Information Processing and Management 000 (2016) 1–10

Contents lists available at ScienceDirect

Information Processing and Management

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

Uncovering mechanisms of co-authorship evolution by multirelations-based link prediction

Jinzhu Zhang

Department of Information Management, Nanjing University of Science and Technology, Nanjing, 210094, China

ARTICLE INFO

Article history:
Received 23 September 2015
Revised 20 May 2016
Accepted 15 June 2016
Available online xxx

Keywords:
Mechanisms
Co-authorship evolution
Multirelations-based link prediction
Heterogeneous network
Research collaboration

ABSTRACT

A single mechanism is insufficient for providing a comprehensive understanding of co-authorship formation and evolution because people choose to co-author with diverse motivations. The ways in which a hybrid mechanism jointly influences co-authorship evolution is not yet very clear, which leads to the following research questions: (1) how does each mechanism leverage with each other and how can multiple mechanisms be combined into the best hybrid mechanism? (2) How can the mechanisms be categorized into different groups and how does each group contribute to co-authorship evolution? This paper addresses these questions by using an improved meta-path based model called multirelations-based link prediction, which denotes every mechanism and their combinations as predictors in heterogeneous networks and quantitatively evaluates predictors via link prediction. Experiments are conducted in Library and Information Science (LIS). The result shows that the most appropriate mechanism is a hybrid mechanism denoted by a combination of predictors with weights. In addition, the contributions of different categorized mechanisms are compared, where the author-based mechanisms are more important than the keyword-based and journal-based mechanisms. The result also indicates that there is information loss when projecting from a heterogeneous bibliographic network to a homogeneous co-authorship network. Our study could add more predictive information into the model and apply the method in other types of heterogeneous networks in the future.

1. Introduction

As science becomes more collaborative, computational and data-intensive, academic researchers have been increasingly involved in multi-disciplinary collaborations (Austin, Hair, & Fullerton, 2012; Katz & Martin, 1997; Wagner, 2005). Research collaboration is a key mechanism for knowledge diffusion within research communities (Heinze & Kuhlmann, 2008), utilizing the expertise from multiple domains and producing high quality research outputs (Börner, Dall'Asta, Ke, & Vespignani, 2005; Egghe, Guns, & Rousseau, 2013; Katz & Martin, 1997).

To facilitate research collaboration, studies have explored several different approaches, among which co-authorship recommendation and prediction is one important line of work. These studies reveal that common interest plays an important role in research collaboration. As a result, scholars have experimented with two different methods to measure common interests: the network-based approach that measures common interests through their shared co-authors (Lü & Zhou, 2010a, 2010b; Liu, Zhang, Lü, & Zhou, 2011), and the content-based approach that measures common interests based on authors'

E-mail address: zhangjinzhu@njust.edu.cn

http://dx.doi.org/10.1016/j.ipm.2016.06.005

associated content, e.g., an author's keywords (Al Hasan, Chaoji, Salem, & Zaki, 2006). However, studies find that research collaboration can be influenced by numerous factors (Beaver & Rosen, 1978; Han, He, Brusilovsky, & Yue, 2013; Wang, Satuluri, & Parthasarathy, 2007). It is not yet clear how to best integrate and utilize multiple features of bibliographic information to form a hybrid mechanism of co-authorship evolution.

The mechanisms of co-authorship evolution are often studied in co-authorship networks, in which nodes represent authors and links indicate co-authorships (Barabási, et al., 2002; Liu, Bollen, Nelson, & Van de Sompel, 2005; Ortega, 2014; Yan & Ding, 2009). These studies denote each mechanism of co-authorship evolution as an indicator of similarity (Lü & Zhou, 2010a; Zhang, Xu, Zhu, & Zhou, 2015). However, the majority of these works study mechanisms in homogeneous networks, i.e., only one type of objects (author type) and one type of links (co-authorship) exist in the network. This cannot reveal the heterogeneity of real-world co-authorship networks, because researchers may collaborate with diverse motivations. Thus, studies have started to investigate collaboration mechanisms in heterogeneous networks. A heterogeneous network usually contains multiple types of objects (e.g., author, paper and keyword) and multiple types of links (e.g., "writes" and "is written by" relations between author and paper; "cites" and "is cited by" relations between papers) among these objects (Grčar, Trdin, & Lavrač, 2013; Sun & Han, 2013). These studies recognize the different paths among authors as mechanisms of co-authorship evolution in heterogeneous networks. However, this method should be applied and verified in other domains, and additional bibliographic information (e.g., keywords and journals) could be added to the network to form new mechanisms. Therefore, our study accounts for additional bibliographic information in order to build a heterogeneous network. This allows for better analyzing and increased understanding of multiple mechanisms of co-authorship evolution, especially regarding how a hybrid mechanism jointly influences co-authorship formation and evolution. Our study aims to address the following questions:

