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

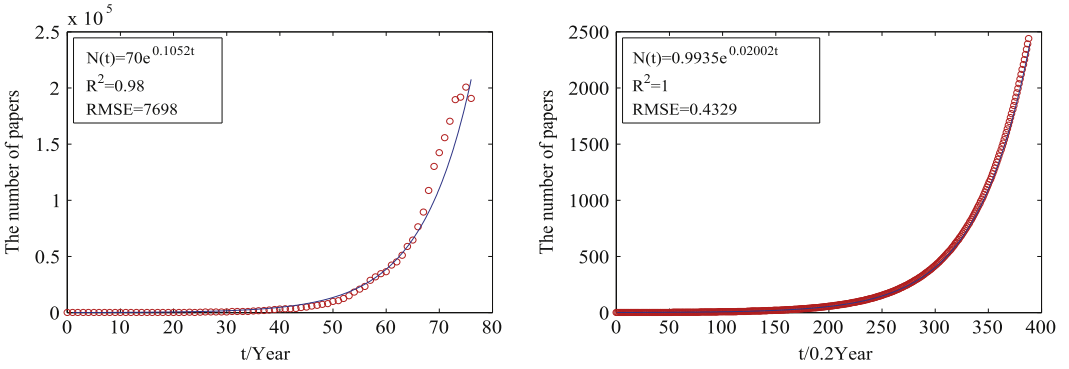
Normally, empirical citation networks of scientific papers are directed acyclic graphs (DAGs): only newer papers can cite older papers. However, some preprinting papers would cite each other, which happens rarely. The geometric DAG proposed here consists of "papers" (nodes) and "citations" (edges) between 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

Table 1
Certain statistical indicators of the analyzed citation networks.

Network	Nodes	Edges	CC	In-AC	Out-AC	PG	MO
DBLP 2013-09-29	2,084,055	2,244,018	0.070	0.036	0.093	0.973	0.771
Cit-HepTh	27,770	352,807	0.165	0.041	0.096	0.987	0.648
Cit-HepPh	34,546	421,578	0.148	0.077	0.111	0.995	0.725
Modeled network	123,008	171,891	0.390	0.120	0.290	0.452	0.992

The indicators are clustering coefficient (CC), in- and out-assortativity coefficients (In-AC, Out-AC), the node proportion of the giant component in the total (PG), and modularity (MO). The first network comes from the papers published before 2013-09-29 in DBLP dataset. The next two networks come from the papers (which are published in the period from 1993-01 to 2003-04) of arXiv in high energy physics phenomenology and in high energy physics theory respectively [30]. The fourth network is generated by our model, the parameters of which are $\alpha = 0.8$, $\beta = 0.05$, $\delta = 5$, $l = 0.02$, $m = 1$, $p = 1$, $q = 0.5$, $r = 0.002$, $T = 390$ and $f(k) \sim k^{-2.5}$, $k \in [3, 60]$.



(a) DBLP 1936–2012. (b) Modeled network.

Fig. 1. Evolutionary trends of annual numbers of papers. Panel (a) shows the trend for the papers of DBLP published in 1936–2012, and Panel (b) for the modeled network in Table 1. The coefficient of determination (R^2) and root mean squared error (RMSE) are used to measure the goodness of fits.

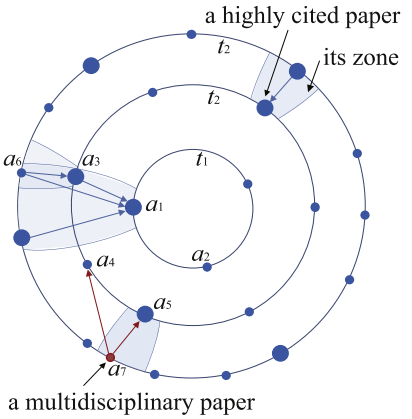


Fig. 2. Illustration of the model. The large nodes illustrate highly cited papers, and their influential zones are represented by light blue areas. Different sizes of academic influences are expressed by different zonal sizes. It demonstrates the core and complement mechanisms of citations respectively: if node a_5 belongs to the zones of nodes a_1 and a_3, then it cites them; if node a_6 is a multidisciplinary paper, it will not only cite the highly cited papers which have zones cover a_6 (e.g. a_5) but also cite a number of existing papers (e.g. a_4) to make its out-degree to be a random variable drawn from a given power-law distribution. The numbers of citations received by newer nodes could be larger than those of older ones, e.g. the times cited of a_4 are larger than that of a_2.

nodes. The time t is related to physical time, which can be explained as the tth unit of time, such as tth week, tth month, etc. Some nodes are selected to attach to specific zones to express their academic influences, which are called "highly cited papers" here. Meanwhile, some nodes are selected to be "multidisciplinary papers", the out-degree of which is drawn from a power-law distribution.

