with which a given journal is cited to create the shortest path (i.e. a bridge) between two other journals (Guns et al. 2011; Hanneman and Riddle 2005; Leydesdorff 2007).

Closeness centrality measures the "independence" of an individual paper with respect to the other papers within the target body of knowledge (Freeman 1979, p. 224). A paper that directly connects with many other papers has a high closeness centrality value while a paper with many indirect connections has a lower value. Articles that reside on the outer edges of a citation network are likely to have low closeness values since they require the presence of other intermediary articles in order to reach other articles (Siler 2013). As such, papers with high closeness centrality values extend their influence throughout the entire network (Li-chun et al. 2006) and can be seen as a measure of how long it will take for information to spread from a given node to the remainder of the network (Liu et al. 2015). Closeness centrality computes the shortest path from one journal to a different journal based on the citations of papers within each journal (Leydesdorff 2007).

Eigenvector centrality provides a centrality value of a node (such as a journal article) based on the summation of its links to other nodes (such as the other journal articles that it cites) weighted by their centralities and the nodes' links are not necessarily weighted equally; therefore, nodes linked to higher value nodes receive a larger benefit than from lower value nodes (Bonacich 1987). However, in the special case where all of the nodes are weighted equally, the eigenvector centrality values may be identical to the betweenness or closeness centrality values (Bonacich 2007). Within a social network, such as a community of connected papers, the eigenvector centrality metric establishes the *keyness* of a paper based on its connections to other papers with high eigenvector centrality values.

Eigenvector centrality computes the importance of a journal by determining how frequently it is cited in other important journals (Bergstrom 2007; Bonacich 2007; West, Bergstrom, and Bergstrom 2010). A benefit of this centrality metric is that it can be used with valued or signed networks which use non-binary relationships between nodes (Bonacich 2007). Eigenvector centrality has been applied to social networks for purposes such as predicting academic positions for faculty members whose positions require publishing (Feeley et al. 2010) as well as to examine the benefits gained within citation networks formed within different institutions on the basis that having knowledgeable coauthors provides a greater benefit to a paper (Liu et al. 2015). Additionally, eigenvector centrality has been applied to numerous other fields: Allesina and Pascual (2009) apply eigenvector centrality to forecast the effects of species' extinctions; Joyce et al. (2010) examine the ability of eigenvector centrality to identify critical nodes within brain networks; Kane (2009) applies eigenvector centrality to examine the influence of editors on the quality of articles within Wikipedia; and Lohmann et al. (2010) demonstrate that this centrality metric is efficient for representing the brain's neural architecture.

The betweenness, closeness, and eigenvector centrality metrics each enumerate a distinct role for interpreting the importance of a node within a given network. Journals with high eigenvector centrality are frequently cited by other important journals and are considered important. Journals that frequently create shortest path bridges between different journals have high betweenness centrality and are considered important. Journals with low closeness centrality scores spread information to other journals quickly and are considered important. These metrics have been applied at the journal–journal level to find key journals; at the journal–journal level to find key authors across journals; at the paper level to find key words and sentences; and for analyzing various topic-specific bodies of knowledge (such as brain networks) to find key papers or authors within the given domain. However, there is a gap in the literature for applying these centrality metrics to determine key papers within an individual journal. This is the gap that we seek to fill by applying these metrics to find key papers within PLOS by examining the co-citation network formed within PLOS for PLOS papers cited by other PLOS papers. The following section provides the algorithms for each of these metrics and provides an example network configuration to better illustrate the determination of key papers for each metric.

What makes a paper key

The word "key" is ambiguous when referring to an article as different criteria are applicable, such as most downloaded, most cited, etc. Neylon and Wu (2009) discuss this issue and the need for measures beyond citation count, download numbers, and comments. Our work addresses this need through the adaption of network centrality measures to identify

different types of key papers within a journal. Here, we define our network representation of the papers within a journal and each of the three centrality measures to quantify the *keyness* of an individual paper within a journal: (1) betweenness centrality, (2) closeness centrality, and (3) eigenvector centrality.

Our network representation of the papers within a journal is constructed as follows. A given paper x in a journal Y is initially represented as unconnected. Then, for all papers y_i within Y that explicitly cite x, an edge is drawn from y_i to x. Next, for all papers z_i that explicitly cite y_i an edge is drawn from z_i to x. This process continues until no more papers in the journal can be linked to paper x and is repeated for each paper. This results in a Journal Citation Network that allows for the application of metrics to determine the status of the journal as a whole (Bollen et al. 2006) or the status, or *keyness*, of individual articles within the journal. The construction of our network assumes that research interest in a paper is conveyed via citation and that information flows from one paper to another along the shortest citation path.

