Regular article

Analysis of reference and citation copying in evolving bibliographic networks

Pradumn Kumar Pandey[a], Mayank Singh[b,*], Pawan Goyal[c], Animesh Mukherjee[c], Soumen Chakrabarti[d]

[a] *Department of Computer Science and Engineering, IIT Roorkee, India*
[b] *Department of Computer Science and Engineering, IIT Gandhinagar, India*
[c] *Department of Computer Science and Engineering, IIT Kharagpur, India*
[d] *Department of Computer Science and Engineering, IIT Bombay, India*

A R T I C L E   I N F O

A B S T R A C T

Extensive literature demonstrates how the copying of references (links) can lead to the emergence of various structural properties (e.g., power-law degree distribution and bipartite cores) in bibliographic and other similar directed networks. However, it is also well known that the copying process is incapable of mimicking the number of directed triangles in such networks; neither does it have the power to explain the obsolescence of older papers. In this paper, we propose REFORCITE, a new model that allows for copying of both the references from (i.e., out-neighbors of) as well as the citations to (i.e., in-neighbors of) an existing node. In contrast, the standard copying model (CP) only copies references. While retaining its spirit, REFORCITE differs from the Forest Fire (FF) model in ways that makes REFORCITE amenable to mean-field analysis for degree distribution, triangle count, and densification. Empirically, REFORCITE gives the best overall agreement with observed degree distribution, triangle count, diameter, h-index, and the growth of citations to newer papers.

1. Introduction

Scholarly repositories and retrieval systems, such as Google Scholar (GS), Microsoft Academic Search (MAS), and Semantic Scholar (SS) play a crucial role in scientific information propagation. Apart from keyword search, they present citing and cited papers, co-author graphs, and related articles. Online scholarly search is now critical to discovery of related work, thanks to explosive growth of online proceedings and manuscript repositories. The presentation bias induced by scholarly search can, therefore, significantly influence the evolution of citation networks (see Fig. 1).

Evolution of citation networks has been under investigation for at least six decades. Many fascinating theories have been proposed to explain the intrinsic forces driving citation. The motive has been to enable newer models mimic more and more salient properties observed in real networks. We review a series of standard properties and corresponding modeling approaches in the rest of this section: tail-heavy degree distribution, plentiful bipartite cores and triangles, densification and reduction of diameter as time passes, and the effects of obsolescence of nodes. In Section 2, we propose REFORCITE,

---

Spatial networks

M Barthélemy - Physics Reports, 2011 - Elsevier
Complex systems are very often organized under the form of netwo
edges are embedded in space. Transportation and mobility network
networks, power grids, social and contact networks, and neural netw
☆ ⁇ Cited by 1333   Related articles   All 17 versions   W

Named data networking

L Zhang, A Afanasyev, J Burke, V Jacobson… - ACM SIGCOMM ..
Abstract Named Data Networking (NDN) is one of five projects fund
Science Foundation under its Future Internet Architecture Program.
earlier project, Content-Centric Networking (CCN), which Van Jacob
☆ ⁇ Cited by 753   Related articles   All 25 versions   Wel

Random graphs and complex networks

R Van Der Hofstad - Available on http://www. win. tue. nl/rhofstad …
These lecture notes are intended to be used for master courses, wh
limited prior knowledge of special topics in probability. Therefore, we
the preliminaries, such as convergence of random variables, probab
☆ ⁇ Cited by 361   Related articles   All 11 versions   Imp

Fig. 1. Sample Google Scholar result page. Each response article is displayed with a link that leads to papers citing that article. This feature promotes two ways of copying references: to papers that the article cites, and papers that cite the article.

a variation based on the Forest Fire (FF) Leskovec, Kleinberg, and Faloutsos (2007) and the RelayCite Singh, Sarkar, Goyal, Mukherjee, and Chakrabarti (2017) models. The key idea in all three models is that new nodes choose and link to a 'base' node, and then nodes in its neighborhood. As the name suggests, FF explores the neighborhood indefinitely (but limited by a geometric distribution on path lengths). In contrast, we argue, and later justify experimentally (Section 3), that indefinite exploration is unnecessary in the small-diameter networks seen in practice; it has also made formal analysis infeasible thus far. In contrast, REFORCITE limits itself to a radius-1 "controlled burn", which not only affords formal analysis, but also fits observed networks *better*.

1.1. Preferential attachment

The initial set of theories (Barabási & Albert, 1999; Caldarelli, 2007; Price, 1976) focused on the "rich gets richer effect", also called the "Matthew Effect", based on the premise of preferential attachment (PA). Although their early popularity was attributed to successful prediction of power-law degree distributions, deviations from power law are well-known. Going further, Brzezinski (2015) showed that, in most citation networks, power-law with exponential cut-off and log-normal distributions fit observed data better than pure power law. Also, if new nodes access the network through a small oligarchy of centralized search engines, degree distribution deviates from power law (Chakrabarti, Frieze, & Vera, 2005).

1.1.1. Aging and oligarchies

PA could not explain obsolescence (loss of popularity over time). This led to another set of elegant theoretical models of *aging* (Dorogovtsev, Goltsev, & Mendes, 2002; Dorogovtsev & Mendes, 2013; Medo, Cimini, & Gualdi, 2011; Wang, Yu, & Yu, 2009). Aging models temper preferential attachment with a temporal decay component, represented by either the log-normal or the exponential distribution. Fitness parameters, modeling a node's competitiveness to attract links, add another dimension to complex growth models (Bianconi & Barabási, 2001; Wang, Song, & Barabási, 2013). Singh et al. (2017) propose an aging model where new node v tentatively chooses a base node u, but then cites a node x that cites u, in case u is "too old".

1.2. Bipartite cores and the copying model (CP)

In an effort to preserve power-law degree distribution *and* explain the formation of dense bipartite cores, researchers (Kleinberg, Kumar, Raghavan, Rajagopalan, & Tomkins, 1999; Kumar et al., 2000a) have proposed a model in which a new node copies all or a fraction of out-degree of the target node (see Fig. 2). At each step, they sample a probability distribution to determine a node v to add edges out of, and a number of edges k that will be added. With probability β, they add k edges from v to nodes V chosen independently and uniformly at random. With probability $1 - \beta$, they copy k edges from a randomly chosen node to v. Stochastic copying (Kumar et al., 2000b) is another variant. Again, at each time step, a new node u enters into the system with d out-links. To generate the out-links, they begin by choosing a "prototype" vertex x. The ith out-link of u is then chosen as follows. With probability α, the destination is chosen uniformly at random from V, and with the remaining probability the out-link is taken to be the ith out-link of x. CP exhibits triangle deficiency (Ren, Shen, & Cheng, 2012), cannot accurately model edge-densification (Leskovec et al., 2007) and cannot adequately model aging and obsolescence.
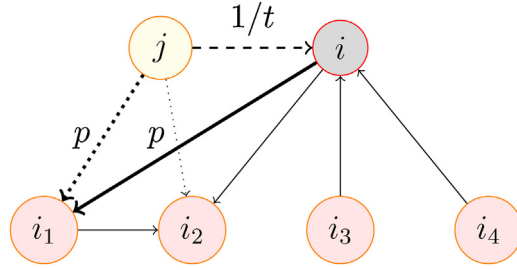
Fig. 2. Copying mechanism of CP model. A node j newly introduced at time t connects to older base node i with probability $1/t$ and then get connected with one of the first neighbors (out-links only) of node i with probability p.
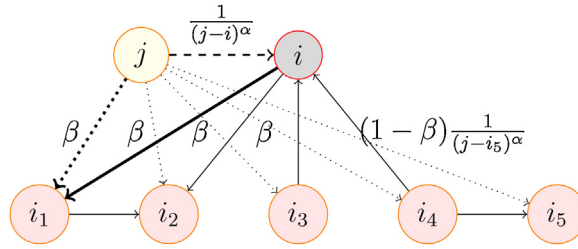
Fig. 3. Copying mechanism of CPT model. A node j newly introduced at time t connects to older base node i with probability $\frac{1}{(j-i)^\alpha}$ and then get connected with one of the first neighbors of node i with probability β.

