

Predicting the Evolution of Scientific Output

Antonia Gogoglou^(✉) and Yannis Manolopoulos

Department of Informatics, Aristotle University, 54124 Thessaloniki, Greece
{agogoglou,manolopo}@csd.auth.gr

Abstract. Various efforts have been made to quantify scientific impact and identify the mechanisms that influence its future evolution. The first step is the identification of what constitutes scholarly impact and how it is measured. In this direction, various approaches focus on future citation count or *h*-index prediction at author or publication level, on fitting the distribution of citation accumulation or accurately identifying award winners, upcoming hot research topics or academic rising stars. A plethora of features have been contemplated as possible influential factors and assorted machine-learning methodologies have been adopted to ensure timely and accurate estimations. Here, we provide an overview of the field challenges, as well as a taxonomy of the existing approaches to identify the open issues that are yet to be addressed.

Keywords: Scientometrics · Bibliographic data · Predictive modeling

1 Introduction

With the extensive recording of scientific progress on the Web and the emergence of large scale open source as well as proprietary databases of bibliographic data (Google Scholar, Web of Science, Scopus, etc.), the quantification and evaluation of scientific impact, the “science of science” [2], has attracted significant attention. In particular, research institutions, universities and even countries have been adopting research policies that emphasize “excellence” and “impact”. In this direction, rigorous efforts have been made to extract meaningful and actionable information from the abundance of bibliometric data to produce rankings, aid decision making and assist peer review. This is evident by the plethora of bibliometric indices that have been proposed in the past decade, since the seminal paper by Hirsch introducing the *h*-index [13], and all attempts to quantify different aspects of scientific impact [39]. The high level of correlation amongst the majority of these indices has been extensively investigated [4,32]. However, the focus is now turning towards the quantification of future impact and rising influence, instead of measuring existing output in different ways.

In his preliminary work, Price deduces that current visibility, publishing venue and age highly influence a publication’s future outreach [35]. In today’s fast paced, ever growing and interdisciplinary research world what determines future influence? Is it possible to provide early estimation using current data? These are

intriguing questions for all stakeholders of the scientific community, from individual scholars to publishers and from funding agencies to hiring committees, as current decisions on tenure, grant allocation and publishing are based on an inherent estimation of future evolution. Identification of future trends, advancing topics and trend shapers or influentials in the science world are examples of efficient utilization of predictive analytics and, therefore, they are attracting attention from both public and private sector, with the number of related publications rising every year. Thompson Reuters, a game changer corporation in publishing, has created the InCite platform (<https://incites.thomsonreuters.com>) for mapping and ranking scientists and their output, as well as identifying “hot” publications or up and coming research avenues. In any case, efficient and meaningful approximation of future trends can provide invaluable tools to stakeholders of the scientific world, to better coordinate research endeavors, utilize funds and create connections that will improve visibility and productivity.

Hitherto, due to difficulties in obtaining reliable and abundant data, the scientific community was largely reliant on judgement of experts to evaluate future potential of a publication or a scholar. However, with the increasing data availability and advances in big data mining, the need for computerised support in decision making has come up, given that peer review can prove to be costly and time-consuming. In addition, peers will use their own knowledge and expertise to formulate judgement leading to more conservative views not receptive to novelty. On the other hand, data intelligence, which has been utilized in various disciplines like marketing, business, security, etc. [9], can overcome personalized criteria and provide evidence based valuable insights to assist in strategy planning. Figure 1 demonstrates a workflow describing the general process for deriving actionable data intelligence from available bibliographic data.

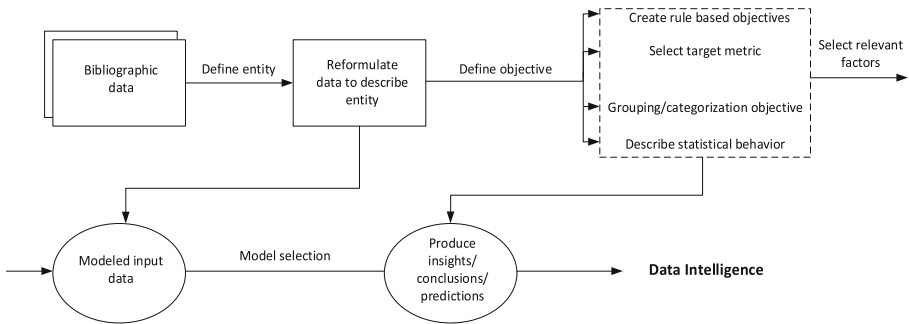


Fig. 1. Workflow from available bibliographic data to actionable data intelligence.

