3c-index: Research Contribution across Communities as an Influence Indicator

Thiago H. P. Silva, Lais M. A. Rocha, Ana Paula C. Silva, Mirella M. Moro

Universidade Federal de Minas Gerais, Brazil
{thps,laismota,ana.coutosilva,mirella}@dcc.ufmg.br

Abstract. This paper proposes a new influence metric (called *3c-index*) derived from bibliographic data and social networks analysis. Given a set of communities defined by publication venues, the goal is to measure the degree of influence of researchers by evaluating the links they establish between communities. Specifically, each researcher has a base community where he/she presents greater influence. Then, when such researcher works on a different community (besides the base community), he/she takes new knowledge to that community and transfers influence, which improves the global quality of the communities. By pondering such transfer, we measure the influence of researchers in their and across communities. We also experimentally evaluate the performance of the new index against well known metrics (volume of publications, number of citations and h-index). The results show 3c-index outperforms them in most cases and can be employed as a complementary metric to assess researchers' productivity.

Categories and Subject Descriptors: [**Information Systems Applications**]: Digital libraries and archives; [**Information Retrieval**]: Retrieval models and ranking

Keywords: Research performance, Bibliometric indicators, Ranking strategy

1. INTRODUCTION

A central task in databases is to extract relevant information from their data. Recently, studies on the data provided by social networks have emerged for understanding properties of relationship dynamics and providing models to characterize their behavior [Barabási 2009]. For instance, Barabási et al. [2002] studied the co-authorship of scientists and proposed a model that captures the complexity of evolving networks, and Procópio Jr et al. [2012] proposed modeling sport social networks over time reinforcing the relative importance of past interactions. Regarding academic social networks, in the last two decades, the scientific community has been trying to define metrics to extract knowledge (usually by rankings) from bibliographic data, called *bibliometrics*. Traditional metrics include volume of produced papers, their citation numbers and the well-known *h-index*, which combines both volume and citation received [Hirsch 2005]. However, all these purely quantitative metrics have been criticized, e.g. [Hicks et al. 2015] and [Zhang 2009].

More recent studies have been changing the evaluation perspective from one researcher (index based on one's own publications [Bollen et al. 2009]) to how the individual (and groups) relates to his/her peers and communities. Specifically, Academic Social Networks studies have evaluated the performance of each researcher individually and grouped in teams, communities, graduate programs and others [Alves et al. 2013; Brandão et al. 2014; Freire and Figueiredo 2011; Lopes et al. 2011; Newman 2004]. More complex analyses explore qualitative aspects such as the influence of publication venues [Garfield 1999], communities [Alves et al. 2013; Silva et al. 2014; Silva et al. 2015b], collaboration networks [Brandão et al. 2014; Freire and Figueiredo 2011], academic profiles [Gonçalves et al. 2014; Lima et al. 2015] and others.

The authors would like to thank CNPq and FAPEMIG for their individual grants and scholarships.

Regarding the social perspective, some studies focus on analyzing the advantage created by location within the social structure. For instance, Granovetter [1973] introduced the idea of *weak ties*; Newman [2004] explored the number of shortest paths between pairs of nodes to measure the influence of information flows among individuals; and Burt [2004] defined as *brokers* the people who build social capital to position themselves at strategic points in the network (*structural holes*). In other words, in academic social networks, researchers who connect different groups should bring more *influence* to those groups. This hypothesis is explored in other contexts (e.g., in economic data), and our study shows that it can also be applied as an influence indicator of researchers in Computer Science (CS).

This article contributes to this discussion in two aspects: (*i*) we show how to expand the concept of building bridges for exploring the *researcher-community relationship* and (*ii*) we propose a new strategy that measures the influence and, at the same time, deals with the problem of discrepancy between communities (e.g., different number of authors, papers and citation rates). To do so, the indicator projects the influence according to a same community. We also analyze the propagation of influence among *communities*, which are formed by researchers who share common interests by publishing papers in the same conferences. We do so by defining that each researcher has a **base community** as the one where he/she has the best performance. We follow the concepts of diversity and novelty [Burt 2004], and we consider that when a researcher works on a different community (besides the base community), he/she transfers new knowledge from the original context to other areas. Therefore, there is a share of new ideas, techniques and methodologies that result in contributions to the development of the overall research quality.

The new proposed indicator, called **3c-index** (*Cross-Community Contribution*), uses a generic function of *score* to evaluate the influence degree of a researcher in different communities; then, it projects such value in the base community. The main advantages are: *flexibility*, because the *3c-index* allows using any score functions (e.g., citation count, publication volume, h-index, sociability, number of co-authors, *rising star*, volume of recent production and complex network metrics); and *equality*, because it reveals different profiles of the communities involved by projecting in a single community. Our analysis shows *3c-index* brings different results in the rankings in relation to traditional indices and it is suitable to identify who the researchers located in structural holes are, i.e., those acting in different parts of the network making it stronger.

Overall, we provide results that cannot be directly measured from a simple SQL query in a bibliographic database. We also argue that quality and influence are not defined by a unique index. Then, our objective is to enable a *complementary* analysis. According to Hicks et al. [2015], a more complete evaluation needs experts and qualitative assessments as well. Finally, this article is an extended version of [Silva et al. 2015] with: (*i*) a more detailed discussion about social concepts; (*ii*) a deeper analysis of social interactions among communities; and (*iii*) a comparison between our results and a cross-area strategy proposed by Lima et al. [2013].

Next, we go over related work in Section 2. Then, Section 3 introduces the new index *3c-index*. Section 4 presents the methodology employed over a broad analysis of 3c-index applicability in Section 5 and an experimental analysis in Section 6. Finally, Section 7 concludes this article.

2. RELATED WORK

Judging people is always hard. In academia, researchers are usually judged by their productivity, which is commonly assessed by different metrics applied to their publication records. In this article, we focus on ranking researchers by considering a more social perspective: we consider data from digital libraries and apply social analysis into the researchers' publications. Specifically, we build social networks by grouping researchers into communities and then measure the knowledge they transfer among those communities. Hence, this section covers related work on the following topics: individual and group evaluations in social network analysis, productivity analysis, evaluating and ranking researchers.

One of the first works dealing with the social perspective of publication records was performed by Granovetter [1973], who proposed the concept of *weak ties* as the relationships that link different parts of the network by building bridges. Newman [2004] considered the role of a researcher in the network and explored the number of shortest paths from a central node to other nodes and, thus, measured the influence of information flows among individuals (*brokering*). Similarly, Burt [2004] defined *brokers* as the people who build social capital when positioning strategically in the network (structural holes). Our work contributes to this discussion by analyzing the influence propagated by researchers among communities, i.e., we show that we can extend the concept of building bridges to explore the *researcher-community relationship*.

