Physics Letters A 380 (2016) 828–836

Contents lists available at ScienceDirect

Physics Letters A

www.elsevier.com/locate/pla

The role of research efficiency in the evolution of scientific productivity and impact: An agent-based model

Zhi-Qiang You [a,b], Xiao-Pu Han [a,b,*], Tarik Hadzibeganovic [c]

[a] Alibaba Research Center for Complexity Sciences, Hangzhou Normal University, Hangzhou 311121, China
[b] Institute of Information Economy and Alibaba Business College, Hangzhou Normal University, Hangzhou 311121, China
[c] Department of Psychology, University of Graz, 8010 Graz, Austria

ARTICLE INFO

Article history:
Received 15 October 2015
Received in revised form 5 December 2015
Accepted 10 December 2015
Available online 29 December 2015
Communicated by C.R. Doering

Keywords:
Agent-based model
Research efficiency
Academic competition
Productivity and impact
Complex networks

ABSTRACT

We introduce an agent-based model to investigate the effects of production efficiency (PE) and hot field tracing capability (HFTC) on productivity and impact of scientists embedded in a competitive research environment. Agents compete to publish and become cited by occupying the nodes of a citation network calibrated by real-world citation datasets. Our Monte-Carlo simulations reveal that differences in individual performance are strongly related to PE, whereas HFTC alone cannot provide sustainable academic careers under intensely competitive conditions. Remarkably, the negative effect of high competition levels on productivity can be buffered by elevated research efficiency if simultaneously HFTC is sufficiently low.

1. Introduction

Considerable effort has been invested in recent years in understanding the mechanisms that govern the evolution of productivity and impact in science, with some of the major contributions originating from the physics community [1–11]. As a result, several quantitative measures have been proposed over the years to assess productivity and scientific influence of individual researchers, research institutions, or whole nations [12–18].

In this fascinating field of *science of science*, two distinct research niches have been built by physicists: a) network-theoretic analyses of scientific collaboration and citation networks [19,20], conducted largely to understand the topological properties as well as mechanisms that lead to the construction of these networks [21,22], and b) soft-modeling of large datasets by using standard statistical physics tools [23,24], mostly to provide theoretical model fits to a variety of publication and citation distributions and to classify their underlying growth patterns [11,25].

To explain, however, in a more detail, how these distributions emerge in the first place, stochastic process (or urn) models have been developed [26,27]. Power law distributions, for example, are typically explained by a stochastic process involving a growth mechanism and a type of cumulative advantage for those who are already rich in publications and citations, ultimately leading to the well-known rich-get-richer dynamics [28]. Such mathematical models have indeed provided much insight into the formation of patterns that are typically discovered in bibliometric data [29].

Nevertheless, to go beyond a rather simplistic picture of how science works, studying the effects of multiple interacting variables on the publication and citation behavior becomes an unavoidable necessity which mathematically may be too demanding or even analytically intractable, such that agent-based or individual-based simulation models [30–34] remain as the only alternative [29,35]. Moreover, to fully understand the complex dynamics behind scientific publication and citation processes, models need to be employed that not only describe the underlying mechanisms and their interactions, but that can also *generate* empirically realistic distributions of publication and citation counts, the evolution of their corresponding growth processes over time, and the associated topological properties as they are observed in real-world collaboration and citation networks [29].

An important limitation of most previous studies in this area is the fact that the usually analyzed datasets did not contain information about the specific intervening variables, factors that can additionally affect the cumulative advantage of individual scientists [36,37], such as their individual or team research efficiency, skill refinement, variable access to resources, or sudden award-driven reputation emergence [3,6,36,38]. This is where generative agent-based models can help in particular, since they can simulate the

* Corresponding author.
E-mail addresses: xp@hznu.edu.cn (X.-P. Han), tarik.hadzibeganovic@gmail.com (T. Hadzibeganovic).

http://dx.doi.org/10.1016/j.physleta.2015.12.022
0375-9601/

relative contributions of many different covariates that may otherwise be unavailable from real datasets.

In other words, agent-based models can easily produce multiple local interactions and their various underlying mechanisms that are ultimately leading to global-level emergent phenomena [39–41], which are thus captured by the model as they gradually unfold over the course of individual publication and citation events [29]. For decades, these and related advantages of the agent-based technology have been studied and successfully applied by physicists in a wide variety of disciplines [35,42].

In the present paper, we employ a multi-agent modeling framework to investigate the effects of production efficiency (PE) and hot field tracing capability (HFTC) of individual researchers on their productivity and scientific impact in competitive research environments. At the initialization stage, we calibrated our agent-based model by employing two real-world citation datasets: The citation network of the American Physical Society (APS) journals and the condensed matter (Cond-mat) citation network of the arxiv.org online preprint repository. After calibrating our model with these bibliometric datasets, we performed a series of simulation experiments by varying the overall levels of research competition as well as the degrees of HFTC and PE of individual agents who competed to occupy the nodes of a citation network characterized by a finite set of possible research topics.

