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 22, 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 impact metrics. A popular and moderately effective approach is to compute metrics based on the number of citations received by a given piece of research and assume that the number of citations is a good proxy for research quality. However, this approach has multiple shortcomings.

First, in the evaluation and bibliometrics research communities, citations are understood as a measure of attention rather than a proxy for impact or quality. Citations measure the attention to a paper of peers in related fields [18], not the quality of the work produced.

Second, citations can take a long time to happen. To illustrate this, in a recent study that looked at first time to citation in a universe of more than a million papers, half the papers only received their first citation 20 months or more after they were published [20].

Third, citations are not simple to compute and are not always broadly available, particularly at the level of individual researchers. Collecting citation counts requires access to the contents of a large number of publications, of which many are not easily available, and most authors do not have up to date citation information in their public internet profiles.

Despite these limitations, citation based metrics continue to be a popular approach to academic research evaluation. Nonetheless, it is desirable to employ alternative and complementary metrics that tackle the exposed problems without loss of effectiveness.

Academic reputation is an individual or group property strongly associated with academic impact. Reputable venues tend to concentrate the most relevant research papers because that is how they acquire and keep their good reputations. At the same time, reputable authors seek to publish on reputable venues because it gives their research more visibility and accreditation. Therefore, a researcher conveys reputation to a venue proportionally to its own reputation and the reputation of a researcher is proportional to the reputation of the venues where she or he publishes. This relationship between authors and venues constitutes a reputation network in which authors influence venue reputations and vice versa.

Pscore is a graph modeling metric that attributes quantitative reputations to venues based on the publication patterns of a reference group of researchers without having to rely on citation information. It deals with the three problems discussed previously.

First, instead of measuring the amount of researcher attention, Pscore measures reputations, which is presumably a better proxy for academic impact as was shown by Ribas.

Second, while citations can take years to happen, reputations tend to be steady over the time, so the quality of new research can be promptly estimated with pscore by simply complementing the reputation flows graph with the newest publication data.

Third, pscore does not require citation data to perform effectively, thus we can avoid the laborious effort of compiling individual researcher citation information.

The Pscore of computer science venues has shown significant correlation with H-index. Yet, there are important cases where the two metrics diverge considerably. Namely, in venues that belong to an intersection between different domains and in venues of less popular subareas of computer science. These differences appeal to the importance of pscore as a complementary metric.

2 Related Work

Quantitative metrics of scientific impact have a long history. An influential approach is Garfield's Impact Factor (1955) [13], a journal-level metric that is defined by the mean number of citations received by articles published in a given journal over a 2-year period. Despite being one of the earliest approaches at measuring scientific impact, it has showed remarkable survivability. Pinski et. al. [24] describe several limitations with the metric. Impact factors also suffered criticism for being misleading [21, 31] and for lacking reliable validation from an independent audit [30]. Riikonen and Vihinen (2008) [29] showed that actual citation and publication counts were better predictors of a scientist's contribution than impact factors in a study that assessed the scientific contribution of Finnish researchers in biomedicine.

The acceptance rate is another important journal-level metric that attempts to quantify the scientific impact. It is defined by the proportion of accepted papers relative to the number of submitted papers. Chen et. al. [9] have showed that highly selective conferences (therefore having low acceptance rates) have higher scientific impact, measured in terms of the number of citations received by the published papers.

The H-index is an author-level metric proposed by Hirsch (2005) [15]. Since its formulation, it has been widely adopted as a measure of invidual scientific research output. An H-index of h means that a researcher has published h papers that were cited at least h times. The H-index takes into account both the number of publications and the number of citations per publication, achieving higher values when a researcher obtains a consistently high number of citations over multiple publications. The H-index has been criticized for not taking into account field specific citation statistics [33]. The G-index, proposed by Egghe (2006) [11], is a less strict variation of the H-index where the score g is the largest number such that the top g articles have received at least g^2 citations.