1.3. Triangle/triad formation

Triangle formation in real-world networks (Eswaran, Rabbany, Dubrawski, & Faloutsos, 2018) is a local event in which a node in the network keeps track of its semi-local structure, for example, neighbours (direct connections) and second-neighbours. It is also more practical due to short visibility of nodes in a large network. In sociology, it is well accepted that the probability of link formation between two nodes would be more if they share more number of common neighbours. There are existing theories and examples which support triangle formation processes in real-world networks. Here, we consider real-world networks which are growing in time but without addition of new edges among older nodes. We do analysis of triangle formation process in citation networks. This is in correspondence to how recommendations of citations by Google Scholar promotes triangle formation process in two ways of copying references: to papers that the article cites, and papers that cites the article. Fig. 2 shows the CP's copying mechanism.

1.3.1. Copying with triad formation (CPT)

Krapivsky and Redner (2005) proposed a variant of CP (see Fig. 3). A newly introduced node randomly selects a target node and links to it, as well as to all out-neighbors (ancestors) of the target node. Thus, if the target node is the first introduced node (root node), no additional links are generated by the copying mechanism. If the newly introduced node were to always choose the root node as the target, a star graph would be generated. On the other hand, if the target node is always the most recent one in the network, all previous nodes are ancestors of the target and the copying mechanism would give a complete graph. In another CPT model (Wu & Holme, 2009), a new node i, having out-degree k_i^{out}, selects one of the old node j with an aging probability proportional to its age $t_j = i - j$ to a power α in the existing network as a base node. The rest of the out-degrees of i are attached to random (in- or out-) neighbors of j with probability β, and otherwise (i.e., with probability $1 - \beta$) attach links to older vertices with similar aging probability as above. If there is no available neighbor to attach to then node i selects a new base node as described above and repeats the entire process. Fig. 3 shows the CPT's copying mechanism.

1.3.2. Forest fire (FF) model

(Leskovec et al., 2007, Section 4.2) proposed the FF model to remedy some of the above limitations. New node v first chooses a base or 'ambassador' node w uniformly at random and links to it. Then it samples two geometrically distributed numbers x and y with means $p_a/(1 - p_a)$ and $bp_a/(1 - bp_a)$, respectively, where p_a is the forward burning probability and b is the backward burning ratio. x and y random unvisited in- and out-neighbors of w are visited and linked from v. This step is then recursively applied to the newly linked nodes, but visited nodes are never revisited. Although FF achieves realistic heavy-tailed degrees, triads, densification and shrinking diameter in experiments, the authors note that "rigorous analysis of the Forest Fire Model appears to be quite difficult". As will become clear, our proposed model, REFORCITE, resembles FF, but is actually simpler. This allows a mean-field analysis of degree distribution and expected number of triangles. Surprisingly, the simplification results in no degradation in the ability to model real networks. In contrast, aggregating over several graph properties, REFORCITE fits real data better than FF, including the effect of aging, which was not measured for FF earlier.

Table 1

Growth equations of different evolution models. Simple closed form equations are not possible in case of FF and CPT.

Models	Degree growth formulation
FF	Simple closed form equation is not possible
CPT	Simple closed form equation is not possible
CP	$\frac{dk_i(t+1)}{dt} = \frac{1}{t} + \sum_{\lfloor N_i^{\text{out}}} \frac{p_1}{t}$
REFORCITE1	$\frac{dk_i(t+1)}{dt} = \frac{1}{t} + \sum_{\lfloor N_i(t)} \frac{p}{t}$
REFORCITE2	$\frac{dk_i(t+1)}{dt} = \frac{1}{t} + \sum_{\lfloor N_i^{\text{in}}(t)} \frac{p_2}{t} + \sum_{\lfloor N_i^{\text{out}}} \frac{p_1}{t}$

1.4. Our contributions

In this paper, we present a new citation network growth model and call it REFORCITE. Like FF and RelayCite Singh et al. (2017), REFORCITE is based on real-life scholarly explorations in citation networks. To trace the origins of an idea, we need to explore back in time, following outlinks. To discover recent improvements on a result, we need to explore forward in time, traversing inlinks in reverse, facilitated by scholarly search — see Fig. 1.

While driven by the same modeling considerations as FF, REFORCITE has important technical differences (see Table 1). Like in FF, new node v chooses a base or 'ambassador' node u, and then expands backward or forward from u, but not recursively. It helps us derive mean-field estimates for degree distribution, triangle count and some other properties. We present two variants of REFORCITE: REFORCITE1, where, after linking v to u, we walk only to newer nodes x that link to u (as in RelayCite Singh et al. (2017)); and REFORCITE2, in which, like FF, walking to both in- and out-links of u is allowed. Accordingly, REFORCITE is characterized by one or two parameters. Comparing the best-fit forward and backward parameters can reveal key behavioral traits of different scholarly communities.

The majority of previous growth models are evaluated by fitting degree distribution against observed data. Only a few attempts have been made to match other structural properties like clustering (Holme & Kim, 2002), bipartite cores (Kumar et al., 2000a), triad formation (Wu & Holme, 2009; Xie, Ouyang, Zhang, Yi, & Kong, 2015), centrality, or temporal bucket signatures (Singh et al., 2017). It is common for a model that mimics one property well to fit other properties poorly. A model that provides *simultaneous* good or reasonable fits for many properties can, therefore, be valuable.

Remarkably, the restrictions above that make REFORCITE amenable to analysis do not impair its ability to fit real network properties. Simulations show that REFORCITE matches observed degree distribution, triangle count, densification, network h-index, and diameter more faithfully than CP and CPT. Moreover, REFORCITE offers a network-driven explanation of obsolescence, gradually shifting citation focus toward newer papers, in excellent agreement with observed data.

2. The proposed growth model

In this section, we present our citation growth model REFORCITE. In REFORCITE, a new node j appears at time $t+1$ and selects an older 'base' node i uniformly randomly from G_t, where G_t is the network at time t and $G_{t+1} = \{j\} \cup G_t$. The base node is not sampled preferentially: that would increase the selection of older nodes beyond what is warranted by obsolescence models Singh et al. (2017). Uniform base node sampling also leads to heavy-tailed degrees and other realistic properties. After introducing the first directed link (j, i) (link is directed towards node i), node j may form additional links with immediate in- and out-neighbours of node i. We propose two possible variants on how j forms each such link to the in- and out-neighbours of i: **REFORCITE1**, with a single probability parameter p; and **REFORCITE2**, with probabilities p_1 and p_2. The next two sections describe the formulations of expected growth of the degree of node for our proposed two variants. Table 1 compares formulations of expected growth of the degree of node for REFORCITE and CP. In case of FF and CPT, simple closed form equation is not possible.