In the present work, we focus on the question: *Is scientific progress quantifiable and predictable?* To riddle this question, we provide an overview of existing approaches and a taxonomy of them based on their common qualities. Additionally, we identify the remaining open research issues and challenges in this area as well as the dangers that arise from the quantification of scientific evolution.

2 Taxonomy of Approaches

There exist multiple approaches to quantify the evolution of scientific impact; they can be categorized with regards to the *scientific entity concerned*, their *modeling approach* and their *target metric* (see Tables 1 and 2 in the next).

Scientific entity: An initial category stems from the entity under evaluation: *publication*, *author*, *venue* or *institution*. Most efforts focus on publications, because for the other three categories one needs to aggregate the respective entire portfolio of publications (author, venue or institution), thus increasing the calculation complexity. Also, complete information about a publication is usually available at several online databases, whereas for the other entities there is a high disambiguation amongst different online entities to ensure that the complete records are retrieved (e.g. names, abbreviations, etc.).

Target variable: Another approach is related to the target variable: e.g. the *citation count* taken as a proxy for impact or a bibliometric index, such as the *h-index*. Due to the exponential distribution of these quantities and the debate regarding their crude limiting nature, a set of works has defined the prediction problem in an alternative way to mitigate the skewness of the predicted output. For instance, in [6] the yearly rise in citation is estimated, in [11] the predictive question is whether a publication will contribute to the rise of the first author's *h-index*, whereas Garner et al. foresee how quickly the first citation of a publication will occur [12]. To avoid the heavy tailed distribution of target variables, which often inhibits the effectiveness of the model, the relative rank of a scientific entity in a network can be predicted instead. In [21] the rank of a publication is compared to all other publications in the same discipline, while in [5] it is compared against the journal publications of the same year. Approaches inspired by network analytics include [30,37], where variations of a future PageRank value are the calculated target. As shown in Table 1, the target of the prediction may also entail an award [33] or a specified position in a scholar's career [36], while in [29] the question at hand is identifying Nobel prize winners.

Modeling approach: Several approaches have been proposed to calculate the evolution of the scientific impact over time.

- *Classification* based models, where a set of predefined categories have been constructed to characterize the current state of a bibliometric quality and measure the changes to occur after a time period. Then, by assigning a new entity into one of the existing categories, its future state is approximated by that of the entire category. Even though this approach manages the diversity of scientific patterns and distinguishes amongst them effectively, placing an entity in a particular cohort only establishes how its current behavior resembles its peers; limited predictability is offered for its future state, which may significantly deviate from the group (e.g. sleeping beauties).
- *Regression* based approaches have been introduced with the seminal work by Acuna [1] and others [20,36]. This methodology has been criticized since

Table 1. Classification of approaches for estimating future impact

Categorization attribute	Examples of each category	Related work
Scientific entity	Publication	[3, 5–8, 10–12, 16, 21, 25, 37, 40]
	Author	[1, 17, 20, 23, 28, 29, 31, 34, 36, 37, 41]
	Venue	[6]
	Institution	[36]
Target variable	h -index	[1, 31]
	Citation count	[7, 12, 16, 20, 23, 28, 40]
	Increase in h -index/citations	[11]
	Shift in impact group	[6, 34]
	Relative ranking	[3, 5, 8, 17, 21, 31, 34]
	Rank position in a network	[21, 25, 37, 41]
	Award or promotion	[29, 33, 36]
Model	Classification	[5, 6, 8, 11, 23]
	Regression	[1, 3, 12, 16, 17, 20, 21, 36]
	Statistical modeling	[29, 34]
	Time series	[28]
	Citation networks	[25, 37, 41]
	Combination of the above	[7, 10, 40]

its predictability depends on the aggregation of career data across multiple age cohorts, leading to unfair models towards young researchers or “late bloomers” [24]. Therefore, recent endeavors combine entities to subsequently calculate regression coefficients individually for each group [7].

