UFMG - ICEx
DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO

UFMG
UNIVERSIDADE FEDERAL DE MINAS GERAIS

Reputation in computer science on a per subarea basis

RT.DCC.001/2017

Alberto Hideki Ueda
Berthier Ribeiro de Araújo Neto
Edmundo de Souza e Silva
Marlon Dias
Nivio Ziviani

MAIO
2017

Reputation in Computer Science on a per Subarea Basis

Alberto Hideki Ueda[1], Berthier Ribeiro de Araújo Neto[1]
Edmundo de Souza e Silva[2], Marlon Dias[1], Nivio Ziviani[1]

[1]CS Dept., UFMG, Belo Horizonte, Brazil

[2]COPPE, UFRJ, Rio de Janeiro, Brazil

{ueda,berthier,nivio}@dcc.ufmg.br, edmundo@land.ufrj.br

Abstract. *In this work, we study the reputation of venues and research groups in Computer Science with focus on its subareas. We adopt the 37 subareas defined by Microsoft Academic Research as starting point and focus on the Computer Science departments in the US. More specifically, we study the impact to the reputation of venues and research groups when subareas are taken into account. For that we extend the usability of a metric called P-Score (for Publication Score), proposed in the literature. The reason we adopted the metric P-Score in this work is because it can be computed without using citation information. P-scores rely on a Markov network which can be used to model relations among researchers and among researchers and the venues they publish in. We run several experiments in which we compare the reputation of venues and research groups per subarea. The results suggest that the extended P-scores yield better results when compared with citation counts. In that matter, the analysis of reputation on a per subarea basis yields additional insights into the reputation of venues and research groups that are useful when deciding how to allocate research funds.*

Introduction

Consider the problem of finding relevant sources of information in the Web about a specific topic of interest, among trillions of documents available. Instances of this problem include: a person looking for great places to visit during her vacation, companies searching for the most compatible candidates to the jobs they are offering, or a research institution interested in knowing how is the public perception about itself in terms of reputation.

In here we concentrate our attention on variants of this problem in the academic research domain. In this case, key entities of interest are researchers, research groups, papers, and publication venues. Regarding these entities, questions of interest include, for instance, retrieving the most popular venues (i.e. publication conferences and journals) in the broad area of Computer Science; or finding the authors whose work is the most related to a specific topic of study (a task generally referred as *expert search*); or deciding which national institutions are the most suitable for the allocation of research funds for the next few years.

To answer these questions a commonly adopted criterion is to define metrics that somehow reflect the reputation of the academic entity of interest. Although ill-defined, the reputation of an entity reflects the perception, the interest, the popularity, the brand name of that entity before the general public or before its academic peers. In principle, the entities with high reputation should receive priority treatment in the allocation of new research funds, grants, academic awards, graduate students, and other incentives, than less reputable entities within the same community.

Several metrics for quantifying the reputation of academic entities have been proposed, including citation-based metrics, machine learning techniques, and Markov processes. Some of these metrics (e.g. H-index [13] and Impact Factor [10]) have been used by prestigious institutions, from academic search engines to government funding agencies to assess the reputation of authors, research groups, and publishing venues from the perspective of the broad areas of knowledge (e.g. Computer Science, Physics, Mathematics, Statistics, Chemistry, and so on).

However, general purpose rankings of Computer Science departments, for instance, may not reflect important information about these departments from the perspective of subareas in the field. For example, if a company wants to invest funds on a reliable research group working on the subarea of Human Computer Interaction, general rankings in the broad area of Computer Science will not facilitate the decision.

Other noteworthy scenario is comparing two different researchers, where the first one works on a subarea A and the second one works on a subarea B. If we assume that it is inherently harder to publish articles in A than in B, it seems natural that the metrics used to rank these researchers should be distinct, or, at least, take those differences between subareas into account. Otherwise, the comparison between them would be unfair.

In this work we suggest methods to get insights about academic entities on a per subarea basis, with focus on the area of Computer Science. In particular, we aim to answer the following research questions:

Q1. How do the reputation of venues and research groups vary per subarea?

Q2. How does the reputation of venues and research groups per subarea compare to their reputation when subareas are not taken into account?

The remainder of this paper is structured as follows. In Section 2 we discuss related work on reputation models and some instantiations of these models in academic search tasks. In Section 3, we present the theoretical concepts supporting our approach, by describing the key ideas of the reputation model we used in this work derived from the metric called P-score from the literature. In that section we also formalize the strategies we adopted to study the academic data on a per subarea basis. The academic dataset and the experimental methodology we adopt are described in detail in Section 4. Our results are discussed in Section 5. In Section 6 we discuss the key contributions of this work and directions for further research.

