Journal of Informetrics 10 (2016) 299–311

Contents lists available at ScienceDirect

Journal of Informetrics

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

A geometric graph model for coauthorship networks

Zheng Xie[a,b,*], Zhenzheng Ouyang[a], Jianping Li[a]

[a] College of Science, National University of Defense Technology, Changsha 410073, China
[b] Centre for Networks and Collective Behaviour, Department of Mathematics, University of Bath, Bath BA2 7AY, UK

ARTICLE INFO

Article history:
Received 8 July 2015
Accepted 1 February 2016
Available online 11 February 2016

Keywords:
Coauthorship network
Modelling
Geometric graph
Hypergraph

ABSTRACT

Modeling coauthorship networks helps to understand the emergence and propagation of thoughts in academic society. A random geometric graph is proposed to model coauthorship networks, the connection mechanism of which expresses the effects of the academic influences and homophily of authors, and the collaborations between research teams. Our analysis reveals that the modeled networks have a range of features of empirical coauthorship networks, namely, the degree distribution made up of a mixture Poisson distribution with a power-law tail, clear community structure, small-world, high clustering, and degree assortativity. Moreover, the underlying formulae of the tail and forepart of the degree distribution, and the tail of the scaling relation between local clustering coefficient and degree are derived for the modeled networks, and are also applicable to the empirical networks.

1. Introduction

Coauthorship networks, an important topic of informetrics, are constructed from articles, using the authors as nodes, and connecting them if they have coauthored. Modeling these networks provides a window on understanding the patterns of collaborations, the emergence and propagation of thoughts in academic society, etc. Many of the empirically observed coauthorship networks are found to be scale-free, small-world, high clustering, and degree assortative (Newman, 2001a, 2001b, 2001c, 2002, 2004). The scale-free property can be modeled by the mechanism of preferential attachment (Barabási et al., 2002; Moody, 2004; Perc, 2010; Tomassini & Luthi, 2007; Wagner & Leydesdorff, 2005; Zhou et al., 2007). The degree assortativity can be modeled by connecting two existing non-connected nodes according to a probability that is a decreasing function of the difference between degrees (Catanzaro, Caldarelli, & Pietronero, 2004).

Besides the preferential attachment, the inhomogeneity of node degree in real networks also arises as a consequence of the inhomogeneous influences of nodes, through which some nodes gain more connections because they are likely to have wider influences than others. A typical example of this phenomenon is the academic influences of researchers. The theory of random geometric graphs (RGG) enables research into complex networks via geometry (Krioukov et al., 2012; Penrose, 2003; Xie, Ouyang, Zhang, Yi, & Kong, 2015; Xie, Zhu, Kong, & Li, 2015). A disassortative graph model, called scale-invariant RGG, models the inhomogeneous influences of nodes found in web-graphs (Xie & Rogers, 2015). In that model, the propensity of nodes to connect only to those that are "similar" (in the sense of spatial location) mirrors the homophily frequently observed in social systems (Papadopoulos, Kitsak, Serrano, Boguná, & Krioukov, 2012; Simşek & Jensen, 2008), e.g. coauthorship networks. The homophily of authors (in the sense of research topics, geographical distance, etc.) is found to have a positive influence on collaborations (Hoekman, Frenken, & Tijssen, 2010).

* Corresponding author at: College of Science, National University of Defense Technology, Changsha 410073, China.
E-mail address: xiezheng81@nudt.edu.cn (Z. Xie).

http://dx.doi.org/10.1016/j.joi.2016.02.001

Analyzing the origins of coauthorship networks helps to understand the aspects of those networks. Collaborations happen in or between scientific research teams. Researchers who appear as authors of an article (who together could be regarded as an "article team") represent "a visible and easily quantifiable manifestation of the collaborations" (Milojević, 2014). Coauthorship networks come from the article teams, so they should have some intrinsic relations. Hence, from a reductionist stance, a sound model of coauthorship networks should capture some basic properties of article teams, e.g. team size distribution.