Suppose a modeled network has $m[e^{lt}]$ papers published in the tth unit of time ($t = 1, 2, \ldots, T$), in which some are highly cited papers and some are multidisciplinary papers, where $l \in \mathbb{R}^+$, $m \in \mathbb{Z}^+$, and $[\cdot]$ is the rounding function. For time $t = 1, 2, \ldots, T \in \mathbb{Z}^+$, the modeled papers and citations are created by following steps.

Step 1. Generate a circle $C_t = ([e^{lt}], \theta, t)$, sprinkle $m[e^{lt}]$ nodes (which are regarded as papers) on C_t randomly and uniformly (namely according to a Poisson point process), and fix nodes with their coordinates, e.g. i with (θ_i, t_i);

Step 2. Select q percent new papers randomly as potential highly cited papers to attach to specific zones: the zone of a such paper i is defined as an interval of angular coordinate with center θ_i and arc-length $\beta(\theta_i)e^{-\alpha l t_i-(1-\alpha)l t_c}$, where t_c is the current time, $\alpha \in (0, 1)$, and β is a positive and piecewise-constant function;

Step 3. Each new paper j cites with a probability p to the potential highly cited papers that have zones covering j and are published δ earlier than j, where $\delta \in \mathbb{Z}^+$;

Step 4. Select r percent new papers randomly as multidisciplinary papers to continually cite a number of existing papers randomly to make the lengths of references (out-degrees) of those papers to be random variables drawn from a power-law distribution $f(k)$.

The parameters p, q and r are properly chosen to make the number of citations generated in Step 3 to be the core while that in Step 4 to be the complement. There are some intuitive explanations for the formula of zonal sizes. First, if two highly cited papers discuss about the same content, the cumulative advantage would make the older one receive more citations than the newer one, so the formula is reasonable to consider to be a decreasing function of t. Second, the number of citations may differ in research contents due to their inhomogeneous attractiveness, so $\beta(\cdot)$ is introduced to the formula. Third, aging of the influences of papers occurs such that the probability of a paper getting cited decreases with the growth of its age, so it is reasonable to consider the formula to be a decreasing function of t_c.

Our model is developed to address following points of particular interest to us, which are the reasons for why we introduce so many parameters. The details about how our model gives a solution to those points are shown in following sections. We discuss how some typical models deal with those points as well.

Firstly, almost existing models focus on power-law tails, e.g. the rules of linear preferential attachment and complete copying predict power-law in-degree distributions with a fixed exponent. We are interested here in more features about the in- and out-degree distributions of empirical data, namely the hook heads in the out-degree distributions and power-law tails (the exponents of which vary from data to data) in both of the in- and out-degree distributions. Through calculations and numerical simulations in Section 4, we reveal the modeled networks capture those features. Especially, the exponents of the predicted in-degree distributions can be tuned by the parameter α in the formula of zonal sizes.

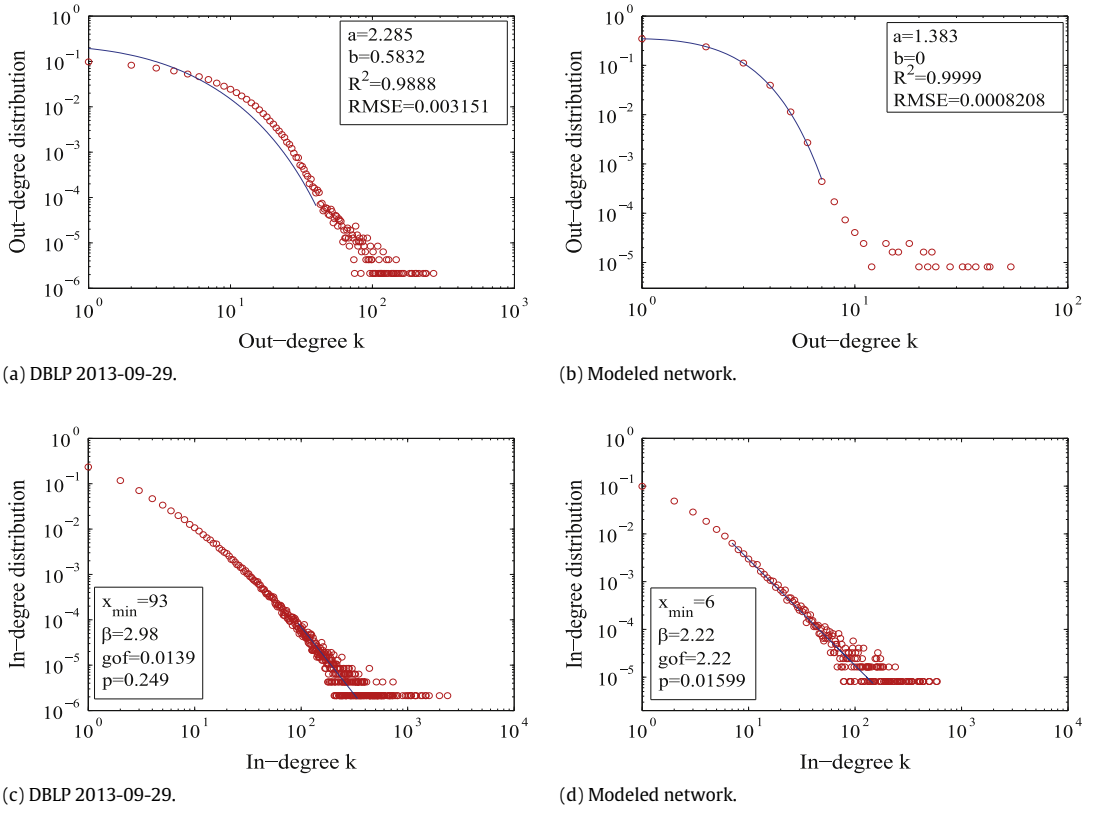
Secondly, the global clustering coefficient (GCC) is the fraction of connected triples of nodes which also form "triangles". Copying citations is a sound explanation for why citation networks have a positive GCC. The values of GCC differ from an empirical network to another (Table 1), which cannot be simulated by the linear attachment (too low), the complete copying (too high), and the CC model (non-tunable). Aging of the influences of papers is a reasonable factor for why the GCCs of empirical data are lower than those predicted by the complete copying, because old papers hardly receive citations. So incomplete copying is more realistic, which is expressed by different ways in some models [20], e.g. the model C in Ref. [14]. The GCCs predicted by those models can be tuned by the proportion of citations generated through incomplete copying. Our model gives a sympathetical expression of aging (expressed by the decreasing sizes of influential zones with the growth of time) and incomplete copying (realized by the parameter p in Step 3), the detail of which is shown in Section 5.