As such, our network representation of a journal is a set of papers, (V) connected via edges (E). Given a network (V, E), Eq. 1 defines the betweenness centrality of a paper:

$$g(v) = \sum_{s \neq t \neq v} \frac{\sigma_{s,t}(v)}{\sigma_{st}} \tag{1}$$

In Eq. 1, σ_{st} is the number of shortest paths from paper s to paper t and $\sigma_{s,t}(v)$ is the number of those paths that pass through paper v. A paper with high betweenness centrality frequently transfers information between two different papers by serving as a bridge between them. Such a paper is key because it acts as a "Glue" paper by connecting many different papers in the publication.

Equation 2 formally defines the closeness centrality of a paper. Once again our network is a set of papers (V) connected via edges (E).

$$g(v) = \sum_{u \in N, u \neq v} \frac{\gamma(u, v)}{N} \tag{2}$$

In Eq. 2, $\gamma(u, v)$ is the minimum number of papers and citations required to traverse the network when moving from (u) to (v) and N is the number of papers in the network. Within the network, papers with low closeness centrality transfer knowledge, on average, to other papers along a very short path of connected edges (Kevin Bacon effect). As a result they are considered key "Kevin Bacon" papers because the knowledge encapsulated in their paper is disseminated efficiently in the network.

Formally, Eq. 3 defines the eigenvector centrality of a paper.

$$Ax = \lambda x, \quad \lambda x_i = \sum_{j=1}^{n} a_{ij} x_j, \quad i = 1, \ldots, n \tag{3}$$

In Eq. 3, x is the eigenvector centrality, A is the adjacency matrix of the network, λ is the largest eigenvalue of A, and n is the number of vertices. Given our network representation, papers with high eigenvector centrality are the most connected to other highly connected papers. Given the number of explicit and implicit citations these papers receive from their other well-cited paper peers, they are considered key (Big Fish) papers in Fig. 1. Eigenvector centrality also measures the *keyness* of a paper based on the overlap of the papers that it cites within its own journal and papers that it cites within other journals

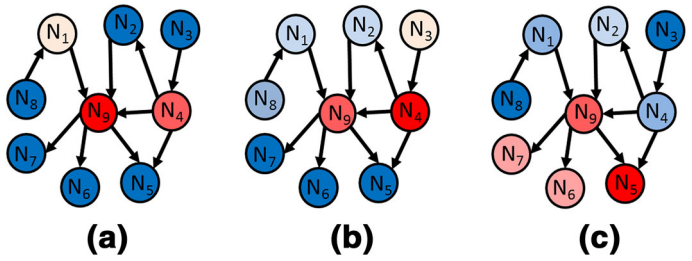

Fig. 1 Example of "key" papers based on the selected centrality metric for directed networks. A color scheme represents the *keyness* of each paper in the network with the most "key" papers shown in *dark red* (maximum value) and the least key papers shown in *dark blue* (minimum value). **a** N_9 falls on the greatest number of shortest paths between other nodes within the network and serves as the Glue paper. **b** N_4 contains the greatest number of connections (direct plus indirect) to the other nodes and serves as the Kevin Bacon paper. **c** N_5 connects to higher valued papers and serves as the Big Fish paper. (Color figure online)

(Bonacich 1972). Bonacich (1972) finds that overlaps with central groups provides a greater benefit than overlaps with isolated groups. Figure 1 provides an example of key papers within a network for each of the three centrality metrics.

Figure 1 displays the same network three times colored differently for each of the centrality metrics. Figure 1a displays the key Glue paper for the network. Node N_9 represents the key paper in this network based on betweenness Centrality since it lies on the greatest number of shortest paths between other nodes within the network. Nodes N_2, N_3, N_5, N_6, N_7, and N_8 all have betweenness values of 0 (dark blue) since they are not located between any other nodes within the network. N_1 has a slightly higher value and is shown in light red since is falls on four shortest paths that each start at N_8 and end at N_5, N_6, N_7, and N_9. N_4 contains a slightly redder color since it falls on five shortest paths that each start at N_3 and end at N_2, N_5, N_6, N_7, and N_9. N_9 has a value of 1 (dark red) and is found on the shortest path between 13 nodes: starting at N_1, N_2, and N_8 and ending at N_5, N_6, and N_7; and starting at N_3 and N_4 and ending at N_6 and N_7.

For the same network configuration, Fig. 1b displays the Kevin Bacon papers based on the closeness centrality metric. N_4 is now the key paper (dark red) with three direct links (distance between nodes equals one) and two indirect links with distance between nodes equal to two. N_9 has the second highest value (medium red) with three direct links and zero indirect links. N_3 is in third (light red) with one direct link, three indirect links with distance between nodes equal to two, and two indirect links with distance between nodes equals three. N_1 and N_2 are both light blue with one direct link and three indirect links each. N_8 is medium blue with one direct link, one indirect with a distance of two, and three order links with a distance of three. N_5, N_6, and N_7 are dark blue with no links to any other nodes.

