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.

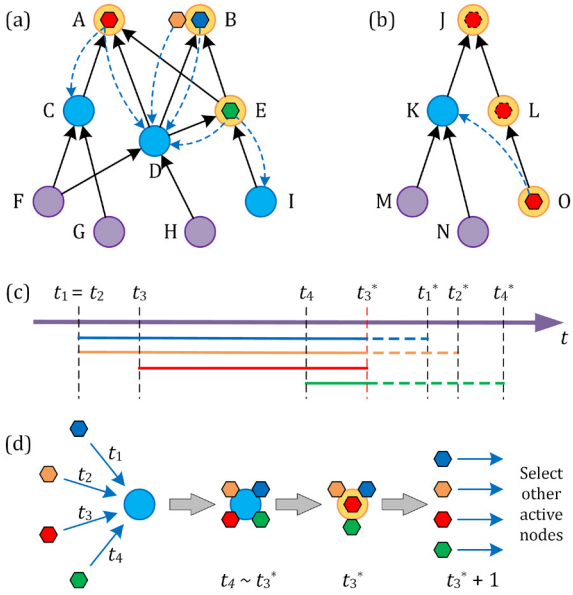

Fig. 1. (Color online.) Node selection and exclusive occupation processes in the model. In panels (a) and (b), yellow, blue and violet circles represent the occupied ("published"), "active" and "inactive" nodes; the black arrows show the citation relationships. The blue, orange, red and green hexagons 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"

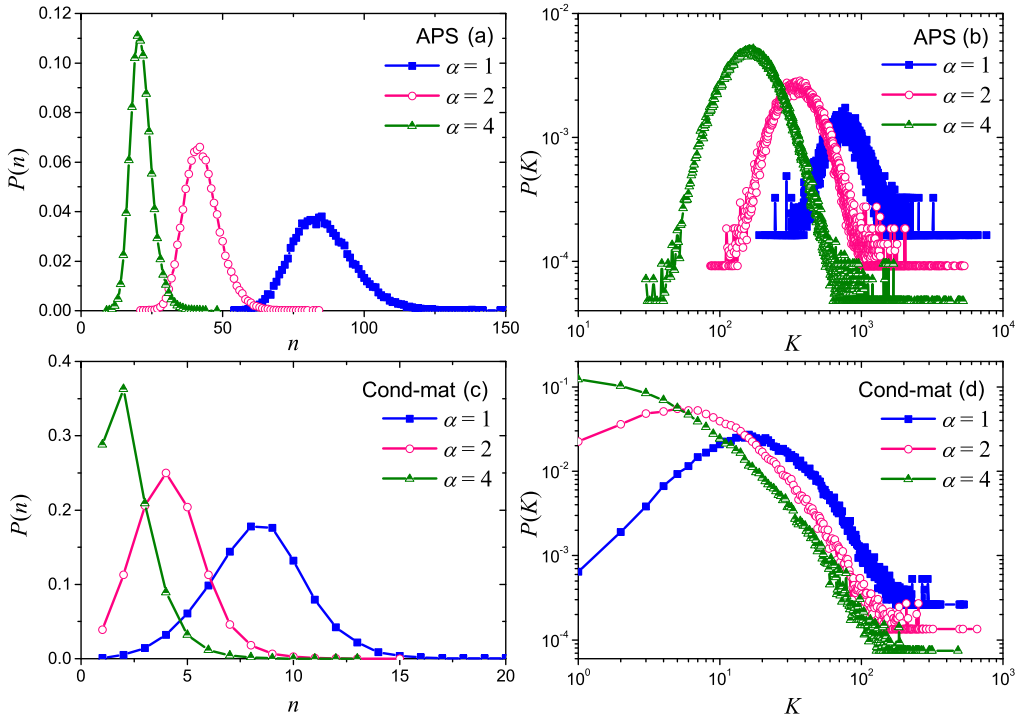

Fig. 2. (Color online.) Distributions of agents' publication outputs and citations. Panels (a) and (b) show the APS case distributions of occupied nodes (publications) $P(n)$ and the corresponding citation numbers (in-degree) $P(K)$, for three different degrees of competition α. Panels (c) and (d) are the corresponding distributions for the Cond-mat dataset, again for three different competition levels.

nodes, once they are occupied by agents (authors), change their status to "published".

At the initialization stage of each simulation, N agents invade the network and compete to occupy its nodes. Each agent i is given a value τ_i describing the typical time needed for completing a given research paper (production efficiency), such that higher τ means lower production efficiency. The values of τ_i are randomly generated and then rounded to integers obeying a Poisson distribution with the average value $\bar{\tau}$. These values are then randomly distributed across agents. Each agent is additionally randomly given a value a ($a > 0$), which is drawn from a Gaussian distribution with the expected value 1 and standard deviation 1. The parameter a describes the ability of an agent to select a currently hot research topic, hereafter denoted as hot field tracing capability (HFTC), with larger a values standing for an agent's ability to select and conduct research on hot and more important research topics.

ii) Node selection process. Each agent selects one active node from its horizon as its future occupation target. At the initial time step, the horizon of each agent includes the set of all active nodes. At time steps $t > 0$, the radius of an agent's horizon includes all active nodes which have cited agent's current node. As illustrated in Fig. 1(a), nodes C and D are both in the horizon of the red hexagon agent occupying the node A. However, if there are no active nodes within an agent's horizon, the horizon is then enlarged by tracing back to the previously selected nodes until there is at least one active node in the citations of the published node. The citations of this published node then become the elements of the enlarged horizon. For example, as shown in Fig. 1(b), the current node O of the red hexagon agent has no active citations, so the agent traces back to its previously selected nodes and finds there the node J which has one active citation (node K); K is then said to be in the enlarged horizon of the agent. Importantly, these backtracking and target selection processes both occur within a single time step of the simulation, irrespective of the length of the total backtracking path.

