Abstract

The notion of reputation in academia is critical for taking decisions on research grants, faculty position tenure, and research excellence awards. And the notion of reputation is always associated with the publication track record of the researcher or research group. Thus, it is important to assess publication track records quantitatively. To quantify a publication record, bibliographic metrics are usually adopted. Among these, citation based metrics, such as H-indices and citation counts, are quite popular. In this paper we study the correlation between P-score, a publication record metric we introduced previously, and H-indices. We show that they are correlated with a Kendall-Tau coefficient that exceeds 0.5. Additionally, we noticed that they have important differences. We were able to identify publication venues with high H-indices and low P-scores, as well as venues with low H-indices and high P-scores. We provide interpretations for these findings and discuss how they can be used by research funding councils and committees to better support their funding decisions.

P-score: a complementary metric to H-index

Edmundo de Souza e Silva

February 9, 2018

1 Introduction

Academic research evaluation is a topic of interest to universities, research groups, research funding institutions and the public at large. The evaluation of research is usually carried out by comparing journals, conferences or papers through the usage of academic performance metrics. A popular and moderately effective approach is to compute metrics based on the number of citations received by a given piece of research, though this approach has multiple shortcomings. First, in the evaluation and bibliometrics research communities, citations are understood as a metric of attention, not of impact or quality. Citations measure the amount of peer attention a paper has gathered in its related fields [25], not the quality of the work produced.

REWRITE (explain the need for additional metrics, such as P-scores)

Evaluating impact of research is of interest to universities, research groups, research funding institutions, governments, and the public at large. And while there are many facets to consider while evaluating the impact of research, a simple approach is to compute metrics based on the number of citations a given piece of research receives. The popularity of citation based metrics derives from their simple definition and consequent intuitive appeal. Despite that, they have shortcomings. First, in the evaluation and bibliometrics research communities, citations are understood as a metric of attention not of impact or quality. That is citations measure the attention to a paper of peers in related fields [25] not the quality of the work. Second, citations might take long to happen. To illustrate, in a recent study that looked at first time to citation in a universe of more than a million papers, half of the papers received their first citation 20 months or more after their publicationËIJ[27]. Third, citations are not simple to compute and are not always broadly available, particularly at the level of individual researchers. To illustrate there are now universities requesting their faculty to create a Google Scholar profile with the objective of making citations of papers written by an individual researcher readily available. However, frequently only a fraction of faculty comply[1].

[1] In a particular instance, a university official complained to one of the authors here that just one third of faculty had created Google Scholar profiles.

Despite these well known limitations, citations continue to be broadly used whenever available. For instance, while citations are not always available at the level of individual researchers, they are widely available at the level of journals and research conferences, that is, at the level of publication venues. This is of interest if one takes a venue-level metric as a good proxy for the quality of individual papers published by that venue, which might be obviously not the case for many papers. Despite that, weighting publications by venue-level metrics is a far simpler and more direct approach than computing citations for individual researchers. Because of that, venue-level metrics are widely used and are of relevance [20, 25].

While venue-level citation metrics are widely available, they do not work well when one is interested on a particular subarea of a major area such as Computer Science. For instance, a research funding agency might create a program aimed at strengthening research focused in particular subareas such as knowledge discovery, data mining, and machine learning. In this case, the question that is difficult to answer is which venues better represent research in each of these subareas. In this particular case, just considering venue-level citation based metrics does no work well because popular journals and conferences in Computer Science frequently span more than one subarea. To illustrate consider the subarea of Information Retrieval in Computer Science and the venues in which their researchers publish frequently. If just a venue-level citation metric is considered, venues such as Jasist (Journal of the American Society in Science and Technology), IPM (Information and Processing Management), and WWW (World Wide Web) will show up among the top 10 venues in Information Retrieval. While these are fine venues, they are not venues on Information Retrieval. This is the first problem we address in this work, how to determine automatically a ranking of venues in a particular subarea of Computer Science.

The second problem of our interest here is the fact that citations might take long to appear, as discussed in [27]. We wonder whether it is possible to notify researchers working on a given subarea of a new venue of interest whose reputation is on the rise. We propose a solution and, without loss of generality, explore its application to a subset of pre-selected venues.