Citation counts and derived metrics are straightforward to calculate, but they represent only a coarse estimate of academic impact, because they are essentialy popularity measures. Some people are popular but not prestigious and vice versa. For example, an author of pulp detectives may sell many books, but may not have earned the respect of literary critics [5]. Citations from prestigious journals are more relevant than citations from peripheral journals, as noted by Pinski et. al. (1976) [24]. To address this issue, some studies distinguish popularity from reputation or prestige with the usage of citation weighting strategies [10, 34] or PageRank based methods [5, 32]. Furthermore, Piwowar (2013) [25] noted that citation based metrics are slow, since the first citation to a scientific article can take years to happen, concluding that the development of alternative metrics to complement citation analysis is not only desirable, but a necessity. As discussed in [17], each metric has advantages and disadvantages deriving from their inherent biases.

Martins et. al. (2009) [19] proposed a method of assessing the quality of scientific conferences through a machine learning classifier that makes use of multiple features in addition to citations. Gonçalves et al. (2014) [14] quantified the impact of various features on a scholar's popularity. 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 between different researchers: the number of publications and the average quality of the scholar's publication venues.

The idea of reputation, without the direct use of citation data, was discussed by Nelakuditi et al. [22]. 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 the selectivity of a research venue than the acceptance ratio.

In this work, we study the correlation between H-index [15] and *P-score* [26], a novel metric for academic reputation. The *P-score* is calculated with a *reputation flows* model [28] instantiantiated in an academic setting. The *reputation flows* model exploits the transference of reputation among entities in order to identify the most reputable ones. Particularly, the reputation flows consist of a random walk model where the reputation of a target set of entities is inferred using suitable sources of reputation.

3 P-score: a network-based metric

Contrary to citation counting metrics such as the Impact Factor [13, 31, 1, 2] and H-index [4, 3, 12, 6, 7], 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 [28]. 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 [8, 16, 16, 23]. 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.

3.1 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 [27]. 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 126 US graduate programs in Computer Science (CS), exactly the same 126 graduate programs considered by NRC in its 2011 evaluation of CS graduate programs in the USA [1], we ended up with a subset of

[1] http://www.nap.edu/rdp/

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.

4 Correlation with H-indices

The H-index [15] is a composite metric of lifelong scientific contribution that takes into account a researcher's productivity and the citation impact of her publications. A researcher has an H-index value of h if she has h publications that have been cited at least h times. For example, if a researcher has published 20 papers that received at least 20 citations each, then her H-index is 20, assuming that 20 is the highest number for which the definition of the H-index holds. A journal's H-index can be computed in the same way by considering all publications and their collective citation data over a definite period as suggested by Braun et. al. (2006) [**?**]. Since the H-index was proposed in 2005, the original paper by Hirsch [15] was cited 7949 times [2] showing the metric's popularity.

[2] According to Google Scholar, up to the end of 2017.

As any scientific impact metric, the H-index has advantages and disadvantages. Some of its advantages are:

- The simplicity and intuitiveness of its formulation.
- The longer citation time window when compared to impact factors.
- The H-index has good predictive power regarding scientific achievement [6, **?**].

The H-index depends on citations to be calculated, thus it suffers from the same problems as other citation-based metrics. Some of the major disadvantages are:

- The H-index does not distinguish citations from prestigious and peripheral journals, thus it measures popularity rather than quality.
- The time to first citation can be very long.
- It is not easy to collect all relevant information.
- The H-index is field-dependent [33].
- Major sources do not agree about its value.
- It allows scientists to rest on their laurels since the number of citations received may increase even if no new paper is published.
- It is useful for comparing best scientists only. Its power for distinguishing amongst average scientists is not acceptable.
- It lacks sensitivity to performance changes: it can never decrease and is only weakly sensitive to the number of citations received.
- The higher the h index the more citations are needed to increase it.11
- It means that the difference between higher h index values (25 and 26, for example) is much greater than between lower values (4 and 5, for example).13

We propose P-score as a complementary metric to deal with the major problems presented by citation-based metrics such as the H-index and offer further insight on venue quality. P-score is a metric of reputation among peers based on the publication pattern of a set of reference groups of researchers, the seeds. It differs from other network-based metrics because the venues' reputations derive exclusively from the reference set of research groups, better reflecting the actual reputation flows. Additionaly, to calculate venue P-scores we do not need citation data, what makes it easier to obtain than the H-index and avoids the problem of lack of data that arises from the long time to first citation.