Related Work

Citation-based metrics have been widely applied to rank computer and information science journals [16, 24]. Also, different approaches using citation data have been proposed to measure the quality of publication venues in the Databases subarea [28] and to rank documents retrieved from a digital library [18].

Garfield's Impact Factor [10] 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 [13]. 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 [9, 7, 35, 31] and studied their impact, advantages, and disadvantages [33, 20].

The idea of reputation, without the direct use of citation data, was discussed by Nelakuditi et al. [23]. 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. [17], 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 [4, 5, 11, 34]. Particularly, Gonçalves et al. [12] 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 scholar's publication venues. In addition, the prediction of scientific success of a researcher is also valuable for several goals [25]. As a result, previous works attempted to predict if a researcher will become a principal investigator [6], her future H-index [8, 26] and the potential number of citations to her publications [22, 3].

Although citation-based metrics are useful, they are not enough to do a complete evaluation of research. In particular, Piwowar [27] 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 [30]. 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* [29].

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[1], Microsoft Academic Search[2], AMiner[3], and CiteSeerX[4]. 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 [15, 19] and in Computer Science [14, 1, 21].

Wainer et al. [32] 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.

Reputation Framework

In this section we provide an overview of the metric P-score [29], a description of modifications in the model to take subareas into account and a discussion about the encroachment problem we discovered while assessing the reputation of research groups in CS, on a per subarea basis.

P-score Overview

The hypotheses supporting the P-score metric are:

1. A research group conveys reputation to a publication venue proportionally to its own reputation.
2. A publication venue conveys reputation to a research group proportionally to its own reputation.

Therefore, the reputation of a research group is strongly influenced by the reputation of its members, which is largely dependent on their publication records. More specifically, the instantiation of P-score in this work is as follows: given a pre-selected set of reference venues (or *seeds*) in a given subarea, the metric finds the n researchers with the largest volume of publications in these reference venues and uses them to assemble a Markov network model which also includes all venues in which they published. The steady state probabilities are interpreted as venue weights which distinguish high

[1] `https://scholar.google.com.br`

[2] `http://academic.microsoft.com`

[3] `http://aminer.org`

[4] `http://citeseerx.ist.psu.edu`

reputation venues from the others. Thus, these weights can be used to rank venues (by considering these weights as venue scores) and also to rank research groups or authors.

More precisely, the steady state probability of an entity (author, research group, or publication venue) can be interpreted as its relative reputation, as transferred from other entities in a reputation graph. We can directly use the value of this probability to rank reputation sources (in the set S) or reputation targets (in the set T). Additionally, this probability can be further propagated to entities we want to compare (in the *collateral set* C). This propagation depends on a matrix $P^{\langle TC \rangle}$ of size $|T| \times |C|$ representing the transitions from reputation targets to collateral entities. More generally, the P-score of an entity e is defined as:

$$\textit{P-score(e)} = \begin{cases} \sum_{t \in T} {P_{te}}^{\langle TC \rangle} \pi_t & \text{if } e \in C \\ \pi_e & \text{otherwise} \end{cases} \tag{1}$$

where ${P_{te}}^{\langle TC \rangle}$ is the transition weight from a target entity $t \in T$ to an entity e, π_e is the reputation of entity e, π_t is the reputation of target entity t. The P-score of all candidate entities (targets or collaterals) can then be used to produce an overall reputation-oriented ranking of these entities. More details on P-score can be obtained from [30].

Venue Ranking

As our primary goal is to perform analysis of academic entities on a per subarea basis, it is crucial to investigate how we could identify suitable publication venues to characterize subareas.

As originally proposed, P-score venue weights do not allow distinguishing venues in a given subarea, even if we choose as seeds venues that are central to that subarea (i.e. venues which are assuredly focused on that subarea). That is expected because P-score is a metric strongly correlated to the volume of publications. In other words, venues with high popularity that are related to the seeds – i.e. have papers written by the authors used as seeds (or references) in the network – are put in the top positions by the raw P-score.

To avoid this problem, we normalize the venue's P-score by the number of publications in the venue's history. The key idea is to obtain an average of the overall reputation of the venue on a per paper basis. This approach penalizes venues with a large volume of publications but with low P-scores (low reputation according to the seeds) and boosts smaller publication venues with good reputation in a given subarea. Thus, the *normalized P-score* for venue v is defined as:

$$\textit{norm-P-score}(v) = \frac{\textit{P-score}(v)}{\textit{number\_of\_publications}(v)} \tag{2}$$

Equation (2) is the starting point for obtaining the ranking of groups on a per subarea basis as presented in the following sections.