Thirdly, the empirical citation networks are in- and out-assortative. Simulations show that the first property can be reproduced by linear preferential attachment, and both by complete copying, the reasons of which could be inferred from their mechanisms. The proposed model can also reproduce those properties, the reason of which is explained by using analytical formulae in Section 6.

Particularly, except the non-tunable global clustering coefficient, the CC model has some other obvious shortcomings, e.g. the out-degree distributions of some empirical data have a short power-law tail. Now the multidisciplinary papers make the out-degree distributions of modeled networks also have a short power-law tail, namely the function $f(k)$ in Step 4. In reality, some newly published papers can also gain more citations than some older papers that discuss about the same content, which cannot happen in the CC model. Due to the selection in Step 2, a paper published at the earliest times could not be a highly cited paper, and a new paper could receive many citations if it is selected to be a highly cited one.

Simulations show modeled networks have giant components (in sense of weak, namely ignore the directions of edges) and clear community structures as the empirical networks do (Table 1). In fact, the papers in the same zone are very likely to belong to the same community, and the zones are loosely connected, or not connected at all (the level of connection can be tuned by the parameter r and the range of $f(k)$ in Step 4) by few multidisciplinary papers. This causes the edges within communities to be significantly more than that between communities. Therefore, there is a reciprocal relationship between the clearness of communities and the node proportion of the giant component in the total.

The fitting function of the annual number of papers in DBLP 2013-09-29 is $N(t) = 70e^{0.1052t}$ (Fig. 1(a)). To make the modeled time span match with, and the annual number of modeled papers proportional to those of the empirical data respectively, we set the unit of the model parameter t to be $l/0.1052 \approx 0.2$ year, which makes the modeled time span to be $Tl/0.1052 \approx 78$ years (Fig. 1(b)). Since our model generates networks by a global selection rule, it is hard to generate a network with a size comparable to DBLP 2013-09-29, which needs $m = 14$. We set $m = 1$ for the modeled network in Table 1, namely there is only one paper in the beginning (because $[e^{0.02}] = 1$) which happens in many classical citation network models, e.g. Price model. We further set other parameters to make the modeled average degree similar to that of the empirical data, and assume that the papers published apart less than a year (namely $\delta = 5$) cannot cite each other.



(a) DBLP 2013-09-29.

(b) Modeled network.

(c) DBLP 2013-09-29.

(d) Modeled network.

Fig. 3. Out- and in-degree distributions. The fitting functions are the generalized Poisson distribution $p(k) = a(a+bk)^{k-1}e^{-a-bk}/k!$ for the hooks in Panels (a,b), and the power-law distribution $p(k) = k^{-\beta}/\sum_{l=0}^{\infty}(l+x_{\min})^{-\beta}$ for the tails in Panels (c,d) (which is fitted by the method in Ref. [32]).

3. Out- and in-degree distributions

The out-degree distributions of empirical citation networks exhibit power-law tails and "hooks" in the foreparts [25] (Fig. 3(a)). The hooks can be adequately fitted by generalized Poisson distributions, which hint at the potential that the predominant citation behaviors are governed by generalized Poisson processes. Many events in real world (in which the probability of occurrence of a single event is small and is affected by previous occurrences) come from generalized Poisson processes [31]. Citation behavior can be viewed as a low-rate event (its realization selects few papers out of a great many papers) and the probability of occurrence of a citation could be affected by previous occurrences (e.g. a citation generated by copying). Now we drive the underlying formulae of the head and tail of the modeled out-degree distribution to show how our model generates a similar hook and power-law tail (Fig. 3(b)).

