Journal of Informetrics 10 (2016) 1207–1223

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Regular article

Identification of milestone papers through time-balanced network centrality

Manuel Sebastian Mariani[a,*], Matúš Medo[a], Yi-Cheng Zhang[a,b]

[a] Department of Physics, University of Fribourg, 1700 Fribourg, Switzerland
[b] Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing 400714, P.R. China

ARTICLE INFO

Article history:
Received 18 June 2016
Received in revised form 14 October 2016
Accepted 14 October 2016
Available online 4 November 2016

Keywords:
Citation analysis
Metrics evaluation
PageRank
Time bias
Milestone papers

ABSTRACT

Citations between scientific papers and related bibliometric indices, such as the *h*-index for authors and the impact factor for journals, are being increasingly used – often in controversial ways – as quantitative tools for research evaluation. Yet, a fundamental research question remains still open: to which extent do quantitative metrics capture the significance of scientific works? We analyze the network of citations among the 449,935 papers published by the American Physical Society (APS) journals between 1893 and 2009, and focus on the comparison of metrics built on the citation count with network-based metrics. We contrast five article-level metrics with respect to the rankings that they assign to a set of fundamental papers, called Milestone Letters, carefully selected by the APS editors for "making long-lived contributions to physics, either by announcing significant discoveries, or by initiating new areas of research". A new metric, which combines PageRank centrality with the explicit requirement that paper score is not biased by paper age, is the best-performing metric overall in identifying the Milestone Letters. The lack of time bias in the new metric makes it also possible to use it to compare papers of different age on the same scale. We find that network-based metrics identify the Milestone Letters better than metrics based on the citation count, which suggests that the structure of the citation network contains information that can be used to improve the ranking of scientific publications. The methods and results presented here are relevant for all evolving systems where network centrality metrics are applied, for example the World Wide Web and online social networks. An interactive Web platform where it is possible to view the ranking of the APS papers by rescaled PageRank is available at the address http://www.sciencenow.info.

1. Introduction

The notion of quantitative evaluation of scientific impact builds on the basic idea that the scientific merits of papers (Narin, 1976; Radicchi, Fortunato, & Castellano, 2008), scholars (Egghe, 2006; Hirsch, 2005), journals (Bollen, Rodriquez, & Van de Sompel, 2006; Liebowitz & Palmer, 1984; Pinski & Narin, 1976), universities (Kinney, 2007; Molinari & Molinari, 2008) and countries (Cimini, Gabrielli, & Labini, 2014; King, 2004) can be gauged by metrics based on the received citations. The respective field, referred to as bibliometrics or scientometrics, is undergoing a rapid growth (Van Noorden, 2010) fueled by the increasing availability of massive citation datasets collected by both academic journals and online platforms, such as

* Corresponding author.
E-mail address: manuel.mariani@unifr.ch (M.S. Mariani).

http://dx.doi.org/10.1016/j.joi.2016.10.005
1751-1577/

Google Scholar and Web of Science. The possible benefits, drawbacks and long-term effects of the use of bibliometric indices are being highly debated by scholars from diverse fields (Hicks, Wouters, Waltman, de Rijcke, & Rafols, 2015; Lawrence, 2008; Van Raan, 2005; Weingart, 2005; Werner, 2015).

Although some effort has been devoted to contrast different metrics with respect to their ability to single out seminal papers (Dunaiski & Visser, 2012; Dunaiski, Visser, & Geldenhuys, 2016; Yao, Wei, Zeng, Fan, & Di, 2014; Zhou, Zeng, Fan, & Di, 2015), differences among the adopted benchmarking procedures and diverse conclusions of the mentioned references leave a fundamental question still open: which metric of scientific impact best agrees with expert-based perception of significance? In agreement with Wasserman, Zeng, and Amaral (2015), the significance of a scientific work is intended here as its enduring importance within the scientific community.

To address this question, we focus on a list of 87 physics papers of outstanding significance – called Milestone Letters – recently made available by the American Physical Society (APS) [http://journals.aps.org/prl/50years/milestones, accessed 25-11-2015]. According to the APS editors' description, the Milestone Letters "have made long-lived contributions to physics, either by announcing significant discoveries, or by initiating new areas of research". These articles have been carefully selected by the editors of the APS, and the choices are motivated in detail in the webpage; the fact that a large fraction of them led to Nobel Prize for some of their authors is an indication of the exceptional level of the selected works.