- *Statistical modeling*, inspired by social networks evolution and Web modeling, attempts to fit bibliometric quantities to existing distributions, thus approximating the mechanism they will continue to evolve over time. Logarithmic and exponential distributions have been fitted to the evolution of productivity and impact over a scholar’s career, whereas parameter thresholds have been utilized to predict impact shifts [29, 40]. In [34] Sinatra et al. produced the random impact model, according to which the highest impact publication may occur randomly at any point of a career and future popularity can be calculated based on a multiplicative process of the impact exponent. Although interpretable, statistical modeling approaches require an abundance of past data to calculate the model parameters, thus discouraging the quantification of future evolution for young researchers or publications. In general they oversimplify when characterizing the complex process of citation dynamics using a distribution model alone even with a plethora of parameters. Thus, the challenge becomes prominent at the author level, where the interactions amongst different models produce the final output.
- *Time series prediction* constitutes another alternative, since citation acquisition is a temporal process. By viewing scientific entities as spatio-temporal

objects [23, 28] one can approximate its future trajectory based on an abundance of data from various past time slots.

- *Network approaches*, where one can consider a citation network, where each link represents a vote of confidence between researchers or publications, and therefore determining the future state of such a network constitutes a *link prediction* problem [30, 41].
- *Combining* two or more of the previous methodologies has proven to yield increased performance, like [10], where a time series classification of publications occurs or [5] where classification of publications is combined with threshold based distribution modeling.

To estimate future evolution, of pivotal importance are the selected features that shape scientific impact. Redundant or irrelevant factors can cause overfitting or add unnecessary complexity to the produced model, while on the other hand failing to account for crucial factors the effectiveness, accuracy and usability of the prediction are threatened. These factors are grouped into six categories: *author centric*, *publication centric*, *content related*, *venue-centric*, *socially derived* and *temporal* ones. The outreach of any scientific contribution is determined in part by who is working to make it, his/her scientific track record, how s/he is trained as scientists and how long s/he is engaged with research. Other features that influence his/her output include the gender, the country of origin and the faculty position. Furthermore, with an increasing productivity, one increases the chances for scientific recognition. The same effect is achieved with interdisciplinary research that merges different domains together [38].

A number of features can determine the future of a publication, ranging from its topic allocation to the number of used keywords or the time of appearance. A high quality work may end up under-appreciated if it gets published in a year that ground-breaking achievements are happening in the same field or analogously if a field has started losing its overall popularity. Based on the “standing on the shoulders of giants” motto, it is expected that high quality works cite other high quality works, thus making the number of cited references a relevant predictor. Also, a high number of co-authors often results in a wider dissemination increasing its probability to be cited. A rising interest has focused recently in mining the actual content of a publication: e.g. the terms used, the position of references, the ordering of author names as well as the originality and diversity of the subject in an effort to provide more detailed and specific predictions.

In today’s prestige-based interconnected world the social characteristics surrounding a publication and its authors are also highly defining factors of future impact, with the authority and networking power of the author being the most popular social features utilized in predictive modeling. It is understood that a well-connected scholar, with a large collaboration network, who also refers to other seminal works or is part of a highly respected institution will be able to better publicize his/her work. The same holds for the venue where a publication is released, with top rated venues attracting usually high quality publications and also providing a broader audience for the released work. However, a large variety of publishing patterns occur in research, as mentioned previously, raising

Table 2. Categories of features to characterize scientific impact and its evolution.

Feature origin	Features
Author-centric	Popularity and productivity
	Academic age and gender
	Affiliation and academic position
	Rank based on bibliometric indices
	Disciplines/domains
Publication-centric	Popularity and age
	Number of references and keywords
	Topic allocation and number of authors
	Relative ranking in portfolio or field
Content-related	Popularity of topic and novelty
	Positioning of references and author ordering
	Diversity of topic and MeSH terms
	Abstract content
Venue-centric	Journal Impact Factor (JIF)
	Popularity of publishing house
	Ranking in international lists (JCR, Scimago, etc.)
	Number of issues/volumes
	Open access
Social	Collaboration network
	Citation network
	Co-citation/bibliographic coupling
Temporal	Differences from past state
	Normalization over time frame
	Rate of activity within a time frame

the need for temporal evaluation of scientific output. To identify rising stars and upcoming trends, there needs to be an accurate prediction based on the timing of the discovery and not only its calculated magnitude. Also, considering the rate at which the status of a scholar or publication rises can provide more insight into the future output, than his/her static current state. The categorization of these features and examples of each category of factors are presented in Table 2.

