Will This Paper Increase Your *h*-index?

Yuxiao Dong, Reid A. Johnson, and Nitesh V. Chawla(✉)

Interdisciplinary Center for Network Science and Applications,
Department of Computer Science and Engineering,
University of Notre Dame, Notre Dame, IN, USA
{ydong1,rjohns15,nchawla}@nd.edu

Abstract. A widely used measure of scientific impact is citations. However, due to their power-law distribution, citations are fundamentally difficult to predict. Instead, to characterize scientific impact, we address two analogous questions asked by many scientific researchers: "How will my h-index evolve over time, and which of my previously or newly published papers will contribute to it?" To answer these questions, we perform two related tasks. First, we develop a model to predict authors' future h-indices based on their current scientific impact. Second, we examine the factors that drive papers—either previously or newly published—to increase their authors' predicted future h-indices. By leveraging relevant factors, we can predict an author's h-index in five years with an R^2 value of 0.92 and whether a previously (newly) published paper will contribute to this future h-index with an F_1 score of 0.99 (0.77). We find that topical authority and publication venue are crucial to these effective predictions, while topic popularity is surprisingly inconsequential. Further, we develop an online tool that allows users to generate informed h-index predictions. Our work demonstrates the predictability of scientific impact, and can help researchers to effectively leverage their scholarly position of "standing on the shoulders of giants."

Scientific impact plays a pivotal role in the evaluation of the output of scholars, departments, and institutions. A widely used measure of scientific impact is citations, with a growing body of literature focused on predicting the number of citations obtained by any given publication. The effectiveness of citation prediction, however, is fundamentally limited by their power-law distribution, whereby publications with few citations are extremely common and publications with many citations are relatively rare. In light of this limitation, we instead investigate scientific impact by addressing two analogous questions [1], both related to the measure of h-index [2] and asked by many academic researchers: "***How will my h-index evolve over time, and which of my previously and newly published papers will contribute to my future h-index?***"

Y. Dong and R.A. Johnson—Provided equal contribution to this work.

This work was published at the 8th ACM International Conference on Web Search and Data Mining (*WSDM'15*) [1]. This extended abstract has been largely extracted from the publication.

A. Bifet et al. (Eds.): ECML PKDD 2015, Part III, LNAI 9286, pp. 259–263, 2015.
DOI: 10.1007/978-3-319-23461-8_26

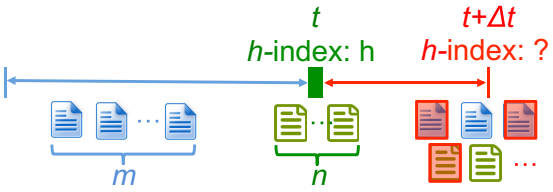


Fig. 1. Illustrative example of scientific impact prediction. Before time t, a scholar published m papers and had an h-index of h. Our prediction problems are targeted at answering two questions: 1) First, what is the scholar's future h-index, h', at time $t+\Delta t$? 2) Second, which of his/her papers, both (a) those m papers previously published before t and (b) those n new papers published at t, will contribute to h'?

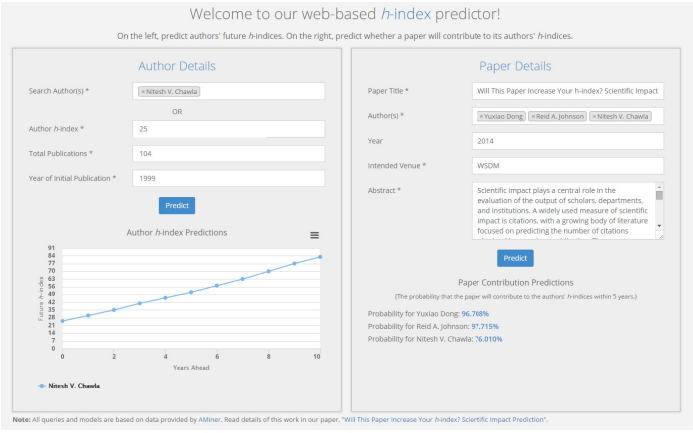
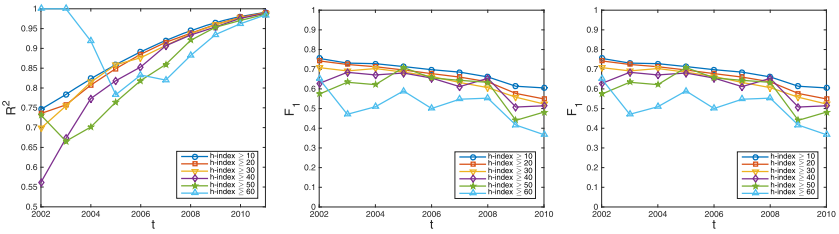


Fig. 2. Prototype h-index prediction tool. The left side may be used to predict the development of authors' h-indices and the right side may be used to predict whether a paper will contribute to its authors' h-indices.

To tackle these questions, we formulate two scientific impact prediction problems, as shown in Figure 1. Our primary problem is to determine whether a given previously or newly published paper will, after a predefined timeframe, influence a particular author's predicted *future* h-index. As a secondary problem, we predict authors' future h-indices based on their current scientific impact. These predicted future h-indices are then used as the future h-indices in our primary task, with the purpose of accounting for the change in the author's h-index over the prediction timeframe. Besides addressing these problems, we have also developed and deployed an online tool (see Figure 2) that allows users to generate h-index predictions informed by our findings.