In its horizon, an active node i is selected by an agent j with probability

$$\Omega_{j\to i} = \frac{(1+k_i)^{a_j}}{\sum(1+k)^{a_j}} \tag{1}$$

where $\sum$ runs over all nodes in the horizon of agent j, and k_i is the real-world citation number (the in-degree) of node i. This particular node selection example describes the node occupation intention of only one agent; however, the case of several agents selecting one and the same target node simultaneously is also allowed.

iii) Exclusive occupation. If agent j selects node i at a time step t, the expected publication time of node i is given by $t^*_{ij} = t + \tau^*_j$, where τ^*_j is a randomly generated integer from a Poisson distribution with average value τ_j. If more than one agent selects the same target node, the one with the earliest t^* will occupy (publish) the target node at time step t^*_{ij} (say agent j). If more than one agent has the same 'earliest' t^*, we randomly select one of them to finally occupy the node. The status of this newly occupied node then turns into "published", and all "inactive" citations of the "published" node will be set to "active". All agents who already selected a given node will repeat the selection procedure to select further active nodes. An example of this process is illustrated in Fig. 1(c) and (d), where four agents compete for one "active" node and the agent with the earliest expected publication time (the red one) wins. The simulation is then further iterated until all network nodes turn their status into "published".

iv) Initial condition settings. In our simulations, the total number of agents is given by $N = \alpha\bar{m}$, where α is the parameter regulating the population density and $\bar{m}$ is the average publication number per year of the employed citation network; for example, $\bar{m} = 5048$ was the average publication number for the APS case from the year 1920, and $\bar{m} = 3663$ for the Cond-mat case from 1995. Since we run our model on two different citation networks (with different sizes and different number of publication years covered), it is better for comparability reasons to use the average

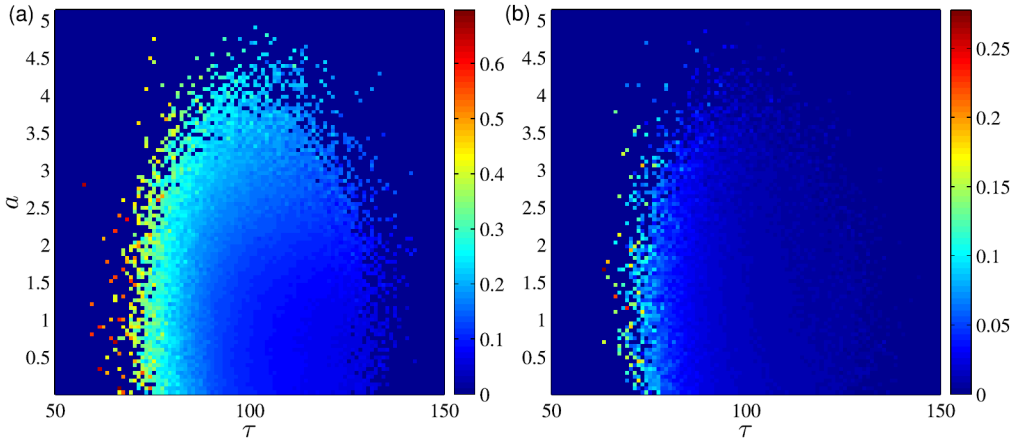

Fig. 3. (Color online.) Node occupation rate patterns for different τ and a (the highest occupation rates are displayed in red color, lowest occupation rates in blue). Panel (a) shows the APS case and panel (b) depicts the Cond-mat case. Simulations for both cases were run with $\alpha = 2.0$, and the results are the averages over 10 independent simulation runs.

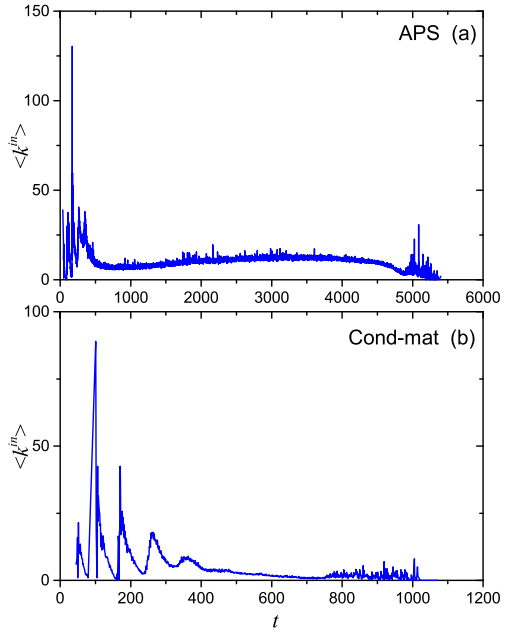

Fig. 4. (Color online.) The average in-degree of occupied nodes as a function of time. The APS case is shown in the panel (a), and the Cond-mat case in panel (b). All simulations were carried out with $\alpha = 2.0$, and the results represent the averages over 10 independent runs.

publication number per year and hence to employ $\bar{m}$ when setting the number of agents in the model. $N = \alpha\bar{m}$ is thus a natural way for setting the total number of agents when running the model on different networks (also see [48,49]). The value of $\bar{\tau}$ is always fixed at 100.