We initially consider the out-degrees of the non-multidisciplinary papers, and then consider those of the multidisciplinary papers. Suppose a paper i is a non-multidisciplinary paper. If i belongs to the influential zone of j, then we have $\Delta(\theta_i, \theta_j) < \beta(\theta_j)e^{-l(\alpha t_j+(1-\alpha t_i))}$ and $t_j < t_i - \delta$. When $\beta(\theta_j)e^{-l(\alpha t_j+(1-\alpha t_i))}$ is small enough, we have $\beta(\theta_j) \approx \beta(\theta_i)$, because $\beta(\cdot)$ is a piecewise constant function. Hence the expected out-degree $k^+(\theta_i, t_i)$ of i is approximately equal to

$$k^+(\theta_i, t_i) = \frac{m}{2\pi}\sum_{t_j=1}^{t_i-\delta}\frac{\beta(\theta_i)pq}{e^{(\alpha t_j+(1-\alpha)t_i)l}} \times [e^{lt_j}] \approx \frac{\beta(\theta_i)mpq}{2\pi}\frac{e^{-(1-\alpha)l} - e^{-(1-\alpha)l(t_i-\delta)}}{(1-\alpha)l}, \tag{1}$$

which is not a constant but increases with the growth of t_i. It is observed that the annual average length of a paper's references is a monotone increasing sequence for some journals [4,28]. So the increasing property of Eq. (1) is reasonable. When $t_i \gg 1$, we can take the expected out-degree $k^+(\theta_i, t_i)$ to be independent of t_i and write $k^+(\theta_i)$ instead

$$k^+(\theta_i) = \frac{\beta(\theta_i)mpq}{2\pi(1-\alpha)l}e^{-(1-\alpha)l}, \tag{2}$$

which holds for the majority papers, because the annual number of papers grows exponentially. It is reasonable to consider that the number of a paper's references cannot grow to infinity and should have an upper bound. Hence, the result given by Eq. (2) is also reasonable, because that a bounded monotonic sequence has a limit.

The out-degrees of papers will not be exactly equal to their expected values because the papers are distributed according to a Poisson point process, and so need to be averaged with the Poisson distribution. Hence the out-degree distribution of non-multidisciplinary papers $P^{+}_{\mathrm{non-m}}(k)$ is approximately to be a mixture Poisson distribution:

$$P^{+}_{\mathrm{non-m}}(k) = \frac{1}{2\pi T}\int_{1}^{T+1}\int_{0}^{2\pi}\left(\frac{k^{+}(\theta_i,t_i)^{k}\mathrm{e}^{-k^{+}(\theta_i,t_i)}}{k!}\right)\mathrm{d}\theta_i\mathrm{d}t_i \approx \frac{1}{2\pi}\int_{0}^{2\pi}\left(\frac{k^{+}(\theta_i)^{k}\mathrm{e}^{-k^{+}(\theta_i)}}{k!}\right)\mathrm{d}\theta_i. \tag{3}$$

The multidisciplinary papers make the out-degree distribution $P^{+}(k)$ have a power-law tail. More precisely, we have

$$P^{+}(k) \approx (1-q)P^{+}_{\mathrm{non-m}}(k) + qf(k), \tag{4}$$

where q is the proportion of multidisciplinary papers, and $f(k)$ is a power-law distribution (both are defined in Step 4 of the model description). As Eq. (4) shows, the model cannot exactly reproduce the hooks of the out-degree distributions of empirical networks, which is a generalized Poisson distribution, but not a mixture Poisson. Overcoming this shortcoming is indicative of the need for further research.

The tails of in-degree distributions of empirical citation networks are well fitted by power-law distributions (Fig. 3(c)), which has been considered as a consequence of cumulative advantage of degrees or complete copying. In our model, the highly cited papers (which have an influential zone) are responsible for the power-law tails of the modeled in-degree distributions (Fig. 3(d)). This hints that the power-law tails may also alternatively be interpreted as a consequence of the inhomogeneous sizes of academic influences.

The expected in-degree $k^{-}(\theta_j,t_j)$ of a highly cited paper j is the probability p multiplied by the expected number of papers belonging to its zone, namely

$$k^{-}(\theta_j,t_j) \approx \frac{m}{2\pi}\sum_{s=t_j+\delta}^{T}\frac{\beta(\theta_j)p}{\mathrm{e}^{(\alpha t_j+(1-\alpha)s)l}} \times [\mathrm{e}^{ls}] \approx \frac{\beta(\theta_j)mp}{2\pi\alpha l}\mathrm{e}^{\alpha(T-t_j)l}. \tag{5}$$

The first approximation is due to the ignorance of the in-degrees contributed by the complement mechanism Step 4, because the order of magnitude of q is set to be smaller than that of p. The second approximation holds for $T \gg 1$. The calculation of underlying formula for the in-degree distribution in the large-k region (namely the tail) $P^{-}_{L}(k)$ is the same as that in Ref. [26] and is briefly listed as follows:

$$P^{-}_{L}(k) = \frac{1}{2\pi k!}\int_{0}^{2\pi}\left(\int_{1}^{T+1}k^{-}(\theta_j,t_j)^{k}\mathrm{e}^{-k^{-}(\theta_j,t_j)}\mathrm{d}t_j\right)\mathrm{d}\theta_j \propto \frac{1}{k^{1+\frac{1}{\alpha}}}, \tag{6}$$

where $k \gg 0$ is needed for the approximation.