2.1. REFORCITE1

Let $\lfloor N_i^{\text{in}}(t)$ and $\lfloor N_i^{\text{out}}$ be the in-neighbours and out-neighbours of node i. $\lfloor N_i(t) = \lfloor N_i^{\text{in}}(t) \cup \lfloor N_i^{\text{out}}(t)$ denote the set of neighbors of node i at time t, $\lfloor N_i^{\text{in}}(t)$ grows with time, whereas $\lfloor N_i^{\text{out}}$ remains fixed after a new node is linked during its introduction into the network. Let $k_i^{\text{out}} = |\lfloor N_i^{\text{out}}|$ and $k_i^{\text{in}}(t) = |\lfloor N_i^{\text{in}}(t)|$ be out- and in-degree of a node i, respectively, with $|\lfloor N_i(t)| = k_i(t) = k_i^{\text{in}}(t) + k_i^{\text{out}}$ being the degree of a node i at time t. New node j joins the network at time $t+1$. The expected growth of the degree of node i is given by

$$\frac{dk_i(t+1)}{dt} = \frac{1}{t} + \sum_{\lfloor N_i(t)} \frac{p}{t}, \tag{1}$$

where $p \in [0, 1]$. The degree of the node i can grow in two ways: either it is selected as the base paper, or it is one of the neighbors of the base paper. In Eq. (1), the first term corresponds to selection of the first paper uniformly at random. At time $t+1$, node i can be the first paper for the new node j with probability $1/t$. Also, any neighbour of the node i can be the base
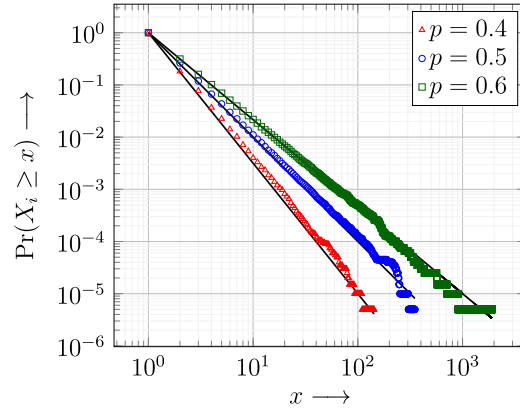
Fig. 4. (Best viewed in color.) Distribution of $X_i = \frac{k_i(t+1)+1/p}{k_i^0+1/p}$ is plotted which is proved to be power-law with exponent $1/p$. Lines correspond to theoretical predictions ($x^{-1/p}$) while markers correspond to simulated results. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

paper for node j with the same probability $1/t$, and then i gets a link from j with probability p as a part of the selection of references/citations of the base paper. This is reflected in the second term.

2.1.1. Degree distribution

In order to obtain an analytical estimate of (expected) degree distribution, we work from Eq. (1):

$$\frac{dk_i(t+1)}{dt} = \frac{1+pk_i(t)}{t}$$

By mean field approximation,

$$\frac{1}{p}\int \frac{dpk_i(t)}{1+pk_i(t)} = \int \frac{dt}{t}$$

Asserting boundary condition $k_i(t_i) = k_i^0$,

$$\ln \frac{k_i(t+1)p+1}{k_i^0 p+1} = p \ln \frac{t+1}{t_i}$$

$$\frac{k_i(t+1)+1/p}{k_i^0+1/p} = \left(\frac{t+1}{t_i}\right)^p$$

For $k_i(t)$ to exceed k, we need

$$t_i < (t+1)(k+1/p)^{-1/p}(k_i^0+1/p)^{1/p}.$$

Since nodes arrive uniformly, we have

$$Pr(k_i > k) \sim (k+1/p)^{-1/p}(k_i^0+1/p)^{1/p}, \tag{2}$$

where $\lim_{t\to\infty} k_i(t) \to k_i$.

Thus, the degree distribution in REFORCITE1 closely follows a power-law with a dependency on initial degree (out degree in citation networks). To work around the initial condition, we consider variable $X_i = \frac{k_i(t+1)+1/p}{k_i^0+1/p}$ instead of degree $k_i(t+1)$, and plot $Pr(X_i > x)$ against x in Fig. 4. The event $X_i > x$ corresponds to $(t/t_i)^p > x$, or $t_i < tx^{-1/p}$, implying that $Pr(X_i > x) \propto x^{-1/p}$, a perfect power law.

2.1.2. Densification

Another interesting property of real networks is the behaviour of average connectivity (densification) which we investigate analytically here. Let $\bar{k}_t$ be the average degree of the network under model Eq. (1) at time t. Then, by counting edges present up to time $t-1$ and adding on k_t^0 edges at time t, we get

$$\bar{k}_t = \frac{(t-1)\bar{k}_{t-1} + 2k_t^0}{t} \tag{3}$$
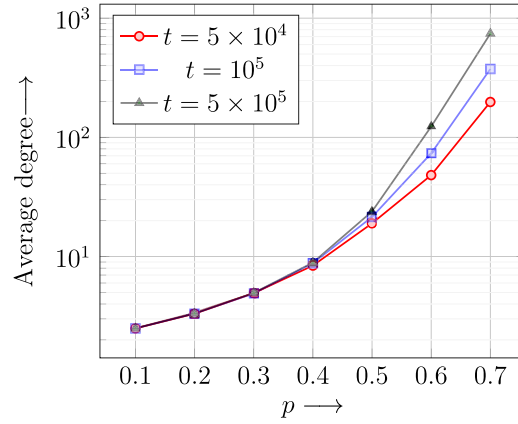
Fig. 5. (Best viewed in color.) Average degree ($\bar{k}_t$) for different values of p for networks of different sizes. For $p < 0.5$, average degree approaches to a constant value ($\approx \frac{2}{1-2p}$) irrespective of the size of the network and depends on p only, but for $p > 0.5$, average degree of the network also depends on the size of the network (t). As the size of the network increases, average degree increases. When $p > 1/2$, a network of larger size has larger average degree (red circles<blue squares<black triangles). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The number of edges added at time t, i.e., k_t^0, can be accounted as one for the link from new node j to the base node, and then $p\bar{k}_{t-1}$ edges to neighbors of the base node. From this we can approximate

$$\bar{k}_t = \frac{(t-1)\bar{k}_{t-1} + 2(1 + p\bar{k}_{t-1})}{t}, \tag{4}$$

$$\bar{k}_t = \bar{k}_{t-1} + \frac{(2p-1)\bar{k}_{t-1} + 2}{t}, \tag{5}$$

$$\frac{d\bar{k}_{t-1}}{dt} = \frac{(2p-1)\bar{k}_{t-1} + 2}{t}, \tag{6}$$

which can be solved as

$$(2p-1)\ln\frac{t}{2} = \ln\frac{(2p-1)\bar{k}_{t-1} + 2}{2} \tag{7}$$

from which we get

$$\bar{k}_{t-1} = \begin{cases} \frac{2}{2p-1}\left(\frac{t}{2}\right)^{2p-1} - 2/(2p-1), & p \neq 1/2. \\ 2\ln(t/2) - 1, & p = 1/2 \end{cases} \tag{8}$$

As the value of t becomes larger

$$\bar{k}_{t-1} \approx \begin{cases} 2/(1-2p), & p < 1/2. \\ \frac{2}{2p-1}\left(\frac{t}{2}\right)^{2p-1}, & p > 1/2. \\ 2\ln(t/2) - 1, & p = 1/2 \end{cases} \tag{9}$$

Thus, average degree follows power-law in the size of the network, and shows a phase transition around $p = 1/2$. In Fig. 5, we plot average degree for different values of p considering networks of $t = 5 \times 10^4$ nodes, $t = 10^5$ nodes, and $t = 5 \times 10^5$ nodes. It is observed that for $p < 0.5$, average degree converges to same values $\left(\approx \frac{2}{1-2p}\right)$ irrespective of the size t of the networks while for $p > 0.5$, average degree of the network also depends on the size of the network. A phase transition in the behaviour of the average degree is observed around $p = 0.5$ in Fig. 5. For $p > 1/2$, it shows densification; for $p < 1/2$ it asymptotically approaches to a constant average degree $\frac{2}{1-2p}$.