The heretofore specified simulation setup indicates that our model explicitly combines two types of dynamics. One is spreading-like process with preferential target selection, in which the agents select and occupy the network nodes. The other type of dynamics involves the competition between agents targeting the same nodes and the resulting exclusive node occupation. For simplicity, competition among agents (and hence the overall degree of competition in a research environment) was operationalized in our model only via population density, while many other realistic covariates such as the availability of funding and research opportunities, research position supply-demand balance, or length of research contracts, are ignored. Finally, quitting academia was not considered in our model; thus, we assume that all agents perpetually continue to compete for publications and citations in the network they occupy, until all network nodes have changed their status to "published".

3. Simulation results

Our simulations revealed two distinct types of output distributions. More specifically, the distribution of agents' occupied nodes (publications) is Poisson-like (Fig. 2(a) and (c)), whereas agents' citations obey log-normal-like distribution (Fig. 2(b) and (d)), which is generally in agreement with previous empirical observations [50]. The distribution peaks for both investigated cases are shifted leftwards in dependence on α.

We were mostly concerned with the publication outputs and their impacts in dependence on different production efficiencies (parameter τ) and varying hot-field tracing capabilities (parameter a) of agents. In the course of competitive node occupation process, agents with small τ would obviously have a higher probability to occupy the designated target nodes. Since agents with higher a are more likely to choose highly-cited nodes, greater competition levels should occur in these node areas. Thus, it is rather trivial that a fraction of agents with both smaller τ and higher a would clearly generate more outputs with greater impacts. However, the question remains who will have a greater potential to attain higher publication outputs and their corresponding impacts if possessing only a single advantage (i.e. either high a or low τ), and what is the actual magnitude of the effects of these advantages on productivity and influence.

To address this question, we investigate the patterns of the node occupation rate ρ for agents with different τ and a under varying levels of competition α. Here, the node occupation rate ρ of an agent is defined as $\rho = n/n_s$, where n is the total number of occupied nodes in the whole evolution process and n_s is the total number of nodes selected by an agent. As shown in Fig. 3, an agent's node occupation rate is mainly influenced by the production efficiency τ.

Moreover, differences between the APS and the Cond-mat cases are visible: Agents with a larger a have higher node occupation rates in the APS case relative to the Cond-mat case (see Fig. 3). Furthermore, relative to the Cond-mat case, the node occupation rate in the APS case remains relatively high or moderate for a wider range of (also lower) production efficiencies τ. These differences could be of relevance to the comparison of detailed structures of the two studied citation networks, and could be due to several factors. For example, such differences could be attributed to a somewhat stronger heterogeneity of the in-degree distribution in

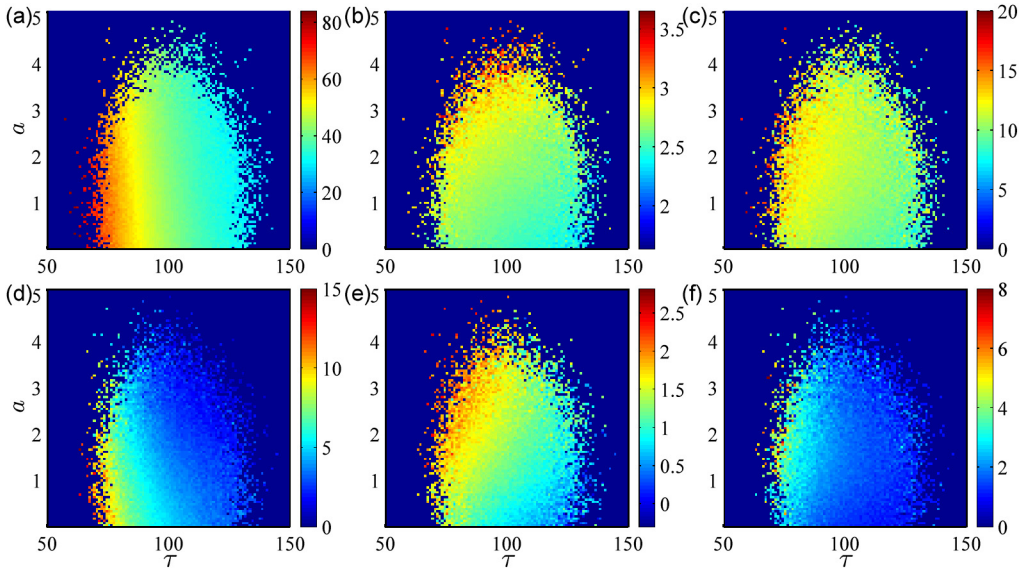

Fig. 5. (Color online.) Patterns of agents' publication outputs n, citations K, and Hirsch indices (h-index) H for different τ and a. The left, middle, and right columns, respectively, show the average number of occupied nodes $\langle n\rangle$ (publication number), the logarithm of the average in-degree of occupied nodes $log\langle K\rangle$ (number of cited records), and the average h-index $\langle H\rangle$. The three upper panels ((a)–(c)) correspond to the simulations calibrated with the APS citation network, whereas the bottom panels ((d)–(f)) show the Cond-mat case. All simulations were carried out with the fixed $\alpha = 2.0$, and the results represent the averages over 10 independent simulations runs.

the Cond-mat citation network relative to the APS case (see Fig. A.7 in the Appendix).