Regarding CS, Freire and Figueiredo [2011] explored the *external* collaborations of individuals and groups. They identified influential individuals and groups through the intensity of their relationships with individuals from outside of their group. Similarly, Silva et al. [2014] explored the concept of *community* to rank publication venues according to the degree of external relationships that connects communities. The results indicate that researchers who publish in others communities are more likely to introduce new ideas from their competence to other contexts. At the end, such researchers are considered highly influential by connecting different parts of the network (acting as *weak ties* or *brokers* in the network). Regarding information flow among communities, Silva et al. [2015a] proposed four metrics to measure the dynamics among researchers in communities (*Permanency*, *Migration*, *Exclusivity* and *Plurality*). Here, we propose to evaluate researchers by exploring their connections with their set of communities and, from there, explore the influence degree that can be transmitted through different communities.

In terms of productivity analysis, considering the existing different publication patterns is crucial for a fairer evaluation. For instance, Gonçalves et al. [2014] characterized the productivity of researchers and concluded that there are indeed well-defined profiles. Research communities have also distinct patterns in relation to the number of members, research subjects, publications rates, etc. To overcome such problem, Silva et al. [2015a] established properties (*Equality*, *Relativity* and *Temporality*) that should be considered in the methodology for a fairer comparison. Here, we define *base community* of a researcher as the one where he/she has the best performance, and the influence is measured on the *perspective of each research in relation to the base community*, resulting in a fairer assessment as well.

Other problem is how to evaluate researchers, as there is no consensus on the ideal metrics for a fairer decision process. To overcome it, some studies use reference groups with members of unquestionable relevance in their assessment. For instance, Hirsch [2005] used the Nobel prize winners to assess the proposed h-index, Lima et al. [2013] considered grant receivers, and Alves et al. [2013] considered awards distributed by the ACM (*Association for Computing Machinery*). Likewise, we use a ranking based on the recognition of contributions and/or innovations from ACM, such as the ACM Awards (given by specifics Special Interest Groups - SIGs[1]) and the distinct member grades to recognize the professional accomplishments of ACM's members (fellow, distinguished and senior)[2].

Regarding ranking strategies, Maroun and Moro [2014] explored a probabilistic approach to ranking researchers in uncertain data scenarios. On a different perspective, Silva et al. [2015b] proposed a strategy based on *similar groups* formed by members with common characteristics regarding the temporal link with their publication venues. Lima et al. [2013] created a generic strategy for researchers from multiple areas by projecting productivity under a single perspective. Similarly, we apply a ranking strategy based on percentiles and map the values from different communities to a *base community* of the researcher. In addition, we correct such values by exploring the influence degree existing among communities by a contribution perspective of each author. Thus, the higher the influence degree, the more likely there is knowledge transfer for different communities.

[1]ACM Special Interest Groups (SIGs): `http://www.acm.org/sigs`
[2]ACM memberships: `http://awards.acm.org/grades-of-membership.cfm`

Table I: Example of influence degree for Christos Faloutsos.

Community	Percentile Rank	Influence Degree
SIGKDD	0.995	0
SIGMOD	0.92	0.07
SIGMETRICS	0.68	0.32
SIGAPP	0.48	0.52

Overall, we contribute to a new ranking strategy based on social perspective. We consider that when a researcher works on a different community (besides the base community), he/she transfers new knowledge from one context to another. Then, we explore brokerage and closure [Burt 2004] to consider both the potential of knowledge acquired and the sharing of information, as explained next. One main advantage of this strategy is that it requires access only to the researchers' publication records, which come straight from a digital library.

3. RESEARCH CONTRIBUTION ACROSS COMMUNITIES

In general, researchers have one or a group of areas of expertise [Lima et al. 2013; 2015]. Indeed, the knowledge transfer among areas is crucial to contribute to the progress of Science, since it enables the application of well-known ideas from one context to solve problems on other research areas [Sun et al. 2013; Silva et al. 2014]. Here, we propose the metric **3c-index** (cross-community contribution) that aims to identify and quantify the transfer of knowledge among different communities. We explore two social concepts [Burt 2004]: **closure** as the potential knowledge acquired and **brokerage** as the potential for sharing information. Then, 3c-index measures the influence of researchers according to their specialties (defined as *degree of influence*) to other contexts.

The 3c-index deals with the unique features of different research areas (for instance, different patterns of publication) establishing the concept of *base community* of a researcher. The *base community* is the one in which the researcher achieves best performance in terms of contribution. Here, the *base community* is defined in terms of percentile ranks. Formally, p_i^c is the percentile ranking for the researcher i in a given community c (from the set of all communities a researcher publishes in), defined by: $p_i^c = \frac{l_i^c + 0.5e_i^c}{N^c}$, where N^c is the number of researchers in the community c, l_i^c and e_i^c the number of researchers with rankings values lower or equal than to the researcher i, respectively. For instance, consider a community with 100 researchers and their different scores (no ties). A particular researcher i at position 10^{th} in the rank has l_i^c and e_i^c equal to 89 and 1, respectively. These values attribute to researcher i a percentile of 89.5%. For all communities the researcher i publishes in, p_i^c is calculated and the *base community* b_i is the one in which the researcher has the highest percentile value, i.e., $b_i = \text{argmax}_{c \in C} p_i^c$. Therefore, the base community designs the potential knowledge acquired, i.e., it contains information (e.g., ideas and techniques) from a closed group of people (defined by closure).

Given the *base community* (b_i) of a researcher, the degree of influence $infd_i^c$ in a community c of a researcher i is defined as the difference between the percentiles obtained by i in the base community and in the community c, i.e., $infd_i^c = b_i - p_i^c$. Here, without loss of generality, we consider as a community the set of authors who have published in a SIG (ACM Special Interest Group). The 3c-index quantifies the influence propagated by its members among different communities. To illustrate our metric, Table I depicts the degree of influence propagated by the researcher *Christos Faloutsos*[3]: his *base community* is SIGKDD (p_i^c highest value) whose value is also similar to SIGMOD, with moderate degrees of contributions for SIGAPP and SIGMETRICS. Note that $infd_i^{b_i} = 0$ for the base community, because we assume there is no knowledge *transfer* in that case. The cases of highest percentile ranks represent the potential knowledge acquired by him (closure) and, in contrast, high influence degrees represent the potential of sharing information (brokerage). Therefore, degree of

[3]The h-index metric is used to rank the author and to get his percentiles.

Table II: Database statistics comprising the time interval [2001,2010].