2.1.3. Triangle formation

We will now get an analytical estimate of the expected **number of triangles**. Let Δ_t^i denote the expected number of triangles attached with node i through time t. Let Δ'_{t+1} be the expected number of triangles generated at time step $t+1$, and $\Delta(t+1)$ be the expected total number of triangles in the network generated through time $t+1$. A new triangle can be generated in two ways:
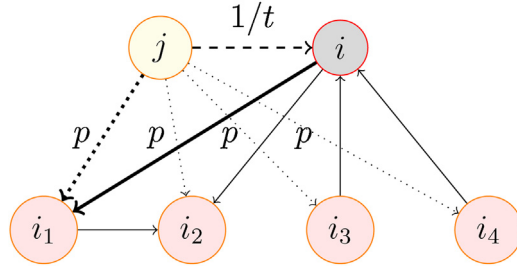
Fig. 6. Triangle formation. A node j newly introduced at time t connects to older base node i with probability $1/t$ and then get connected with one of the first neighbors of node i with probability p. One possible resulting triangle (j, i, i_1) is shown with thicker lines.

1. The new node j gets connected with an older node i with probability $1/t$ and one of its neighbors with probability p, for example, triangle (i, j, i_1) in Fig. 6.
2. During the same growth process new triangles get formed due to existing triangles, for example, triangle (j, i_1, i_2) is formed due to triangle (i, i_1, i_2) in Fig. 6.

Accordingly, we can write:

$$\Delta'_{t+1} = \frac{1}{t}\sum_i pk_i(t) + \frac{1}{t}\sum_i \Delta_t^i p^2 = p\bar{k}_t + p^2\sum_i \frac{\Delta_t^i}{t} \tag{10}$$

Because each triangle is counted thrice in $\sum_i \Delta_t^i$, we can see that $(1/t)\sum_i \Delta_t^i = 3\Delta(t)/t = 3\bar{\Delta}_t$. Thus we can write

$$\Delta'_{t+1} = p\bar{k}_t + 3p^2\bar{\Delta}_t \tag{11}$$

$$\Delta(t+1) = \sum_{\tau=2}^{t+1} \Delta'_\tau \tag{12}$$

$$= \sum_{\tau=1}^{t} \left(p\bar{k}_\tau + 3p^2\bar{\Delta}_\tau\right) = 3p^2\sum_{\tau=1}^{t}\bar{\Delta}_\tau + p\sum_{\tau=1}^{t}\bar{k}_\tau \tag{13}$$

$$\Delta(t+1) = 3p^2\sum_t\bar{\Delta}_t + p\sum_t \left(\frac{2}{2p-1}\left(\frac{t}{2}\right)^{2p-1} - \frac{2}{2p-1}\right) \tag{14}$$

$$\Delta(t+1) = 3p^2\sum_t\bar{\Delta}_t + \frac{4p}{2p-1}\left(\frac{t}{2}\right)^{2p} - \left(\frac{2p}{2p-1}\right)t. \tag{15}$$

Δ'_{t+1}, initially, starts with the existence of non-zero value of average degree. This dependency and multiplier p^2 results in slower growth of $3p^2\sum_{\tau=1}^t\bar{\Delta}_\tau$ as compared to $p\sum_{\tau=1}^t\bar{k}_\tau$. So, for large value of t, the triangle count can be approximated in the following way:

$$\Delta(t+1) \approx \begin{cases} \dfrac{4p}{2p-1}\left(\dfrac{t}{2}\right)^{2p}, & p > 1/2. \\[3mm] \dfrac{2p}{1-2p}t, & p < 1/2 \end{cases} \tag{16}$$

At $p = 0.5$, $\bar{k}_{t-1} = 2\ln(t/2) - 1$

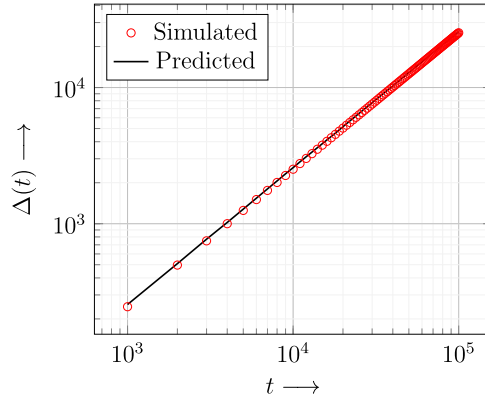$$\Delta(t+1) = 3p^2\sum_t\bar{\Delta}_t + p\sum_t (2\ln(t/2) - 1). \tag{17}$$

$$\Delta(t+1) = 3p^2\sum_t\bar{\Delta}_t + p(2\ln t! - 2t\ln 2 - t). \tag{18}$$

Using Stirling's formula

$$\Delta(t+1) \approx 3p^2\sum_t\bar{\Delta}_t + p(2t\ln t - 2t\ln 2 - 3t + 2). \tag{19}$$

$$\Delta(t+1) \approx 2pt\ln t. \tag{20}$$

Fig. 7. (Best viewed in color.) Actual number of triangles for the value of $p = 0.1$ are plotted in red circles and expected number of triangles are plotted in black line. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The (analytically obtained) expected number of triangles and the actual number of triangles obtained from a simulated network are shown in Fig. 7. The theory and the simulation exhibit near-perfect agreement.

2.2. REFORCITE2

In the model REFORCITE2, copying probabilities of references (out-links) and citations (in-links) are different while in REFORCITE1, both are same (p). Similar to Eq. (1), here the expected growth of the degree of node i is given by

$$\frac{dk_i(t+1)}{dt} = \frac{1}{t} + \sum_{\lfloor N_i^{in}(t) \rfloor} \frac{p_2}{t} + \sum_{\lfloor N_i^{out} \rfloor} \frac{p_1}{t}, \tag{21}$$

where $p_1, p_2 \in [0, 1]$ are the probabilities of copying in- and out-links of the base node.

2.2.1. Degree distribution

The procedure followed to compute the degree distribution of REFORCITE2 is similar to REFORCITE1.

$$\frac{dk_i(t+1)}{dt} = \frac{1 + p_2 k_i^{in}(t) + p_1 k_i^{out}}{t} = \frac{F_i^0 + p_2 k_i(t)}{t}, \tag{22}$$

where $F_i^0 = 1 + (p_1 - p_2)k_i^{out}$, and F_i^0 is a constant for node i which depends only on the out-degree.

$$Pr(k_i > k) \sim (k + F_i^0/p_2)^{-1/p_2} (k_i^{out} + F_i^0/p_2)^{1/p_2}, \tag{23}$$

where $\lim_{t \to \infty} k_i(t) \to k_i$.