Importantly, Fig. 3 also clearly shows that the effect of a on the occupation rate ρ is generally weaker than that of τ. To explain this somewhat unexpected result, we studied the average in-degree (citations) of newly occupied nodes (publications) as a function of time. As shown in Fig. 4, since highly-cited paper nodes usually attract a greater number of agents (authors), they are occupied much earlier in evolutionary time than other 'less-cited' nodes. The competition for highly-cited papers is obviously much more pronounced, and the competition outcome is consequentially mainly decided by the efficiency level of agents.

We used three measures to calculate agents' productivity and impact in the model: The total number of publications n, total citation records K, and the corresponding Hirsch index H. The calculation of H in our model, that is analogous to the original h-index, is described in Appendix C. Since publications in our model are produced by the occupation of network nodes, the occupation rate ρ directly affects an agent's productivity. Consequentially, the productivity patterns (Fig. 5) show a similar behavior as those observed for the node occupation (see Fig. 3). Moreover, the total number of occupied nodes (total number of publications) for each agent mainly depends upon the value of τ (see the patterns shown in Fig. 5(a) and (d)).

Even though agents with a larger a have a higher probability to select nodes with greater in-degrees (i.e. highly-cited publications), the effect of τ on an occupied node's in-degree K (total citation records) is still clearly evident, as shown in Fig. 5(b) and (e), indicating that agents with both high degrees of research efficiency and research direction selectivity are more likely to obtain stronger citation records. Interestingly enough, the patterns of agents' H-indices display the mixed features of both other pattern types separately shown for n and K (see Fig. 5(c) and (f)), but again, they reveal that τ plays the main role also in determining the value of H of an individual agent.

For further clarity, the effects of τ and a on the measures n, K, and H are plotted functionally for different degrees of competition (controlled by the parameter α). We calculate $\langle n\rangle$, $\langle K\rangle$ and $\langle H\rangle$ for different τ, a and α, where the operator $\langle\rangle$ denotes the value averaged over the agents with the same α, same τ and the same range of a. As shown in Fig. 6(a) and (g), $\langle n\rangle$ as a function of τ rapidly decreases with an increasing τ (i.e. lower efficiency), especially for the agents with better research topic insights ($a > 2$) who select hot research topics and exhibit an intense competition in seizing the nodes with high in-degrees. This decreasing function is steeper in the Cond-mat citation network case relative to the APS case.

In contrast, there is only a slight and insignificant decrease of $\langle n\rangle$ with an increasing a (Fig. 6(d) and (j)), indicating that HFTC cannot strongly affect the total number of publications. Remarkably, we can see in the APS case (Fig. 6(a)) that at a very high efficiency level (i.e. the lowest values of τ), the typically negative effect of high competition α on $\langle n\rangle$ can be buffered if HFTC is sufficiently low (i.e. $a \leq 2$), such that agents can attain very high productivity levels even under fiercely competitive conditions.

The effects of the studied variables on the total citation records $\langle K\rangle$ show some pertinent differences. As shown in Fig. 6(b), (e), (h) and (k), the pattern of the effects of τ and a on $\langle K\rangle$ is similar, since a directly affects agents' selection preferences for target node citation numbers. The calculated H-index also mainly depends upon the efficiency level τ (it significantly decreases with lower efficiency, i.e. higher τ), while it only slightly increases with growing a, with stronger fluctuations observed in the simulated Cond-mat case (Fig. 6(c), (f), (i) and (l)).

The curves corresponding to different values of α in the APS case are almost parallel over much of the τ range with rather small fluctuations at very low and very high efficiency levels (Fig. 6(c)), indicating here that the global intensity of competition actually has a rather trivial effect on $\langle H\rangle$, without interactions with other variables. Under conditions with higher α, due to the more intense competition, some agents would have to select more marginally influential nodes because the highly-cited ones are usually already occupied earlier in evolutionary time, and thus the impact of the intense competition in highly-cited nodes is weakened.

In summary, we can conclude that an agent's production efficiency plays the dominant role in determining both its scientific output and impact measures. Thus, even slight reductions of the total paper completion time could significantly boost agents' performance, implying that a fast completion of a given research assignment is of great importance to all involved collaborators.