2 P-score: a network-based metric

Contrary to citation counting metrics such as the Impact Factor [37, 15, 34, 1] and H-index [4, 3, 13, 5, 6], P-scores are a graph modeling metric that takes into account the relations among researchers, papers they published and their publication venues. They are based on a framework of reputation flows we introduced previously [36]. Let us review them briefly.

A *reputation graph* in academia is a graph with three node types: (a) *reputation sources* representing groups of selected researchers, (b) *reputation targets* representing venues of interest, and (c) *reputation collaterals* representing entities we want to compare such as research groups and academic departments. Figure 1 provides a generic illustration of our reputation graph and introduces the following notation: S is the set of reputation sources,

T is the set of reputation targets, and C is the set of reputation collaterals. We first propagate the reputation of source nodes to target nodes, which allows assigning weights to the target nodes. Following, we propagate these weights to the collaterals. In here we adopt research groups as source nodes and publication venues as target nodes. Further we focus on the weights assigned to the publication venues and do not consider potential collaterals of interest (such as individual researchers).

Figure 1: Structure of the reputation graph.

Our reputation graph for academia can be naturally associated with paper authors, modelled as reputation sources, and papers, modelled as reputation targets. The interaction between reputation sources and reputation targets is inspired by the notion of *eigenvalue centrality* in complex networks [7, 21, 30]. In the reputation graph, if we consider only sources and targets, it is easy to identify reputation flows from sources to sources, from sources to targets, from targets to sources, and from targets to targets. These reputation flows can be modeled as a stochastic process. In particular, let P be a *right stochastic* matrix of size $(|S| + |T|) \times (|S| + |T|)$ with the following structure:

$$P = \left[\begin{array}{c|c} (d^{\langle S \rangle}).P^{\langle SS \rangle} & (1 - d^{\langle S \rangle}).P^{\langle ST \rangle} \\ \hline (1 - d^{\langle T \rangle}).P^{\langle TS \rangle} & (d^{\langle T \rangle}).P^{\langle TT \rangle} \end{array}\right] \tag{1}$$

where each quadrant represents a distinct type of reputation flow, as follows:

$P^{\langle SS \rangle}$: right stochastic matrix of size $|S| \times |S|$ representing the transition probabilities between reputation sources;

$P^{\langle ST \rangle}$: matrix of size $|S| \times |T|$ representing the transition probabilities from reputation sources to targets;

$P^{\langle TS \rangle}$: matrix of size $|T| \times |S|$ representing the transition probabilities from reputation targets to sources;

$P^{\langle TT \rangle}$: right stochastic matrix of size $|T| \times |T|$ representing the transition probabilities between reputation targets.

The parameters $d^{\langle S \rangle}$, the fraction of reputation one wants to transfer among the source nodes themselves, and $d^{\langle T \rangle}$, the fraction of reputation one wants to transfer among the target nodes themselves, control the relative importance of the reputation sources and targets. Assuming that the transition matrix P is ergodic, we can compute the steady

state probability of each node and use it as a reputation score. More formally, we can write:

$$\gamma = \gamma P \tag{2}$$

where γ is a row matrix with $|S| + |T|$ elements, where each row represents the transition probabilities of a node in the set $S \cup T$.

We should note that, while our network model allows modeling citations in the fourth quadrant, it is possible to compute steady state probabilities for the network without consideration to citations. This is accomplished by setting the parameter $d^{\langle T \rangle} = 0$. Thus, it should be clear, in all of our experiments here P-scores are computed without taking citations into account.

2.1 Reputation Sources

The choice of the reputation sources is an important part of the method since its composition has a direct impact in the final rankings. There is no definitive way to make it. This choice depends on what we want to measure. In here we use the top CS departments in the US as reputation sources. It is a simple procedure that allows assigning P-scores to publication venues, P-scores that reflect the publication patterns of top CS departments in the US. We then compare how these scores compare with H-indices assigned to the same publication venues.