Figure 1c displays the Big Fish paper based on the eigenvector centrality metric, again using the same network configuration as the previous two networks. Assuming that N_3 and N_8 (dark blue) provide very low-scoring nodes, they provide very little benefit to N_4 and N_1 (medium blue), respectively. N_2 (light blue) receives a slight benefit from N_4 which it conveys to N_9 along with small benefits from N_1 and N_4. Therefore, due to the culmination of benefits from its incoming nodes N_9 is moderately important (medium red) within the network. N_9 provides its benefit to N_6 and N_7 (each shown in light red) as well as N_5. N_5 is the key node based on its eigenvector centrality value due to the cumulative benefits from N_4 and N_9. Therefore, each centrality metric identifies a different key node within the same

network. There are network configurations that can lead to the same node being the key paper for different metrics (Bonacich 2007). In the next section, we present how these centrality metrics fit into our methodology for examining papers within PLOS.

Methodology

We employ a two-step approach for filtering and finding key papers within a publication. We start by modeling the publication as a co-citation network consisting of the papers within the publication as nodes and the relationships (the citations) between each of the papers as links. To generate this network, we first retrieve the list of citations for each paper within the journal. Next, we check each of the cited papers to identify if each cited paper comes from the same publication. For each of these papers, we then check their citations to see if they also originate from the same publication. We repeat this process until there are no additional citations that originate from the same journal as the initial paper. This procedure results in a directed network with the flow of information originating from the cited paper into the paper doing the citing. Therefore, we form a dataset of papers A_n citing papers B_{Ax}, where n is the number of papers within the publication and B_{AX} is the set of papers within the publication that A_i cites (where i is the current paper). Figure 2 displays the methodology for creating the network and for analyzing the papers within the network.

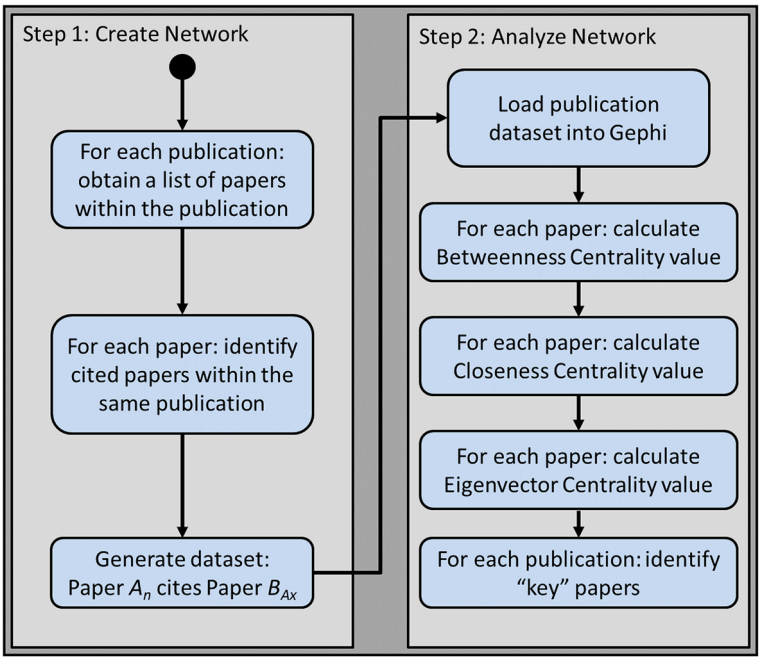

Fig. 2 Two-step process for determining the "key" papers within a publication based on the three centrality metrics. *Step 1* involves creating the citation network for all of the papers within each publication. *Step 2* involves calculating the centrality values for each paper within the network in order to identify the "key" papers based on each centrality metric

In Step 2, we load the set of papers for each publication into an open source graph and network analysis software tool named Gephi. Gephi provides a means to explore networks through "spatializing, filtering, navigating, manipulating and clustering" nodes under a variety of visualizations (Bastian et al. 2009, p. 1). Among other types of network analysis metrics, Gephi can calculate the eigenvector, betweenness, and closeness centrality metrics of each node within a directed network. Therefore, we use Gephi to read in the dataset of papers and to calculate these three centrality values for every paper within a publication. We then sort the papers from high to low values in order to determine the key papers for each of these centrality metrics. The best filter is one that returns the smallest non-zero number of papers.

In order to determine which centrality metric can be used to identify key papers, we analyze the papers within the network to determine the centrality metric that best correlates with the citation counts for each paper. Our assumption is that since citation count is the standard metric, any good metric should correlate well with the number of citations. The approach is as follows. For each key paper obtained we identify (1) the total number of times the paper has been cited in any journal, (2) the total number of times the paper was cited within the journal it was published, and (3) the number of times the paper was cited by other papers within its publishing journal in the first five years of its publication. We use the Spearman correlation coefficient to measure the degree of association between each metric and these three citation counts for each key paper in order to compare the rank of a paper's centrality to its respective citation count. Spearman correlation functions by sorting the papers by high to low centrality value and assigning them a numerical value based on their ranked position after being sorted. Then, the papers are sorted by high to low value based on their citation counts and assigned a numerical value indicating their ranking after being sorted. These two sets of ranking values are then checked for correlation. This correlation approach normalizes the non-linearities between the two sets of data and also accounts for ties within the data sets (i.e. multiple papers with 1 citation). In the following section we apply this methodology to find key papers within PLOS.

