Physica A 419 (2015) 29–39

Contents lists available at ScienceDirect

Physica A

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

A new method to construct co-author networks

Jie Liu [a], Yunpeng Li [a,e], Zichan Ruan [a], Guangyuan Fu [a], Xiaowu Chen [b], Rehan Sadiq [c], Yong Deng [a,d,*]

[a] School of Computer and Information Science, Southwest University, Chongqing, 400715, China
[b] School of Computer Science, BeiHang University, Beijing, 100191, China
[c] School of Engineering, University of British Columbia Okanagan, 3333 University Way, Kelowna, BC, Canada
[d] School of Engineering, Vanderbilt University, Nashville, TN, 37235, USA
[e] Faculty of Information, Toronto University, Toronto, ON, M5S 3G6, Canada

HIGHLIGHTS

- The published time of the papers is taken into consideration to identify their influence.
- The importance of the author itself as well as that of its community status are combined to evaluate its influence.
- The proposed method is robust to different network sizes and community structure.

ARTICLE INFO

Article history:
Received 25 February 2014
Received in revised form 28 August 2014
Available online 15 October 2014

Keywords:
Co-author network
PageRank
Effective distance
Erdos number

ABSTRACT

In this paper, we propose a new method to evaluate the importance of nodes in a given network. The proposed method is based on the PageRank algorithm. However, we have made necessary improvements to combine the importance of the node itself and that of its community status. First, we propose an improved method to better evaluate the real impact of a paper. The proposed method calibrates the real influence of a paper over time. Then we propose a scheme of evaluating the contribution of each author in a paper. We later develop a new method to combine the information of the author itself and the structure of the co-author network. We use the number of co-authorship to calculate the *effective distance* between two authors, and evaluate the strength of their influence to each other with the law of gravity. The strength of influence is used to build a new network of authors, which is a comprehensive topological representation of both the quality of the node and its role in network. Finally, we apply our method to the Erdos co-author community and AMiner Citation Network to identify the most influential authors.

1. Introduction

With the development of modern science, more and more researches are completed by the cooperation of multiple scientists. These cooperations can prompt the efficiency and fertilize new ideas. As a result, the academic articles as the main form of achievements of scientific research, are composed by more and more authors. Almost all of the academic articles are published by several co-authors now, especially for those inter-discipline subjects.

Such co-authorship forms a network. One of the most famous academic co-author network is the 20th-century mathematician Paul Erdos's, who has over 500 co-authors and published over 1400 technical research papers [1–3]. Erdos's role as a

* Corresponding author at: School of Computer and Information Science, Southwest University, Chongqing, 400715, China. Tel.: +86 023 68254555; fax: +86 023 68254555.
E-mail addresses: ydeng@swu.edu.cn, prof.deng@hotmail.com (Y. Deng).

http://dx.doi.org/10.1016/j.physa.2014.10.006
0378-4371/

collaborator was so significant in the field of mathematics that *E number* is defined to measure the closeness to Erdos through Erdos's amazingly large and robust co-author network. This phenomenon is interesting, which provides us a good sample of collaborative network of scientists. Studying co-author network is useful, with which we can evaluate one's status in a certain field, and improve relations to get closer to the core community through identifying the more influential researchers [4].

There has been many researches about analyzing co-author network and identifying the influential authors in co-author network [5]. In Refs. [6–8] the authors construct the co-author network by extracting the authors in a same paper, and then analyze the network with the common methods of graph theory and complex network. In Refs. [9,10], the authors analyze the co-author network with some other algorithm like PageRank. Yet in some of these papers, the methods to construct the co-author network are not comprehensive. The co-author network is built according to the citation times of the authors' papers without considering the year of the publications of the authors, which to our consideration is an important feature of the papers. To a certain degree, the earlier the papers were published, the more opportunities the papers have to be cited. Thus the earlier published papers are unfairly preferred. We also have found many other methods to identify influencers in the network [5,11–19,12,20–34], kinds of methods are proposed to identify the influencers in social network [35–37,11], microblogging network [38], customer network [39], or other certain area. In Ref. [40], authors propose ClusterRank to consider the neighborhoods' influence to the node. In Ref. [22], the authors use K-shell to identify the influential spreaders. In Ref. [20], the authors proposed a leaderRank algorithm and Qian Li improved it in Ref. [23]. Most of them indicate that figuring out the mechanism of network can help identifying the influential nodes in the network [41,42,33,21,43,24,32,44,40]. So, in this paper, we consider the time sequence to better construct the co-author network and identify the influential author.

With the calibrated influence of papers, a model based on PageRank is proposed in this paper to analyze the Erdos co-author network and measure the influence or impact of the authors. We use k-shell [45,46], betweenness centrality [18,47], and PageRank [48–51] to validate the result. We have made necessary improvements of the PageRank algorithm to combine the importance of the author itself and that of its community status. We use the time factor to obtain the calibrated importance of the authors. Then we propose a scheme of evaluating the contribution of each author in a paper. Then we use the time of co-authoring to calculate the *effective distance* between two authors, and evaluate the strength of their influence to each other with the law of gravity. The strength of influence is used to build a new network of authors, which is a comprehensive topological representation of both the quality of the node and its role in network. The PageRank algorithm is later applied to the transformed network, which gives a rank of importance of the authors.

2. Preliminary

2.1. Co-author network

Co-author network is a significant part of social networks. It has been widely used to identify the key researchers in a specific area and analyze the collaborations among the researchers. The sources of co-authorship the authors got usually from the bibliographic records [52], the classical dataset like Erdos co-author network [53,54] and extracting co-authorship from lots of papers from database, for example, in Ref. [8], Newman collected co-authorship from four databases: MEDLINE, the Los Alamos e-Print Archive, SPIRES, and NCSTRL.

An early example of co-author network is the network including about 511 mathematicians collaborate with the Erdos and their co-authorships [3]. Newman has constructed and analyzed the structure of the scientific collaboration networks in Refs. [6,7]. In Ref. [55], Stephen P. Borgatti identified the key nodes in network by removing the kp-set would result in a residual network with the least possible cohesion and identifying the kp-set which is maximally connected to all other nodes. According to Borgatti, the nodes of degree k are iteratively removed from the network until the remaining nodes all process a higher degree than k. The removed nodes are called the kp-set of a network. By decompensating the network into kp-sets, we can exclude peripheral nodes and identify the core of a network. And in Ref. [9], the authors proposed a new method based on PageRank algorithm called author rank to analyze co-author network in the digital library research community. Lots of methods have been constantly proposed to analyze the co-author network, but no single method is suited for all application, each one has it own specific utility.

2.2. Network analysis

Generally, there are three basic centrality metrics, namely degree centrality, closeness centrality and betweenness centrality [42]. Degree centrality reflects numbers of co-authors' collaborators, betweenness centrality represents different extents of co-authors' operationalization of centrality and closeness centrality describes how close a co-author is to others [56,57].

2.3. PageRank

PageRank is a method for rating web pages effectively by measuring the human interest and attention devoted to them [58]. It is the ranking mechanism at the heart of google [48]. In PageRank, a hyperlink is understood as an "endorsement" relationship. PageRank works by counting the number and quality of links to a page to determine a rough estimate