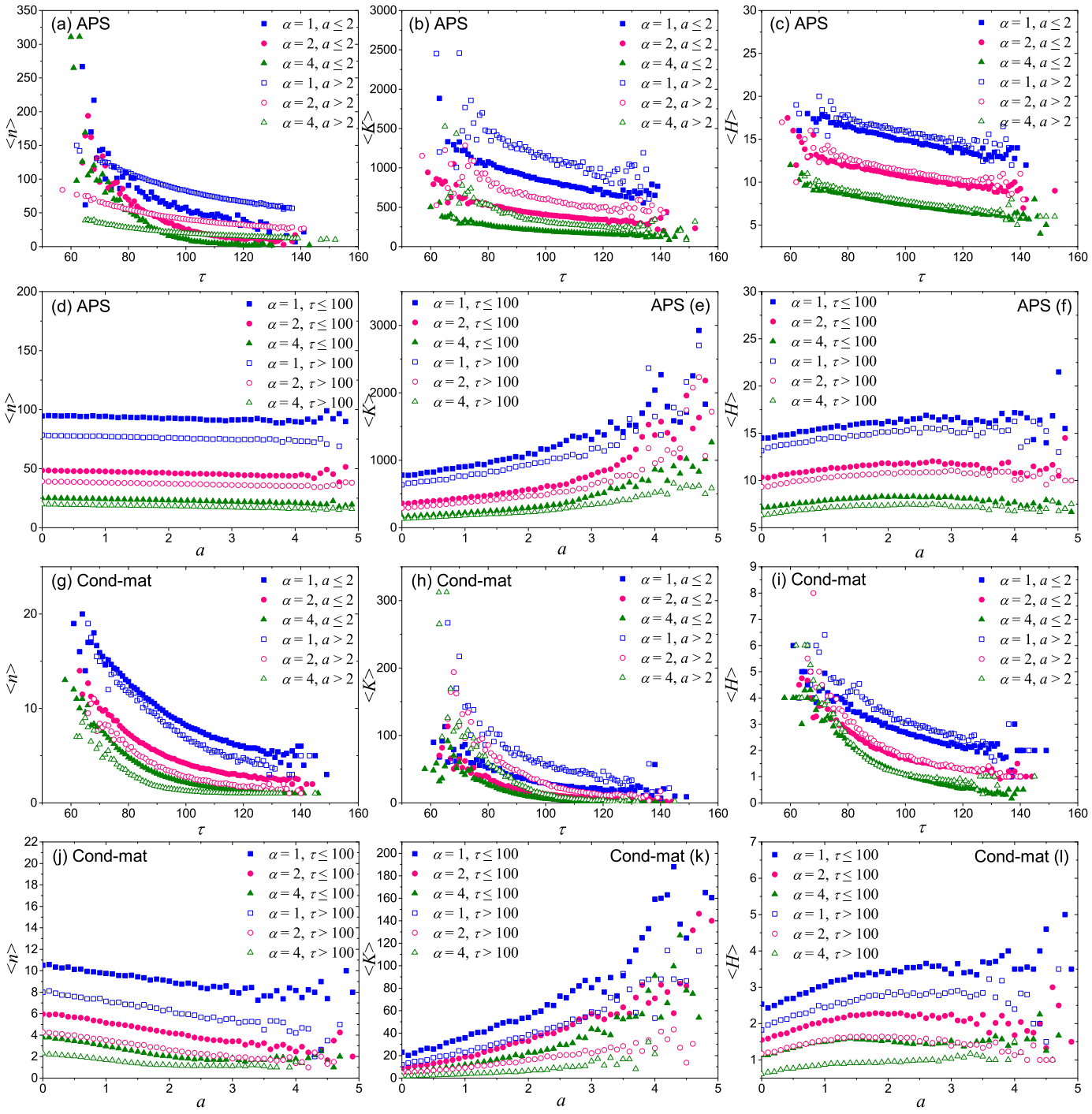

Fig. 6. (Color online.) Average productivity and impact as a function of τ, a, and α. Panels in the left, middle and right columns are the average numbers of occupied nodes $\langle n \rangle$, the average in-degree of occupied nodes $\langle K \rangle$, and the average H-index $\langle H \rangle$, respectively. The upper six panels (a)–(f) show the APS citation network case, and the bottom panels (g)–(l) correspond to the Cond-mat case.

4. Discussion

In this paper, we have studied the effects of research efficiency and research topic selectivity on scientific productivity and impact of individual agents embedded in variously competitive research environments. To this effect, we employed an agent-based computational model that was calibrated by using publication and citation records of the American Physical Society journals and the condensed matter (Cond-mat) citation network of the arxiv.org preprint repository. In our model, agents with different research efficiencies and research topic selection capabilities invaded the initially calibrated artificial citation network, and then iteratively competed to occupy its nodes in order to become increasingly published and cited agent-authors.

Our simulation experiments revealed that the work efficiency strongly affects agents' academic outputs and impacts under a wide variety of conditions. Research direction selectivity (HFTC), on the other hand, plays a less important role, since our findings indicate that a selection of hot research topics alone cannot provide sustainable academic careers under intensely competitive conditions. Remarkably, we observed that the negative effect of competition on productivity can be buffered by higher research ef-

ficiency if simultaneously HFTC is sufficiently low, indicating that agents with different HFTC levels do not seem to equally benefit from work efficiency in a highly competitive environment.

Overall, these findings suggest that even scientists without perceptive insights into which research direction to choose would still be able to attain high-level career achievements by adhering strongly to hard but efficient work. These findings are additionally corroborated by a recent empirical report [43], suggesting that to become influential, it may be much better for a scientist to produce the first instead of the best paper in a given research area. However, in the ever-growing competition in academia, to be the first-mover it is necessary not only to be creative and have interesting novel ideas, but also to be able to publish on time and before the other competing peers.

In our study, we addressed three widely and perhaps most frequently employed measures of scientific productivity and impact: Total number of publications, total citation counts, and the well-known h-index, which combines productivity and impact of a scientist into a single metric [12]. However, we note that even though such indices have progressively been used to compare individual scientists [51], e.g. when assessing their grant or job applications, it has increasingly been advised against an unreflexive use of such quantitative measures as the only tool for judging scientific achievement [52,53]. Indeed, consensus has yet to be achieved on the extent to which a single measure should be administered to evaluate the actual quality of science and to influence policy decisions [54]. Consequentially, as the debate around science performance indicators and their wide-spread usage continues to thrive, much further research [55] and critical reflection [56] will be necessary to arrive at meaningful and cross-disciplinarily valid [7] conclusions.

In a sense, our model is similar to the epidemic spreading process [57] on directed weighted networks, but the difference is that all agents in our model are always active and can jump to nonadjacent nodes in some cases (e.g. in the backtracking process), mimicking to some extent repeated medium- and long-range travels which are also important in the study of epidemics [48]. However, differently from a typical spreading model, we do not include the reproduction-like process of agents whose number thus remains fixed throughout the simulation. Moreover, traditional spreading models usually do not consider multiple types of pathogens, whereas in our model, each agent can be interpreted as an independent 'pathogen' type. Importantly, in standard models of contagion transmission, an infected node can typically infect multiple neighbors synchronously. In contrast, agents in our model can select and occupy network nodes only asynchronously, one by one.