SIG	Authors	Articles	Authors / Articles	Citations (10^{-3})	Citations / Authors	Citations / Articles	#Awards
SIGACCESS	849	418	2.03	7.9	9.25	18.79	2
SIGAda	169	135	1.25	0.7	4.21	5.27	15
SIGAPP	6,732	3,078	2.19	45.2	6.71	14.68	7
SIGCOMM	816	338	2.41	104.7	128.35	309.86	14
SIGCSE	1,957	1,143	1.71	23.1	11.78	20.17	25
SIGDOC	460	316	1.46	3.0	6.61	9.62	9
SIGIR	2,313	1,528	1.51	85.2	36.86	55.79	5
SIGKDD	2,172	1,075	2.02	109.3	50.33	101.69	17
SIGMETRICS	1,053	449	2.35	31.2	29.63	69.49	5
SIGMOBILE	748	276	2.71	66.9	89.41	242.31	7
SIGMOD	2,196	1,098	2	103.9	47.32	94.65	29
SIGUCCS	838	655	1.28	1.7	2	2.56	5
All	18,511	10,509	1.76	582.8	31.49	55.46	137

influence captures researchers who act as *bridges among communities* regarding the probability of sharing information, i.e., it is more likely that Faloutsos can bring knowledge from SIGKDD and SIGMOD to other contexts in SIGMETRICS and SIGAPP.

Given that scientific communities have different profiles (e.g., number of articles, publications and citation rates, etc), a normalization factor is required to attenuate such differences. Following the rank model proposed by Lima et al. [2013], our strategy considers the percentile ranks of researchers mapped to their base community. The *3c-index* of a researcher i is then defined as

$$3c\text{-}index(i) = f_b(b_i) + \sum_{c \in C} (b_i - p_i^c) f_b(p_i^c),$$

where $f_b(x)$ is the projection function that maps the percentile x to the respective value ranking on the researcher's base community b. The index value is the combination of the base community information (potential knowledge acquired) and the contributions made to external communities (potential of sharing information). It is worth noting that the ranking strategy allows different metrics to be used in the projection function, such as number of citations, volume of publications, h-index, number of students and co-authors, complex network metrics, and others, giving flexibility to 3c-index.

4. METHODOLOGY

This section presents the methodology underlying the analysis and experimental validation of 3c-index.

Publication Dataset. The 3c-index is based on the concept of community. Without loss of generality, we consider the communities from the ACM SIGs. Each SIG organizes one or more scientific events focusing on specific topics of interest. Besides promoting big events, SIGs grant awards to the members for their contributions and innovations (*ACM Awards*). We consider 12 SIG conferences as communities, which are top-tier in CS as top researchers publish in them. Table II presents the communities and their statistics. We collected their sets of publications from the DBLP[4], within the time frame [2001, 2010]. Citations were collected from *Google Scholar*[5], by matching 4-tuples formed by {title, year, authors, venue} in January 2015[6]. Finally, the dataset contains more than 18 thousand authors, over 100 thousand papers with more than half a million citations received. The average number of authors per paper is 1.8, whereas the average number of citations is 55.5.

[4]DBLP: `http://www.informatik.uni-trier.de/~ley/db`

[5]GS: `http://scholar.google.com`

[6]We use publications until 2010 so that all have at least 4 years to be cited (i.e., the goal is to avoid high levels of biased when considering freshly published papers that have had not enough time to be cited.

Ranking Baseline. We consider three well-known bibliometric indices as both base for comparison and score function: publication volume, citations count and h-index. Such indices are widely used in online research-oriented search engines as *Google Scholar*, *Microsoft Academic Search* and *AMiner*. Our ranking strategy is generic (flexible) and allows using any *score function* (Section 3). We named as 3c-citations, 3c-volume and 3c-h-index the 3c-index metric with *score functions* for number of citations received, volume of publications and h-index, respectively.

Ground-truth. Ranking researchers is a challenging task and there is no consensus on the ideal metrics for a fairer decision process. Alternatives for evaluation include of determining reference groups containing members with unquestionable relevance. For example, Hirsch [2005] used the winners of Nobel Prizes to evaluate his metric (h-index), Lima et al. [2013] used sets of grant receivers, and Alves et al. [2013] considered authors awarded by ACM. Likewise, we use a ranking based on the recognition of contributions and/or innovations from ACM: the ACM Awards (given by specific SIGs) and the distinct member grades that recognize the professional accomplishments of its members (fellow, distinguished and senior). We built a *ground-truth* of influential researchers formed by 137 (out of 18,511) winners of at least one *ACM Award*. We assume that those distinguished researchers have unquestionable relevance and impact on the communities that awarded them.

Evaluation Metrics. We evaluate the ranking metrics by comparing their results to the set of researchers who were awarded by ACM. Then, we consider two metrics: one to compare the ranking result and another to validate the ranking. Specifically, we consider the Spearman's correlation and the Jaccard coefficient to measure the dissimilarity between ranks. We aim at verifying the independence between our approach and the baselines, i.e., we want to verify if our index brings novelty to the ranking[7]. Overall, such behavior should allow using 3c-index as a complementary metric to evaluate researcher's influence.

To further investigate the efficacy of 3c-index, we use the DCG metric (*Discounted Cumulative Gain*). In summary, it measures the quality of the ranking and applies a log-based discount factor to penalize relevant items in lower positions [Järvelin and Kekäläinen 2002]. Formally, the DCG at ranking position k is defined as

$$DCG@k = g_1 + \sum_{i=2}^{k} \frac{g_i}{log_2(i)},$$

where g_i denotes a binary relevance (i.e., 1 if the researcher is winner of an *ACM Award*, 0 otherwise). We use the normalized version (nDCG), which is obtained by dividing the DCG@k with the best possible ranking in the same cutoff k.

5. ANALYSIS

This section presents a broad analysis of the 3c-index applicability in the context of scientific communities. We first verify how external contributions impact a particular community, which allows applying the proposed index (Section 5.1). Using 3c-index, we rank the most influential researchers in SIGs (Section 5.2) and then, we compare our proposal with well-known indices to show the independence between them (Section 5.3).

5.1 Knowledge Transfer

By definition, 3c-index quantifies the external member participation in a specific community. Then, a crucial point to be firstly checked is the actual existence of knowledge transfer among the communities under analysis. Figure 1 shows the proportion of researchers that have a specific SIG as their base

[7]These same metrics are also employed to measure dissimilarity by Silva et al. [2015b].

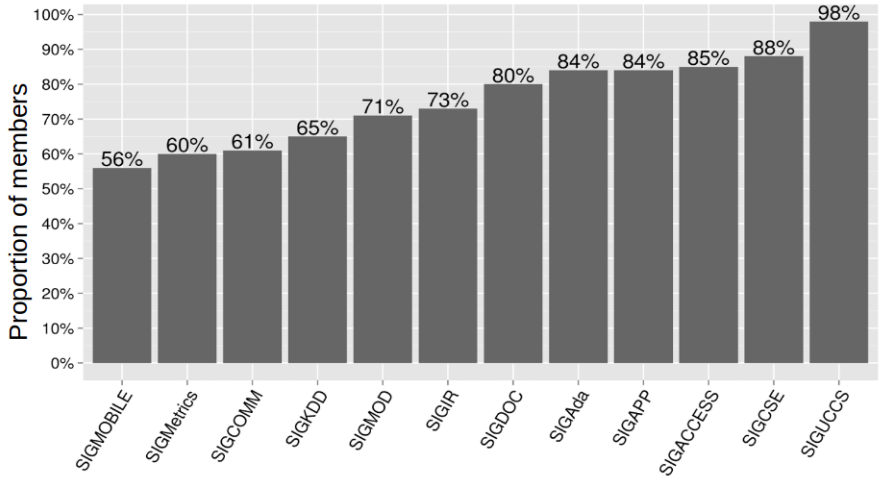


Fig. 1: Percentage of authors (with more than one article) having the own community as their *base community* (in accordance with the h-index).

