Measuring Science in our Highly Digitized World

Yannis Manolopoulos
Open University of Cyprus
Nicosia, Cyprus
yannis.manolopoulos@ouc.ac.cy

Dimitrios Katsaros
University of Thessaly
Volos, Greece
dkatsar@e-ce.uth.gr

ABSTRACT

During the past two decades the availability of scholarly data repositories such as Google Scholar, Elsevier Scopus offered tremendous opportunities to the field of scientometrics to analyze the dynamics of science and help design indicators to evaluate the scientific impact of scholars, journals and academic institutions. Thus, scientometrics transformed fast into the exciting new discipline of Science of Science, which seeks to understand, quantify and predict research activities and the corresponding outcomes. In this article we provide a brief coverage of very recent work related to scientific ranking and impact prediction, aim at deepening our faith to Science of Science and stimulating further efforts in this transdisciplinary field.

CCS CONCEPTS

• **General and reference** → **Metrics**; **Evaluation**; • **Human-centered computing** → **Collaborative and social computing design and evaluation methods**; **Social network analysis**; *Social engineering (social sciences)*;

KEYWORDS

Scientometrics, h-index, citation prediction, Science of Science

ACM Reference Format:
Yannis Manolopoulos and Dimitrios Katsaros. 2018. Measuring Science in our Highly Digitized World. In *22nd Pan-Hellenic Conference on Informatics (PCI '18), November 29-December 1, 2018, Athens, Greece.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3291533.3291547

1 MEASURING MODERN SCIENCE

It is common ground that modern science progresses in a very rapid and much different way that it used to during the past few decades. A significant factor that facilitated this evolution is the development of digital infrastructures for the storage and dissemination of scientific results. These infrastructures record a variety of metadata related to science 'products' (i.e., patents, articles), such as their creators and the affiliations of them, their impact as a function of time, their funding and so on. The easy, immediate access – e.g., from one's desktop computer – to this wealth of information created a growing wave of interest to methods for analyzing this vast amount of data. This interest was strongly motivated by the growing need to develop automated methods for evaluating (ranking) entities related to the 'production' of science. The main such entities are individual scientists, publication fora (journals, conferences), organizations (universities, research institutions) or even specialized academic programs such as doctoral studies. These rankings are used for decision support when it comes to hire or promote academic personnel, to allocate grants, to choose universities for undergraduate or graduate studies, to select to which journal to submit to scientific work, and so on.

The availability of these voluminous scholarly data and the development of computerized analysis methods are driving the area of Science of Science (SoS), which seeks to understand, quantify and even predict research activities and the corresponding outcomes [7, 38]. Even though Science of Science (SoS) draws concepts and methodologies from scientometrics, informetrics, network science, sociology, we can somehow consider it as a dramatic evolution of Scientometrics, which studies the impact of publications, researchers, journals and tries to model the scientific collaboration, to establish concrete comprehension of innovation and predict future evolution of science. If we try to provide a synopsis of what is modern scientometrics about, then we would list the following topics:

- (Static) Data: scientific publication data, funding data, patent data, collaboration network data, citation network data.
- (Dynamics of) publications/citations/collaborations: preferential attachment growth, aging effect, sleeping beauties, team formation/assembly.
- Scientific significance: citation measures, spectral centralities measures, scientist/journal/university/country ranking, credit allocation.
- Prediction in science: impact prediction, link prediction.
- Success in science: interdisciplinary research, funding, collaborators.
- Innovation and knowledge diffusion: patents, co-authorships, researchers mobility.

This article aims at presenting recent progress in scientometrics, in particular from the perspective of measuring and predicting scientific impact (third and fourth bullets). The rest of this article is structured as follows: Section 2 will briefly cover the issues and advances in quantifying scientific significance; Section 3 will survey the topic of predictive modeling in science, and finally Section 4 will conclude this article.

2 QUANTIFICATION OF SCIENTIFIC IMPACT

Quantification of scientific impact concerns individual articles, scientists, journals, study programs, institutions, and whole countries. Almost all these impact measurement methods are based on the

PCI '18, November 29-December 1, 2018, Athens, Greece

ACM ISBN 978-1-4503-6610-6/18/11…$15.00
https://doi.org/10.1145/3291533.3291547

evaluation of the impact of individual articles which are related to the entity being evaluated. Therefore, we will focus on article's impact quantification methods and we will present the most established methods and also some recent promising techniques.