Group Ranking

Evaluating a group, such as the researchers in Databases of a given CS department, requires weighting the contributions of its members who are responsible for the reputation of the group. We say that the group's reputation is the sum of the reputation of the venues the members of the group published in, taken on a per author basis. Thus:

$$\textit{P-score(g)} = \sum_{p \in \delta(g)} \frac{\textit{P-score(venue(p))}}{\textit{number\_of\_authors}(p)} \tag{3}$$

where $\delta(g)$ are the publications of group g and *venue(p)* is the venue at which paper p was published.

According to Equation (3), the score of a group is based on the papers it published. Each paper has a P-score, which is given by the venue where the paper was published. Usually, researchers combine efforts so they can produce a paper. In consequence, we normalize the paper's score by the number of authors.

The members of a research group are professors, postdoctoral fellows, doctoral students, among others. Normally, professors are responsible for the research groups at which postdoctoral fellows and doctoral students work. At some point, the group publishes a paper and the responsible professor is typically involved. This allows us to use professors as the anchors for transferring reputation to the research group. Notice that the indication of the author's affiliation is key to estimate reputation for groups. Hence, Equation (3) may be rewritten as:

$$\textit{P-score(g)} = \sum_{a \in \varphi(g)} \sum_{p \in \delta(a)} \frac{\textit{P-score(venue(p))}}{\textit{number\_of\_ authors}(p)} \tag{4}$$

where $\varphi(g)$ are the researchers associated with group g and $\delta(a)$ are the publications of author a.

The Encroachment Problem

P-score ranks groups according to Equation (4). Still, this approach may not be appropriate when considering subareas because venues cover multiple subareas and thus their reputation need to be split among its component subareas. One example of such venue is the ACM Conference on Information and Knowledge Management (CIKM). It is a high reputation venue that covers three subareas: Databases (DB), Information Retrieval (IR), and Knowledge Management (KM). If we are interested in the subarea of IR in particular, we need to find a way to discount or weigh down the contributions of CIKM papers that are not on IR. If we do not, we might end with large P-score contributions to a given subarea, such as IR, from papers that are really on another subarea, such as DB.

This is what we call the *encroachment problem*. We illustrate with an example. Elisa Bertino is a well known and respected researcher who has published over 800 papers. Her interests cover many areas with focus on the fields of Information Security and DB systems. She has papers on CIKM and other venues that also accept papers on IR. Because of that she appears on the list of authors that publish frequently on IR related venues. And, because of the large number of papers she publishes her P-score on IR is high, which leads to a high rank of her group at Purdue University on the subarea of IR.

Given CIKM does not distinguish in its proceedings which papers are on IR, on DB or on KM, determining whether a given CIKM paper is on IR, for instance, would require examining its text contents. However, P-score is a metric that does not rely on paper contents – one of its inherent advantages given it is much simpler to compute than citation-based metrics. Thus, imposing the need to have access to the contents of papers is a constraint we purposely want to avoid. Therefore, we look for a different solution.

The solution we propose for the encroachment problem is to examine the main subareas of interest of each researcher and produce weights for the pairs $[researcher, subarea]$. We do so by examining the publications of the researchers on venues that are specific to a single subarea such as SIGIR and SIGMOD, for instance. Our rational is that a researcher that publishes eight SIGIR papers and two SIGMOD papers is focused on IR 80% of the time and on DB 20% of the time. In other words, this researcher interest factor on IR is 0.8 and on DB is 0.2. We then use this *subarea interest factor* (hereinafter presented as γ_f) to weigh the papers of this author in venues that cover multiple subareas, such as CIKM. That is, instead of solving a classification problem (determine the subarea of each CIKM paper) which would require access to paper contents, we propose a ranking solution that ranks CIKM papers on each of its subareas based on their authors interest factors. Our ranking solution simplifies the implementation and leads to good results, as discussed in Section 5.

Let S_f be a set of venues closely associated with subarea f and $\delta(a)$ be the set of publications of author a, as before. Then:

$$\gamma f(a) = \frac{1}{\delta(a)} \sum_{p \in \delta(a)} \begin{cases} 1 & \text{if } p \in S_f \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where $\gamma_f(a)$ is the weight of author a in subarea f,i.e., a measure of how much the author belongs to that subarea.

Factor γ quantifies the relation of authors to a given subarea, but does not take into account the history of publications by a given author. If an author changes their field of study, we should factor in that the author's relation to the subarea of interest has weakened. We do so by introducing a publication age penalty, as follows:

$$\gamma_f(a) = \frac{1}{\delta(a)} \sum_{p \in \delta(a)} \begin{cases} \frac{1}{\log_2 (y(0) - y(p) + 2)} & \text{if } p \in S_f \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where $y(p)$ is the year in which the paper p was published and $y(0)$ is the current year, or the year of the most recent paper in the collection.