The two independent variables, PE and HFTC, are generally known as relevant career-enhancing strategies which, to our knowledge, have not been investigated previously in the context of agent-based models, and have generally received very little attention in the studies of publication and citation networks [37]. For example, even though research efficiency is known to play an important role in the evolution of academic careers, the actual magnitude of its effect on productivity and impact as well as its relationship to other career-influencing factors are still unknown.

Individual scientists can plausibly work at different efficiency levels and can be first-movers [43], by publishing the first paper in a relevant discipline (resulting in a great cumulative advantage), and/or followers [44], by tracing and extending already established works from hot fields in science. Nevertheless, as most other innovative and income-driven activities, scientific research is an unabating competition for success and reputation among researchers, communities, and whole nations, which can have positive [45] but also negative consequences [6,46,47].

From our computational experiments, we expected that scientific productivity and impact are influenced by a scientist's research efficiency level (PE), whereas the ability to trace and follow hot research topics (HFTC) alone should not provide sustainable academic careers under fiercely competitive conditions. Moreover, our simulations should lead to a better understanding of efficiency-based inter-individual differences in scientific output and influence, and how these differences can be modified by competition and research topic selection.

2. The model

Our agent-based model captures several aspects of behaviors that naturally emerge in real-world publication and citation networks such as competition, inheritance, directedness, and asymmetry. Agents (authors or research teams) competitively occupy the nodes (publications) of citation networks in which the inheritance process is manifested through the spread of citation relationships among publications and the gradual activation of nodes along the direction of citation relationships, forming thereby directed citation links (e.g. paper A cites paper B, but B may not cite A in return). As a result, the local asymmetry in citation behavior and the global asymmetry in the distribution of publications and citations across agents yields a cumulative advantage for authors who have already published and were already cited in the past.

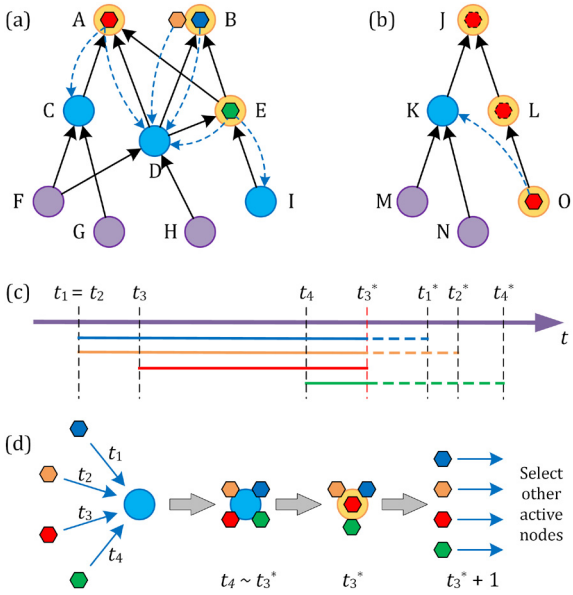

Fig. 1. (Color online.) Node selection and exclusive occupation processes in the model. In panels (a) and (b), yellow, blue and violet circles represent the occupied ("published"), "active" and "inactive" nodes; the black arrows show the citation relationships. The blue, orange, red and green hexagons denote four different agents and the blue dashed arrows represent their possible selections of "active" nodes. Notice that in this example, two agents (the blue and the orange one) compete for the node B and the blue agent finally occupies it. In panel (b), the yellow nodes J and L hosting the red dashed hexagons show the previously occupied nodes of the red hexagon agent currently occupying the node O. Since this agent cannot find any active nodes linked to node O, it has to trace back to its previously published nodes (L and J) until it finds an active node; in this example, one such active node is found in the node J's citation network (node K), and K therefore becomes the occupation target of the red agent in node O (depicted by the blue dashed arrow). Panel (c) depicts the timeline of the "active" node occupation process where the four agents shown in panel (a) all select the node D as their target. The blue, orange, red and green lines (with the dashed regions) show the length of τ^* of the four competing agents. t_1, t_2, t_3 and t_4, respectively, represent the elapsed time steps of the node selection time, and t_1^*, t_2^*, t_3^* and t_4^*, respectively, are the corresponding expected publication times. Since t_3^* is the earliest one of all expected publication times, the red agent occupies the targeted node at this time step and the remaining agents have to terminate their procedures. Finally, all agents then again initiate the selection procedure of new target nodes. These selection and occupation processes are also illustrated in panel (d).

Our model runs on two real-world citation networks of academic publications: The APS and Cond-mat citation networks, with a total of 450,084 and 40,421 of published articles (network nodes) respectively; the detailed description of these citation networks can be found in the Appendix A. The detailed definitions and evolutionary rules of our model are given as follows:

i) Definitions. In our model, each network node can assume one out of three possible status types: An "inactive" status indicates that the node is currently unoccupied and cannot be selected as a research target node of agents, whereas an "active" status signals that the unoccupied node can be selected; all occupied nodes have the "published" status and cannot be selected again.

First, we randomly choose several citation network nodes to be the initial points of the exclusive node occupation process. These initial points are the earliest "foundational" papers (for the APS dataset) or the papers with longest citation chains (for the Cond-mat dataset). The detailed algorithm describing the selection of the initial points is given in Appendix B.

The initial points of a citation network in our model are set as "active", while others remain as "inactive" nodes. All "active"