Again, the degree distribution in REFORCITE2 closely follows a power-law with a dependency on initial degree (out degree in citation networks) with power-law exponent $1/p_2$.

2.2.2. Average degree as a function of time

The procedure followed to compute average degree of REFORCITE2 is similar to REFORCITE1. Let $\bar{k}_{t-1}^{in}$ be the average in-degree of REFORCITE2 at time $t - 1$. Consider,

$$\bar{k}_t^{in} = \frac{(t-1)\bar{k}_{t-1}^{in} + k_t^{out}}{t}, \tag{24}$$

which leads to the following result:

$$\bar{k}_{t-1}^{in} = \begin{cases} \ln(t/2) - 1/2, & p_1 + p_2 = 1 \\ \left(\frac{1}{p_1 + p_2 - 1} + \frac{1}{2}\right)\left(\frac{t}{2}\right)^{p_1 + p_2 - 1} - \frac{1}{p_1 + p_2 - 1}, & \text{o.w.} \end{cases} \tag{25}$$

Thus, average in-degree follows power-law in the size of the network, and shows a phase transition around $p_1 + p_2 = 1$. For $p_1 + p_2 < 1$ and $t \to \infty$, the average in-degree of the networks produced under REFORCITE2 approach to the fixed point $1/(1 - p_1 - p_2)$, asymptotically. For $p_1 + p_2 > 1$, REFORCITE2 shows densification.

Later, we utilize the relation between model parameters p_1 and p_2, and average in-degree to simulate the model networks under REFORCITE2 corresponding to the given real data. For a given real data, from Eq. (25), we evaluate $p_1 + p_2$ numerically.

Table 2
Brief descriptions and salient properties of data sets.

Networks	Description	Nodes	Edges
Biomedical	Consists of biomedical papers indexed in NCBI (2001–2008).	43,937	162,404
Supreme court	US Supreme Court cases (1754–2002). Judgements refer to previous judgements.	25,417	446,490
ArXiv-HepTH	*High Energy Physics – Theory* papers from arXiv.org (1992–2002).	27,770	352,807
ArXiv-HepPH	*High Energy Physics – Phenomenology* papers from arXiv.org (1992–2002).	34,546	421,578

Table 3
L1 error (smaller is better) between in-degree distributions estimated from simulated networks and corresponding real networks. Each simulation result is reported for the optimal choice of its parameters, which are also shown in the table.

Networks	CP Model	CPT Model	FF Model	REFORCITE1	REFORCITE2
Biomedical	1.73	3.26	**0.42**	0.52	0.75
	$p = 0.55$	$\alpha = -1,\ \beta = 0.99$	$b = 1,\ p_a = 0.001$	$p = 0.41$	$p_1 = 0.25,\ p_2 = 0.50$
Supreme court	4.67	4.18	2.37	0.95	**0.83**
	$p = 0.57$	$\alpha = -1,\ \beta = 0.99$	$b = 1,\ p_a = 0.03$	$p = 0.47$	$p_1 = 0.80,\ p_2 = 0.17$
ArXiv-HepTH	7.84	8.16	3.93	1.28	**0.65**
	$p = 0.58$	$\alpha = -1,\ \beta = 0.99$	$b = 10,\ p_a = 0.05$	$p = 0.51$	$p_1 = 0.40,\ p_2 = 0.65$
ArXiv-HepPH	9.18	7.68	3.92	**0.61**	**0.61**
	$p = 0.61$	$\alpha = -1,\ \beta = 0.99$	$b = 2,\ p_a = 0.04$	$p = 0.52$	$p_1 = 0.60,\ p_2 = 0.43$

Let us say, for a dataset, $p_1 + p_2 = c$, then during the simulation we select $p_1 \in [0, c]$ (or p_2) as free parameter and $p_2 = c - p_1$, accordingly.

In Table 3, in some cases error for REFORCITE2 is more as compared to REFORCITE1 when $p_1 + p_2 < 2p$. The reason is that REFORCITE2 is not able to explore $p_1 = p_2 = p$ condition in these cases due to the constraint noted in Eq. (25). We point out that if $p_1 + p_2 \geq 2p$ then in almost all cases error(REFORCITE2) $\leq$ (REFORCITE1).

3. Experimental evaluation

3.1. Datasets

Investigating the questions raised in this work requires rich trajectories of time-stamped network snapshots. However, such intricately detailed datasets are rare, even while there are an increasing number of new repositories being built and updated regularly.[1] We conduct empirical analysis on four citation networks constructed from (i) Biomedical papers,[2] (ii) US Supreme Court cases (Fowler & Jeon, 2008), (iii) ArXiv's High Energy Physics Theory papers (Leskovec & Krevl, 2014), and (iv) ArXiv's High Energy Physics – Phenomenology papers (Leskovec & Krevl, 2014), respectively. As evident from the description, three networks represent scientific article citation networks and one legal document citation network. Biomedical citation network contains papers indexed in PMC Open Access (OA) Subset.[3] The articles in the OA Subset are made available under a Creative Commons that generally allows more liberal redistribution and reuse than a traditional copyrighted work. The U.S. Supreme Court citation network contains opinions written by the U.S. Supreme Court and the cases they cite from 1754 to 2002 in the United States Reports. The ArXiv citation datasets were originally released as a part of 2003 KDD Cup. It begins within a few months of the inception of the ArXiv, and thus represents essentially the complete history of Physics Theory and Phenomenology papers, respectively. A brief description of each of these networks, along with some bulk statistics, are provided in Table 2. We consider degree distribution, number of triangles, average diameter and obsolescence to compare the real networks against those obtained by the various proposed models. Note that, we keep the same number of nodes in the simulated networks as the corresponding real networks.