Our dataset consists of 882 conferences in the area of computer science for which the H-index was available on Google Scholar [3]. The P-score was calculated with data obtained from DBLP [4]. The reference set of research groups was selected through the randomization process described in section 3.1, yielding the 12 departments listed in table 12.

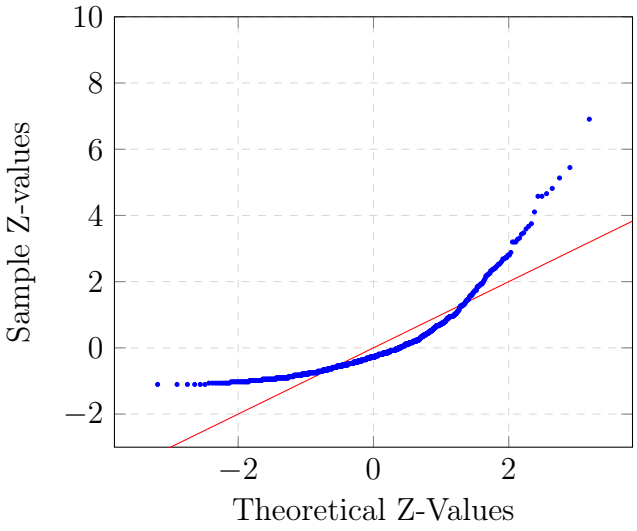

Figure 2: H-index Q–Q plot.

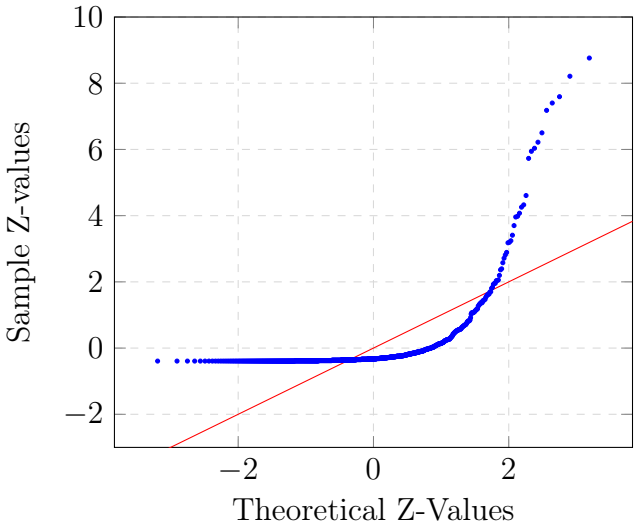

Figure 3: P-score Q–Q plot.

[3]https://scholar.google.com.br/
[4]http://dblp.uni-trier.de/

First, we conducted a normality test to verify if P-scores and H-indexes follow normal distributions. We plotted Q-Q (quantile-quantile) plots for both metrics comparing the sample Z-scores for all 882 quantiles to theoretical Z-scores in figures **??** and **??**. The red line is the identity line for which sample Z-scores and theoretical Z-scores converge. As it is clear from the plots, both distributions are significantly skewed, indicating that both P-scores and H-indexes do not follow normal distributions. This prevents the application of correlation measures such as Pearson's Product Moment Correlation in our analysis, since it requires that variables be approximately normally distributed. For that reason, we chose the Kendall-Tau coefficient to study the correlation between H-index and P-score. It is a measure of rank correlation that assumes a value between -1 and 1. It is higher when when observations have similar ranks, assuming a value of 1 when all ranks are identical and -1 when they are opposite. A Kendall-Tau close to 0 indicates that both ranks are completely uncorrelated.