Table III: Proportions of knowledge transfer among Communities (ACM SIGs).

	SIGKDD	SIGAPP	SIGIR	SIGMOD	SIGCOMM	SIGMETRICS	SIGMOBILE	SIGCSE	SIGACCESS	SIGDOC	SIGAda	SIGUCCS	Transfered
SIGKDD	-	3.38%	7.06%	9.05%	0.25%	0.80%	0.21%	0.25%	0.00%	0.04%	0.00%	0.00%	**21.07%**
SIGAPP	3.64%	-	5.58%	3.85%	0.42%	0.72%	0.47%	1.14%	0.47%	1.06%	0.13%	0.04%	**17.51%**
SIGIR	7.57%	3.89%	-	2.54%	0.34%	0.63%	0.17%	0.25%	0.30%	0.04%	0.00%	0.04%	**15.78%**
SIGMOD	6.64%	2.37%	2.66%	-	0.59%	0.55%	0.68%	0.21%	0.08%	0.00%	0.00%	0.04%	**13.83%**
SIGCOMM	0.21%	0.38%	0.25%	1.06%	-	4.74%	2.71%	0.34%	0.04%	0.00%	0.00%	0.04%	**9.77%**
SIGMETRICS	0.55%	0.51%	0.21%	0.72%	5.41%	-	1.95%	0.34%	0.00%	0.00%	0.00%	0.00%	**9.69%**
SIGMOBILE	0.30%	0.30%	0.08%	0.55%	1.95%	1.95%	-	0.04%	0.00%	0.00%	0.00%	0.00%	**5.16%**
SIGCSE	0.17%	1.31%	0.21%	0.25%	0.47%	0.25%	0.13%	-	0.51%	0.08%	0.21%	0.08%	**3.68%**
SIGACCESS	0.04%	0.25%	0.21%	0.08%	0.04%	0.00%	0.04%	1.23%	-	0.30%	0.00%	0.00%	**2.20%**
SIGDOC	0.00%	0.47%	0.04%	0.04%	0.00%	0.00%	0.00%	0.00%	0.04%	-	0.00%	0.00%	**0.59%**
SIGAda	0.00%	0.04%	0.00%	0.00%	0.00%	0.00%	0.04%	0.30%	0.00%	0.00%	-	0.00%	**0.38%**
SIGUCCS	0.00%	0.04%	0.04%	0.00%	0.00%	0.04%	0.00%	0.21%	0.00%	0.00%	0.00%	-	**0.34%**
Received	**19.12%**	**12.94%**	**16.37%**	**18.15%**	**9.48%**	**9.69%**	**6.39%**	**4.31%**	**1.44%**	**1.52%**	**0.34%**	**0.25%**	**100%**

community. Here, we consider as score function the percentile rank according to the h-index[8] and members with more than one publication. Some interesting results arise. For instance, 56% of SIGMOBILE members have the SIGMOBILE itself as their base community, leading to a 44% of contribution coming from outside the community. In contrast, SIGUCCS has only 2% of external contribution, suggesting that this conference tends to be a much closer community (in fact, the focus of this SIG is giving logistical support to information technology services to educational institutions[9]). A range of 12% to 40% of external members contribute to the other communities. These results enforce the existence of knowledge transfer among different communities, an important premise of 3c-index.

Table III shows how researchers' influence spreads over SIGs communities. Each cell has the global proportion of knowledge transfer from one SIG to another. Each row shows the knowledge transferred; for example, row SIGKDD shows the propagation of knowledge from SIGKDD, which is bigger to SIGMOD (9.05%), SIGIR (7.06%) and SIGAPP (3.38%). Each column shows the knowledge received; for example, column SIGIR shows bigger proportions from SIGKDD (7.06%), SIGAPP (5.58%) and SIGMOD (2.66%) to SIGIR. The last column represents the total proportion transferred (for example, SIGKDD transferred a total of 21.07%), and the last line represents the total received (for example,

[8]Similar results are obtained by using the number of citations received and volume indices.
[9]SIGUCCS: `http://www.siguccs.org/`

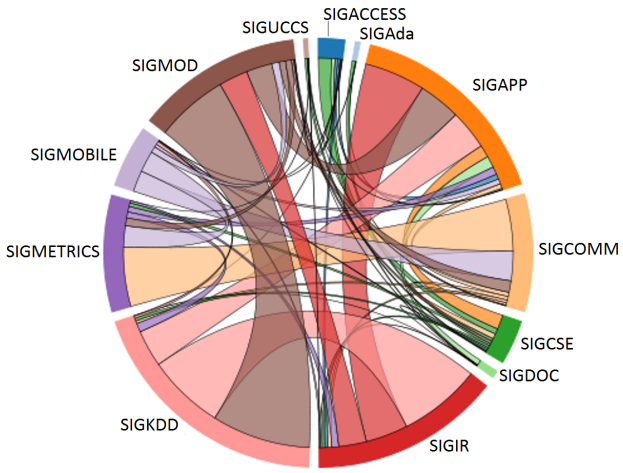


Fig. 2: Knowledge transfer among Communities (ACM SIGs).