A straightforward way to measure impact is by merely counting number of received citations. The famous *Impact Factor* [9] and its Scopus' alternative *CiteScore index* are estimators of the average impact of an article published in a specific journal measured during a time window (two years for the IF versus three years for CiteScore). IF-type of indicators are the basis for evaluating journals, and they are used by librarians to decide subscriptions, or perspective authors to submit their works. Newer methods have been based on centrality measures; for instance, Scimago Journal Rank [1] (SJR) is based on the famous PageRank spectral centrality [17] and it accounts for both the number of citations received by a journal and the importance or prestige of the journals where such citations come from.

Naive measures for scientists' performance quantification include the number of published articles, number of citations received, average number of citations per published article, and so on. A significant deviation from that line of indicators was the introduction of the now famous h-index [13]; h-index is actually a proxy for both productivity and impact, and lower bounds the total number of citations received by a scientist. h-index was the first indicator in a row of many similar in spirit indices, such as our own *contemporary h-index* [26], *f index* [14], *Perfectionism index* [27], the fractal-based index [11], and the *skyline-based* indicators [25, 30]. The observation that author centrality in collaboration networks is strongly correlated to author impact led to proposing the estimation of author impact by applying variations of (personalized) PageRank such as the AuthorRank [18]. All the aforementioned indicators treat the citation network as a single layer network.

Recently, network science introduced novel network types with richer modelling capabilities, where the interacting entities are assumed to belong to more than one network, called *layer*. These networks are called multiplex [6], multisliced [19], multilevel [33], interdepedent [5], or in general multilayer [4, 16]. Despite the fact that this modeling approach offers huge potential to better exploit articles' metadata, only a handful of works addressed the issue of developing scientometrics indicators for multilayer networks. AuthorPaper rank (APrank) index [40] – similar to the idea of Hyperlinked Induced Topic Search (HITS) algorithm [17] – interweaved paper and author quality into a recursive definition: a paper is of high quality if it is cited by prestigious scientists and that high-quality papers raise the prestige of their authors. Another notable effort is the P-Rank [36]. Observing that none of these methods employed native spectral centralities for multilayer networks as impact indicators, we proposed the use of Biplex PageRank [21] to measure scientist's impact. In that context, we introduced a new index, namely C^4-index [24] aiming at identifying scientists with the most outstanding articles, while at the same time paying attention to their consistent focus on producing excellent results and also showing the biggest potential. There are even proposals for extending h-index for multilayer networks, the so-called *Minimal-layers PCI* [3].

[1] https://www.scimagojr.com/

With respect to ranking institutions, there exist three well-respected lists, namely ARWU[2], The Times Higher Education[3], and QS World University Rankings[4]. These ranking systems take into account apart from citations per faculty, the graduate employability, faculty/student ratios, proportion of international faculty, etc. Finally, there exist efforts to rank university doctoral programs. Ranking lists for this purpose based on peer review are highly reputable for the USA education system, i.e., the *US News* ranking of computer science doctoral programs[5]. Recently, a study developed a computerized methodology for the same goal, i.e., produce rankings of doctoral programs [31]. The ranking results of this methodology correlated strongly with that of peer assessment, using a new h-index variant, namely the $t10$-index which is simply the number of citations of a faculty member's 10-th most cited paper. Similar efforts but at a smaller scale employed scientometrics indices to evaluate the impact of PhD graduates of departments of computer science/engineering of Greek universities [15].

3 IMPACT PREDICTION IN SCIENCE

Forecasting future impact of a scientist is a very interesting problem because it allows for making effective hiring/promotion decisions and research fund allocation. Impact prediction has mainly focused on paper citation prediction. Since a citation network is an always expanding network, theories developed for network growth in the context of network science have been used to attack this prediction problem. The famous preferential attachment model along with linear extrapolation was applied [39] for citation prediction. The accuracy of that model was good for the short term, because fast aging of papers and diversity in their quality turned it inadequate for the long term. In a famous paper [32], a preferential attachment model along with fitness and aging parameters was introduced that exhibited extraordinary prediction accuracy.

Scientist impact prediction is a more complicated task because there exist proposals for several hundred scientometric indicators and even though they are based on the standard impact measure, i.e., citations, only a couple of them have gained acceptance by the scientific community. The h-index – despite the critique against it [37] – seems to have gained adequate acceptance. Thus, several works attempted predicting future h-index; the first work was that described in [2], which developed a linear regression model and established a set of equations (Equations 1–3) for predicting h-index value for one, five and ten years ahead, based on the following parameters: n is the number of articles written by the author, h is the current h-index of the author, y is years since the author published the first article; j is the number of distinct journals in which the author has published, and q is the number of the author's top journal articles.

