Journal of Informetrics 8 (2014) 295–309

Contents lists available at ScienceDirect

Journal of Informetrics

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

Predicting and recommending collaborations: An author-, institution-, and country-level analysis

Erjia Yan[a,*], Raf Guns[b]

[a] College of Computing and Informatics, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104, USA
[b] University of Antwerp, Institute of Education and Information Sciences, IBW, Venusstraat 35, 2000 Antwerpen, Belgium

ARTICLE INFO

Article history:
Received 13 September 2013
Received in revised form 8 January 2014
Accepted 13 January 2014
Available online 4 February 2014

Keywords:
Collaboration
Link prediction
Coauthorship
Networks
Dynamics

ABSTRACT

This study examines collaboration dynamics with the goal to predict and recommend collaborations starting from the current topology. Author-, institution-, and country-level collaboration networks are constructed using a ten-year data set on library and information science publications. Different statistical approaches are applied to these collaboration networks. The study shows that, for the employed data set in particular, higher-level collaboration networks (i.e., country-level collaboration networks) tend to yield more accurate prediction outcomes than lower-level ones (i.e., institution- and author-level collaboration networks). Based on the recommended collaborations of the data set, this study finds that neighbor-information-based approaches are more clustered on a 2-D multidimensional scaling map than topology-based ones. Limitations of the applied approaches on sparse collaboration networks are also discussed.

1. Introduction

Social networks have the propensity to evolve over time. Every second, new friendships are established and old friendships are updated in Facebook connections, new collaborations are formed and populated in academic databases, and Twitter follower-following relationships are constantly subject to changes and updates. To capture such evolving features, earlier studies mainly employed a macro-perspective to model network growths and simulate network behaviors (e.g., Albert & Barabási, 2000; Barrat, Barthélemy, & Vespignani, 2004; Jeong, Néda, & Barabási, 2003; Sakaki, Okazaki, & Matsuo, 2010). Later on, studies that focused on individuals' growth patterns and behaviors in social networks were also introduced (e.g., Kretschmer, 2004; Liu, Bollen, Nelson, & Van de Sompel, 2005; Yan & Ding, 2009). These micro-level analyses have complemented the scholarship of social network analysis. They are specialized in examining individuals' power, stratification, ranking, and inequality in various sociological settings (Wasserman & Faust, 1994).

Micro-level analyses have been one of the foci in informetric research. Studies in this field have typically employed authors, research communities, and institutions as the unit of analysis. Informetric studies have applied various indicators to collaboration networks. These studies have revealed the most "central" authors through centrality measures (e.g., Fiala, Rousselot, & Ježek, 2008; Liu et al., 2005; Yan & Ding, 2011; Yin, Kretschmer, Hanneman, & Liu, 2006), identified factors that are associated with collaboration and citation (e.g., Yan & Sugimoto, 2011), and examined the relationship between geographic location and collaboration (e.g., Ponds, Van Oort, & Frenken, 2007). However, these studies mainly used static

* Corresponding author.
E-mail addresses: erjia.yan@drexel.edu (E. Yan), raf.guns@uantwerpen.be (R. Guns).

1751-1577/$ – see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.joi.2014.01.008

approaches, and consequently did not inform the dynamic characteristics of collaborations. The goal of this study is to fill this gap by probing into collaboration dynamics using a ten-year data set on library and information science publications.

Specifically, we aim to predict and recommend collaborations based on the structure of current collaboration networks. This topology-based prediction is also known as link prediction (Liben-Nowell & Kleinberg, 2007). Link prediction recommends collaborations purely based on the intrinsic collaboration topology. This method does not rely on any data concerning the complex social, cognitive, institutional, or geographical factors (e.g., Ponds et al., 2007; Yan & Sugimoto, 2011). These factors are indirectly accounted for, because they may influence the network topology through mechanisms like homophily or the Matthew effect.