A geometric graph model built on a circle is proposed here, which generates hypergraphs and then networks. As usual, nodes are created by a unit intensity Poisson process. The novel aspect is the connection mechanism of the model, which expresses the effects of the academic influences and homophily of authors on collaborations, and the communications between research teams. This mechanism gives the modeled networks (which are simple graphs obtained from the modeled hypergraphs) a range of properties found in coauthorship networks, viz. the degree distribution made up of a mixture Poisson distribution with a power-law tail, clear community structure, small-world, high clustering, and degree assortativity. Meanwhile, the hyperedge cardinality distributions of the modeled hypergraphs are similar to those of the empirical data.

The proposed model is suitable for modeling the coauthorship networks, the article team size distributions of which have hooks with peak values larger than two. Although it cannot exactly reproduce some behaviors of the empirical networks without such team size distributions, the model captures their main features. To show this, we derive the underlying formulae of the tail and forepart of the degree distribution, and the tail of the scaling relation between local clustering coefficient and degree for the modeled networks. The applicability of those formulae to the empirical data, together with the ability to capture the above addressed features, hint at the potential that this model has given a new geometric view for understanding the aspects of coauthorship networks.

This report is organized as follows: the model and empirical data are described in Sections 2 and 3 respectively; the degree distribution, clustering, assortativity, and small world property of the modeled networks are analyzed in Sections 4–7; and the conclusion is drawn in Section 8.

2. The model

The coauthorships of an article set can be represented as a hypergraph by regarding authors as nodes, and article teams as hyperedges (Estrada & Rodrŕguez-Velázquez, 2006). We first lay out a hypergraph model, the constructing mechanism of which considers certain factors engendering collaborations, viz. academic influences, homophily, and the communications between research teams. We next generate simple graphs from the modeled hypergraphs by treating hyperedges as cliques (connecting every two nodes in each hyperedge), removing isolated nodes, and combining the edges between the same pair of nodes. The coauthorship networks are also generated in this way. The proposed model is built on a circle, so it is literally a modified RGG.

We require certain properties of the modeled networks, such as the distributions of hyperedge cardinalities and degrees, the scaling relation between local clustering coefficient/average degree of neighbors and degree, to be similar to those of the empirical data. Moreover, some properties of the empirical data differ from field to field, e.g. the team size distribution of certain mathematical articles and that of certain biological articles (Fig. 1). We require the proposed model to be universal and not field dependent at certain levels. To achieve these requirements, we introduce many parameters, which makes the model relatively complex. The motivation of each used parameter will be explained.

The distributions of the article team sizes of the empirical data approximately follow a generalized Poisson distribution with a power-law tail, and the Poisson parts are predominant (Fig. 1). Express the probability mass function (PMF) of such distribution by $f(x) = qf_1(x) + (1 - q)f_2(x)$, where $q \in [0, 1]$, $f_1(x) = a(a + bx)^{x-1}e^{-a-bx}/x\,!$, $f_2(x) = cx^d$ are the PMFs of the distributions of generalized Poisson and power-law respectively, $a, b, c, d \in \mathbb{R}$, and $x \in \mathbb{R}^+$.

In reality, articles are produced by research teams. Normally, each team has a leader who is responsible for putting together the team and producing articles as a corresponding author with some team members. The set of researchers who appear as authors of an article forms an article team.

The proposed hypergraph model consists of "authors" (nodes) and their "articles" (hyperedges). Articles always have at least two authors. Each hyperedge forms an "article team". Some nodes are selected to attach to specific zones to express their academic influences. Those nodes are regarded as "leaders" (Fig. 2). The set of nodes in a zone is regarded as the "research team" of the corresponding leader, and each team member coauthors with the leader.

Suppose a modeled hypergraph has N_1 "potential authors", N_2 leaders with the scientific age $T - t + 1$ ($t = 1, 2, \ldots, T$) and N_3 articles between research teams, where $N_1, N_2, N_3, T \in \mathbb{Z}^+$. In reality, the leaders are a small part of researchers (potential authors). So, in the model, the total number of leaders TN_2 is supposed to be far less than N_1. The modeled hypergraph is constructed by following four processes. The first three are

Process 0. Sprinkle nodes:
Sprinkle N_1 nodes uniformly and randomly on a circle S^1;