Using a large-scale academic dataset with over 1.7 million authors and 2 million papers from the premier online academic service ArnetMiner [3], we demonstrate a high level of predictability for scientific impact as measured by our two problems in Figure 3. Accordingly, we find strong performance for our first task of predicting an author's future h-index. We can predict an author's h-index



(a) h-index prediction (b) Predicting for new papers (c) Predicting for old papers

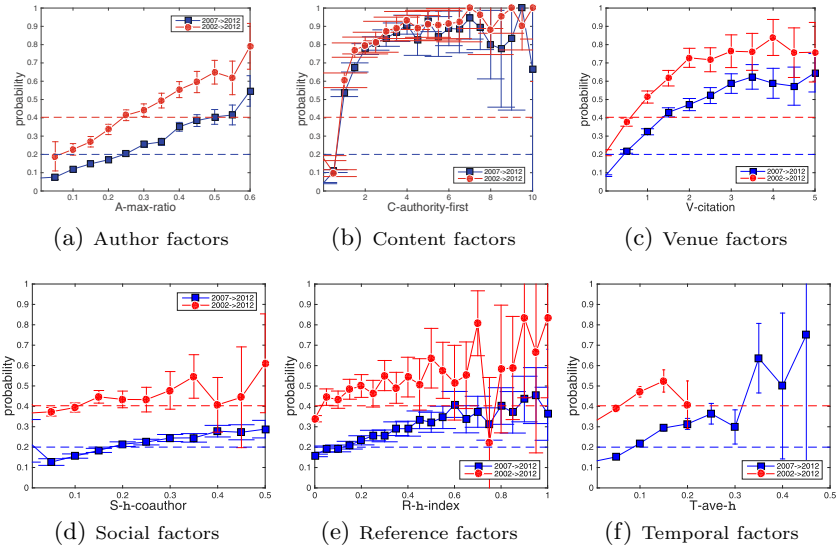
Fig. 3. Scientific Impact Predictability. (a) Predicting authors' future h-indices; (b) Predicting whether newly published papers can increase their primary authors' future h-indices; (c) Predicting whether previously published papers can increase their primary authors' future h-indices.

in five years with an R^2 value of 0.9197. This performance generally increases as the prediction timeframe is shortened, with a prediction of ten years achieving an R^2 of 0.7461. We can predict whether a previously or newly published paper will contribute to an author's future h-index in five years with respective F_1 scores of 0.99 and 0.77, improvements of +130% and +160% over random guessing. Predictive performance for newly published papers generally decreases as the prediction timeframe is shortened, but is consistently high for previously published papers. Our results also indicate that authors with low h-indices are easier to predict for than those with high ones.

We also assess the factors that influence our predictive results. For our secondary problem, we find that the author's current h-index is most telling, followed by the number of publications and co-authors. For our primary problem, we investigate six groups of factors that drive a paper's citation count to become greater than its primary author's h-index, including the paper's author(s), content, published venue, and references, as well as social and temporal effects related to its author(s). Figure 4 shows the response curve of the most important factor (as evaluated by correlation coefficients) for each group of factors. We find that topical authority is the most telling factor for newly published papers, while the existing citation information is most telling for previously published ones, followed by the authors' influence and the publication venue. We also find that publication venue and the author collaborations are moderately significant factors for longer prediction periods, but inconsequential for shorter ones. Finally, we are surprised to find that topic popularity is insignificant for both previously and newly published papers.

Overall, our findings unveil the predictability of scientific impact, deepen the understanding of scientific impact measures, and provide scholars with concrete suggestions for expanding their scientific influence. Salient points include:

– A scientific researcher's authority on a topic is the most decisive factor in facilitating an increase in his or her h-index. This coincides with the fact that the society fellows or lifetime honors are typically awarded for contributions to a topic or domain.



(a) Author factors (b) Content factors (c) Venue factors

(d) Social factors (e) Reference factors (f) Temporal factors

Fig. 4. Factor response curves with $\Delta t = 5$ or 10 ($t + \Delta t = 2012$). x-axis: factor value; y-axis: probability that a given paper published in t increase its primary author's h-index in 2012. *A-max-ratio*: the ratio between max-h-index and #papers attributed to the primary author; *C-authority-first*: the consistence between the first author's authority and this paper; *V-citation*: the #average-citations of papers published in this venue; *S-h-coauthor*: the average h-index of co-authors of the paper's authors; *R-h-index*: the references' h-index; *T-ave-h*: the average Δh-indices of the authors between now and three years ago. All response probabilities are observed at a 95% confidence interval.

- The level of the venue in which a given paper is published is another crucial factor in determining the probability that it will contribute to its authors' h-indices. The suggestion here lies in the every scholar's aim: *Target and publish influential scientific results in top venues.*
- Publishing on an academically "hot" but unfamiliar topic is unlikely to further one's scientific impact, at least as measured by an increase in one's h-index. This reminds us that *one should not turn to follow the vogue topics that are beyond his or her expertise.*

We strongly believe that our findings can help lead to the improved use of scientific impact measures, though we caution that *in no way should our research be construed as advocating the use of the h-index or any other measure as a deciding factor in one's research pursuits.*

Acknowledgments. This work is supported by the Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053, the U.S. Air Force Office of Scientific Research (AFOSR) and the Defense Advanced Research Projects Agency (DARPA) grant #FA9550-12-1-0405, and the National Science Foundation (NSF) Grant OCI-1029584.

References

1. Dong, Y., Johnson, R.A., Chawla, N.V.: Will this paper increase your h-index?: Scientific impact prediction. In: WSDM 2015, pp. 149–158. ACM, New York (2015)
2. Hirsch, J.E.: An index to quantify an individual's scientific research output. PNAS **102**(46), 16569–16572 (2005)
3. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: ArnetMiner: Extraction and mining of academic social networks. In: KDD 2008, pp. 990–998 (2008)