JOURNAL OF APPLIED CRYSTALLOGRAPHY

ISSN 1600-5767

The Science of Science. By Dashun Wang and Albert-László Barabási. Cambridge University Press, 2021. Pp. 308. Hardback price GBP 64.99, ISBN 9781108492669. Paperback price GBP 22.99, ISBN 9781108716956.

John R. Helliwell*

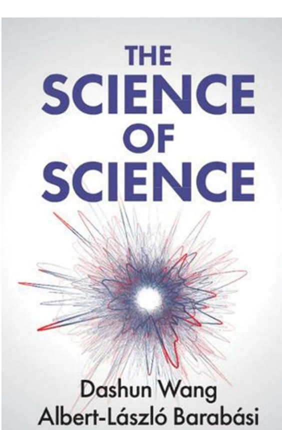
Department of Chemistry, University of Manchester, Manchester M13 9PL, United Kingdom. *Correspondence e-mail: john.helliwell@manchester.ac.uk

Keywords: book reviews; scientific impact; productivity; artificial intelligence; citation analyses; scientists and teams.

This book is in four parts: Part I *The Science of Career*, Part II *The Science of Collaboration*, Part III *The Science of Impact* and Part IV *Outlook*. Its introduction opens with 'Scientific revolutions are often driven by the invention of new instruments – the microscope, the telescope, genome sequencing – each of which have radically changed our ability to sense, measure and reason about the world. The latest instrument at our disposal? A windfall of data that traces the entirety of the scientific enterprise, helping us to capture its inner workings at a remarkable level of detail and scale.' The back cover proclaims that 'This is the first comprehensive overview of the 'science of science' . . . The book relies on data to draw actionable insights' for individual scientists and science policy makers.

Part I Science and Career. Chapters 1, 2 and 3 analyse metrics such as the h index and the i10 index to assess their correctness for measuring the productivity of scientists as well as asking, interestingly, how these measures can predict future productivity of a scientist. A caveat is made about the difficulty of being sure, through a name alone, of the identity of an author (page 16). Chapter 4 asks a deep question: is there a correlation between a person's age and scientific achievement? So, the Nobel Prizes' data show (Fig. 4.3) that the increasing interval between discovery and formal recognition projects that (page 45) 'most candidates will not live long enough to attend their Nobel Prize ceremonies'. Chapter 5 moves on from laureates to the 'everyday Joes and Janes' of science via big data archives. Here the authors' 'big data approach' shows that the most likely cause of 'best works' is quite simply productivity, so keep publishing and the best can occur at any time. That scientists tend to publish more earlier in their career drives the (apparent) link of best work being done early in the career, but actually best work is equally likely at any career point. Chapter 6 analyses if the ultimate impact of a paper is the product of whether an idea is brilliant multiplied by a person's ability to turn that idea into a discovery. The authors show that in Fig. 6.1. Part I concludes with Chapter 7's analysis of what the authors call hot streaks, periods of exceptional impact. This analysis is seen as relevant, as Section 7.3 explores, be it for faculty hiring committees or research grant proposal committees. These try to link track record and a person's future success and the quality of the research proposal and its future success. I would have liked to see these data analyses compared with the Leiden Manifesto on research assessment, which stresses combining analytics with judgement (Hicks *et al.*, 2015).

Part II The Science of Collaboration. Chapter 8, entitled *The Increasing Dominance of Teams in Science*, presents numerous data to prove the chapter title with various graphs spanning the past 40, and in one case 100, years. Chapter 9, on *The Invisible College*, analyses the impact that the arrival or departure of a star or average colleague has on a researcher. Chapter 10 is on *Coauthorship Networks*. Unfortunately this chapter treats equally the categories of coworkers and collaborators. I would say that a laboratory has a laboratory leader and the assistants and students who work on a study. Two laboratories, one team each, can also collaborate. To be sympathetic, however, presumably the difficulty for metrics specialists in analysing coauthorship trends is that cases such as two laboratories with the same departmental address may end up being counted as one corresponding author rather than two or more laboratory leaders. Chapter 11 is entitled *Team Assembly*. Like Chapter 10 it is marred by conflating team building with an organization, *i.e.* employer and place of work. It starts off airing the truism that a team of



all-stars is not necessarily a top-performing team. Equality, diversity and inclusion monitoring may be essential for monitoring a large employer, but for a collaborating team of only two, or even three or four, members, it is difficult to see how it can be so monitored. A pattern seems to be building up of the authors of the book having difficulty stepping into the actual practice of doing laboratory science, and its many permutations and combinations. Anyway, they do go on to analyse citation impact gains and offer results such as diversity on a team being associated with 'an impact gain of 10.63%'. This is stated without an uncertainly estimate and I simply do not believe it is precise to two decimal places. Chapter 12 analyses *Small and Large Teams*. This chapter confuses the Large Hadron Collider at CERN, which published a paper with 5154 authors, and a marvellous management collaboration entity, with an average team of 5.24 authors in 2013 (note again the two decimal place precision); it is not clear whether this was a single laboratory or more. Fig. 12.1 graphs the growing size of teams, which I do not doubt; but what is the cause? It could be any number of things: the very low success rates of research proposals, for instance. The authors conclude (page 127) that 'in science, teams are growing in size' (ok), 'hence' (really?) 'innovation increasingly happens in larger team settings'. This assertion, in my opinion, they have not proved. Section 12.3 is much improved in its assertions: Large teams develop science; small teams disrupt it. Section 12.4 is again sound, as the authors, of course properly, argue that science needs both small and large teams. Chaper 13, *Scientific Credit*, and Chapter 14, *Credit Allocation*, are meaty chapters, getting close to important ethical matters. Box 13.1 is an excellent description of author order and credit allocation challenges within team science, especially addressing interdisciplinary research involving more than one specialist laboratory. This box answers my criticism above on defining a team, but why is it so late in the book and with so little emphasis? Section 14.2 features a credit allocation algorithm involving one of the book's authors. This is used to analyse where prize-winning topics have largely been awarded correctly while a significant number, with very interesting examples such as that given in Box 14.3 on page 157, were not. The algorithm looks a useful way to inform (improve) judgement of prize-awarding committees and a supplementary way to avoid issues such as gender-based biases.