The multidisciplinary papers can slightly affect the forepart of in-degree distribution, because those papers cite some existing papers randomly, which makes the forepart have a small component of Poisson distribution. Hence, the forepart of in-degree distribution of the modeled network slightly deviates from the power-law distribution.

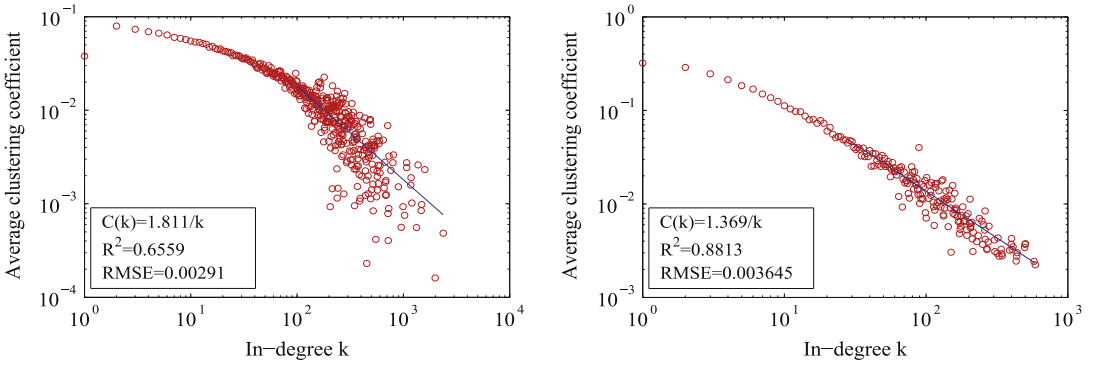
4. Local clustering coefficients and in-degrees

The local clustering coefficient (LCC) of a node is the connecting probability of two neighbors of the node. In empirical citation networks, the large in-degree papers have low LCCs. A sound explanation is that some highly cited papers have influences in many topics, and the papers in different topics may not cite each other even if they cite a common paper. Moreover, considering the average LCC of in-degree k nodes $C(k)$, the tails of $C(k)$ are roughly proportional to $1/k$. We illustrate this property for the DBLP 2013-09-29 (Fig. 4(a)). The other two empirical networks also have this property, which is not illustrated here. To show how the model generates a similar tail (Fig. 4(b)), we next derive the underlying formula of the tails of $C(k)$ for modeled networks.

Suppose i is a highly cited paper and t_i is small enough. The large in-degree nodes of modeled networks only count for the papers like i. As a highly cited paper with many citations, the out-degree of i can be ignored, compared with in-degree. So we only consider the papers that cite i. Consider any two papers j, l belong to the zone of i and cite i as well. Further suppose $t_j < t_l - \delta$. If l is not a multidisciplinary paper, the probability of l citing j is $pq\beta(\theta_j)\mathrm{e}^{-\alpha l t_j-(1-\alpha)lt_l}/(\beta(\theta_i)\mathrm{e}^{-\alpha l t_i-(1-\alpha)lt_l}) \approx pq\mathrm{e}^{-\alpha l(t_j-t_i)}$, where we use the piecewise constant property of $\beta(\cdot)$ and make the assumption that the overlap of the zone j (having a zone with probability q) is fully contained by that of i (this is justified if t_j is large, which is common in the modeled networks with $T \gg 1$). If l is a multidisciplinary paper, the probability of l citing j should add $\max(0, E(f)-k^{+}(\theta_l,t_l))/N(t_l-\delta)$, where $E(f)$ is the expected value of $f(k)$, and $N(t_l-\delta)$ is the number of papers published until time $t_l-\delta$. The increment is very small and could be ignored, so the probability of l citing j is approximately equal to $pq\mathrm{e}^{-\alpha l(t_j-t_i)}$.

Summing over all possible values of t_j with the weight: the proportion of papers published in the t_jth time unit and citing i to all papers citing i, we obtain the approximate value of the expected LCC of i as follows

$$C(\theta_i,t_i) \approx \frac{\int_{t_i+\delta}^{T}pq\mathrm{e}^{\alpha l(t_i-t_j)}\left(\frac{\beta(\theta_j)mp}{2\pi}\right)\mathrm{e}^{-\alpha l t_i-(1-\alpha)lt_j}\mathrm{e}^{lt_j}\mathrm{d}t_j}{k^{-}(\theta_i,t_i)} = \frac{\beta(\theta_i)(T-t_i-\delta)mp^2q}{2\pi k^{-}(\theta_i,t_i)}, \tag{7}$$



(a) DBLP 2013-09-29. (b) Modeled network.

