Physica A 513 (2019) 325–332

Contents lists available at ScienceDirect

Physica A

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

Age preference of metrics for identifying significant nodes in growing citation networks

Zhuo-Ming Ren

Alibaba Research Center for Complexity Sciences, Alibaba Business School, Hangzhou Normal University, Hangzhou 311121, PR China

HIGHLIGHTS

- Identifying significant works of Arts and Sciences should avoid age preference.
- The time-balanced variant of PageRank and degree could eliminate the age preference.
- The age preference is investigated through two time-aggregated citation networks.

ARTICLE INFO

Article history:
Received 28 June 2018
Received in revised form 29 August 2018
Available online xxxx

Keywords:
Age preference
Significance identification
Growing citation networks

ABSTRACT

Identifying significant works in the field of arts and sciences should avoid age preference. Recent research has shown that the time-balanced variant of the popular Google's PageRank and degree for identifying significant works could provide an objective approach to eliminate the age preference in growing networks. However, a fundamental question remains open: How much performance capability do time-balanced metrics expect when they identify significant nodes in the growing networks? Through investigating of two large time-aggregated citations networks between movies procured from the Internet Movie Database and papers published on the journals of American Physical Society respectively, we analyze the age preference of several time-balanced metrics of PageRank and degree for identifying significant nodes in comparison.

1. Introduction

As we known, the world overflows with creative works. In reality, it is difficult to measure the significance of works, because the evaluation of the true significance of the work depends on the historical moment, and very much "in the eye of the beholder". Fortunately, Complex networks have emerged as one of the leading frameworks to describe complex social, economic and information systems [1–4]. The network approach to complex systems has provided novel insights into various real-world problems [5,6], including understanding the growth of information systems [7,8], identifying influential spreaders [9–12], predicting the hitting time of an infectious disease [13] and so on. Thanks to the increasing availability of massive citation datasets collected by both academic journals and online platforms, we can build complex networks to evaluate significance of the work independently, and can predict the work who will potentially win big awards like "Oscar" or "Nobel". Network centrality metrics are widely used and rely on specific assumptions on what an important node should be. Degree centrality (indegree if we consider directed networks) assumes that a node is central, or important, if it has received many connections [14,15]. A natural criticism to the use of degree centrality comes from the fact that this simple metric entirely neglects the centrality of neighbors, whereas it may be plausible to assume that links coming from important nodes should be given larger weight. To implement this idea, Google's PageRank algorithm [16–18] assigns a score to each node

E-mail address: zhuoming.ren@hznu.edu.cn.

https://doi.org/10.1016/j.physa.2018.09.001
0378-4371/© 2018 Elsevier B.V. All rights reserved.

which linearly depends on the scores of the nodes that cited the node, enforcing the basic idea that a node is important if it is pointed to by other important nodes.

Many popular ranking algorithms like Google's PageRank and degree are static in nature which exhibit main shortcomings when applied to real networks that rapidly evolve over time. The time-balanced metrics like CiteRank [19] and long gap degree [20,21] consider time effect to provide a solution. Mariani et al. [22] analyzed the relationship between the algorithm's efficacy and properties of the network and showed that realistic temporal effects make PageRank fail in individuating the most valuable nodes for a broad range of model parameters. Mariani et al. [23] also developed a rescaled PageRank centrality with the explicit requirement that paper score is not biased by paper age and identified the Milestone papers [24] and predicted significant papers [25] according to the network of citations published by the American Physical Society (APS) journals and from the Microsoft Academic Graph. In addition, Medo et al. [26] introduced discoverers as the users in data from real systems who significantly outperform the others in the rate of making discoveries, i.e. in being among the first ones to collect items that eventually become very popular. Statistical null models serve this purpose by producing random networks whilst keeping chosen network's properties fixed. While there is increasing interest in networks that evolve in time, we still lack a robust time-aware framework to assess the statistical significance of their observed structural properties. Ren et al. [27] proposed a dynamic null model that preserves both the network's degree sequence and the time evolution of individual nodes' degree values. The proposed model can be used to explore the significance of widely studied network properties such as degree–degree correlations and the relations between popular node centrality metrics. Recently, the review [28] surveyed the existing ranking algorithms both static and time-aware and their applications to evolving networks, and deep understanding of how existing ranking algorithms perform.

Simultaneously, recent advances in predicting the significance of the node in evolving networks have enabled the development of a wide and diverse range of ranking algorithms that take the temporal dimension into account. However, a fundamental question remains open: how about time-balanced metrics on identifying significant nodes in the growing networks? In this paper, we argue that the answer to this question strongly depends on the time-balanced variant during the evolution of the network. Here, we use two growing networks, the network of citations between papers published in American Physical Society journals and the network of citations between US movies to analyze the five metrics.

2. Materials and methods

2.1. Materials

2.1.1. Citation networks

Scientific citation networks: Our citation database includes all papers published in journals of the American Physical Society (APS) from 1893 to 2009 [15,22]. There are 4672812 citations among 449937 papers. We can think of the set of all APS articles and their citations as a network, with nodes representing articles and a directed link between two nodes representing a citation from a citing article to a cited article.

Movie citation networks: Like scientists, artists are also often influenced or inspired by prior works. For instance, the famous flying bicycle scene in E.T.: The Extra-Terrestrial (1982) is similar to a sequence in The Thief of Bagdad (1924) where characters also fly in front of the moon. The movie citations come in the form of similar quotes, similar settings, or similar movie techniques and so on. Using the movie citations between movies, we can construct a directed network where a node is a movie, and a direct link is a citation. As the above example, we can build a directed link from E.T.: The Extra-Terrestrial (1982) to The Thief of Bagdad (1924). This network consists of 15,425 movies connected by 42,794 citations. The whole movies produced in the United States from 1894 to 2011. The detailed description of the movie citation network can be seen in Refs. [20,21].

From these datasets, we separate whole time-aggregated networks into different rating snapshots with the interval of the time unit. The time unit is one year in movie and six months in scientific citation networks. Then we can construct new time-aggregated citation networks from the initial time to a time point x, with $x = t_0 + unit, t_0 + 2unit, \ldots$.

2.1.2. Significant works

In agreement with expert-based perception of significance for creative works, a list of works of outstanding significance selected by the community of sciences or arts are regarded as benchmarks. We then use these benchmarks to analyze different metrics with respect to their ability to single out significant works.

PRL milestones The collection of milestone Physical Review Letters contains Letters that have made long-lived contributions to physics, either by announcing significant discoveries, or by initiating new areas of research [24]. The collection of milestone Physical Review Letters is chosen by the Physical Review Letters editors.

NFR The National Film Registry (NFR) highlights "culturally, historically, or aesthetically significant", who selects 25 films each year showcasing the range and diversity of American film heritage to increase awareness for its preservation [29].

Oscar The Academy Award is commonly known as simply Oscars, which is an annual American award ceremony honoring cinematic achievements in the film industry [30]. The awards are overseen by the Academy of Motion Picture Arts and