3 Challenges

With the rising abundance and complexity in bibliographic data the science of science is focusing on measurable quantities regarding scientific output: for instance citations to past work, timing of scientific discoveries and events in career trajectories such as promotions, awards, reaching top percentile of impact

amongst a group and many others. Using computational tools one can identify quantitative patterns in these events that present a straightforward metric to predict, but also raise controversy regarding fair and meaningful predictions. Many of these quantities, like the number of citations or the h -index, are heavily subject to *preferential attachment*, meaning that the majority of the scientific community achieves low scores in these metrics, with a selected few attracting significant attention. Due to the Matthew effect in citation counting [22] a number of scientists have altered the definition of the prediction problem at hand, aiming for example to predict whether a publication will contribute to a scholar's rise in h -index values [11] or his/her relative ranking amongst a group of peers [21, 37], instead of addressing the future citation count prediction per se. Scientists also argue that the inert property of citations and citation based metrics to be always accumulating creates false self-fulfilling predictive models [31]. Consequently, adjusted metrics have been utilized, like the number of citations each publication receives every year which tends to be a decreasing quality [30].

Another challenge is the prediction of the timing in which a shift in the citation pattern will occur (e.g. a scientific discovery, a seminal publication, etc.) as opposed to predicting the magnitude of one's impact. Studies have concluded that the timing of scientific output is rather random compared to the more predictable impact of this output [34]. Given that extraordinary temporal patterns appear, like "sleeping beauties" [27] indicating publications that receive recognition after a long period of time or "premature discoveries" [18], the provocative issue of the ageing of scientific output rises: Does an abrupt boost in citations mean recognition and how long before scientific work becomes obsolete? In [15, 27] threshold based and parameter free methods are proposed respectively to differentiate different ageing patterns for publications, while in [8] six categories of publication trajectories are identified with individual predictive models trained for each one of them and achieving different levels of prediction performance [7].

Studies [14, 17, 34] also suggest that the early years in a scientist's career provide the appropriate circumstances for one's seminal publication. However, this pattern could be highly related to the tendency of younger researchers to be more productive compared to more mature ones. While most of the proposed models address future impact of existing work, the problem of predicting future impact of future works that have not yet been published, is a real controversy [20]. Similarly to the link prediction problems in complex networks [19], predicting a link to an existing node is a challenging but addressable issue, while predicting the introduction of a new node and its connections is progressively harder [16, 25]. In the same direction, predicting highly cited scholars or publications is often a different problem from identifying the truly innovative ones that will break new ground in a field or will shape a new research domain. It has been pointed out that publications that conform to the mainstream within a field get cited more often than novel original works [16].

In general, many different publishing patterns are present across various disciplines, countries and academic levels, thus a predictive model that would be

fair towards all groups of scientists is hard to create [24]. A set of works has undertaken the creation of different models for individual groups, e.g. publications of a specific journal [5], scholars of a given domain, academic age or position [17, 36] to limit the variety of social processes that can lead to increased scientific output. These publishing patterns do not remain steady over time and a model created on a specific dataset in a limited time frame may contain significant bias.

The identification of such patterns becomes increasingly difficult given the interaction of the complex networks formulated in the scientific life, collaboration amongst scholars, affiliation of scholars, topics of a publication, citation links between publications and authors to name a few. These networks are interconnected and their evolution is co-dependent, affecting the evolution of science in non-obvious ways [26]. An additional challenge arises when considering long term vs. short term impact, with [3, 23] explaining that different factors influence early predictions, whereas long term success is more complicated. However, it has been argued that short term impact is more important since it may influence the whole career trajectory of a scholar or the fate of a new publication [3].

4 Discussion and Future Research Directions

Despite the wide range of proposed approaches to quantify the future of science, a unified framework for all levels and patterns of scholarly impact is still missing. Instead of focusing on a single metric, the various aspects of scientific output need to be taken into account to produce an overall fair framework accounting for young scientists, for truly novel out of the ordinary research ideas and under-represented fields. Additionally, such a predictive framework needs to mitigate the drawbacks of the aforementioned approaches by combining them effectively and creating a resulting model that is robust to manipulation, meaning it should not encourage greed and strategic networking over the advancement of science.

There is a rising belief in the scientometric community that more timely and accurate predictions can occur from incorporating context-specific data, such as the length and content of papers, the terms used and their relationship to the topic. Also, integration with online presence and social media dissemination (posts, number of downloads, number of views, etc.) has given birth to the rising field of Altmetrics, which measures scientific outreach in today's digital era more effectively compared to accumulated citations. Most importantly, the data bias introduced by each online database, with its different properties and coverage range, significantly hinders the comparisons amongst introduced approaches. The need for a detailed diverse ground truth dataset that is widely accepted by the scientific community is imminent.