- To what extent does each mechanism contribute to co-authorship evolution, and how can multiple mechanisms be combined into a hybrid mechanism? Previous studies have mainly examined one pure mechanism to depict co-authorship network evolution, e.g., common neighbours, preferential attachment and random walks. This is usually insufficient because people collaborate with each other with different motivations. Therefore, studies mix different mechanisms in order to achieve better simulation, which proves that networks are likely to be driven by multiple mechanisms (Papadopoulos, Kitsak, Serrano, Boguñá, & Krioukov, 2012). However, quantitatively measuring the contribution of each mechanism in one fair platform and combining them into a best hybrid mechanism still need to be explored.
- How can the mechanisms be categorized into different groups, and how does each group contribute to co-authorship evolution? Current studies mainly focus on the differences among all mechanisms that influence co-authorship evolution, while in fact, the mechanisms could be categorized into groups via multiple perspectives. In addition, every group of mechanisms contributes differently to co-authorship evolution, and this needs to be clarified.

2. Literature review

2.1. Mechanisms of co-authorship evolution

Prior studies organize the motivations of research collaboration into different categories (Beaver & Rosen, 1978; Beaver, 2001), including access to facilities, skills, materials and expertise, efficiency to learn tacit knowledge, satisfaction of curiosity and intellectual interest, etc. These diverse collaboration motivations form multiple mechanisms of co-authorship evolution that encourage the generation of new collaboration. They are summarized as similar or related interests among researchers, which are often quantitatively measured in three ways: by common co-authors in a co-authorship network (Lü & Zhou, 2010a, 2010b; Liu, et al., 2011), by overlapping keywords between two authors' publications (Al Hasan, et al., 2006) and by shared references in a citation network (Ding, 2011). We will discuss the related methods and indices in homogeneous networks and heterogeneous networks, respectively.

For homogeneous networks, similar or related interests are usually computed on top of topological features, which can be further classified into node-dependent and path-dependent similarity indices. Typical node-dependent indicators are common neighbours (CN) (Lorrain & White, 1971), Jaccard coefficient (Hamers, et al., 1989), Adamic–Adar Index (Adamic & Adar, 2003), Preferential Attachment (Jeong, Néda, & Barabási, 2003) and Resource Allocation (RA) (Zhou, Lü, & Zhang, 2009). Typical path-dependent indicators are Katz Index (Katz, 1953), Leicht–Holme–Newman Index (Leicht, Holme, & Newman, 2006) and SimRank (Jeh & Widom, 2002). Studies have examined and compared these similarity indicators for predicting and recommending research collaborations (Guns & Rousseau, 2014; Yan & Guns, 2014; Zhao, et al., 2015). In our paper, we choose two representative indicators in each group, i.e., CN as a node-dependent indicator and Katz as a path-dependent indicator for our later experiments, as well as RA and SimRank. We choose these because of their simple computation and relatively better performances.

For heterogeneous networks, Davis, Lichtenwalter, and Chawla (2011) introduce a probabilistically weighted extension of the Adamic/Adar. It demonstrates the potential benefits of incorporating diverse relations among authors, particularly in the cases where homogeneous relationships are sparse. However, multiple relationships among different objects in heterogeneous networks have not been systematically represented and compared. Later, Sun et al. defined the relationships between authors by the different paths from one author to another (Sun, Barber, Gupta, Aggarwal, & Han, 2011; Sun & Han, 2013), i.e., the meta path. They call this model the meta-path based model. Experiments on the DBLP bibliographic network show that

Please cite this article as: J. Zhang, Uncovering mechanisms of co-authorship evolution by multirelations-based link prediction, Information Processing and Management (2016), http://dx.doi.org/10.1016/j.ipm.2016.06.005