For simplicity, we only kept the very few and basic elements of a real-world academic competition process in our model while excluding several important factors that can also affect publication and citation behaviors, such as team work efficiency, the availability of financial and human capital resources, changes in individual reputation, scientists' relocations, or sudden shifts in institutional research policy [36,37]. Moreover, other aspects of research efficiency such as its sustainability over longer periods of time (relative to shorter, burst-like productivity phases), or the effects of the number of topics covered in the produced work and the associated degree of interdisciplinarity of a paper, rightfully deserve special attention in subsequent studies. These factors will be considered in more realistic future modifications of the present model.

In addition, since earlier theoretical and empirical reports were typically limited to the analysis of only modestly large datasets (and usually only within a single discipline), it remains a grand challenge for further research to employ large-scale agent-based models [58–60] and calibrate them with much greater amounts of data in order to enable comparisons with real-world citation networks originating from many different scientific disciplines. Similarly to what is envisioned in the prestigious Human Brain Project [61], which employs exascale computer simulations of the human brain, investing related efforts in the area of scientometrics could eventually lead to the development of predictive large-scale agent-based models that would enable us to peek into the much farther stages of the evolution of science. Such models could help us gain a multi-level trans-disciplinary understanding of how science works and how its future may look like in the world of the ever growing scientific [37] and economic inequality [62].

Future extensions of our model could also contain algorithms with more explicitly detailed evolutionary features, such as the population of reproducing authors who continuously leave and enter academia, mutations of research strategies, and various selection pressures on working agent-scientists. Furthermore, to make the agent-based publication and citation process models more realistic, one could also include sophisticated rules for collaborator and research topic selection, performance measures beyond mere publication and citation numbers (e.g. recency, paper length, or number of coauthors), the artificial peer review process, and special reviewer selection rules [29].

In summary, we proposed an agent-based modeling framework to investigate two factors affecting academic outputs of individual researchers – production efficiency and hot field tracing capability. The majority of researchers is knowingly aware of the significance of production efficiency, consequentially making great efforts in their daily research works; however, they still have to face many extra tasks and administrative duties causing frequent project completion delays and efficiency shocks. Our results highlight the importance of maintaining and optimizing the work efficiency in academia, such that reducing scientists' efficiency gaps could have profound effects on boosting their academic success.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grants Nos. 11205040, 11205042, and 11305043), the research startup fund of Hangzhou Normal University, the European Commission FP7 Project GROWTHCOM (Grant No. 611272), and the CCF-Tencent Open Research Fund.

Appendix A. Data description

Two datasets of real-world citation networks are used in our model: The data of APS journals (available online: https://publish.aps.org/datasets) and the Cond-mat dataset of arxiv.org preprint repository (available online: http://www-personal.umich.edu/~mejn/netdata/).

The APS dataset covers all publications in Physical Review, Physical Review Letters, and Reviews of Modern Physics. It includes the information on each paper's date of acceptance, publication time, author information, citations, citation relationships, etc. The dataset is comprised of a total of 450,084 articles with 4,692,056 citation relationships from 1893 to 2009. The distributions of degree, in-degree and out-degree of citation relationships are shown in the top panel of Fig. A.7.

Cond-mat dataset contains all preprints uploaded between January 1, 1995, and March 31, 2005, to the condensed matter electronic print archive arxiv.org, containing a total of 40,421 papers and 175,692 citations. The citation network degree distributions of the Cond-mat dataset are shown in the bottom panel of Fig. A.7.

Appendix B. Initialization stage

At the initialization stage of our model, several nodes were set as "active" in the first time step $t = 0$. For the model calibration

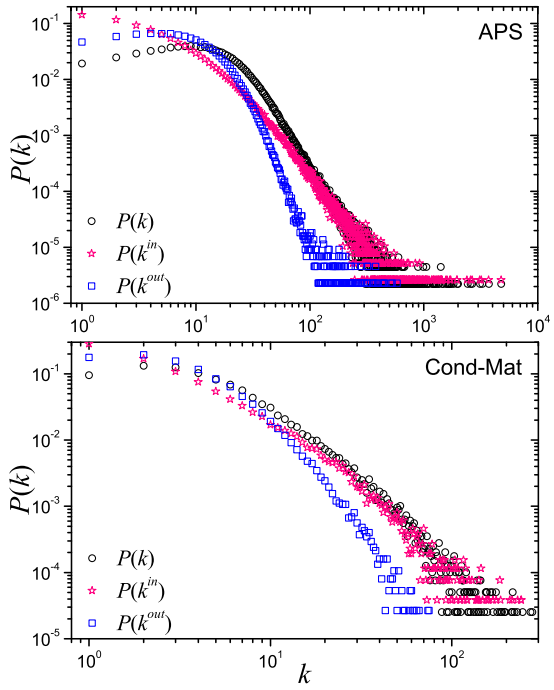

Fig. A.7. (Color online.) Degree distributions of the two real-world citation networks: (a) APS and (b) Cond-mat.

with the APS dataset, we selected the papers published before 1920 (a total of 809 articles) and their citations to represent the initially active nodes.