Finding key papers in PLOS

We apply the eigenvector, betweenness, and closeness centrality metrics to identify key papers within each of the discipline-specific PLOS periodicals: PLOS Biology, PLOS Computational Biology, PLOS Genetics, PLOS Medicine, PLOS Pathogens, and PLOS Neglected Tropical Diseases. To accomplish this task, we first create a model of PLOS as a co-citation network that displays the papers as nodes and the links (via citations) as the connections between the nodes. The inception dates of these journals vary from 2003 to 2005, but each journal has continued to be published through 2014. For each of these journals, we construct a network representation of the papers using the approach described in the "Methodology" section, resulting in six unique co-citation networks. The process of creating each of these domain-specific networks starts by taking only the papers within the current periodical and then linking each paper to each other paper that it cites within any of the PLOS journals. This process repeats for each of the cited papers, and then for each of the cited papers' cited PLOS papers, and so on until there are no more cited papers that cite a PLOS paper. Additionally, while we do not create a co-citation network for PLOS One (since PLOS One is not domain specific) the cited papers that reside in PLOS One are included in the co-citation network.

As an example, to create a co-citation network for PLOS Biology, we start with a paper within PLOS Biology PB_1, extract its citation list, and identify that three of its cited papers are contained within a PLOS journal: two from PLOS Biology; and one from PLOS One. These three papers are added as nodes to the network and linked to PB_1 using directed links that point towards PB_1. Then we take the citation lists for each of these three papers and add them as nodes with links pointing towards the papers that they cite. This process repeats until no more citations exist within PLOS. Then, we take the next paper within PLOS Biology that is not already contained within the network and start this process again, each time adding to the existing network for PLOS Biology. Using these six co-citation networks, we compute the betweenness, closeness, and eigenvector centrality values of each paper within the six PLOS periodicals. Table 1 provides the characteristics for each of the created co-citation networks.

We observe that the networks for all journals are scale-free and the paper to citation ratio (nodes to links in Table 1) follows a Power law. This means that that the citation counts do not increase proportionally with the increasing number of papers. This finding is consistent with Redner (1998) and Price (1965) about scientific papers in general and indicates that PLOS papers that are highly cited tend to get more cited over time (Price 1965) which might overvalue some papers when relying solely on citation count. We summarize our findings in Table 2 which shows the *Glue*, *Kevin Bacon*, and *Big Fish* papers within PLOS for each of the six journals. As expected, there are many Kevin Bacon papers for each of the periodicals; therefore, the total number of Kevin Bacon papers is given in place of the paper titles.

Discussion

In terms of filtering, we observe that betweenness centrality (Glue Paper) is not an effective filter especially in cases where there are no distinct communities of papers. In the case of the journals that we analyzed, this filter returns zero papers. Closeness centrality (Kevin Bacon Paper) is also not a good filter because it yields a large number of papers. This finding confirms the recency bias discussed in Price (1965) and points to the lack of existence of the "super classic" paper that was posited in Price (1965). In other words, there is no foundational set of papers that are cited at a relatively higher frequency in the journals that we analyzed. Eigenvector centrality (Big Fish Paper) yields a small set of papers and therefore is the best filter that can be used as an initial starting point. A key observation is that the papers we discover as Big Fish are relatively recent (2007-2014).

Table 1 Network configuration of each of the discipline-specific PLOS periodicals

PLOS Periodical	Graph type	Number of nodes	Number of links
PLOS Biology	Directed	3215	1926
PLOS Computational Biology	Directed	4877	2191
PLOS Genetics	Directed	4593	464
PLOS Medicine	Directed	2541	894
PLOS Pathogens	Directed	5818	2620
PLOS Neglected Tropical Diseases	Directed	3150	226

The number of links shown in Column 4 are the number of links connecting any PLOS paper to another PLOS paper within the dataset