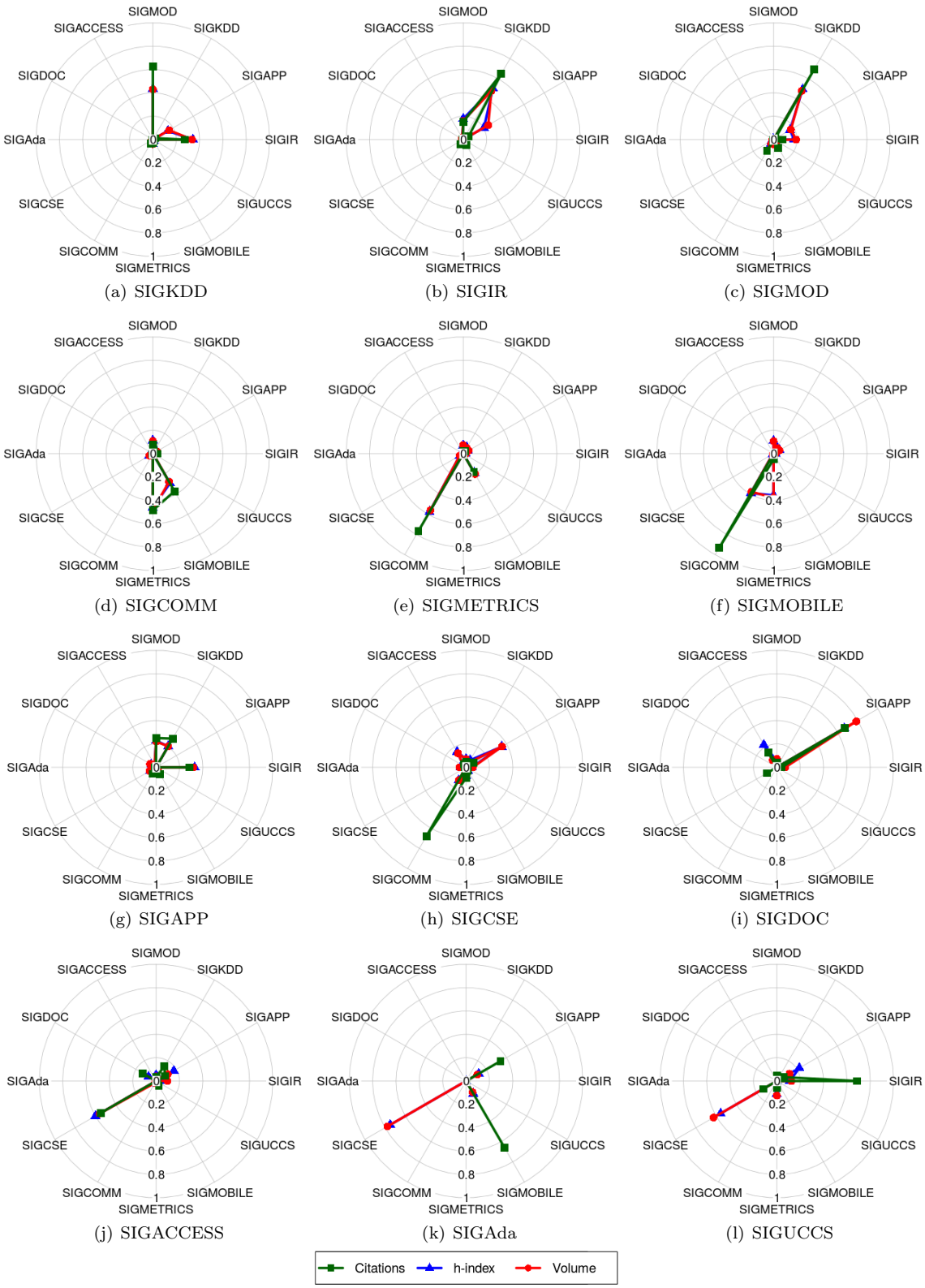
SIGKDD received a total of 19.12%). As expected, there is more proportion of transfer for those with high number of papers (see Table II), since those values are global.

Figure 2 illustrates the same proportions through a circular chart for a better and more intuitive visualization. Each circular sector is proportional to both transferred and received contribution. For instance, if the researcher's base community is SIGMOD and he also publishes at SIGKDD, his 3c-index migrates from SIGMOD to SIGKDD. SIGUCCS, SIGAda and SIGDOC communities have the lowest knowledge transfer values, whereas SIGKDD, SIGAPP, SIGIR and SIGMOD communities have the highest values. Overall, there is a non-negligible knowledge transfer among communities, where bigger proportions are between SIGs with similar research tracks as {SIGKDD, SIGMOD and SIGIR} and {SIGCOMM, SIGMETRICS and SIGMOBILE}.

Figure 3 shows the proportions among communities normalized by each one (avoid bias due to the number of papers). For instance, Figure 3a shows three variables (citations, h-index and volume) and their values (0 to 100%) of knowledge transferred from SIGKDD to others SIGs. In this case, there are bigger flows to SIGMOD (around 60% of external citations) and SIGIR. Note that, by definition, there is no knowledge transfer to the same SIG. The first line (Figures 3a, b and c) shows the results of SIGKDD, SIGIR and SIGMOD with high interactions between them. A similar interaction behavior appears in the second line for SIGCOMM, SIGMETRICS and SIGMOBILE (Figures 3d, e and f). Thus, SIGs with similar techniques and research tracks tend to have high reciprocal contributions. Another possible analysis is to check the strength of variables. For instance, SIGAda (Figure 3k) shows almost 80% of volume and h-index to SIGCSE, whereas more than 60% of its citations received was made in SIGMOBILE. Overall, all figures show specific patterns and confirm the existence of knowledge transfer among communities.

5.2 Influential Researchers in SIGs

We rank the most influential researchers for SIG communities in accordance with 3c-index by using score functions as the number of citations received, volume and h-index. We then contrast the top-positions to the winners of ACM SIG Awards and ACM member distinction in Table IV. Note that 3c-index ranks well-known researchers in CS at the first positions. The 3c-volume and 3c-h-index rank first researchers (names in bold) that are recognized in their communities: five for 3c-volume and four for 3c-h-index. Instead, 3-citation ranks two outstanding researchers. Overall, 3c-index successfully measures the importance of the knowledge transfer to promote the most outstanding researchers careers.



(a) SIGKDD (b) SIGIR (c) SIGMOD

(d) SIGCOMM (e) SIGMETRICS (f) SIGMOBILE

(g) SIGAPP (h) SIGCSE (i) SIGDOC

(j) SIGACCESS (k) SIGAda (l) SIGUCCS

Fig. 3: Proportion of knowledge transfer by SIGs normalized by number of papers.

Table IV: Researchers best ranked according to 3x-index in SIGs. In bold, the winners of ACM Awards, ♯*ACM fellow*, ‡*ACM distinguished scientist*, †*ACM senior member*.

Position	3c-citation	3c-volume	3c-h-index
1^{st}	**Scott Shenker**♯	**W. Bruce Croft**♯	**Jiawei Han**♯
2^{nd}	Ion Stoica♯	**Christos Faloutsos**♯	**Christos Faloutsos**♯
3^{rd}	M. Frans Kaashoek♯	**Surajit Chaudhuri**♯	**Scott Shenker**♯
4^{th}	David R. Karger♯	**Jiawei Han**♯	**W. Bruce Croft**♯
5^{th}	Sylvia Ratnasamy	**Scott Shenker**♯	ChengXiang Zhai‡
6^{th}	Mark Handley	ChengXiang Zhai‡	**Surajit Chaudhuri**♯
7^{th}	Paul Francis	Philip S. Yu♯	Wei-Ying Ma‡
8^{th}	Richard M. Karp♯	Zheng Chen†	Zheng Chen†
9^{th}	**Jon M. Kleinberg**♯	Divesh Srivastava♯	Philip S. Yu♯
10^{th}	Dina Katabi♯	Leif Azzopardi	Divesh Srivastava♯

Table V: Spearman correlations for the set of the top-x rank positions between citations (C), volume (V), h-index (H) and their 3c-index counterparts ($3c$). Rankings ordered by each original metric.