Finally, the performance of predictive frameworks relies heavily on what constitutes proven high impact, an award, a high h -index or a tenure position? A set of selected criteria accepted by the computing community and peer review that efficiently evaluate recognition needs to be utilized as a target variable for proposed frameworks. Given that a great deal of controversy has surrounded

both bibliographic data and impact metrics during the past decades, the challenge to create a common basis for evaluation of predictive efforts in the science of science remains open.

References

1. Acuna, D.E., Allesina, S., Kording, K.P.: Future impact: predicting scientific success. *Nature* **489**(7415), 201–202 (2012)
2. Börner, K., Dall’Asta, L., Ke, W., Vespignani, A.: Studying the emerging global brain: analyzing and visualizing the impact of co-authorship teams. *Complexity* **10**(4), 57–67 (2005)
3. Bornmann, L., Leydesdorff, L., Wang, J.: How to improve the prediction based on citation impact percentiles for years shortly after the publication date? *J. Inf.* **8**(1), 175–180 (2014)
4. Bornmann, L., Mutz, R., Hug, S.E., Daniel, H.P.: A multilevel meta-analysis of studies reporting correlations between the h index and 37 different h index variants. *J. Inf.* **5**(3), 346–359 (2011)
5. Brizan, D.G., Gallagher, K., Jahangir, A., Brown, T.: Predicting citation patterns: defining and determining influence. *Scientometrics* **108**(1), 183–200 (2016)
6. Cao, X., Chen, Y., Liu, K.R.: A data analytic approach to quantifying scientific impact. *J. Inf.* **10**(2), 471–484 (2016)
7. Chakraborty, T., Kumar, S., Goyal, P., Ganguly, N., Mukherjee, A.: Towards a stratified learning approach to predict future citation counts. In: Proceedings 14th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL), pp. 351–360 (2014)
8. Chakraborty, T., Kumar, S., Goyal, P., Ganguly, N., Mukherjee, A.: On the categorization of scientific citation profiles in computer science. *Commun. ACM* **58**(9), 82–90 (2015)
9. Chaudhuri, S., Dayal, U., Narasayya, V.: An overview of business intelligence technology. *Commun. ACM* **54**(8), 88–98 (2011)
10. Davletov, F., Aydin, A.S., Cakmak, A.: High impact academic paper prediction using temporal and topological features. In: Proceedings 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM), pp. 491–498 (2014)
11. Dong, Y., Johnson, R.A., Chawla, N.V.: Can scientific impact be predicted? *IEEE Trans. Big Data* **2**(1), 18–30 (2016)
12. Garner, J., Porter, A.L., Newman, N.C.: Distance and velocity measures: using citations to determine breadth and speed of research impact. *Scientometrics* **100**(3), 687–703 (2014)
13. Hirsch, J.E.: An index to quantify an individual’s scientific research output. *Proc. Natl. Acad. Sci.* **102**(46), 16569–16572 (2005)
14. Jones, B.F., Weinberg, B.A.: Age dynamics in scientific creativity. *Proc. Natl. Acad. Sci.* **108**(47), 18910–18914 (2011)
15. Ke, Q., Ferrara, E., Radicchi, F., Flammini, A.: Defining and identifying sleeping beauties in science. *Proc. Natl. Acad. Sci.* **112**(24), 7426–7431 (2015)
16. Klimek, P.S., Jovanovic, A., Egloff, R., Schneider, R.: Successful fish go with the flow: citation impact prediction based on centrality measures for term-document networks. *Scientometrics* **107**(3), 1265–1282 (2016)
17. Laurance, W.F., Useche, D.C., Laurance, S.G., Bradshaw, C.J.: Predicting publication success for biologists. *Bioscience* **63**(10), 817 (2013)