In this work, we analyze the network of citations between the $N = 449{,}935$ papers published in APS journals from 1893 until 2009 to compare five article-level metrics with respect to the ranking position they assign to the Milestone Letters. A reliable expert-based evaluation of the significance (intended as enduring importance, as in Wasserman et al., 2015) of a paper necessarily requires a time lag between the paper's publication date and the expert's judgment. For example, there is a time interval of 14 years between the most recent Milestone Letter (from 2001) and the year at which the list of Milestone Letters was released (2015). However, we show that a well-designed quantitative metric offers us the opportunity to detect potentially significant papers relatively short after their publication – an aspect often neglected in the evaluation of bibliometric indicators. To show this, we study how the ability of the different metrics to identify the Milestone Letters changes with paper age.

A plethora of quantitative metrics exist and could be studied in principle. Our focus here is narrowed to metrics that rely on a diffusion process on the underlying network of citations between papers and their comparison with simple citation count. The five metrics considered in this work are thus: the citation count, PageRank (introduced by Brin & Page, 1998), CiteRank (introduced by Walker, Xie, Yan, & Maslov, 2007), rescaled citation count (introduced by Newman, 2009), and novel rescaled PageRank. PageRank is a classical network centrality metric which combines a random walk along network links with a random teleportation process. The metric has been applied to a broad range of real-world problems (Ermann, Frahm, & Shepelyansky, 2015; Franceschet, 2011; Gleich, 2015 for a review), including ranking academic papers (Chen, Xie, Maslov, & Redner, 2007; Yao et al., 2014), journals (Bollen et al., 2006; González-Pereira, Guerrero-Bote, & Moya-Anegón, 2010) and authors (Nykl, Ježek, Fiala, & Dostal, 2014; Radicchi, Fortunato, Markines, & Vespignani, 2009; Yan & Ding, 2009) (see Waltman & Yan, 2014 for a review of the applications of PageRank-related methods to bibliometric analysis).

To overcome the well-known PageRank's bias toward old nodes in citation data (detailedly studied by Chen et al., 2007; Mariani, Medo, & Zhang, 2015), the CiteRank algorithm introduces exponential penalization of old nodes, resulting in a node score that well captures the future citation count increase of the papers and, for this reason, can be considered as a reasonable proxy for network traffic, as shown by Walker et al. (2007). However, we show below that CiteRank score does not allow one to fairly compare papers of different age. Rescaled citation count and rescaled PageRank are derived from citation count and PageRank score, respectively, by explicitly requiring that paper score is not biased by age – the adopted rescaling procedure is conceptually close to the methods recently developed by Radicchi et al. (2008), Newman (2009), Radicchi and Castellano (2011), Newman (2014), Radicchi and Castellano (2012b), Radicchi and Castellano (2012a), Crespo, Ortuño-Ortín, and Ruiz-Castillo (2012) and Kaur, Ferrara, Menczer, Flammini, and Radicchi (2015) to suppress biases by age and field in the evaluation of academic agents. We find that the rankings produced by the rescaled scores are indeed consistent with the hypothesis that the rankings are not biased by age.

We find that PageRank can compete and even outperform rescaled PageRank in identifying *old* milestone papers, but completely fails to identify recent milestone papers due to its temporal bias. CiteRank can compete and even outperform rescaled PageRank in identifying *recent* milestone papers, but markedly underperforms in identifying old milestone papers due to its built-in exponential penalization for older papers. Indicators based on simple citation count are outperformed by rescaled PageRank for papers of every age. This leads us to the conclusion that rescaled PageRank is the best-performing metric overall. With respect to previous works by Chen et al. (2007), Dunaiski and Visser (2012), Fiala (2012) and Dunaiski et al. (2016) that claimed the superiority of network-based metrics in identifying important papers, our results clarify the essential role of paper age in determining the metrics' performance: rescaled PageRank excels and PageRank performs poorly in identifying MLs short after their publication, and the performance of the two methods becomes comparable only 15 years after the MLs are published. Qualitatively similar results are found for an alternative list of APS outstanding papers which only includes works that have led to Nobel prize for some of the authors (the list is provided in the Table S2).

Our results indicate that network centrality and time-balance are two essential ingredients – though neglected by popular bibliometric indicators such as the *h*-index for scholars (Hirsch, 2005) and impact factor for journals (Garfield, 1972) – for an effective detection of significant papers. This sets a new benchmark for article-level metrics and quantitatively support the paradigm that considering the whole network instead of simple citation count can bring substantial benefits to the ranking of academic agents. In a broader context, our results show that a direct rescaling of PageRank scores is an effective