One way to determine the top CS departments is to adopt the randomization procedure we first described in [36]. A run of that procedure works as follows:

1. randomly select 10 entities from the set of reputation targets and use them as the set S of reputation sources
2. compute steady state probabilities for all nodes
3. using the steady state probabilities of reputation targets as a score, select the 10 entities with highest scores and use them as a new set S_{new} of reputation sources
4. if $S_{new} \not\equiv S$ then $S \leftarrow S_{new}$ and go back to step 1
5. $S_{auto} \leftarrow S_{new}$
6. take S_{auto} as the set of automatically selected reputation sources
7. exit

By applying this randomization procedure 100 times to a set of 125 US graduate programs in Computer Science (CS), exactly the same 125 graduate programs considered by NRC in its 2010 evaluation of CS graduate programs in the USA [31], we ended up with a subset of 12 CS programs to be considered as reputation sources. These are the CS programs that appeared as least once in the set S_{auto} of automatically selected reputation sources and they are as follows:

1. Carnegie Mellon University
2. Georgia Institute of Technology
3. Massachusetts Institute of Technology
4. Stanford University
5. University of California-Berkeley
6. University of California-Los Angeles
7. University of California-San Diego
8. University of Illinois at Urbana-Champaign
9. University of Maryland College Park
10. University of Southern California
11. University of Michigan-Ann Arbor
12. Cornell University

All the 12 departments above are among the top 5th percentile in the ranking produced by NRC. This suggests that our recursive procedure is able to take advantage of patterns in the publication streams of the various CS departments to determine the most reputable ones in fully automatic fashion. We further observe that this was done while setting the parameter $d^{\langle T \rangle} = 0$. That is, we did not use information on citation counts in the model.

3 Correlation with H-indices

TO WRITE: explain that distributions are not Normal, show graphs, propose Kendall-Tau, show macro values

4 Assessing conferences in CS

Discuss the cone strategy, justify the 10 degrees angle, discuss the problem of positioning the cone vertex at the origin, introduce the (-100,-100) pivot, discuss items above and below the cone

5 Related Work

REWRITE: adjust for theme of this paper

Citation-based metrics have been widely applied to rank computer and information science journals [19, 29]. Also, different approaches using citation data have been proposed to measure the quality of publication venues in the Databases subarea [33] and to rank documents retrieved from a digital library [22].

Garfield's Impact Factor [14] is one of the first metrics proposed to quantify research impact. In a nutshell, it indicates the average number of citations per publication of a journal, in the last two years. One of the most widespread citation-based metric, the H-index, was proposed by Hirsch [18]. It has been mainly used to rank researchers both in terms of productivity and scientific impact. The key idea behind the H-Index is to detect the quantity of publications of high impact an author has in her research career – for instance, penalizing authors with a large volume of articles but with a low number of citations for the majority of them. Additionally, several works proposed different uses of citation data [12, 11, 40, 38] and studied their impact, advantages, and disadvantages [26, 23].

The idea of reputation, without the direct use of citation data, was discussed by Nelakuditi et al. [28]. They proposed a metric called *peers' reputation* for research conferences and journals, which ties the selectivity of the publication venue based upon the reputation of its authors' institutions. The proposed metric was shown to be a better indicator of selectivity of a research venue than acceptance ratio. In addition, the authors observed that, in the subarea of Computer Networks, many conferences have similar or better peers' reputation than journals. This result is similar to the conclusions obtained by Laender et al. [20], who show that conference publications are important vehicles for disseminating CS research, while in other areas such as Physical Sciences and Biology the most relevant venues are arguably the scientific journals.

Regarding the assessment of individual researchers' influence and expertise, many approaches have been introduced [2, 9, 10, 16, 39]. Particularly, GonÃğalves et al. [17] quantified the impact of various features on a scholar popularity throughout her career. They concluded that, even though most of the considered features are strongly correlated with popularity, only two features are needed to explain almost all the variation in popularity across different researchers: the number of publications and the average quality of the scholarâĂŹs publication venues. In addition, the prediction of scientific success of a researcher is also valuable for several goals [**?**]. As a result, previous works attempted to predict if a researcher will become a principal investigator [**?**], her future H-index [**?**, **?**] and the potential number of citations to her publications [**?**, 8].

Although citation-based metrics are useful, they are not enough to do a complete evaluation of research. In particular, Piwowar [32] showed that metrics as the H-Index are slow, as the first citation of a scientific article can take years. He concludes that the development of alternative metrics to complement citation analysis is not only desirable, but a necessity.