Fig. 4. Local clustering coefficient as a function of in-degree, compared with the theoretical prediction of Eq. (7). The panels show the average local clustering coefficient of k-degree nodes of two networks in Table 1 respectively.

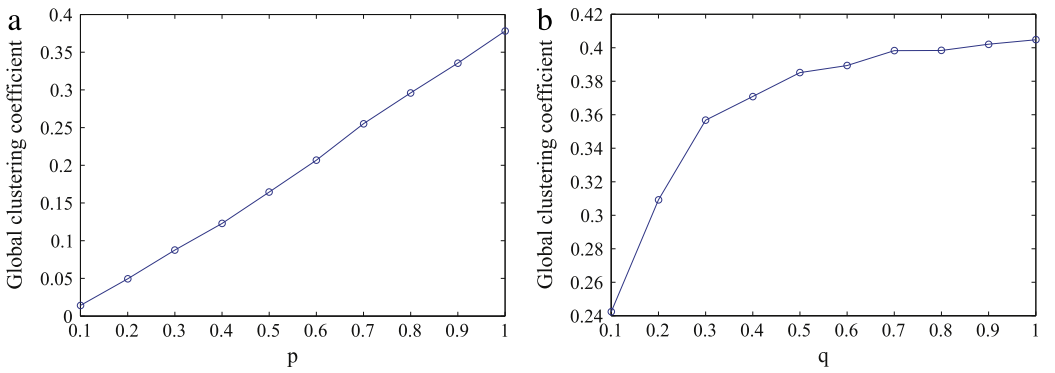


Fig. 5. The tunable global clustering coefficients of modeled networks. Increasing the values of parameters p and q can increase the value of GCC. When varying p, we set $q = 0.5$, and varying q, $p = 1$. The other parameters are listed in Table 1.

which is inversely proportional to the in-degree $k^-(\theta_i, t_i)$. Eq. (7) also shows the parameters p and q can tune the value of LCC, and so that of GCC (Fig. 5).

5. In- and out-assortativity

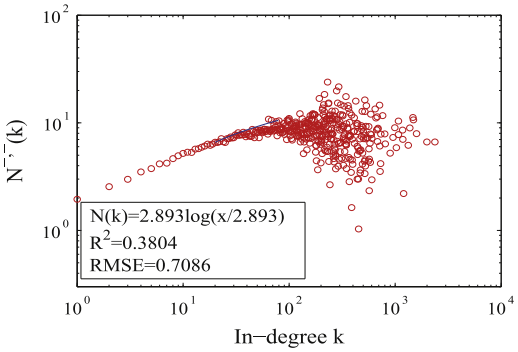
The empirical data show the highly cited papers tend to cite other papers also highly cited (Fig. 6(a)), which means the citation networks are in-assortative. Tracking hot research topics is a reason for the in-assortativity. If a paper is highly cited, it means that the topic discussed in this paper is a hot topic. Many researchers would focus on the topic, publish a lot of papers about the topic, create a lot of citations between those papers, and then make those papers to be highly cited. The empirical data are also out-assortative (Table 1, Fig. 6(c)). In fact, the annual average reference length of a paper increases yearly, and papers usually tend to cite some papers newly published.

Simulations show modeled networks by our model are also in- and out-assortative (Table 1). To show how the model does (Fig. 6(b), (d)), we derive the underlying formulae of the scaling relations between the published time of a paper i and the average in- and out-degree of the papers citing and cited by i, $N^{-,-}(\theta_i, t_i)$ and $N^{+,+}(\theta_i, t_i)$, respectively.

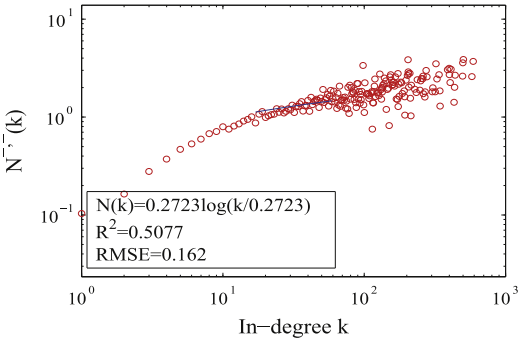
Firstly, we consider $N^{-,-}(\theta_i, t_i)$. If i is a highly cited paper, the average in-degree of the papers citing i is mainly contributed by some highly cited papers and can be calculated as follows

$$N^{-,-}(\theta_i, t_i) \approx \frac{\int_{t_i+\delta}^{T} qk^-(\theta_i, s)\left(\frac{\beta(\theta_i)mp}{2\pi}\right)\mathrm{e}^{-\alpha l t_i-(1-\alpha)ls}\mathrm{e}^{ls}\mathrm{d}s}{k^-(\theta_i, t_i)} \approx \frac{\beta(\theta_i)mpq}{2\pi}(T - t_i - \delta), \tag{8}$$