Position	ρ_{3cc}	ρ_{3cv}	ρ_{3ch}
10	0.04	0.57	0.67
20	0.56	0.67	0.66
40	0.45	0.27	0.12
60	0.27	0.13	0.06
80	0.30	0.09	0.13
100	0.34	0.21	0.16

Table VI: Jaccard distance for the set of the top-x researchers between citations (C), volume (V), h-index (H), and their 3c-index counterparts ($3c$).

Top-x researchers	$d_J(3cc)$	$d_J(3cv)$	$d_J(3ch)$
10	0.33	0.46	0.18
20	0.10	0.26	0.40
40	0.33	0.40	0.43
60	0.29	0.48	0.38
80	0.24	0.43	0.40
100	0.28	0.35	0.40

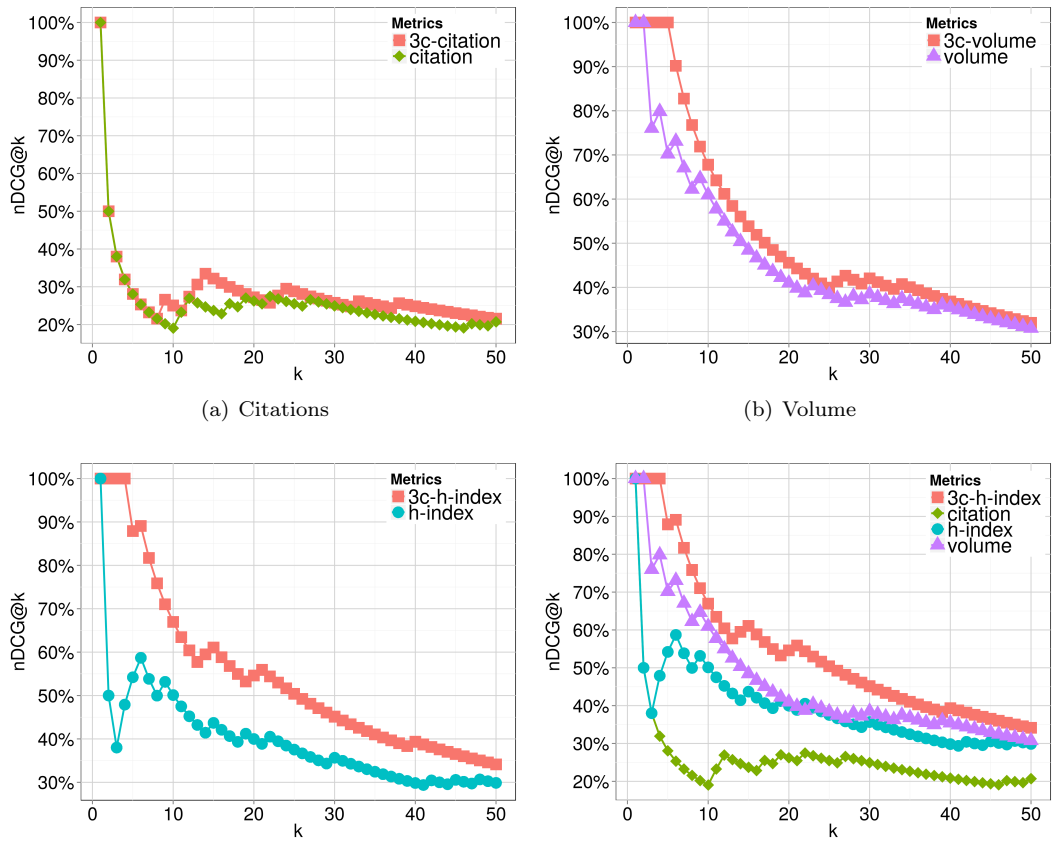
5.3 Ranking Similarities

In this section, we verify if 3c-index brings new researchers (novelty) to its resulting raking, i.e., if the sets of researchers and their positions at the top rankings are different from other indices. We use Spearman's and Jaccard coefficients for measuring dissimilarity with respect to rank position and to the set of researchers ranked, respectively. Our goal is to show if our proposed index captures new information and, thus, if it can be used in a complementary way with other influence indices. Table V shows the Spearman's correlation between citations, volume, h-index indices and their counterparts using 3c-index. Correlations are strong and very strong for $\rho \geq 0.7$, moderate for $0.4 \leq \rho < 0.7$, and weak for $\rho < 0.4$.

For the top-10 rank positions, correlation between 3c-index and total number of citations (ρ_{3cc}) is almost zero[10]. Moderate values are found when the rank considers the top-40. Rankings above 40 have weak correlation values. Correlations between 3c-index and volume (ρ_{3cv}) and 3c-index and h-index (ρ_{3ch}) have moderate values to top-10 and top-20 rank positions, decreasing to weak correlation values for rankings above 40 positions.

To investigate if the rank produced by 3c-index is dissimilar regarding the set of researchers (diversity), Table VI shows the Jaccard distance between the indices. Except for the top-20 rank given by the citation index ($d_J(3cc)$) and the top-10 rank given by the h-index ($d_J(3ch)$), dissimilarity values are at least 24%. Overall, 3c-index differs from the baselines by bringing different positioning and new researchers to the top-ranking.

[10]The first 10 researchers in citation index are ranked between the top-16 by 3c-index, with only two researchers with the same position in both rankings.



(a) Citations (b) Volume

(c) h-index (d) 3c-h-index

Fig. 4: Comparison between 3c-index and the *baselines* according to the nDCG.