Table 2 Comparative view of key papers based on each centrality metric

Journal	Glue paper	Kevin Bacon paper	Big Fish paper	Eigenvector value
PLOS Biology (Established 2003)	None	452 papers originating from all PLOS journals except for PLOS Disease	Wong, O. K., Guthold, M., Erie, D. A., & Gelles, J. (2008). Interconvertible Lac repressor–DNA loops revealed by single-molecule experiments. *PLOS Biology*, *6*(9), e0060232	1.0
			Knoops, K., Kikkert, M., van den Worm, S. H., Zevenhoven-Dobbe, J. C., van der Meer, Y., Koster, A. J., … & Snijder, E. J. (2008). SARS-coronavirus replication is supported by a reticulovesicular network of modified endoplasmic reticulum. *PLOS Biology*, *6*(9), e0060226	0.50
PLOS Computational Biology (Established 2005)	None	1271 papers originating from all PLOS journals	Rougier, N. P., Droettboom, M., & Bourne, P. E. (2014). Ten simple rules for better figures. *PLOS Computational Biology*, *10*(9), e1003833	1.0
			Staff, T. P. C. B. (2014). Correction: The Self-Limiting Dynamics of TGF-β Signaling In Silico and In Vitro, with Negative Feedback through PPM1A Upregulation. *PLOS Computational Biology*, *10*(8), e1003832	0.37
PLOS Disease (Established 2007)	None	184 papers originate from all PLOS journals except for PLOS Biology	Simarro, P. P., Diarra, A., Postigo, J. A. R., Franco, J. R., & Jannin, J. G. (2011). The human African trypanosomiasis control and surveillance programme of the World Health Organization 2000–2009: the way forward. *PLOS Neglected Tropical Diseases*, *5*(2), e0001007	1.0
			Teixeira, A. R., Gomes, C., Nitz, N., Sousa, A. O., Alves, R. M., Guimaro, M. C., … & Hecht, M. M. (2011). Trypanosoma cruzi in the chicken model: Chagas-like heart disease in the absence of parasitism. *PLoS Neglected Tropical Diseases*, *5*(3), e0001000	0.17

Table 2 continued

Journal	Glue paper	Kevin Bacon paper	Big Fish paper	Eigenvector value
PLOS Genetics (Established 2005)	None	345 papers originate from all PLOS journals except for PLOS Disease and PLOS Medicine	Wojciechowski, R., & Hysi, P. G. (2013). Focusing in on the complex genetics of myopia. *PLOS Genetics*, *9*(4), e1003442	1.0
			Connallon, T., & Clark, A. G. (2013). Sex-differential selection and the evolution of X inactivation strategies. *PLOS Genetics*, *9*(4), e1003440	0.27
PLOS Medicine (Established 2004)	None	188 papers originate from all PLOS journals	Gong, Y., Somwar, R., Politi, K., Balak, M., Chmielecki, J., Jiang, X., & Pao, W. (2007). Induction of BIM is essential for apoptosis triggered by EGFR kinase inhibitors in mutant EGFR-dependent lung adenocarcinomas. *PLOS Medicine*, *4*(10), e0040294	1.0
			Unger, A., & Riley, L. W. (2007). Slum health: from understanding to action. *PLOS Medicine*, *4*(10), e0040295	0.47
PLOS Pathogens (Established 2005)	None	1331 papers originate from all PLOS journals except for PLOS Disease	Lamb, E. W., Walls, C. D., Pesce, J. T., Riner, D. K., Maynard, S. K., Crow, E. T., … & Davies, S. J. (2010). Blood fluke exploitation of non-cognate CD4 + T cell help to facilitate parasite development. *PLOS Pathogens*, *6*(4), e1000892	1.0
			Bánki, Z., Posch, W., Ejaz, A., Oberhauser, V., Willey, S., Gassner, C., … & Wilflingseder, D. (2010). Complement as an endogenous adjuvant for dendritic cell-mediated induction of retrovirus-specific CTLs. *PLoS Pathogens*, *6*(4), e1000891	0.39

Column 1 provides each journal within PLOS along with the year that the journal was established. Columns 2, 3, and 4 present the *Glue*, *Kevin Bacon*, and *Big Fish* papers within each journal, respectively. Column 3 also indicates which journals the Kevin Bacon papers originate from including PLOS One. Column 5 provides the Eigenvector values for the Big Fish papers. Due to the large number of Kevin Bacon papers within each journal the total number of Kevin Bacon papers is provided in place of the titles

This finding fits the metric since eigenvector centrality looks at the importance of a paper based on the papers that it cites. The transitive property of eigenvector centrality dictates that newer papers mentioning several important works gains more importance regardless of length of time between the publications. We also observe that although the papers identified as key have a low number of citations, the number of downloads is very high (average 1704 and median 1082 downloads). Table 3 provides additional information on the Big Fish papers for each of the six domain specific PLOS journals.