The performance of link predictors determines the effectiveness of collaboration recommendations. In the past, various link predictors were proposed and applied (e.g., Guns & Rousseau, 2013; Liben-Nowell & Kleinberg, 2007; Sharan & Neville, 2008). These studies, however, focused largely on author collaborations. Consequently, we have limited understanding of collaboration dynamics of other major collaborative entities, such as institutions and countries. These collaborative entities should not be neglected. Rather, they should be systematically examined. They deliver unique perspectives to examine collaborations that author-level analysis may be inadequate to afford. For instance, institutions can be used as proxies to delve into authors' collective collaboration behaviors (Hoekman, Frenken, & van Oort, 2009). Country-level collaboration analysis can provide "a tool for high-level scrutiny of the quality and quantity of the research enterprise" (Holton, 1978, p. 200). Both institution- and country-level analysis can signify spatial-temporal discoveries of knowledge production and innovation (Havermann, Heinz, & Kretschmer, 2006; Yan & Sugimoto, 2011).

This study is thus motivated to further our understanding of collaboration dynamics. It investigates collaboration prediction and recommendation at author-, institution-, and country-levels. Through the application of several link predictors, the following research questions are addressed:

- To what extent do different levels of aggregation, i.e., author-, institution-, or country-level, affect the performance of link predictors? Previous studies have mainly examined the dynamics of author collaborations. A systematic analysis of all three levels of collaboration (i.e., author, institution, and country) has not yet been carried out. To fill this gap, the current study conducts an integrated examination of collaboration dynamics at the levels of authors, institutions and countries using link prediction methods.
- Based on true/false positive and true/false negative statistics, what are the between-object distances of different link predictors? A set of evaluation methods (i.e., precision–recall, receiver operating characteristic (ROC) curve, area under the ROC curve (AUC), and top k evaluation) is triangulated to ensure the highest level of validity. Specifically, the current study also uses multidimensional scaling to visualize the between-object distances among eight predictors on a two-dimensional map.
- Starting from past collaboration relations in library and information science, what new collaborations are most probable to establish at author-, institution-, and country-levels? And what approach can be used to integrate the recommended collaborations obtained from multiple link predictors? Previous efforts on link prediction mainly relied on one predictor to recommend collaborations (e.g., Guns, 2009, 2011). Nonetheless, different link predictors may capture different collaboration characteristics. We propose a straightforward approach to merge the collaboration recommendations obtained from different predictors.

Note that these research questions are addressed using a data set on library and information science publications. Thus, findings of this study do not necessarily generalize to other research fields. Nevertheless, this study should inform dynamic analyses of collaborations in general and assist scholars trying to discern collaboration characteristics at different collaboration levels. This study also contributes to micro-level informetric research by providing ways to assess individuals' collaboration potentials.

2. Literature review

2.1. Collaboration networks

Studies of collaboration networks have a long standing in information science. Collaboration networks furnish an important medium to examine scholarly communication (e.g., Logan & Shaw, 1991; Luukkonen, Persson, & Sivertsen, 1992). Although early studies of coauthorship networks helped information scientists gain an in-depth understanding of the socio-cognitive structure of several author communities, these studies were limited to a small scale and the employed approaches were largely constrained to descriptive statistics.

In the last decade, we have witnessed a new movement in network analysis. The focus has shifted to large-scale statistical properties of graphs (Newman, 2001a, 2001b). In particular, the discoveries of small-world (Watts & Strogatz, 1998) and scale-free (Barabási & Albert, 1999) properties have promoted studies of collaboration networks. Recently, collaboration networks have been used to evaluate various clustering techniques, such as modularity-based techniques (Newman & Girvan, 2004), Clique Percolation Method (Farkas, Ábel, Palla, & Vicsek, 2007), link communities (Ahn, Bagrow, & Lehmann, 2010), and community kernel (Wang, Lou, Tang, & Hopcroft, 2011). These meso-level techniques have reshaped the research landscape of scientific collaboration and have propelled its analysis toward a more granular level. They have provided insights