The reputation model we use in this work was proposed in [**?**]. This model, called *reputation flows*, exploits the transference of reputation among entities in order to identify

the most reputable ones. Particularly, the reputation flows consist in a random walk model where the reputation of a target set of entities is inferred using suitable sources of reputation. To evaluate this model, they instantiated the reputation flows in an academic setting, proposing a novel metric for academic reputation, the *P-score* [35].

By and large, the aforementioned works or variations of them are commonly used in assessments of academic output and also by modern search engines for scientific digital libraries, e.g. Google Scholar[2], Microsoft Academic Search[3], AMiner[4], and CiteSeerX[5]. However, none of the referred metrics take into account the different publication patterns in the subareas. Studies suggesting those differences and the negative impact of uniform evaluation metrics have been discussed in the field of Economics [**?**, **?**] and in Computer Science [**?**, **?**, 24].

Wainer et al. [**?**] presented the first attempt to quantify the differences in publication and citation practices between the subareas of Computer Science. Their key findings were: i) there are significant differences in productivity across some CS subareas, both in journals (e.g. Bioinformatics has a significantly higher productivity than Artificial Intelligence) and in conferences (e.g. Image Processing and Computer Vision has a significantly higher productivity than Operational Research and Optimization); ii) the mean number of citations per paper varies depending on subarea (e.g. Management Information Systems has significantly higher citation rates per paper than Computer Architecture); and iii) there are significant differences in emphasis on publishing in journals or in conferences (e.g. Bioinformatics are clearly journal oriented while Artificial Intelligence are conference oriented). However, they do not focus on modeling a new productivity metric for academic domain taking into account those differences between the subareas.

To the best of our knowledge, this is the first work that tackles the problem of both identifying the most important venues of a subarea in Computer Science and rank research groups based on this information, in a semi-automatic fashion.

6 Conclusions

ZZZ Rewrite entirely.

ACKNOWLEDGEMENTS

Omitted for blind review.

[2] `https://scholar.google.com.br`
[3] `http://academic.microsoft.com`
[4] `http://aminer.org`
[5] `http://citeseerx.ist.psu.edu`

References

[1] A. T. Balaban. Positive and negative aspects of citation indices and journal impact factors. *Scientometrics*, 92(2):241–247, 2012.

[2] K. Balog. Expertise retrieval. *Found. Trends Inf. Retr.*, 6(2-3):127–256, 2012.

[3] J. Bar-Ilan. Which h-index? - a comparison of wos, scopus and google scholar. *Scientometrics*, 74(2):257–271, 2008.

[4] F. Benevenuto, A. H. Laender, and B. L. Alves. The h-index paradox: your coauthors have a higher h-index than you do. *Scientometrics*, 106(1):469–474, 2016.

[5] L. Bornmann and H.-D. Daniel. Does the h-index for ranking of scientists really work? *Scientometrics*, 65(3):391–392, 2005.

[6] L. Bornmann and W. Marx. The h-index as a research performance indicator. *Eur Sci Ed*, 37(3):77–80, 2011.

[7] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Comput. Netw.*, 30(1-7):107–117, 1998.

[8] C. Castillo, D. Donato, and A. Gionis. Estimating number of citations using author reputation. In *Proc. of String processing and information retrieval*, pages 107–117, 2007.

[9] G. Cormode, S. Muthukrishnan, and J. Yan. People like us: mining scholarly data for comparable researchers. In *Proc. of WWW*, pages 1227–1232, 2014.

[10] H. Deng, J. Han, M. R. Lyu, and I. King. Modeling and exploiting heterogeneous bibliographic networks for expertise ranking. In *Proc. of JCDL*, pages 71–80, 2012.

[11] Y. Ding and B. Cronin. Popular and/or prestigious? measures of scholarly esteem. *Information Processing & Management*, 47(1):80–96, 2011.

[12] L. Egghe. Theory and practise of the g-index. *Scientometrics*, 69(1):131–152, 2006.

[13] L. Egghe. The influence of transformations on the h-index and the g-index. *J. Am. Soc. Inf. Sci. Technol.*, 59(8):1304–1312, 2008.

[14] E. Garfield. Citation indexes for science. *Science*, 122(3159):108–111, 1955.