3.2. Degree distribution

We adopt the method explained by Pandey and Adhikari (2017) to compute the values of parameters under considered models corresponding to each data set. The method is as follows: we discretize $p \in (0, 1)$ and simulate a model network corresponding to the considered model. L_1 distance is computed between degree distribution of the real network and corresponding model network. Value of p is selected corresponding to the minimum L_1 distance. Minimized L_1 distances are reported in Table 3 corresponding to our proposed models (REFORCITE1 and REFORCITE2), CP model, CPT model, and FF model.
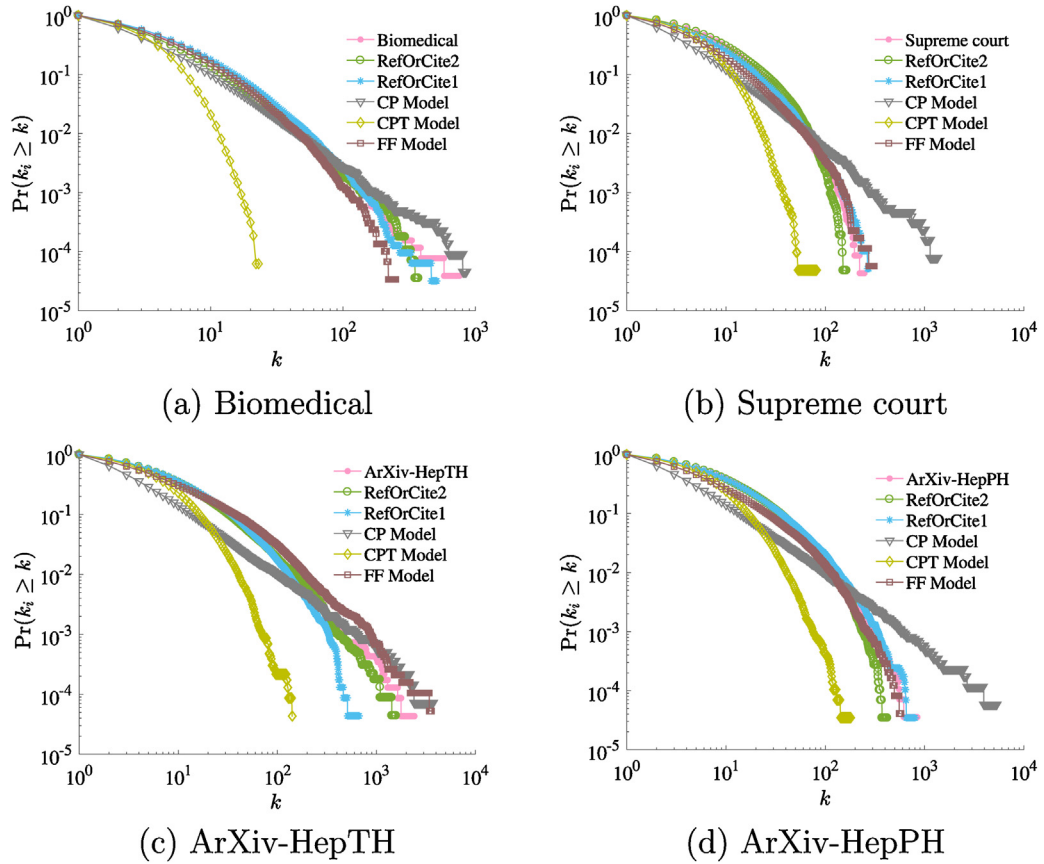
Fig. 8. (Best viewed in color.) In-degree distributions for (a) Biomedical, (b) Supreme court, (c) ArXiv-HepTH, and (d) ArXiv-HepPH. Observed data, REFORCITE2 and REFORCITE1, CP, CPT, and FF predictions are plotted in pink dots, green circles, cyan stars, gray triangles, yellow diamonds, and violet, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

In Fig. 8, we plot the degree distributions of the four real networks (described in Section 3.1). We compare these with the degree distributions predicted by REFORCITE1, REFORCITE2, as well as FF, CPT and CP. Clearly, REFORCITE1 and REFORCITE2 show much better agreement with real data compared to CPT and CP. The CPT model performs the worse since this model is most suitable for networks that show a slow increase in degree over time; the data sets that we consider on the other hand exhibit faster growth in node degrees. Both REFORCITE variants fit real data as well as the (more complex) FF model.

We also experiment with an ensemble of 100 model realizations to understand the sensitivity associated with the model outputs. Fig. 9 shows degree distribution of real-network ArXiv-HepPH compared against REFORCITE. Small standard deviations over 100 model realizations demonstrate low variability in the generated outputs. Similar observations are obtained for CP, CPT, and FF. CP, CPT, and FF with standard deviations for 100 model realizations are 9.9×10^{-5}, 3.4×10^{-4}, and 2.1×10^{-3}, respectively. Similar results were obtained for other real-networks.
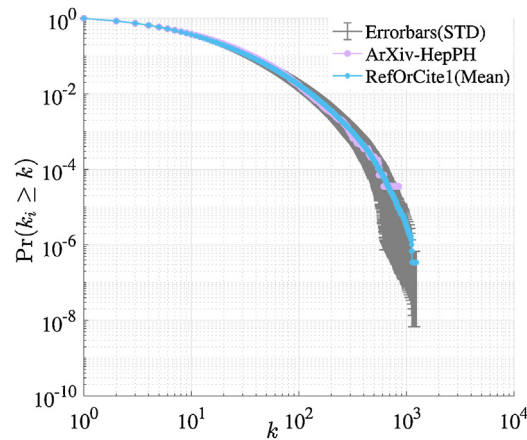
3.3. Triangle counts

In Table 4 we report the number of triangles present in the real datasets. In addition, we also report the ratio between the number of triangles obtained from the simulated networks (CP model, CPT model, FF model, REFORCITE1 and REFORCITE2 models) and the real networks for each of the datasets. We observe that our models match the real data much better than all the other three models. Table 5 shows mean and standard deviation of ratio of simulated and real triangle counts for 100 realizations of above models. As expected, we observe significantly low standard deviation values.

---

[1] http://snap.stanford.edu/ is a prominent example.
[2] http://www.ncbi.nlm.nih.gov/pmc/tools/ftp.
[3] https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/.

Fig. 9. Degree distribution of real-network ArXiv-HepPH compared against our model. Mean of the degree distributions over 100 model runs obtained under REFORCITE1 is plotted in cyan colored stars, and error bars (standard deviations (STD)) are plotted in gray color. Mean STD (mSTD) over all networks is 0.001 and mean variance (mVAR) is 7.9×10^{-6}. For same real data set, CP, CPT, and FF have values of (mSTD, mVAR) (9.9×10^{-5}, 5.6×10^{-8}), (3.4×10^{-4}, 4.2×10^{-7}), and (2.1×10^{-3}, 2.6×10^{-5}), respectively.

Table 4
Real and simulated triangle counts, average diameter and h-index over the lifetime of four networks. The third column (row 2–5) shows the number of triangles observed in each data network. Subsequent columns show the ratio between the simulated and observed numbers of triangles. A ratio close to 1 indicates a better model driving the simulation. REFORCITE1 and REFORCITE2 generally achieve the ratios closest to 1. The third column (row 6–9) shows average diameter for the real data set. Subsequent columns show the ratio between the simulated and observed average diameters. A ratio close to 1 indicates a better model. Similarly, the third column (row 10–13) shows h-index of real networks compared against simulated h-index (column 4–8).

Statistic	Networks	Observed	CP	CPT	FF	REFORCITE1	REFORCITE2
Triangles	Biomedical	6.2×10^6	0.45	0.10	0.60	**1.05**	0.45
	Supreme court	7.1×10^6	0.21	0.24	0.59	0.73	**0.98**
	ArXiv-HepTH	3.4×10^7	0.09	0.35	1.94	0.68	**0.76**
	ArXiv-HepPH	2.3×10^7	0.09	0.56	0.83	0.48	**0.96**
Diameter	Biomedical	57.8	0.51	**1.3**	0.36	0.57	0.56
	Supreme court	10.3	1.14	3.0	**0.87**	1.45	1.7
	ArXiv-HepTH	17.8	0.59	0.66	0.43	**0.92**	1.22
	ArXiv-HepPH	15.0	0.81	0.82	0.74	1.23	**1.01**
H-index	Biomedical	84	82	20	81	94	**84**
	Supreme court	89	85	39	**87**	94	**87**
	ArXiv-HepTH	170	115	61	193	155	**175**
	ArXiv-HepPH	158	125	67	143	175	**160**