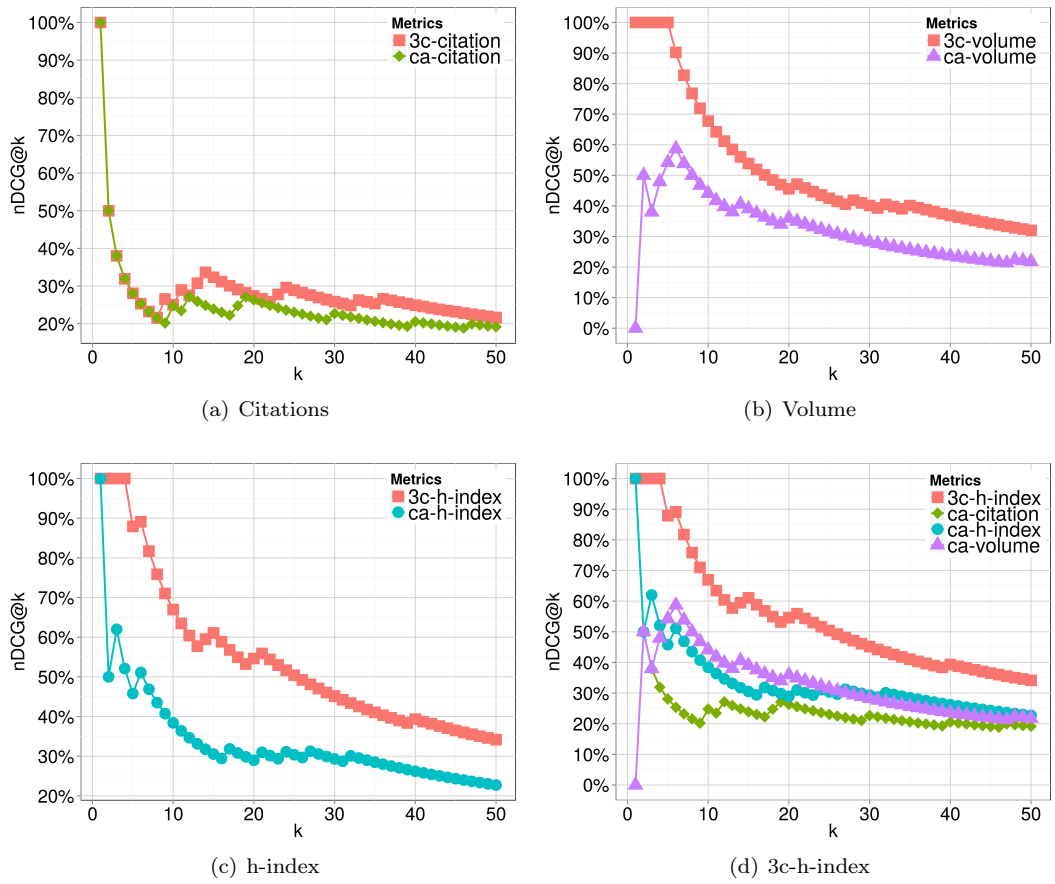
6. EXPERIMENTAL VALIDATION

After comparing 3c-index to citation, volume and h-index, we now validate 3c-index against a ground truth composed by award winning researchers. Out of the 18,511 researchers, the goal is to have the metrics ranking the 137 ACM award winners at the top. We compare the ranking results to those given by citations, volume and h-index in Section 6.1 and ca-index (cross-area index) proposed by Lima et al. [2013] in Section 6.2.

6.1 Comparison with citations, volume and h-index rankings

Figure 4 compares the ranking produced by 3c-index to those by citation, volume and h-index in terms of attained *nDCG@k*, with $1 \leq k \leq 50$. The results show 3c-index outperforms the other indices in most cases. Note that both 3c-index and number of citations are tied in the top eight positions in Figure 4(a). Then, 3c-index outperforms or has equal performance with the citation index, except for the 22^{nd} position. Regarding volume of publications in Figure 4(b), there is a tie in the top two positions, and 3c-index outperforms volume from the 3^{rd} position on.

Indices that rely on citations and publication volume tend to be sensitive to *outliers*: few publications may attract many citations, then introducing noise in the evaluations that consider sets of papers. Publication volume is also sensitive to cases where an author is ranked in the top positions due to a large number of publications at specific venues (for instance, venues with higher acceptance rate



(a) Citations (b) Volume

(c) h-index (d) 3c-h-index

Fig. 5: Comparison between 3c-index and the *baselines* according to the nDCG.

and/or lower quality). To overcome such weaknesses, h-index attempts to control the sensitivity by considering both volume and number of citations received. Figure 4(c) depicts 3c-index results using h-index as the score function. Our proposed index outperforms h-index in all positions. Figure 4(d) shows the comparison between 3c-index (with h-index as its score function) and three standard metrics summarizing the good performance of our ranking strategy.

Finally, regarding flexibility and equality, 3c-index allows flexibility for any score function can be applied on its definition. Here, we consider three well known metrics, but others could be easily applied as well. It also allows equality by revealing different profiles of the communities (SIGs) through projecting them into a single one (base community). In a sense, it comes from an overall normalization by the strongest community of each individual, then allowing a fairer comparison.

6.2 Comparison with ca-index ranking

We now compare our approach against the ca-index proposed by Lima et al. [2013]. Recall that ca-index addresses the problem of ranking researchers across multiple areas by projecting the researchers' scores in their base area (Section 2). In our context, a SIG is considered as an *area*, given that each SIG represents different research interests (e.g., SIGMOD for database management systems and technology[11], and SIGCOMM for communications and computer networks[12]).

[11]SIGMOD: `http://www.sigmod.org/`
[12]SIGCOMM: `http://www.sigcomm.org/`

Figure 5 shows the results of the ranking produced by 3c-index and ca-index using score given by the number of citations, volume of publications and h-index. In most cases, 3c-index outperforms its counterparts. Results related to number of citations in Figure 5(a) have similar behavior to results shown in Figure 4(a). Regarding the volume of publications in Figure 5(b) and h-index in Figure 5(c), 3c-volume and 3c-h-index outperform ca-volume and ca-h-index, respectively. Figure 5(d) summarizes the good performance of our proposed index.

The ranking strategy of ca-index has three properties to follow [Lima et al. 2013]: (*i*) plurality indicating that researcher's productivity should be assessed in all areas of his/her publications; (*ii*) diversity indicating that the profile of each research area should be considered; and (*iii*) equality indicating that all research areas should be regarded as equally important and, thus, deserving of scientific merit. Analyzing these properties in our dataset, we note that *not all* CS research areas are considered, because we use just a subset of SIGs (Table II). In fact, the *equality* property puts SIGUCCS (focus on logistical support to information technology services to educational institutions) with equally importance as SIGMOD. Thus, ca-index is more susceptible to rank at the top researchers from both conferences because, in terms of productivity, the premise of each *research area* as deserving the same scientific merit has negatively affected this behavior. Therefore, ca-index is not indicated to be used in a *closed community perspective*; however, it still has good results when assessing *research in multiple area scenario* as reported by Lima et al. [2013].

7. CONCLUSION

In this article, we presented a new influence index (3c-index) based on the social perspective as a combination of base community information (closure, potential knowledge acquired) and the contributions made to external communities (brokerage, potential of sharing information). One main advantage is to require only the publication records of sets of researchers. We have also performed an analysis whose results showed independence between 3c-index and traditional indices in relation to: overall rank positioning and ranking diversity at the top positions. Such results endorse using 3c-index as a complementary information towards a fairer judging of researchers. Furthermore, 3c-index is flexible as it can be applied with any score function (volume, citation, metrics of complex networks, etc). The experimental evaluation results showed 3c-index outperformed traditional metrics and a cross-area index in the task of ranking influential people in a real dataset (ACM awards winners according to the SIGs). We show the importance of knowledge and expertise of the researcher in a base community, building cohesive groups (exploring the closure concept) and the gain given by influential researchers with good position in the network by linking different communities (exploring the brokerage concept). Overall, measuring the impact of knowledge transfer among communities (through brokerage and closure concepts) and dealing with the different communities patterns (projection function) form a good strategy to rank influential researchers in scientific communities.