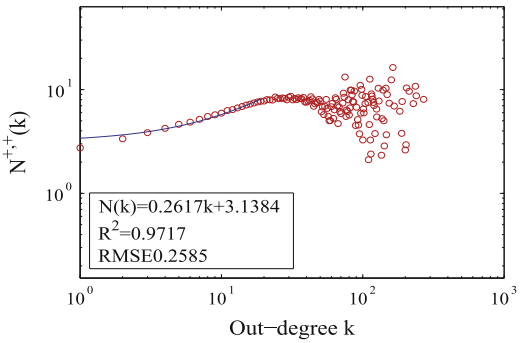
which is a decreasing function of t_i so an increasing function of the in-degree of i. If i is not a highly cited paper (so has a small in-degree) the papers citing i come from multidisciplinary papers. The probability of a paper simultaneously to be a multidisciplinary and highly cited paper is small (which is qr) so the probability of $N^{-,-}(\theta_i, t_i)$ having a large value is also small. To sum up, the modeled networks are in-assortative. Secondly, we consider $N^{+,+}(\theta_i, t_i)$. If i is a non-multidisciplinary



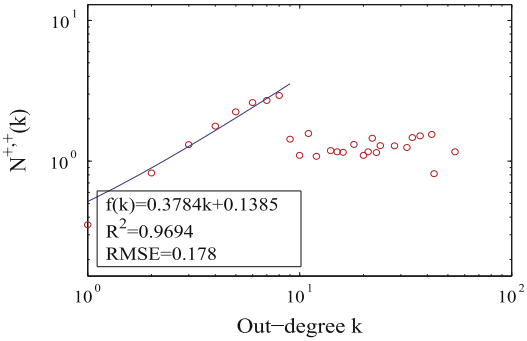
(a) DBLP 2013-09-29.



(b) Modeled network.



(c) DBLP 2013-09-29.



(d) Modeled network.

Fig. 6. The means of the average in- and out-degree of the neighbors citing and cited by k in- and out-degree papers, $N^{-,-}(k)$ and $N^{+,+}(k)$, respectively. Panels (a,b) show those for DBLP 2013-09-29, and Panels (c,d) for the modeled network in Table 1.

paper, the average out-degree of the papers cited by i is

$$N^{+,+}(\theta_i, t_i) \approx \frac{\int_1^{t_i-\delta} k^+(\theta_i, s)\left(\frac{\beta(\theta_i)mpq}{2\pi}\right)\mathrm{e}^{-\alpha ls-(1-\alpha)lt_i}\mathrm{e}^{ls}\mathrm{d}s}{k^+(\theta_i, t_i)} \approx \frac{\beta(\theta_i)mpq}{2\pi}\left(\frac{\mathrm{e}^{-(1-\alpha)l}}{(1-\alpha)l} - \frac{t_i-1-\delta}{\mathrm{e}^{(1-\alpha)l(t_i-\delta-1)}-1}\right), \tag{9}$$

which is an increasing function of t_i so an increasing function of the out-degree of i. If i is a multidisciplinary paper, $N^{+,+}(\theta_i, t_i)$ does not satisfy the result given by Eq. (9), because it cites some papers randomly. Simulations also show the average out-degree of the papers cited by a paper with large out-degree fluctuates around a constant (Fig. 6(d)). In the model, the large out-degree papers only count for multidisciplinary papers. Meanwhile, the number of multidisciplinary papers is very small (which is achieved by giving a small value to r, e.g. $r = 0.002$) so the modeled networks are still out-assortative.

6. Conclusion

A directed geometric graph is proposed to model citation networks with exponentially growing nodes, which overcomes certain shortcomings of our prior proposed model, and provides better reproductions of some typical statistical features of the empirical data, e.g. out-degree distribution. The model potentially paves a geometric way to understand some aspects of citation networks, e.g. it abstractly quantifies the academic influences of papers by geometric zones, expresses aging of papers by decreasing the sizes of influences with the growth of time, and then interprets the power-law tails of in-degree distributions of citation networks by the inhomogeneous influences of papers. Some shortcomings of the model are indicative of the need for further research: how to geometrically realize that the in-degree distributions of papers published in the same year are power-law; how to design a more realistic strategy of localizing the copying mechanism rather than the randomly and uniformly selecting strategy used here. The last but most important problem that needs further research is that, if to infer proper coordinates of a part of empirical data, how to infer the coordinates for the others and predict the future, which needs the techniques in statistics and mapping knowledge domains.

Acknowledgments

We thank the editor and anonymous reviewer for their valuable suggestions and great help, and the hospitality of Professor Alastair Spence in the centre for networks and collective behaviour in the University of Bath. This work is supported

by the open fund from Key Laboratory of High Performance Computing (No. 2 01403-01), and the National University of Defense Technology Graduate Teaching Reform Project (No. yjsy2013012).