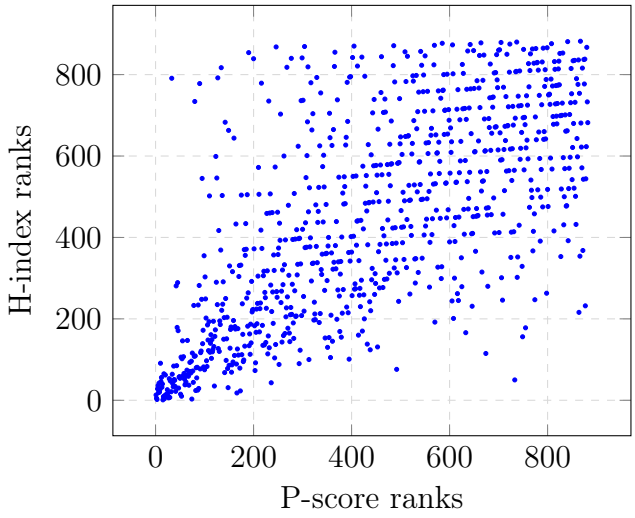

Figure 4: P-score Q–Q plot.

Figure **??** shows the ranks of all data points according to P-score on the X-axis and H-index on the Y-axis. Points closer to the origin have better scores, therefore better ranks. The plot shows an evident linear correlation between the ranks that is confirmed by a very high Kendall-Tau of 0.5200259445, with a p-value smaller than 0.000001, rejecting the null hypothesis (that is, the rankings are uncorrelated) by a good margin.

5 Assessing conferences in CS

Since P-score and H-index are strongly correlated, most datapoints are clustered close to the identity line, showing an agreement between the metrics on these datapoints. However, there are noticeable differences particularly in the cases where the P-score rank is high and

the H-index rank is low (top left corner of figure **??**) or where the H-index rank is high and the P-score rank is low (bottom right corner of figure **??**).

We propose a simple strategy to delimit the 3 groups of datapoints: the center group, consisting of the highly correlated datapoints the top group, consisting the ones with a high P-score and low H-index the bottom group, consisting the ones with a high H-index and low P-score

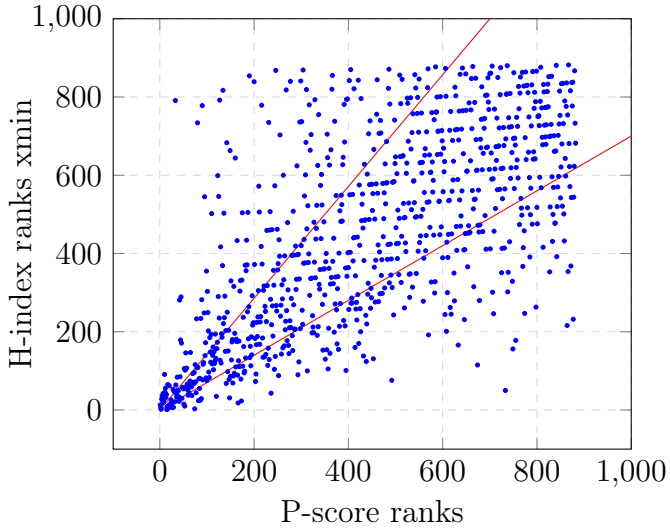

Figure 5: P-score Q–Q plot.

Angle	Kendall-Tau	Items in center group
90	0.5003	882
80	0.5027	877
70	0.5127	862
60	0.5304	836
50	0.5510	795
40	0.5946	734
30	0.6334	637
20	0.7129	483
10	0.8295	281

The cone strategy consists of two lines crossing at the origin that delimit the boundaries between the 3 groups. The datapoints become more scattered in lower positions because venues with low reputation or few citations become increasingly harder to rank because of a lack of signals. The cone strategy tries to deal with this problem by imposing a softer criterion for cut off (from the center group) of venues with lower ranks. Figure **??** shows the cone strategy with an angle of 20 degrees between the boundary lines. Table

?? describes the Kendall-Tau values for varying angles between boundary lines and the number of datapoints in the center group.

We elected the 20 degrees angle in our further analysis.

We propose a strategy

Despite the strong correlation between H-index and P-score.

What are the top case and bottom case?

How to separate the outliers? Cone strategy Pivot

What are the internal cases, the highly correlated conferences? What makes them highly correlated?