As future work, we plan to apply 3c-index to investigate the special committees of the Brazilian Computer Society[13] – although collecting information on the awards distributed may be challenging. In addition, we plan to investigate how to apply the 3c-index from ranking researchers that publish in multiple research areas. Other studies include verifying the potential of using 3c-index in: academic recommendation systems [Brandão et al. 2014], learning-to-rank techniques (similar to [Canuto et al. 2013]), and as a feature to be extracted from author profiles (similar to [Weren et al. 2014]).

REFERENCES

Alves, B. L., Benevenuto, F., and Laender, A. H. F. The Role of Research Leaders on the Evolution of Scientific Communities. In *Proceedings of the International World Wide Web Conferences, Workshop*. Rio de Janeiro, Brazil, pp. 649–656, 2013.

[13]SBC: `http://www.sbc.org.br/`

Barabási, A.-L. Scale-free Networks: a decade and beyond. *Science* 325 (5939): 412–413, 2009.

Barabási, A.-L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., and Vicsek, T. Evolution of the Social Network of Scientific Collaborations. *Physica A: statistical mechanics and its applications* 311 (3): 590–614, 2002.

Bollen, J., Van de Sompel, H., Hagberg, A., and Chute, R. A Principal Component Analysis of 39 Scientific Impact Measures. *PLoS ONE* 4 (6): e6022:1–e6022:11, 2009.

Brandão, M. A., Moro, M. M., and Almeida, J. M. Experimental Evaluation of Academic Collaboration Recommendation Using Factorial Design. *Journal of Information and Data Management* 5 (1): 52–63, 2014.

Burt, R. S. Structural Holes and Good Ideas. *American Journal of Sociology* 110 (2): 349–399, 2004.

Canuto, S. D., Belém, F. M., Almeida, J. M., and Gonçalves, M. A. A Comparative Study of Learning-to-rank Techniques for Tag Recommendation. *Journal of Information and Data Management* 4 (3): 453–468, 2013.

Freire, V. P. and Figueiredo, D. R. Ranking in Collaboration Networks Using a Group based Metric. *Journal of the Brazilian Computer Society* 17 (4): 255–266, 2011.

Garfield, E. Journal Impact Factor: a brief review. *Canadian Medical Association Journal* 161 (8): 979–980, 1999.

Gonçalves, G. D., Figueiredo, F., Almeida, J. M., and Gonçalves, M. A. Characterizing Scholar Popularity: a case study in the Computer Science research community. In *Proceedings of the ACM IEEE Joint Conference on Digital Libraries*. London, UK, pp. 57–66, 2014.

Granovetter, M. S. The Strength of Weak Ties. *American Journal of Sociology* 78 (6): 1360–1380, 1973.

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., and Rafols, I. Bibliometrics: the Leiden Manifesto for research metrics. *Nature* 520 (7548): 429–431, 2015.

Hirsch, J. E. An Index to Quantify an Individual's Scientific Research Output. *Proceedings of the National Academy of Sciences of the United States of America* 102 (46): 16569–16572, 2005.

Järvelin, K. and Kekäläinen, J. Cumulated Gain-based Evaluation of IR Techniques. *ACM Transactions on Information Systems* 20 (4): 422–446, 2002.

Lima, H., Silva, T. H., Moro, M. M., Santos, R. L., Meira Jr, W., and Laender, A. H. Aggregating Productivity Indices for Ranking Researchers Across Multiple Areas. In *Proceedings of the ACM IEEE Joint Conference on Digital Libraries*. Indianapolis, USA, pp. 97–106, 2013.

Lima, H., Silva, T. H. P., Moro, M. M., Santos, R. L. T., Meira Jr, W., and Laender, A. H. F. Assessing the Profile of Top Brazilian Computer Science Researchers. *Scientometrics* 103 (3): 879–896, 2015.

Lopes, G. R., Moro, M. M., da Silva, R., Barbosa, E. M., and de Oliveira, J. P. M. Ranking Strategy for Graduate Programs Evaluation. In *Proceedings of the International Conference on Information Technology and Applications*. Sydney, Australia, pp. 59–64, 2011.

Maroun, L. B. and Moro, M. M. Ranking de Pesquisadores em Cenários Ambíguos: uma abordagem probabilística. In *Proceedings of the Brazilian Symposium on Databases*. Curitiba, Brazil, pp. 157–166, 2014.

Newman, M. E. J. Who Is the Best Connected Scientist? A Study of Scientific Coauthorship Networks. In E. Ben-Naim, H. Frauenfelder, and Z. Toroczkai (Eds.), *Complex Networks*. Lecture Notes in Physics, vol. 650. Springer, pp. 337–370, 2004.

Procópio Jr, P. S., Gonçalves, M. A., Laender, A. H. F., Salles, T., and Figueiredo, D. Time-aware Ranking in Sport Social Networks. *Journal of Information and Data Management* 3 (3): 195–210, 2012.

Silva, T. H. P., Moro, M. M., and Silva, A. P. C. Authorship Contribution Dynamics on Publication Venues in Computer Science: an aggregated quality analysis. In *Proceedings of the ACM Symposium on Applied Computing*. Salamanca, Spain, pp. 1142–1147, 2015a.

Silva, T. H. P., Moro, M. M., and Silva, A. P. C. tc-index: a new research productivity index based on evolving communities. In S. Kapidakis, C. Mazurek, and M. Werla (Eds.), *Research and Advanced Technology for Digital Libraries*. Lecture Notes in Computer Science, vol. 9316. Springer, pp. 209–221, 2015b.

Silva, T. H. P., Moro, M. M., Silva, A. P. C., Meira Jr, W., and Laender, A. H. F. Community-based Endogamy as an Influence Indicator. In *Proceedings of the ACM IEEE Joint Conference on Digital Libraries*. London, UK, pp. 67–76, 2014.

Silva, T. H. P., Rocha, L. M. A., Moro, M. M., and Silva, A. P. C. Contribuição de Pesquisa entre Comunidades como um Indicador de Influência. In *Proceedings of the Brazilian Symposium on Databases*. Petrópolis, Brazil, pp. 65–76, 2015.

Sun, X., Kaur, J., Milojević, S., Flammini, A., and Menczer, F. Social Dynamics of Science. *Scientific Reports, Nature* 3 (1069): 1–6, 2013.

Weren, E. R. D., Kauer, A. U., Mizusaki, L., Moreira, V. P., de Oliveira, J. P. M., and Wives, L. K. Examining Multiple Features for Author Profiling. *Journal of Information and Data Management* 5 (3): 266–279, 2014.

Zhang, C.-T. The e-index, Complementing the h-index for Excess Citations. *PLoS ONE* 4 (5): e5429:1–e5429:4, 2009.