References

[1] M.H. MacRoberts, B.R. MacRoberts, Problems of citation analysis, Scientometrics 36 (3) (1996) 435–444.
[2] D. Wang, C. Song, A.L. Barabási, Quantifying long-term scientific impact, Science 342 (2013) 127–131.
[3] T.A. Brooks, Evidence of complex citer motivations, J. Am. Soc. Inf. Sci. Technol. 37 (1) (1986) 34–36.
[4] S. Rednes, How popular is your paper? An empirical study of the citation distribution, Eur. Phys. J. B 4 (2) (1998) 131–134.
[5] F. Radicchi, S. Fortunato, A. Vespignani, Citation networks, in: A. Scharnhorst, K. Börner, P.V.D. Besselaar (Eds.), Models of Science Dynamics, Springer, 2012, pp. 233–257.
[6] F. Radicchi, C. Castellano, Understanding the scientific enterprise: citation analysis, data and modeling, in: Social Phenomena, Springer, 2015, pp. 135–151.
[7] Price D.J. de Solla, Networks of scientific papers, Science 149 (3683) (1965) 510–515.
[8] Price D.J. de Solla, A general theory of bibliometric and other cumulative advantage process, J. Am. Soc. Inf. Sci. 27 (5) (1976) 292–306.
[9] D.J. Price, The scientific foundations of science policy, Nature 206 (1965) 233–238.
[10] F. Radicchi, S. Fortunato, B. Markines, A. Vespignani, Diffusion of scientific credits and the ranking of scientists, Phys. Rev. E 80 (5) (2009) 056103.
[11] L. Bornmann, H.D. Daniel, Universality of citation distributions-A validation of Radicchi et al.'s relative indicator $c_f = c/c_0$ at the micro level using data from chemistry, J. Am. Soc. Inf. Sci. 60 (8) (2009) 1664–1670.
[12] T.S. Evans, N. Hopkins, B.S. Kaube, Universality of performance indicators based on citation and reference counts, Scientometrics 93 (2012) 473–495.
[13] J.R. Clough, J. Gollings, T.V. Loach, T.S. Evans, Transitive reduction of citation networks, J. Complex. Netw. 3 (2) (2015) 189–203.
[14] S.R. Goldberg, H. Anthony, T.S. Evans, Modelling citation networks, Scientometrics 105 (2015) 1577–1604.
[15] K.B. Hajra, P. Sen, Modeling aging characteristics in citation networks, Physica A 368 (2) (2005) 575–582.
[16] S.N. Dorogovtsev, J.F.F. Mendes, Evolution of networks with aging of sites, Phys. Rev. E 62 (2000) 1842–1845.
[17] Y.H. Eom, S. Fortunato, Characterizing and modeling citation dynamics, Plos One 6 (2011) e24926.
[18] P.L. Krapivsky, S. Redner, Network growth by copying, Phys. Rev. E 71 (2005) 036118.
[19] M.V. Simkin, P.V. Roychowdhury, Stochastic modeling of citation slips, Scientometrics 62 (2005) 367–384.
[20] Z.X. Wu, P. Holme, Modeling scientific-citation patterns and other triangle-rich acyclic networks, Phys. Rev. E 80 (3) (2009) 037101.
[21] G.J. Peterson, Steve Pressé S, K.A. Dill, Nonuniversal power law scaling in the probability distribution of scientific citations, Proc. Natl. Acad. Sci. USA 107 (2010) 16023–16027.
[22] M. Penrose, Random Geometric Graphs, in: Oxford Studies in Probability, 2003.
[23] M. Barthélemy, Spatial networks, Phys. Rep. 499 (2011) 1–101.
[24] Z. Xie, J. Zhu, D.X. Kong, J.P. Li, A random geometric graph built on a time-varying Riemannian manifold, Physica A 436 (2015) 492–498.
[25] Z. Xie, T. Rogers, Scale-invariant geometric random graphs, arXiv:1505.01332.
[26] Z. Xie, Z.Z. Ouyang, J.P. Li, A geometric graph model for coauthorship networks, J. Informetr. 10 (2016) 299–311.
[27] D. Krioukov, M. Kitsak, R.S. Sinkovits, D. Rideout, D. Meyer, M. Boguñá, Network cosmology, Sci. Rep. 2 (2012) 793.
[28] Z. Xie, Z.Z. Ouyang, P.Y. Zhang, D.Y. Yi, D.X. Kong, Modeling the citation network by network cosmology, Plos One 10 (2015) e0120687.
[29] J. Tang, J. Zhang, R.M. Jin, Z. Yang, K.K. Cai, L. Zhang, et al., Topic level expertise search over heterogeneous networks, Mach. Learn. J. 82 (2011) 211–237.
[30] KDD Cup 2003: Network mining and usage log analysis. http://www.cs.cornell.edu/projects/kddcup/datasets.html.
[31] P.C. Consul, G.C. Jain, A generalization of the Poisson distribution, Technometrics 15 (4) (1973) 791–799.
[32] A. Clauset, C.R. Shalizi, M.E.J. Newman, Power-law distributions in emprical data, SIAM Rev. 51 (2009) 661–703.