What are the top outliers? What makes them top? What are the bottom outliers? What makes them bottom?

What is the conclusion? What is the benefit of using p-score as a complementary metric?

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

6 Conclusions

ZZZ Rewrite entirely.

ACKNOWLEDGEMENTS

Omitted for blind review.

References

[1] K. M. Ahmed. Thomson Reuters Impact Factor 2017. *Current Contents*, 2017.

[2] A. T. Balaban. Positive and negative aspects of citation indices and journal impact factors. *Scientometrics*, 92(2):241–247, 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] J. Bollen, M. A. Rodriquez, and H. Van de Sompel. Journal status. *Scientometrics*, 69(3):669–687, dec 2006.

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

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

[8] S. Brin and L. Page. The anatomy of a large scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1/7):107–17, 1998.

[9] J. Chen and J. a. Konstan. Conference paper selectivity and impact. *Communications of the ACM*, 53(6):79, 2010.

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

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

[12] L. Egghe. The Influence of Transformations on the h-Index and the g-Index. *Journal of the American Society for Information Science and Technology*, 59(8):1304–1312, 2008.

[13] E. Garfield. Citation Indexes for Science. *Science*, 122:108–111, 1955.

[14] G. D. Gonçalves, F. Figueiredo, J. M. Almeida, and M. A. Gonçalves. Characterizing scholar popularity: a case study in the computer science research community. *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 57–66, 2014.

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

[16] A. N. Langville and C. D. Meyer. Google's pagerank and beyond: The science of search engine rankings. *Princeton University Press*, 2008.

[17] L. Leydesdorff. How are new citation-based journal indicators adding to the bibliometric toolbox? *Journal of the American Society for Information Science and Technology*, 60(7):1327–1336, 2009.

[18] T. V. Loach and T. S. Evans. Ranking Journals Using Altmetrics. *ISSI 2015, the 15th International Society of Scientometrics and Informetrics conference*, 6, jul 2015.

[19] W. S. Martins, M. A. Gonçalves, A. H. Laender, and G. L. Pappa. Learning to assess the quality of scientific conferences: a case study in computer science. *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 193–202, 2009.

[20] T. Nane. Time to First Citation Estimation in the Presence of Additional Information. In *The 15th International Conference on Scientometrics & Informetrics*, pages 249–260, 2015.

[21] Nature. This week editorials. *Nature*, 535:466, 2016.

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

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

[24] G. Pinski and F. Narin. Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. *Information Processing and Management*, 12(5):297–312, 1976.

[25] H. a. Piwowar. Value all research products. *Nature*, 493:159, 2013.

[26] 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 *Proceedings of the 24th International Conference on World Wide Web - WWW '15 Companion*, pages 603–608, New York, New York, USA, 2015. ACM Press.

[27] S. Ribas, B. Ribeiro-Neto, E. de Souza e Silva, and N. Ziviani. On the reputation of venues in Computer Science. *Unpublished*, 2015.

[28] S. Ribas, B. Ribeiro-Neto, R. Santos, E. Souza e Silva, A. Ueda, and N. Ziviani. *Random Walks on the Reputation Graph*. ACM Press, New York, New York, USA, 2015.

[29] P. Riikonen and M. Vihinen. National research contributions: A case study on Finnish biomedical research. *Scientometrics*, 77(2):207–222, 2008.

[30] M. Rossner, H. Van Epps, and E. Hill. Show me the data. *The Journal of Cell Biology*, 179(6):1091–1092, dec 2007.

[31] S. Saha, S. Saint, and D. a. Christakis. Impact factor: a valid measure of journal quality? *Journal of the Medical Library Association : JMLA*, 91(1):42–46, 2003.

[32] Y. Sun and C. L. Giles. Popularity weighted ranking for academic digital libraries. *Advances in Information Retrieval*, pages 605–612, 2007.

[33] M. C. Wendl. H-index: However ranked, citations need context [2]. *Nature*, 449(7161):403, 2007.

[34] 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 and Technology*, 62(3):467–477, 2011.