Table 3 Analysis of key Big Fish papers for each PLOS journal

Paper title	Journal	Year journal established	Year paper published	Times cited within PLOS	Total times downloaded from PLOS
Interconvertible Lac repressor–DNA loops revealed by single-molecule experiments	PLOS Biology	2003	2008	4	1082
SARS-coronavirus replication is supported by a reticulovesicular network of modified endoplasmic reticulum	PLOS Biology	2003	2008	12	2158
Correction: The Self-Limiting Dynamics of TGF-β Signaling In Silico and In Vitro, with Negative Feedback through PPM1A Upregulation	PLOS Computational Biology	2005	2014	0	49
Ten simple rules for better figures	PLOS Computational Biology	2005	2014	0	8672
Focusing in on the complex genetics of myopia	PLOS Genetics	2005	2013	1	508
Sex-differential selection and the evolution of X inactivation strategies	PLOS Genetics	2005	2013	0	473
Slum health: from understanding to action	PLOS Medicine	2004	2007	2	3009
Induction of BIM is essential for apoptosis triggered by EGFR kinase inhibitors in mutant EGFR-dependent lung adenocarcinomas	PLOS Medicine	2004	2007	5	2238
The human African trypanosomiasis control and surveillance programme of the World Health Organization 2000–2009: the way forward	PLOS Neglected Tropical Diseases	2007	2011	42	1976
Trypanosoma cruzi in the chicken model: Chagas-like heart disease in the absence of parasitism	PLOS Neglected Tropical Diseases	2007	2011	3	866
Complement as an endogenous adjuvant for dendritic cell-mediated induction of retrovirus-specific CTLs	PLOS Pathogens	2005	2012	2	527
Blood fluke exploitation of non-cognate CD4 + T cell help to facilitate parasite development	PLOS Pathogens	2005	2010	3	345

Column 1 provides the title of the key paper. Column 2 displays the journal containing the key paper. Column 3 shows the year that the journal was established. Column 4 displays the year that the key paper was published. Column 5 shows the number of times that the paper is cited by other papers within PLOS as of February 19, 2015. Column 6 shows the total number of times that the paper has been downloaded from the PLOS website in either PDF or XML formats as of February 20, 2015

In terms of using eigenvector centrality as a metric, papers with high eigenvector values have an advantage over papers with low values since an increasing citation count for a paper in a network causes the eigenvector value of all of its connected papers to also increase. However, since it has been shown that papers that get cited tend to get cited more (scale-free property), papers connected to those with high eigenvector centrality value will get an indirect boost in relative importance. As a result, eigenvector centrality is a good filter but we need to investigate more in order to determine if it also serves as a good metric for determining which papers are key.

As stated in the "Finding key papers in PLOS" section, we explore the correlation between centrality metrics and citation counts to determine which metrics serve as good filters. We will focus only on eigenvector centrality since it is the best filter according to our criteria. Specifically, we explore the extent to which eigenvector centrality explains: (1) the number of times the paper was cited within PLOS in the first 5 years of its publication; (2) the total number of times that the paper was cited within PLOS; and (3) the total number of times the paper has been cited in any publication according to Google Scholar. The numbers of citations for (1) and (2) will be equal for papers that are not yet 5 years old. We use the Spearman correlation coefficient to measure the degree of association between the eigenvector centrality of a paper and the three collected citation counts to compare the rank of a paper's centrality to the rank of its respective citation count. This is preferable to comparing the value of a paper's centrality measure to the value of its respective citation counts because each of the three citation counts follows a logarithmic distribution. Since we know that the citation count follows a power law, ranking each of the variables normalizes these non-linearities between the data sets. Table 4 shows that for all six discipline-specific PLOS journals there is a positive statistically significant Spearman correlation between the eigenvector centrality of a paper within our network and each of the three citation counts. Furthermore, almost all the correlations are of moderate strength (>0.50) and many related to the number of citations within the journal are strong

Table 4 Correlation of eigenvector centrality metrics for papers in each publication within PLOS

Publication	Correlation between eigenvector centrality of publication and its discipline-specific 5 Year PLOS Cites	Correlation between eigenvector centrality of publication and its total citations within PLOS at the end of 2014	Correlation between eigenvector centrality of publication and its total overall citations as of 2014
PLOS Biology	0.9206309	0.9883407	0.448269
PLOS Computational Biography	0.9537765	0.9801788	0.5124322
PLOS Disease	0.9667521	0.9740426	0.6196576
PLOS Genetics	0.9626788	0.9825487	0.5449447
PLOS Medicine	0.9492422	0.9624285	0.3817552
PLOS Pathogens	0.9613983	0.9788194	0.5681881

Column 1 displays the name of the PLOS journal. Column 2 displays the correlation between the eigenvector centrality value and the number of citations that the paper has accumulated by other papers within PLOS after 5 years. Column 3 displays the correlation between the eigenvector centrality value and the number of citations that the paper has accumulated by other papers within PLOS by the year 2014. Column 4 displays the correlation between the eigenvector centrality value and the total number of citations from any source that the paper has accumulated by the end of year 2014 according to Google Scholar

(>0.80). The eigenvector centrality values correlate well within discipline key papers within 5 years of publication and at least moderately predictive for all other citations. As a result, we conclude that eigenvector centrality is the most appropriate centrality metric for filtering papers and measuring the relative importance of a paper within a journal.

