Physica A 456 (2016) 167–175

Contents lists available at ScienceDirect

Physica A

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

A geometric graph model for citation networks of exponentially growing scientific papers

Zheng Xie [a,b,*], Zhenzheng Ouyang [a], Qi Liu [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

HIGHLIGHTS

- A geometric graph is proposed to model the citation networks of exponentially growing papers.
- The model expresses certain factors engendering citations, e.g. the relativity of contents.
- The model predicts certain features of the citation networks, e.g. in-degree assortativity.

ARTICLE INFO

Article history:
Received 22 August 2015
Received in revised form 6 January 2016
Available online 1 April 2016

Keywords:
Geometric graph
Citation network
Assortativity
Modelling
Bibliometric
Causal network

ABSTRACT

In citation networks, the content relativity of papers is a precondition of engendering citations, which is hard to model by a topological graph. A geometric graph is proposed to predict some features of the citation networks with exponentially growing papers, which addresses the precondition by using coordinates of nodes to model the research contents of papers, and geometric distances between nodes to diversities of research contents between papers. Citations between modeled papers are drawn according to a geometric rule, which addresses the precondition as well as some other factors engendering citations, namely academic influences of papers, aging of those influences, and incomplete copying of references. Instead of cumulative advantage of degree, the model illustrates that the scale-free property of modeled networks arises from the inhomogeneous academic influences of modeled papers. The model can also reproduce some other statistical features of citation networks, e.g. in- and out-assortativities, which show the model provides a suitable tool to understand some aspects of citation networks by geometry.

1. Introduction

Citation networks constructed from scientific papers are important research objects of scientometrics, in which each node represents a paper, and each edge represents a citation of one paper by another. Modeling these networks provides a window on understanding hot topics of research, the emergence and propagation of academic thoughts in scientific society, etc. [1–4]. Many of the empirically observed citation networks are found to be scale-free (their in-degree distributions have a power-law tail), clustering, and assortative (in the sense of in- and out-degrees respectively). Seeking mechanisms to illustrate one or more of those properties has attracted extensive attention [5,6].

There have existed several important studies of the scale-free property of citation networks. Price noted the "cumulative advantage" of citation behavior: highly cited scientific papers accumulate additional citations more quickly than papers that

* 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.physa.2016.03.018
0378-4371/

have fewer citations. He abstractly expressed this phenomenon by a rule: the probability that a paper receives a citation is proportional to the number of citations it has received, which successfully predicts the scale-free property [7–9]. In network science, cumulative advantage is also called preferential attachment. Price model has been generalized to illustrate other properties of citation networks in various contexts [10–13], e.g. Goldberg et al. set the number of citations given by new papers to be random variables drawn from a lognormal distribution, which fits the out-degree distributions of empirical data well (the model A in Ref. [14]).

It is empirically observed that the probability for a paper to get cited decreases as its age increases, which is called aging phenomenon of citation behavior. Some models introduced time decay to the cumulative advantage, namely the probability of an existing paper to be cited is proportional to its current in-degree multiplied by a decay factor dependent on its age [15,16], e.g. exponential decay factor (the model B in Ref. [14]). Aging makes citation bursts typically occur in the early life of a paper. Eom et al. generalized the Price model to simulate the bursts, in which a new paper cites a constant number of existing papers by a linear preferential attachment with time dependent initial attractiveness [17].

Empirical data have a positive clustering coefficient, which is zero in theory for the networks generated by aforementioned models. Krapivsky et al. noted that the authors of a new paper may not only cite a paper, but also could cite some references of the cited paper. They called this phenomenon copying, and mimicked it by a rule: a new node connects to a randomly selected node, as well as all the ancestors of the selected node, which successfully predicts the scale-free property of citation networks, and clustering as well [18]. In reality, copying is incomplete: a paper unlikely cites all references of the papers it cited. In the misprints propagation model [19] and the model C in Ref. [14], incomplete copying is realized by a one-step random walk from a cited paper. Incomplete copying can also be added to the cumulative advantage with time decay to address the scale-free, clustering and aging simultaneously [20].

The exponents of in-degree distributions of citation networks vary from data to data, but the predicted power-law exponents (if provided) given by all above models are fixed. Peterson et al. proposed a model to address this problem, which involves a direct mechanism: a new paper cites an old paper randomly; an indirect mechanism, which is a kind of incomplete copying [21]. The model can generate networks with similar full in-degree distributions of empirical citation networks (the exponents of in-degree distributions of which can be tuned) namely not only similar on the power-law tails but also on the foreparts. The indirect mechanism makes the modeled networks have a positive clustering coefficient as well.

The theory of random geometric graphs (RGGs) enables research into networks via geometry [22–26]. The nodes of some RGGs are points chosen at random in the space time, e.g. through a Poisson point process, and they are connected by edges if they are causally connected [27]. The causal relationship is induced by light cones in the space-time. Meanwhile, the idea of a paper is inspired by its references at certain levels, so citation behavior could be regarded as a causal relationship. We have proposed a RGG built on a cluster of concentric circles (so called it CC model) for citation networks [28], in which the influences of modeled papers are expressed by geometric zones liking light cones, and a paper i cites a paper j if the influential zone of i contains j. The model can capture the scale-free and clustering properties of empirical networks, but has a range of shortcomings, e.g. the out-degree distributions of empirical data cannot be well simulated.

Besides the causal property of citations, using RGGs can also illustrate an important precondition of engendering citations, namely the content relativity of papers, by spatial coordinates of nodes: diversities of research contents between papers are illustrated by geometric distances between nodes. Here, we continue to use RGG built on the space time of the CC model to predict certain statistical features of the citation networks with exponentially growing papers, and address some shortcomings of the CC model as well. We seek a geometric mechanism to simultaneously express certain factors engendering citations, namely academic influences, the aging of those influences, relativity of contents, and incomplete copying of references. The model shows, besides the cumulative advantage, the scale-free property of citation networks can also be explained as a consequence of the inhomogeneous academic influences of scientific papers, through which some papers gain more citations because they are likely to have wider influences than others. We also examined how the model predicts some other statistical features of empirical networks, namely the out-degree distribution, the scaling relation between local clustering coefficients and in-degrees, assortativities for in- and out-degrees.

This report is organized as follows. The model is described in Section 2. The degree distributions, clustering and assortativity of the modeled networks are analyzed in Sections 3–5 respectively. The conclusion is drawn in Section 6.

2. The model

Normally, empirical citation networks of scientific papers are directed acyclic graphs (DAGs): only newer papers can cite older papers. However, some preprinting papers would cite each other, which happens rarely. The geometric DAG proposed here consists of "papers" (nodes) and "citations" (edges) between those papers. In some citation networks, the annual numbers of papers grow exponentially, e.g. the citation network DBLP 2013-09-29 (Table 1) collected by Tang et al. for the papers in DBLP dataset, which are published in the period from 1936-01-01 to 2013-09-29 [29] (Fig. 1(a)). We focus on simulating the exponentially growing case and compare some properties of a network generated by our model to those of DBLP 2013-09-29.

In our model, the nodes are sprinkled on a cluster of concentric circles in a $(2+1)$-dimensional spacetime with circumference polar coordinates $\{r, \theta, t\}$ (Fig. 2). The angular coordinates of nodes could be regarded as research contents of papers. So diversities of research contents between papers could be abstractly expressed by geometric distances between