Part III The Science of Impact. Here the authors take the bull by the horns, with respect to their data analyses, and ask 'Are citations meaningful at all?'. Chapter 15 is entitled *Big Science*: not the usual meaning, of CERN or the Hubble Telescope, but the exponential growth in the number of publications, doubling every 12 years. It goes on to analyse the match of number of PhDs to jobs suitable for them, in academe or industry. This is a deep analysis including various countries and various disciplines. Similarly careful is Chapter 16 on *Citation Disparity*, analysing variations by discipline and by year. The authors show that, when internally normalized, cross-discipline comparisons are viable and each follow a very similar power-law distribution. So, the authors document that objective impact comparisons between disciplines, papers and individuals are viable whilst acknowledging (minor) shortcomings. Chapter 17 is on *High-Impact Papers* and Chapter 18 is on *Scientific Impact*, which seems an odd separation. These chapters mention various variables that might decide impact, such as novelty, an existing high number of citations attracting yet more or a paper which cites widely different fields in its reference list. Another cause and effect on impact that they consider is media coverage. The example chosen is *The New York Times*, where articles covered in the newspaper received '72.8%' more citations (I ignore the false precision) and during a 12 week strike, where they still logged articles they would have covered, the citation boost disappeared. University media office staff will be pleased to learn this no doubt. Chapter 19 is *The Time Dimension of Science* and explores such erudite questions as when a paper's citations stop. But also, for the assiduous citation hunter, we learn that, by a decent factor of two, papers that cite a good mix of recent and older work do better than papers engulfed only in the recent present or those, now rather unlikely for science, which only cite older papers. Box 19.2 on page 207 importantly reiterates that a journal 'impact factor cannot predict an individual paper's impact'. Chapter 20 is *Ultimate Impact*. This is an effort to parameterize the predicted total number of citations of any paper. It is illuminating as to the work of science metrics specialists. At this point I wonder if the book's title should be *The Science of Science Metrics Modelling of Publications*.

Part IV Outlook. The opening of this part states that the authors will examine 'how to generate causal insights with actionable policy implications' by considering 'how knowledge is discovered, hypotheses are raised, and experiments are prioritized'. Chapter 21, *Can Science Be Accelerated*, enthusiastically expands on the *Outlook* with examples of machines capable of generating their own hypotheses. But this chapter veers off into the safer ground of pre-registration of studies to avoid making hypotheses after the data are measured and how to encourage researchers to take risks. The latter misses the point that researchers do propose risky projects but funding agencies do not necessarily wish to take too much risk. Chapter 22, *Artificial Intelligence*, focuses on DeepMind and its protein-fold predictive success [for details see Helliwell (2020)]. Suffice to say, DeepMind did not 'beat all scientists at predicting the 3D structure of proteins'; rather it surpassed other predictor algorithms. Indeed, is deep learning/artificial intelligence yet satisfying the Turing test? Namely, is a machine able to exhibit intelligent behaviour equivalent to a human? Or is it still number crunching, very impressive though that that may be? They do conclude Chapter 22 with circumspection and recognize that the future is a 'strategic partnership between humans and machines'. Chapter 23 is *Bias and Causality in Science*. This presents interesting anecdotes but sensibly concludes that 'Small, limited questions can be answered with confidence, but bigger questions are subject to much more uncertainty' (in their conclusions, if any): my words added in parentheses to complete the sentence. The final section, of two pages, is entitled *Last Thought: All the Science of Science*. It is a meander, suggesting that the authors are unsure how to conclude their book; maybe those answers to bigger questions are elusive. There then follow two appendices. One is on *Modelling Team Assembly* and the other on *Modelling Citations*. These divulge details of the modelling analyses.

Overall, I found this book very stimulating. It made me wonder whether in-depth metrics analyses of ‘only’ the subjective narratives of authors, such as the references list they select, actually creates a foundation on which to form judgement rather than opinion? Namely, what fraction of these publications analysed for their metrics were actually underpinned by their data? As well as provoking thought, this book offers a feast of references, 424 in all. There are such further enticing reads as reference 396, *Life3.0: Being Human in the Age of Artificial Intelligence*. To conclude, I recommend this book for your library, and maybe even take it for your summer beach reading.

References

Helliwell, J. R. (2020). *IUCr Newslett.* **28**(4), 6.

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S. & Rafols, I. (2015). *Nature*, **520**, 429–431.