18. Li, J., Shi, D., Zhao, S.X., Ye, F.Y.: A study of the “heartbeat spectra” for “sleeping beauties”. *J. Inf.* **8**(3), 493–502 (2014)
19. Lü, L., Zhou, T.: Link prediction in complex networks: a survey. *Phys. A Stat. Mech. Appl.* **390**(6), 1150–1170 (2011)
20. Mazloumian, A.: Predicting scholars’ scientific impact. *PLoS ONE* **7**(11), 1–5 (2012)
21. McNamara, D., Wong, P., Christen, P., Ng, K.S.: Predicting high impact academic papers using citation network features. In: Li, J., Cao, L., Wang, C., Tan, K.C., Liu, B., Pei, J., Tseng, V.S. (eds.) *PAKDD 2013. LNCS*, vol. 7867, pp. 14–25. Springer, Heidelberg (2013). doi:[10.1007/978-3-642-40319-4_2](https://doi.org/10.1007/978-3-642-40319-4_2)
22. Merton, R.K.: The Matthew effect in science. *Science* **159**(3810), 56–63 (1968)
23. Nezhadbiglari, M., Gonçalves, M.A., Almeida, J.M.: Early prediction of scholar popularity. In: *Proceedings 16th ACM/IEEE-CS on Joint Conference on Digital Libraries (JCDL)*, pp. 181–190 (2016)
24. Penner, O., Pan, R.K., Petersen, A.M., Fortunato, S.: The case for caution in predicting scientists’ future impact. *Phys. Today* **66**(4), 8 (2013)
25. Pobiedina, N., Ichise, R.: Citation count prediction as a link prediction problem. *Appl. Intell.* **44**(2), 252–268 (2016)
26. Pradhan, D., Paul, P.S., Maheswari, U., Nandi, S., Chakraborty, T.: C3-index: revisiting author’s performance measure. In: *Proceedings 8th ACM Conference on Web Science (WebSci)*, pp. 318–319 (2016)
27. van Raan, A.F.J.: Sleeping beauties in science. *Scientometrics* **59**(3), 467–472 (2004)
28. Revesz, P.Z.: A method for predicting citations to the scientific publications of individual researchers. In: *Proceedings 18th International Database Engineering and Applications Symposium (IDEAS)*, pp. 9–18 (2014)
29. Revesz, P.Z.: Data mining citation databases: a new index measure that predicts Nobel prizewinners. In: *Proceedings 19th International Database Engineering and Applications Symposium (IDEAS)*, pp. 1–9 (2015)
30. Sayyadi, H., Getoor, L.: Futurerank: ranking scientific articles by predicting their future pagerank. In: *Proceedings SIAM International Conference on Data Mining (SDM)*, pp. 533–544 (2009)
31. Schreiber, M.: How relevant is the predictive power of the *h*-index? A case study of the time-dependent Hirsch index. *J. Inf.* **7**(2), 325–329 (2013)
32. Sidiropoulos, A., Gogoglou, A., Katsaros, D., Manolopoulos, Y.: Gazing at the skyline for star scientists. *J. Inf.* **10**(3), 789–813 (2016)
33. Sidiropoulos, A., Manolopoulos, Y.: A citation-based system to assist prize awarding. *ACM SIGMOD Rec.* **34**(4), 54–60 (2005)
34. Sinatra, R., Wang, D., Deville, P., Song, C., Barabási, A.: Quantifying the evolution of individual scientific impact. *Science* **354**(6312), aaf5239 (2016)
35. de Solla Price, D.J.: Networks of scientific papers. *Science* **149**(3683), 510–515 (1965)
36. Vieira, E.S., Cabral, J.A., Gomes, J.A.: How good is a model based on bibliometric indicators in predicting the final decisions made by peers? *J. Inf.* **8**(2), 390–405 (2014)
37. Wang, S., Xie, S., Zhang, X., Li, Z., Yu, P.S., He, Y.: Coranking the future influence of multiobjects in bibliographic network through mutual reinforcement. *ACM Trans. Intell. Syst. Technol.* **7**(4), 64:1–64:28 (2016)
38. Way, S.F., Morgan, A.C., Clauset, A., Larremore, D.B.: The misleading narrative of the canonical faculty productivity trajectory. *CoRR* abs/1612.08228 (2016)

39. Wildgaard, L., Schneider, J.W., Larsen, B.: A review of the characteristics of 108 author-level bibliometric indicators. *Scientometrics* **101**(1), 125–158 (2014)
40. Xiao, S., Yan, J., Li, C., Jin, B., Wang, X., Yang, X., Chu, S.M., Zhu, H.: On modeling and predicting individual paper citation count over time. In: Proceedings 25th International Joint Conference on Artificial Intelligence (IJCAI), pp. 2676–2682 (2016)
41. Zhang, J., Ning, Z., Bai, X., Wang, W., Yu, S., Xia, F.: Who are the rising stars in academia? In: Proceedings 16th ACM/IEEE-CS on Joint Conference on Digital Libraries (JCDL), pp. 211–212 (2016)