[15] E. Garfield. How can impact factors be improved? *British Medical Journal*, (313):411–413, 1996.

[16] S. D. Gollapalli, P. Mitra, and C. L. Giles. Ranking authors in digital libraries. In *Proc. of JCDL*, pages 251–254, 2011.

[17] G. D. Gonçalves, F. Figueiredo, J. M. Almeida, and M. A. Gonçalves. Characterizing scholar popularity: A case study in the computer science research community. In *Proc. of JCDL*, pages 57–66, 2014.

[18] J. Hirsch. An index to quantify an individual's scientific research output. *Proc. Nat. Acad. Sciences*, pages 16569–16572, 2005.

[19] P. Katerattanakul, B. Han, and S. Hong. Objective quality rankings of computing journals. *Commun. ACM*, 45, 2003.

[20] A. H. F. Laender, C. J. P. de Lucena, J. C. Maldonado, E. de Souza e Silva, and N. Ziviani. Assessing the research and education quality of the top brazilian computer science graduate programs. *ACM SIGCSE Bulletin*, 40(2):135–145, 2008.

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

[22] B. Larsen and P. Ingwersen. Using citations for ranking in digital libraries. In *Proc. of JCDL*, pages 370–370, 2006.

[23] L. Leydesdorff. How are new citation-based journal indicators adding to the bibliometric toolbox? *J. Am. Soc. Inf. Sci. Technol.*, 60(7):1327–1336, 2009.

[24] H. Lima, T. H. P. Silva, M. M. Moro, R. L. T. Santos, W. M. Jr., and A. H. F. Laender. Aggregating productivity indices for ranking researchers across multiple areas. In *Proc. of JCDL*, pages 97–106, 2013.

[25] T. Loach and T. Evans. Ranking journals using altmetrics. In *Proc. of ISSIzzz*, page zzz, 2016.

[26] W. S. Martins, M. A. Gonçalves, A. H. F. Laender, and G. L. Pappa. Learning to assess the quality of scientific conferences: a case study in computer science. In *Proc. of JCDL*, pages 193–202, 2009.

[27] T. Nane. Time to first citation estimation in the presence of additional information. In *Proc. of ISSIzzz*, page zzz, 2016.

[28] S. Nelakuditi, C. Gray, and R. R. Choudhury. Snap judgement of publication quality: how to convince a dean that you are a good researcher. *Mobile Computing and Commun. Review*, 15(2):20–23, 2011.

[29] S. Nerur, R. Sikora, G. Mangalaraj, and V. Balijepally. Assessing the relative influence of journals in a citation network. *Commun. ACM*, 48(11):71–74, 2005.

[30] M. Newman. *Networks: An Introduction*. Oxford University Press, 2010.

[31] NRC. United States National Research Council, National Academy of Sciences, 2010.

[32] H. Piwowar. Altmetrics: Value all research products. *Nature*, 493(7431):159–159, 2013.

[33] E. Rahm and A. Thor. Citation analysis of database publications. *ACM Sigmod Record*, 34(4):48–53, 2005.

[34] T. Reuters. The thomson reuters impact factor, 2011.

[35] S. Ribas, B. Ribeiro-Neto, E. de Souza e Silva, A. H. Ueda, and N. Ziviani. Using reference groups to assess academic productivity in computer science. In *Proc. of WWW*, pages 603–608, 2015.

[36] S. Ribas, B. Ribeiro-Neto, E. de Souza e Silva, and N. Ziviani. On the reputation of venues in computer science. Submitted, 2015.

[37] S. Saha, S. Saint, and D. Christakis. Impact factor: a valid measure of journal quality? *J. Med. Lib. Assoc.*, 91(1):42–46, 2003.

[38] Y. Sun and C. L. Giles. *Popularity weighted ranking for academic digital libraries*. Springer, 2007.

[39] H. Wu, Y. Pei, and J. Yu. Detecting academic experts by topic-sensitive link analysis. *Front. Comp. Science in China*, 3(4):445–456, 2009.

[40] E. Yan, Y. Ding, and C. R. Sugimoto. P-rank: An indicator measuring prestige in heterogeneous scholarly networks. *J. Am. Soc. Inf. Sci. Technol.*, 62(3):467–477, 2011.