For the model calibration with the Cond-mat dataset, since it does not include temporal information of articles, the initial active nodes were selected in accordance with the following two conditions: The node is in the head of a citation chain, i.e. it does not cite any papers in the dataset, and the total length of the chain is longer than 20 papers (max. 28). Under these conditions, 124 articles in Cond-mat dataset were selected as the initially active nodes in our simulated network.

Appendix C. Calculation of H-index

The h-index is a widely employed measure for quantification of a scientist's academic performance [12]. This index can be defined in the following way: An individual scientist has an index H if H of her/his N_p published articles have at least H citations each, while the rest of $(N_p - H)$ articles have no more than H citations each [12]. The value H is then the h-index of a scientist.

In our model, the in-degrees (i.e. real-world citation numbers) of an agent's occupied nodes are treated as citations of agent's papers, and thus in the calculation of the agent's h-index we firstly rank all of its occupied nodes in a descending order of their in-degree. We then find a rank H of an agent by satisfying the following conditions: The in-degree of an occupied node with rank H satisfies $k_H^{in} \geq H$, and for the node with rank $H+1$, its in-degree $k_{H+1}^{in} < H+1$. This value H is then the h-index analogue of an agent in our model.

References

[1] D.J. de S. Price, Networks of scientific papers, Science 149 (1965) 510–515.
[2] C. Tsallis, M.P. de Albuquerque, Are citations of scientific papers a case of nonextensivity?, Eur. Phys. J. B 13 (2000) 777–780.
[3] A. Mazloumian, Y.-H. Eom, D. Helbing, S. Lozano, S. Fortunato, How citation boosts promote scientific paradigm shifts and Nobel prizes, PLoS ONE 6 (2011) e18975.
[4] D. Wang, C. Song, A.-L. Barabási, Quantifying long-term scientific impact, Science 342 (2013) 127–132.
[5] R.K. Pan, K. Kaski, S. Fortunato, World citation and collaboration networks: uncovering the role of geography in science, Sci. Rep. 2 (2012) 902.
[6] A. Petersen, M. Riccaboni, H.E. Stanley, F. Pammolli, Persistence and uncertainty in the academic career, Proc. Natl. Acad. Sci. USA 109 (2012) 5213–5218.
[7] F. Radicchi, S. Fortunato, C. Castellano, Universality of citation distributions: toward an objective measure of scientific impact, Proc. Natl. Acad. Sci. USA 105 (2008) 17268–17272.
[8] R. Carvalho, M. Batty, The geography of scientific productivity: scaling in U.S. computer science, J. Stat. Mech. (2006) P10012.
[9] A. Petersen, F. Wang, H.E. Stanley, Methods for measuring the citations and productivity of scientists across time and discipline, Phys. Rev. E 81 (2010) 036114.
[10] M.E.J. Newman, Prediction of highly cited papers, Europhys. Lett. 105 (2014) 28002.
[11] K. Matia, L.A.N. Amaral, M. Luwel, H.F. Moed, H.E. Stanley, Scaling phenomena in the growth dynamics of scientific output, J. Am. Soc. Inf. Sci. 56 (2005) 893–902.
[12] J.E. Hirsch, An index to quantify an individual's scientific research output, Proc. Natl. Acad. Sci. USA 102 (2005) 16569–16572.
[13] M. Ausloos, Assessing the true role of coauthors in the h-index measure of an author scientific impact, Physica A 422 (2015) 136–142.
[14] L. Egghe, Theory and practise of the g-index, Scientometrics 69 (2007) 131–152.
[15] B. Jin, L. Liang, R. Rousseau, L. Egghe, The R- and AR-indices: complementing the h-index, Chin. Sci. Bull. 52 (2007) 855–863.
[16] M. Ausloos, A scientometrics law about co-authors and their ranking: the co-author core, Scientometrics (2013) 1–15.
[17] J.-F. Molinari, A. Molinari, A new methodology for ranking scientific institutions, Scientometrics 75 (2008) 163–174.
[18] R.M. May, The scientific wealth of nations, Science 275 (1997) 793–795.
[19] M.E.J. Newman, The structure of scientific collaboration networks, Proc. Natl. Acad. Sci. USA 98 (2001) 404–409.
[20] E. Mones, P. Pollner, T. Vicsek, Universal hierarchical behavior of citation networks, J. Stat. Mech. (2014) P0502.
[21] A.L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, T. Vicsek, Evolution of the social network of scientific collaborations, Physica A 311 (2002) 590–614.
[22] M. Tomassini, L. Luthi, Empirical analysis of the evolution of a scientific collaboration network, Physica A 385 (2007) 750–764.
[23] L.A.N. Amaral, P. Gopikrishnan, K. Matia, V. Plerou, H.E. Stanley, Application of statistical physics methods and concepts to the study of science and technology systems, Scientometrics 51 (2001) 9–36.
[24] Y.H. Eom, S. Fortunato, Characterizing and modeling citation dynamics, PLoS ONE 6 (2011) e24926.
[25] M. Medo, G. Cimini, S. Gualdi, Temporal effects in the growth of networks, Phys. Rev. Lett. 107 (2011) 238701.
[26] H.A. Simon, On a class of skew distribution functions, Biometrika 42 (1955) 425–440.
[27] Q.L. Burrell, Hirsch's h-index: a stochastic model, J. Informetr. 1 (2007) 16–25.
[28] R.K. Merton, The Matthew effect in science, Science 159 (1968) 56–63.
[29] C. Watts, N. Gilbert, Does cumulative advantage affect collective learning in science? An agent-based simulation, Scientometrics 89 (2011) 437–463.
[30] R. Conte, M. Paolucci, On agent-based modeling and computational social science, Front. Psychol. 5 (2014) 668.
[31] C.-X. Yang, R. Wang, S. Hu, Modeling and analysis of an agent-based model for Chinese stock market, Phys. Lett. A 377 (2013) 2041–2046.
[32] E. Bruch, J. Atwell, Agent-based models in empirical social research, Sociol. Methods Res. 44 (2015) 186–221.
[33] G. Yang, C.G. Zhu, K.N. An, J.P. Huang, Overall fluctuations and fat tails in an artificial financial market: the two-sided impact of leveraged trading, Phys. Lett. A 379 (2015) 1857–1863.
[34] X.-K. Meng, C.-Y. Xia, Z.-K. Gao, L. Wang, S.-W. Sun, Spatial prisoner's dilemma games with increasing neighborhood size and individual diversity on two interdependent lattices, Phys. Lett. A 379 (2015) 767–773.
[35] E. Bonabeau, Agent-based modeling: methods and techniques for simulating human systems, Proc. Natl. Acad. Sci. USA 99 (2002) 7280–7287.
[36] P. Azoulay, T. Stuart, Y.M. Wang, Mathew: effect or fable?, Manag. Sci. 60 (2014) 92–109.
[37] A.M. Petersen, O. Penner, Inequality and cumulative advantage in science careers: a case study of high-impact journals, EPJ Data Sci. 3 (2014) 24.
[38] J. Duch, X.H.T. Zeng, M. Sales-Pardo, F. Radicchi, S. Otis, T.K. Woodruff, L.A.N. Amaral, The possible role of resource requirements and academic career-choice risk on gender differences in publication rate and impact, PLoS ONE 7 (2012) e51332.
[39] R.L. Goldstone, M.E. Roberts, T.M. Gureckis, Emergent processes in group behavior, Curr. Dir. Psychol. Sci. 17 (2008) 10–15.
[40] T. Hadzibeganovic, D. Stauffer, C. Schulze, Agent-based computer simulations of language choice dynamics, Ann. N.Y. Acad. Sci. 1167 (2009) 221–229.
[41] R.L. Goldstone, M.A. Janssen, Computational models of collective behavior, Trends Cogn. Sci. 9 (2005) 424–430.
[42] J.D. Farmer, D. Foley, The economy needs agent-based modelling, Nature 460 (2009) 685–686.
[43] M.E.J. Newman, The first-mover advantage in scientific publication, Europhys. Lett. 86 (2009) 68001.