$$h_{+1} = 0.76 + 0.37\sqrt{n} + 0.97h - 0.07y + 0.02j + 0.03q \quad (1)$$

$$h_{+5} = 4 + 1.58\sqrt{n} + 0.86h - 0.35y + 0.06j + 0.2q \quad (2)$$

$$h_{+10} = 8.73 + 1.33\sqrt{n} + 0.48h - 0.41y + 0.52j + 0.82q \quad (3)$$

[2] http://www.shanghairanking.com/
[3] https://www.timeshighereducation.com/
[4] https://www.topuniversities.com/qs-world-university-rankings
[5] https://www.usnews.com/best-graduate-schools/top-science-schools/computer-science-rankings

Subsequent research work questioned the validity of these equations [8, 22]. A more successful prediction model was developed in [28] based on the earlier model in [32]. However, the main conclusion of these and of contemporary works is that it is pretty difficult to make predictions for young scholars for the long term. This is a disappointing result since this is exactly the essence of the scientist's future impact prediction problem.

More recent works focused on predicting rising stars [20], on incorporating the identify of early citers into the estimation of future impact [29], or providing more elaborate machine learning models for performance prediction [10, 34, 35].

In recent years, deep artificial neural networks [23] have won numerous contests in pattern recognition and machine learning. These neural networks equipped with powerful training algorithms developed in the context of deep learning [12] and trained with huge volumes of data have started to be used as tools for long term citation prediction [1]. Despite the fact that simple feedforward neural networks are not able to realize in an effective manner the input sequence order resulting in less prediction accuracy, recurrent neural networks are far more promising alternatives.

4 CONCLUSIONS

The deluge of scholarly data offers unprecedented opportunities for identifying the drivers of modern science, for studying career paths of scientists, and for measuring research performance. These massive data and analysis methodologies created tidal effects to modern scientometrics transforming it into a whole new field named Science of Science. In this article, we provided a comprehensive coverage of recent advances in SoS related to impact-based ranking indicators and to impact prediction. We recognized fertile research work in the boundary of scientometrics and network science in terms of exploiting multilayer Scientometrics networks, and in the boundary between scientometrics and machine learning in terms of applying deep learning algorithms for impact prediction.

REFERENCES

[1] A. Abrishami and S. Aliakbary. NNCP: A citation count prediction methodology based on deep neural network learning techniques, 2018. Available at https://arxiv.org/abs/1809.04365.

[2] D. E. Acuna, S. Allesina, and K. P. Kording. Future impact: Predicting scientific success. *Nature*, 489(7415):201–202, 2012.

[3] P. Basaras, G. Iosifidis, D. Katsaros, and L. Tassiulas. Identifying influential spreaders in complex multilayer networks: A centrality perspective. *IEEE Transactions on Network Science & Engineering*, 2018. To appear.

[4] S. Boccaletti, G. Bianconi, R. Criado, C. I. del Genio, J. Gómez-Gardeñes, M. Romance, I. Sendina-Nadal, Z. Wang, and M. Zanin. The structure and dynamics of multilayer networks. *Physics Reports*, 544:1–222, 2014.

[5] S. V. Buldyrev, R. Parshani, P. Gerald, S. H. Eugene, and H. Shlomo. Catastrophic cascade of failures in interdependent networks. *Nature*, 464:1025–1028, 2010.

[6] A. Cardillo, J. Gómez-Gardeñes, M. Zanin, M. Romance, D. Papo, F. del Pozo, and S. Boccaletti. Emergence of network features from multiplexity. *Nature Scientific Reports*, 3:1344, 2013.

[7] S. Fortunato, C. T. Bergstrom, K. Borner, J. A. Evans, D. Helbing, S. Milojevic, A. M. Petersen, F. Radicchi, R. Sinatra, B. Uzzi, A. Vespignani, L. Waltman, D. Wang, and A. L. Barabasi. Science of science. *Science*, 359(6379), 2018.

[8] M. A. Garcia-Perez. Limited validity of equations to predict the future h index. *Scientometrics*, 96(3):901–909, 2013.

[9] E. Garfield. Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, 122:108–111, 1955.

[10] A. Gogoglou and Y. Manolopoulos. Predicting the evolution of scientific output. In *Proceedings of the 9th International Conference on Computational Collective Intelligence (ICCCI)*, volume 1, pages 244–254, 2017.

[11] A. Gogoglou, A. Sidiropoulos, D. Katsaros, and Y. Manolopoulos. The fractal dimension of a citation curve: Quantifying an individual's scientific output using the geometry of the entire curve. *Scientometrics*, 111(3):1751–1774, 2017.

[12] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[13] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46):16569–16572, 2005.

[14] D. Katsaros, L. Akritidis, and P. Bozanis. The f index: Quantifying the impact of coterminal citations on scientists' ranking. *Journal of the American Society for Information Science & Technology*, 60(5):1051–1056, 2009.