Eigenvector centrality identifies the papers frequently cited by other important papers within other disciplines of the same journal for papers within a specific discipline. A high Spearman ranking correlation between eigenvector centrality values and citation counts indicates that papers with high eigenvector centrality values also have high citation counts relative to the other papers within the journal. Recall that Spearman ranking correlation accounts for the ranking of eigenvector values against the ranking of the citation counts for the papers. This might indicate that people are more likely to cite (1) papers within the outlet that they are trying to publish in because the authors are aware of the existing research or they feel pressured to cite papers within this outlet as discussed in Wilhite and Fong (2012) and (2) papers within their first 5 years of publication since they are on the cutting edge of research as observed in Price (1965).

The strength of the Spearman Correlation coefficient decreases (despite remaining statistically significant) when we examine the rank of the eigenvector values against the rank of the total Google Scholar citation counts. This may occur due to the authors being less aware of the papers, being less inclined to search for papers outside of the journal that they are publishing in, or feeling less pressured to cite papers within the current outlet thereby denoting a real non-spurious relationship between the two measures. The papers with low eigenvector centrality values are less likely to be cited by researchers going forward as these papers may not be at the forefront of the research within that discipline, as evidenced by the high correlation between low eigenvector values and low citation counts. However, these papers remain likely to be cited since the correlations remain strong. These finding is also in line with those of Price (1965). In short, Eigenvalue is a good metric for in-journal article ranking with the limitation that the value of a paper can be inflated as the papers it relates grows in citation count.

Conclusions

In this paper, we examine the extent to which the metrics are useful for identifying important papers within a journal and we identify 12 Big Fish papers within PLOS. We show that the co-citation network within PLOS follows a power law similar to that of the citation count of a paper over time. Consequently, closeness centrality is not very useful for filtering the papers within a journal because it yields too many important papers and betweenness centrality is not a useful metric when there are no communities of papers to connect. Results show that eigenvector centrality is a good metric for identifying important papers in a journal. Our results are critical in that eigenvector centrality is time-independent and allows for the identification of key papers at the time of their publication. While a newly published paper may be classified as a key paper with respect to eigenvector centrality, the paper will not yet have had a chance to be read by the scientific community; therefore, the eigenvector centrality metric serves as a starting hypothesis that the paper is indeed a key paper. These Big Fish papers link with highly regarded papers and having passed through a peer reviewed process may indeed serve as a key paper within the same networks as the papers that they cite.

We recommend that journals calculate and include the eigenvector centrality of each paper within the publication to provide both researchers and students an idea of the relative importance of the paper within its own publication or within a larger body of knowledge. We also recommend that journals publish and update the top 10 or top 25 papers in decreasing order of eigenvector centrality each time a new issue of the journal is published. This top 10 or top 25 can constitute the initial set of papers that the journal recommends a researcher use as a starting point for exploring the journal. The peer-review process remains critical in preventing unworthy papers from being published even if they cite a large number of key papers. Chan et al. (2015) note that the quality of the scientific contribution of a work can be biased due to factors such as institutional affiliation and the time of publication. Therefore, the eigenvector centrality metric identifies papers that are reachable from other key papers, but it does not serve as an absolute metric on the quality of identified Big Fish papers. Furthermore, in terms of proposal submission and promotion, researchers can show the importance of new publications before they accumulate citations by employing eigenvector centrality. Finally, each researcher can create a co-citation network of their publication and measure the relative importance of their work based on the eigenvector centrality rather than the total citation count. By doing so, researchers can identify other important work that is related to their most important work and thus explore fruitful lines of work and areas of collaboration with other researchers working in the same domain.

Future work includes further exploring the connections between Big Fish papers and their download counts over time. Of particular interest to eigenvector centrality is the potential to explore the usefulness of a weighted-eigenvector metric that accounts for the total download counts or frequency of downloads over time for each paper within the network. Additionally, the centrality values of the papers obtained using citations within the journals that they are published in can be compared against their centrality values using citations from all sources to examine the effect on the determination of *keyness* for a paper between these two sources.

Acknowledgments We acknowledge our colleagues at the Virginia Modeling Analysis and Simulation Center (VMASC) and the numerous discussions and feedback they provided for this paper.

References

Abt, H. A. (2000). Do important papers produce high citation counts? *Scientometrics, 48*(1), 65–70.

Allesina, S., & Pascual, M. (2009). Googling food webs: Can an eigenvector measure species' importance for coextinctions. *PLoS Computational Biology, 5*(9), e1000494.

Barnett, G. A., Huh, C., Kim, Y., & Park, H. W. (2011). Citations among communication journals and other disciplines: A network analysis. *Scientometrics, 88*(2), 449–469.

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. *In Proceedings of the Third International AAAI Conference on Weblogs and Social Media (ICWSM), 8*, 361–362.

Beliga, S., Meštrović, A., & Martinčić-Ipšić, S. (2015). An overview of graph-based keyword extraction methods and approaches. *Journal of Information and Organizational Sciences, 39*(1), 1–20.