[44] T. Wei, M. Li, C. Wu, X.Y. Yan, Y. Fan, Z.R. Di, J.S. Wu, Do scientists trace hot topics?, Sci. Rep. 3 (2013) 2207.
[45] P. Stephan, How Economics Shapes Science, Harvard University Press, 2012.
[46] M.S. Anderson, E.A. Ronning, R. De Vries, B.C. Martinson, The perverse effects of competition on scientists' work and relationships, Sci. Eng. Ethics 13 (2007) 437–461.
[47] J. Couzin-Frankel, Chasing the money, Science 344 (2014) 24–25.
[48] X.-P. Han, Z.D. Zhao, T. Hadzibeganovic, B.H. Wang, Epidemic spreading on hierarchical geographical networks with mobile agents, Commun. Nonlinear Sci. Numer. Simulat. 19 (2014) 1301–1312.
[49] N. Masuda, Effects of diffusion rates on epidemic spreads in metapopulation networks, New J. Phys. 12 (2010) 093009.
[50] R.K. Pan, S. Fortunato, Author Impact Factor: tracking the dynamics of individual scientific impact, Sci. Rep. 4 (2014) 4880.
[51] P. Ball, Achievement index climbs the ranks, Nature 448 (2007) 737.
[52] R. Costas, M. Bordons, The h-index: advantages, limitations and its relation with other bibliometric indicators at the micro level, J. Informetr. 1 (2007) 193–203.
[53] A. Abbott, D. Cyranoski, N. Jones, B. Maher, Q. Schiermeier, R. Van Noorden, Metrics: do metrics matter?, Nature 465 (2010) 860–862.
[54] S. Lehmann, A.D. Jackson, B.E. Lautrup, Measures for measures, Nature 444 (2006) 1003–1004.
[55] J. Lane, Let's make science metrics more scientific, Nature 464 (2010) 488–489.
[56] R. Van Noorden, Metrics: a profusion of measures, Nature 465 (2010) 864–866.
[57] X.-P. Han, Disease spreading with epidemic alert on small-world networks, Phys. Lett. A 365 (2007) 1–5.
[58] F. Boulaire, M. Utting, R. Drogemuller, Dynamic agent composition for large-scale agent-based models, Complex Adapt. Syst. Model. 3 (2015) 1.
[59] T. Hadzibeganovic, F.W.S. Lima, D. Stauffer, Evolution of tag-mediated altruistic behavior in one-shot encounters on large-scale complex networks, Comput. Phys. Commun. 183 (2012) 2315–2321.
[60] H.R. Parry, Agent based modeling, large scale simulations, in: R.A. Meyers (Ed.), Computational Complexity, Springer Verlag, New York, 2012, pp. 76–87.
[61] E.R. Kandel, H. Markram, P.M. Matthews, R. Yuste, C. Koch, Neuroscience thinks big (and collaboratively), Nat. Rev. Neurosci. 14 (2013) 659–664, https://www.humanbrainproject.eu/.
[62] T. Piketty, Capital in the Twenty-First Century, Harvard University Press, 2014.