[15] D. Katsaros and Y. Manolopoulos. Impact and productivity of PhD graduates of computer science/engineering departments of Hellenic universities, 2017. Available at https://arxiv.org/abs/1707.05801.

[16] M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter. Multilayer networks. *Journal of Complex Networks*, 2:203–271, 2014.

[17] A. N. Langville and C. D. Meyer. *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton University Press, 2006.

[18] X. Liu, J. Bollen, M. L. Nelson, and H. van de Sompel. Co-authorship networks in the digital library research community. *Information Processing & Management*, 41(6):1462–1480, 2005.

[19] P. J. Mucha and M. A. Porter. Communities in multislice voting networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 20(4):041108, 2010.

[20] Z. Ning, Y. Liu, J. Zhang, and X. Wang. Rising star forecasting based on social network analysis. *IEEE Access*, 5:24229–24238, 2017.

[21] F. Pedroche, M. Romance, and R. Criado. A biplex approach to PageRank centrality: From classic to multiplex networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 26(6), 2016.

[22] O. Penner, R. K. Pan, A. M. Petersen, K. Kaski, and S. Fortunato. On the predictability of future impact in science. *Nature Scientific Reports*, 3(10), 2013.

[23] J. Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.

[24] G. Sideris, D. Katsaros, A. Sidiropoulos, and Y. Manolopoulos. The Science of Science and a multilayer network approach to scientists' ranking. In *Proceedings of the 22nd International Database Engineering & Applications Symposium (IDEAS)*, pages 5–11, 2018.

[25] A. Sidiropoulos, A. Gogoglou, D. Katsaros, and Y. Manolopoulos. Gazing at the skyline for star scientists. *Journal of Informetrics*, 10(3):789–813, 2016.

[26] A. Sidiropoulos, D. Katsaros, and Y. Manolopoulos. Generalized Hirsch h-index for disclosing latent facts in citation networks. *Scientometrics*, 72(2):253–280, 2007.

[27] A. Sidiropoulos, D. Katsaros, and Y. Manolopoulos. Ranking and identifying influential scientists versus mass producers by the Perfectionism index. *Scientometrics*, 103(1):1–31, 2015.

[28] R. Sinatra, D. Wang, P. Deville, C. Song, and A. L. Barabasi. Quantifying the evolution of individual scientific impact. *Science*, 354(6312), 2016.

[29] M. Singh, A. Jaiswal, P. Shree, A. Pal, A. Mukherjee, and P. Goyal. Understanding the impact of early citers on long-term scientific impact. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 59–68, 2017.

[30] G. Stoupas, A. Sidiropoulos, A. Gogoglou, D. Katsaros, and Y. Manolopoulos. Rainbow Ranking: An adaptable, multidimensional ranking method for publication sets. *Scientometrics*, 116(1):147–160, 2018.

[31] S. Vucetic, A. K. Chanda, S. Zhang, T. Bai, and A. Maiti. Peer assessment of CS doctoral programs shows strong correlation with faculty citations. *Communications of the ACM*, 61(9):70–76, 2018.

[32] D. Wang, C. Song, and A. L. Barabasi. Quantifying long-term scientific impact. *Science*, 342(6154):127–132, 2013.

[33] P. Wang, G. Robins, P. Pattison, and E. Lazega. Exponential random graph models for multilevel networks. *Social Networks*, 35:96–115, 2013.

[34] L. Weihs and O. Etzioni. Learning to predict citation-based impact measures. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 49–58, 2017.

[35] S. Xiao, J. Yn, C. Li, B. Jin, X. Wang, X. Yang, S. M. Chu, and H. Zha. On modeling and predicting individual paper citation count over time. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2676–2682, 2016.

[36] E. Yan, Y. Ding, and C. R. Sugimoto. P-rank: An indicator measuring prestige in heterogeneous scholarly networks. *Journal of the American Society for Information Science & Technology*, 62(3):467–477, 2011.

[37] A. Yong. Critique of Hirsch's citation index: A combinatorial Fermi problem. *Notices of the AMS*, 61(9):1040–1050, 2014.

[38] A. Zeng, Z. Shen, J. Zhou, J. Wu, Y. Fan, Y. Wang, and H. E. Stanley. The science of science: From the perspective of complex systems. *Physics Reports*, 714–715:1–73, 2017.

[39] A. Zeng and C. H. Yeung. Predicting the future trend of popularity by network diffusion. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 26(6):063102, 2016.

[40] Y.-B. Zhou, L. Lu, and M. Li. Quantifying the influence of scientists and their publications: Distinguishing between prestige and popularity. *New Journal of Physics*, 14(3):033033, 2012.