Bergstrom, C. (2007). Measuring the value and prestige of scholarly journals. *College & Research Libraries News, 68*(5), 3146.

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

Bonacich, P. (1972). Technique for analyzing overlapping memberships. *Sociological Methodology, 4*, 176–185.

Bonacich, P. (1987). Power and centrality: A family of measures. *American Journal of Sociology, 92*(5), 1170–1182.

Bonacich, P. (2007). Some unique properties of eigenvector centrality. *Social Networks, 29*(4), 555–564.
Chan, H. F., Guillot, M., Page, L., & Torgler, B. (2015). The inner quality of an article: Will time tell? *Scientometrics, 104*(1), 19–41.
Erkan, G., & Radev, D. R. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research, 22*, 457–479.
Feeley, T. H., LaVail, K. H., & Barnett, G. A. (2010). Predicting faculty job centrality in communication. *Scientometrics, 87*(2), 303–314.
Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry, 40*(1), 35–41.
Freeman, L. C. (1979). Centrality in social networks conceptual clarification. *Social Networks, 1*(3), 215–239.
Garfield, E. (2006). The history and meaning of the journal impact factor. *The Journal of the American Medical Association, 295*(1), 90–93.
Griffin, D. J., Bolkan, S., Holmgren, J. L., & Tutzauer, F. (2015). Central journals and authors in communication using a publication network. *Scientometrics,*. doi:10.1007/s11192-015-1774-4.
Gross, P. L. K., & Gross, E. M. (1927). College libraries and chemical education. *Science, 66*(1713), 385–389.
Guns, R., Liu, Y. X., & Mahbuba, D. (2011). Q-measures and betweenness centrality in a collaboration network: A case study of the field of informetrics. *Scientometrics, 87*(1), 133–147.
Hanneman, R. A., & Riddle, M. (2005). *Introduction to social network methods*. Riverside, CA: University of California, Riverside.
Joyce, K. E., Laurienti, P. J., Burdette, J. H., & Hayasaka, S. (2010). A new measure of centrality for brain networks. *PLoS ONE, 5*(8), e12200.
Kane, G. C. (2009). It's a network, not an encyclopedia: A social network perspective on wikipedia collaboration. In *Proceedings of the academy of management* (Vol. 2009, No. 1, pp. 1–6).
Khan, G. F., & Wood, J. (2015). Information technology management domain: Emerging themes and keyword analysis. *Scientometrics,*. doi:10.1007/s11192-015-1712-5.
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. (2011). "Structuration" by intellectual organization: the configuration of knowledge in relations among structural components in networks of science. *Scientometrics, 88*(2), 499–520.
Li-chun, Y., Kretschmer, H., Hanneman, R. A., & Liu, Z. Y. (2006). Connection and stratification in research collaboration: An analysis of the COLLNET network. *Information Processing and Management, 42*(6), 1599–1613.
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.
Liu, X., Jiang, S., Chen, H., Larson, C. A., & Roco, M. C. (2015). Modeling knowledge diffusion in scientific innovation networks: an institutional comparison between China and US with illustration for nanotechnology. *Scientometrics,*. doi:10.1007/s11192-015-1761-9.
Lohmann, G., Margulies, D. S., Horstmann, A., Pleger, B., Lepsien, J., Goldhahn, D., et al. (2010). Eigenvector centrality mapping for analyzing connectivity patterns in fMRI data of the human brain. *PLoS ONE, 5*(4), e10232–e10232.
Neylon, C., & Wu, S. (2009). Article-level metrics and the evolution of scientific impact. *PLoS Biology,*. doi:10.1371/journal.pbio.1000242.
Price, D. (1965). Networks of scientific papers. *Science, 149*, 510–515.
Redner, S. (1998). How popular is your paper? An empirical study of the citation distribution. *The European Physical Journal B-Condensed Matter and Complex Systems, 4*(2), 131–134.
Sikorav, J. L. (1991). The utility of scientific papers. *Scientometrics, 21*(1), 49–68.
Siler, K. (2013). Citation choice and innovation in science studies. *Scientometrics, 95*(1), 385–415.
Stringer, M. J., Sales-Pardo, M., & Amaral, L. A. N. (2008). Effectiveness of journal ranking schemes as a tool for locating information. *PLoS ONE,*. doi:10.1371/journal.pone.001683.
West, J. D., Bergstrom, T. C., & Bergstrom, C. T. (2010). The eigenfactor metrics™: A network approach to assessing scholarly journals. *College & Research Libraries, 71*(3), 236–244.
Wilhite, A. W., & Fong, E. A. (2012). Coercive citation in academic publishing. *Science, 335*(6068), 542–543.
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.
Zhu, X., Wu, Q., Zheng, Y., & Ma, X. (2004). Highly cited research papers and the evaluation of a research university: A case study: Peking University 1974–2003. *Scientometrics, 60*(2), 237–347.