Table 5
Mean values of the ratio of simulated to real triangle counts, average diameter and h-index over the lifetime of ArXiv-HepPH network is reported in the table along with standard deviation (STD). A ratio close to 1 indicates a better model driving the simulation. Our proposed models are performing well as compared to CP, CPT, and FF. (Mean ± STD) means 68% models networks would have ration between (Mean − STD) and (Mean + STD), (Mean ± 2STD) means 95% models networks would have ration between (Mean + STD), and (Mean ± STD) means 99.7% models networks would have ration between (Mean − STD) and (Mean + STD). From the discussion, it is observed that in case of CP and CPT it is very rare to get ratio 1, because mean and standard deviation both are very less that results: For CP, (Mean ± 3STD) in to (0.08 to 0.2), (0.6 to 0.84), and (0.454 to 0.506) for triangles, average diameter, and h-index, respectively, For CPT, (Mean ± 3STD) in to (0.584 to 0.596), (0.77 to 0.89), and (0.391 to 0.469) for triangles, average diameter, and h-index, respectively. FF has 68% model networks which have ratio for triangles, average diameter, and h-index in the ranges (0.45 to 0.93), (0.67 to 0.85), and (0.658 to.902), respectively. REFORCITE1 has 68% model networks which have ratio for triangles, average diameter, and h-index in the ranges (0.83 to 1.27), (1.08 to 1.28), and (0.917 to 1.13), respectively. REFORCITE2 has 68% model networks which have ratio for triangles, average diameter, and h-index in the ranges (0.78 to 1.16), (1.01 to 1.17), and (0.845 to 1.075), respectively.

Statistic	Metric	CP	CPT	FF	REFORCITE1	REFORCITE2
Triangles	Mean	0.14	0.59	0.69	1.05	**0.97**
	STD	0.02	0.002	0.24	0.22	0.19
Diameter	Mean	0.79	0.83	0.76	1.18	**1.09**
	STD	0.05	0.03	0.09	0.10	0.08
H-index	Mean	0.49	0.43	0.78	**1.03**	0.96
	STD	0.012	0.013	0.122	0.113	0.115

The bold values represent models that are closest to 1.
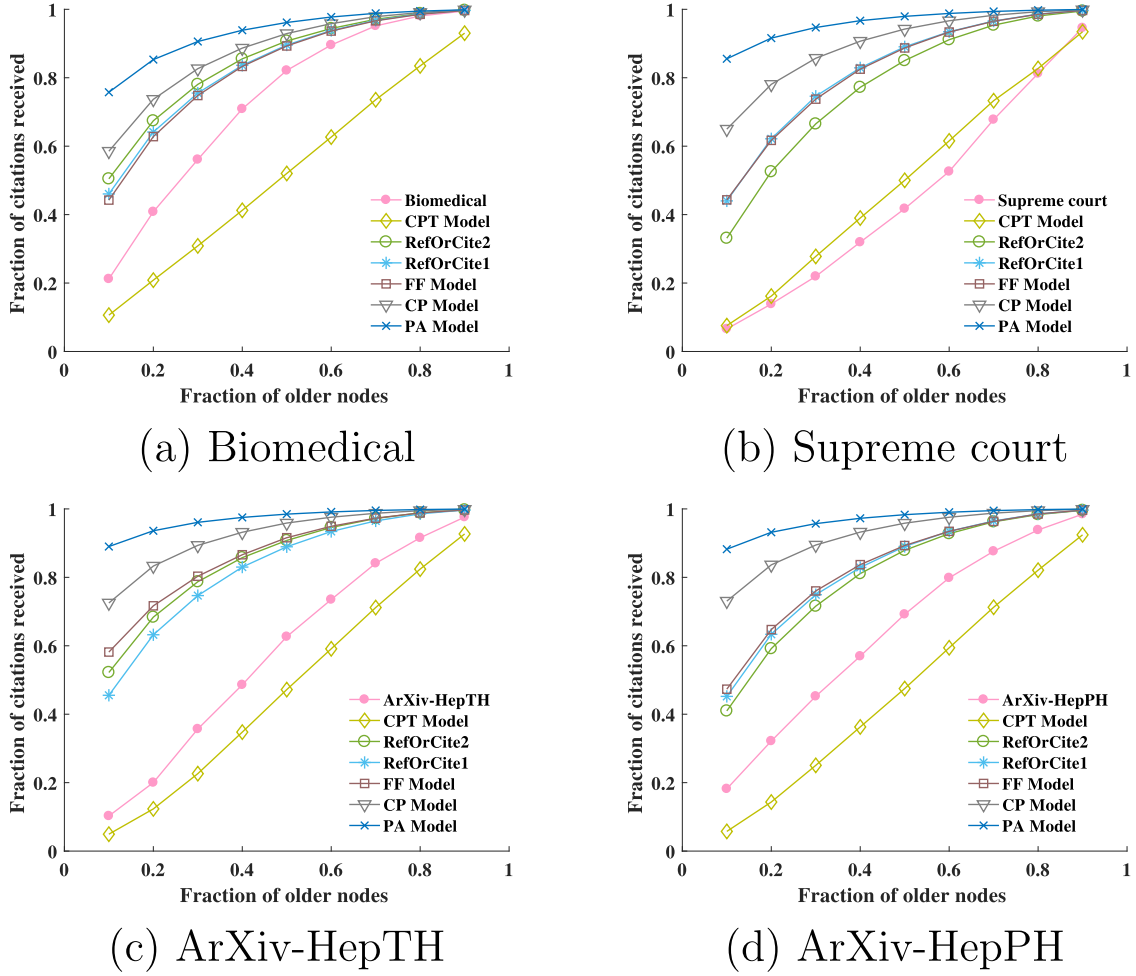
Fig. 10. (Best viewed in color.) Fraction of citations received by the oldest *o* fraction of nodes: (a) Biomedical, (b) Supreme court, (c) ArXiv-HepTH, and (d) ArXiv-HepPH. A comparison among real data (pink), CPT model (yellow), CPT model (gray), BA model (blue), REFORCITE1 (cyan), REFORCITE2 model (green) and FF (violet). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.4. Average diameter

Table 4 reports the average diameter over the lifetime for various real networks in the second column (Observed), where the step size is 5000 in terms of the number of nodes. Similarly, for all the competing models, we compute the diameter of the network at each step, and take an average. The ratio of the average value of the diameter obtained by the model and observed value is shown in the table. In two of the four data sets, REFORCITE variants provide simulated diameters closest to the observed diameters. Note that the CPT model has advantage over the other models because it uses degree sequence from the real dataset. Table 5 shows mean and standard deviation of ratio of simulated and real average diameter for 100 realizations of above models. As expected, we observe significantly low standard deviation values.

3.5. H-index

In Table 4, we report h-index of the real datasets. Additionally, we also compute h-index of the networks obtained under different network models considered in this paper (CP model, CPT model, FF model, REFORCITE1 and REFORCITE2 models). We observe that our models match the real data much better than all the other three models. This result indicates that our model is able to much better replicate the h-index of the network and can therefore find an application in predicting the h-index of authors and journals in future. Table 5 shows mean and standard deviation of ratio of simulated and real h-index for 100 realizations of above models. As expected, we observe significantly low standard deviation values.

3.6. Obsolescence

It is well known that the number of citations to a randomly sampled article does not keep growing over time (Singh et al., 2017; Wang et al., 2009, 2013; Waumans & Bersini, 2016). The rate of acquisition of citations is known to rise to a peak between three and five years (for most communities) and then decline, sometimes sharply. Nevertheless, PA, CP and related models favor the growth of citations to older nodes; age is always an asset and never a liability. This sharply contradicts observed data, where a vast majority of old papers are eventually forgotten. Thus, PA-style models overestimate the popularity of old papers and underestimate the popularity of younger papers.

Singh et al. (2017) proposed a temporal sketch of a network's evolution history that captures obsolescence dynamics. We adapt it slightly for our use here. Consider the $o\%$ oldest nodes, and count their total degree at the end of time. Divide by the total degree over all nodes at the end of time. This ratio r grows with o to a maximum of 1 when 100% of the nodes are included. The more quickly r grows with o, the closer the situation is to PA. In contrast, slower growth of r with increasing o indicates a strong effect of obsolescence.

Fig. 10 shows r-against-o plots for different data sets. For each data set, the real network shows one trajectory. A model is faithful to the obsolescence behavior of the real network if its trajectory is close to the real trajectory. With the exception of CPT, the models closest to the real trajectory are REFORCITE1 and REFORCITE2. By allowing links from new nodes to (more recent) inlinks of the base node, they naturally model obsolescence. In contrast, PA and CP, as expected, confer undue popularity to older nodes (very large r for small o). Curiously, REFORCITE is as good as FF in most cases. Although CPT models obsolescence better than REFORCITE, since it is the only model that incorporates a link probability that depends on paper age, its match with degree distribution and triangle count are far worse than REFORCITE. This under performance of the CPT model can be attributed to the fact that this model is most suitable for networks where the node degrees grow very slowly over time (Krapivsky & Redner, 2005), as opposed to the data sets we study, where the node degrees increase relatively fast.

We do not compare our model with other mechanistic growth models (Singh et al., 2017; Wang et al., 2009) that are primarily composed of PA with an age-based decay component (often exponential), because this mechanism inherently limits triangle formation. Wang et al. (2009) multiply PA's linking probability with an age-based exponential decay term. It suffers from similar limitations of clustering and triangle formation as PA. The exponential decay factor only restricts the growth of degrees of the nodes to incorporate aging.

4. Conclusion and future work

Idealized network evolution models that explain preferential attachment in citation networks are abundant, but only a few analyze citation and reference copying. We present REFORCITE: novel network-driven models to explain triangle formation and obsolescence in real bibliographic networks. We conduct formal analysis of various properties of REFORCITE to establish behavior expected from real networks. Traditional growth models do not fit the real data well, but our REFORCITE models do. Overall, REFORCITE fits the largest number of important network properties better than other proposals.

However, a number of potential limitations remain to be addressed. First, the current study employs relatively small bibliographic datasets. Therefore, we do not claim generic applicability on very large bibliographic networks. In future, we plan to extend this study to other citation networks, for example, patent citation networks. Second, our proposed REFORCITE models do not consider topic or author (collaboration) information which might be relevant in copying citations and references.

Author contributions

Pradumn Kumar Pandey: Conceived and designed the analysis; Collected the data; Contributed data or analysis tools; Performed the analysis; Wrote the paper.

Mayank Singh: Conceived and designed the analysis; Collected the data; Contributed data or analysis tools; Performed the analysis; Wrote the paper.

Pawan Goyal: Conceived and designed the analysis; Wrote the paper; Interpretation of results.

Animesh Mukherjee: Conceived and designed the analysis; Wrote the paper; Interpretation of results.

Soumen Chakrabarti: Conceived and designed the analysis; Wrote the paper; Interpretation of results.

References

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*, 509–512.
Bianconi, G., & Barabási, A. L. (2001). Competition and multiscaling in evolving networks. *EPL (Europhysics Letters), 54*, 436.
Brzezinski, M. (2015). Power laws in citation distributions: Evidence from scopus. *Scientometrics, 103*, 213–228.
Caldarelli, G. (2007). *Scale-free networks: Complex webs in nature and technology*. Oxford University Press.
Chakrabarti, S., Frieze, A. M., & Vera, J. (2005). The influence of search engines on preferential attachment. *SODA*, 293–300.
Dorogovtsev, S. N., & Mendes, J. F. (2013). *Evolution of networks: From biological nets to the internet and WWW*. OUP.
Dorogovtsev, S. N., Goltsev, A. V., & Mendes, J. F. F. (2002). Pseudofractal scale-free web. *Physical Review E, 65*, 066122.
Eswaran, D., Rabbany, R., Dubrawski, A. W., & Faloutsos, C. (2018). Social-affiliation networks: Patterns and the SOAR model. *Joint European conference on machine learning and knowledge discovery in databases*, 105–121.
Fowler, J. H., & Jeon, S. (2008). The authority of supreme court precedent. *Social Networks, 30*, 16–30.

Holme, P., & Kim, B. J. (2002). Growing scale-free networks with tunable clustering. *Physical Review E, 65*, 026107.

Kleinberg, J. M., Kumar, R., Raghavan, P., Rajagopalan, S., & Tomkins, A. S. (1999). The web as a graph: Measurements, models, and methods. *International computing and combinatorics conference*, 1–17.

Krapivsky, P. L., & Redner, S. (2005). Network growth by copying. *Physical Review E, 71*, 036118.

Kumar, R., Raghavan, P., Rajagopalan, S., Sivakumar, D., Tomkins, A., & Upfal, E. (2000). Random graph models for the web graph. *FOCS*, 57–65. http://dlib.computer.org/conferen/focs/0850/pdf/08500057.pdf

Kumar, R., Raghavan, P., Rajagopalan, S., Sivakumar, D., Tomkins, A., & Upfal, E. (2000). Stochastic models for the web graph. *41st annual symposium on foundations of computer science, 2000, Proceedings*, 57–65.

Leskovec, J., & Krevl, A. (2014). *SNAP Datasets: Stanford large network dataset collection*. http://snap.stanford.edu/data

Leskovec, J., Kleinberg, J., & Faloutsos, C. (2007). Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data (TKDD), 1*, 2.

Medo, M., Cimini, G., & Gualdi, S. (2011). Temporal effects in the growth of networks. *Physical Review Letters, 107*, 238701.

Pandey, P. K., & Adhikari, B. (2017). A parametric model approach for structural reconstruction of scale-free networks. *IEEE Transactions on Knowledge and Data Engineering, 29*, 2072–2085.

Price, D. D. S. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science, 27*, 292–306.

Ren, F. X., Shen, H. W., & Cheng, X. Q. (2012). Modeling the clustering in citation networks. *Physica A: Statistical Mechanics and its Applications, 391*, 3533–3539.

Singh, M., Sarkar, R., Goyal, P., Mukherjee, A., & Chakrabarti, S. (2017). Relay-linking models for prominence and obsolescence in evolving networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. pp. 1077–1086. New York, NY, USA: ACM. http://dx.doi.org/10.1145/3097983.3098146

Wang, M., Yu, G., & Yu, D. (2009). Effect of the age of papers on the preferential attachment in citation networks. *Physica A: Statistical Mechanics and its Applications, 388*, 4273–4276.

Wang, D., Song, C., & Barabási, A. L. (2013). Quantifying long-term scientific impact. *Science, 342*, 127–132.

Waumans, M. C., & Bersini, H. (2016). Genealogical trees of scientific papers. *PLOS ONE, 11*, e0150588.

Wu, Z. X., & Holme, P. (2009). Modeling scientific-citation patterns and other triangle-rich acyclic networks. *Physical Review E, 80*, 037101.

Xie, Z., Ouyang, Z., Zhang, P., Yi, D., & Kong, D. (2015). Modeling the citation network by network cosmology. *PLOS ONE, 10*, e0120687.