Using the subarea interest factor we can rewrite Equation (4) and present the *weighted P-score* of a group *g* in subarea *f* as:

$$\textit{weighted-P-score(g)}_f = \sum_{a \in \varphi(g)} \gamma_f(a) \times \sum_{p \in \delta(a)} \frac{\textit{P-score(venue(p))}}{\textit{number\_of\_authors}(p)} \tag{7}$$

Experimental Setup

In this section, we describe the academic dataset we used on our experiments. Subsequently, we present the definition of subareas in CS we adopted in this paper and also the process of creating our venue ground-truth with the help of experts on each subarea.

Academic Search Dataset

We compiled a collection of academic publications records extracted from DBLP,[5] an online reference for bibliographic information on major CS publications. DBLP data has been used in related studies on CS research communities [32, 2, 14]. The database is publicly available in XML format and contains more than three million publication records from more than 1.5 million authors over the last 50 years, albeit the data before 1970 is rather irregular. Each publication record includes a title, list of authors, year of publication, and publication venue. Publication records do not include the contents of the papers neither information related to citation.

Our collection is actually an extension of the DBLP repository. While it contains all publication venues and authors from DBLP, we have enriched it adding information regarding research groups. To do so, we manually collected information about the top 126 CS graduate programs evaluated in the 2011 assessment conducted by the US National Research Council (NRC).[6] In particular, for each of these groups we retrieved the list of group members, which were then manually reconciled against the repository.

Despite our efforts, there were still imprecisions related to the affiliation of the authors. To address them, we combined our dataset with the one provided by the *csranking* project,[7] which ranks CS departments based purely on their publications. They do so by collaboratively collecting information on authors, such as their homepage and affiliation. Therefore, we used that information to enhance our repository. Salient statistics on our dataset are shown in Table 1.

Table 1. Salient statistics of the dataset used in our evaluation.

Type of Entity	Total Number of Instances
Papers	2,931,849
Authors	1,595,771
Venues	5,765
Departments	126

Computer Science Subareas

The definition of subareas in CS vary on time and also on the institution responsible for the classification. For instance, ACM[8] (through *Special Interest Groups*) and IEEE[9] (through *Technical Committees*) divide CS in subareas in rather distinct ways. Further, some of these divisions reflect historical decisions that may be less relevant nowadays. For this reason, previous works have attempted to automatically identify such subareas [32] or use other source of information [14].

In this work, we adopt the classification of CS subareas provided by Microsoft Academic Research.[10] This classification divides CS in 37 subareas, including relatively

[5] http://dblp.uni-trier.de

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

[7] http://csrankings.org/

[8] http://acm.org/sigs

[9] http://computer.org/web/tandc/technical-committees

[10] http://academic.research.microsoft.com

new subareas such as Knowledge Management and Natural Language Processing. The full list of subareas are presented in Table 2.

Table 2. The Microsoft 37 Subareas of Computer Science

CS Subareas	
Algorithm	Internet privacy
Artificial intelligence	Knowledge management
Bioinformatics	Machine learning
Cognitive science	Management science
Computational biology	Mathematical optimization
Computational science	Multimedia
Computer architecture	Natural language processing
Computer graphics	Operating system
Computer hardware	Operations research
Computer network	Parallel computing
Computer security	Pattern recognition
Computer vision	Programming language
Data mining	Real-time computing
Data science	Simulation
Database	Speech recognition
Distributed computing	Telecommunications
Embedded system	Theoretical computer science
Human-computer interaction	World Wide Web
Information retrieval	

While the subareas in Table 2 are not perfect, nor exhaustive, they are detailed enough to allow us to gain insight on how the reputation of authors, venues and research groups vary from one subarea to another.

Venues Ground-Truth

To evaluate the effectiveness of normalized P-scores from Equation (2) on the task of finding venues in a subarea, we considered as ground-truth the opinion of experts. Specifically, we asked reputable CS researchers and their graduate students, working on subareas of IR, DB and Data Mining (DM) to asses the relevance to their subarea of venues included in a pre-selected list. This list consists of the venues at the top 50 positions in the P-score ranking, when we use as seeds two publication venues only: a journal and a conference closely associated with that subarea. For examples of seeds, see Table 3.

Table 3. Seeds of publication venues for P-score used in this work.

	Subareas		
Type of venue	Database	Data Mining	Information Retrieval
Conference	SIGMOD	KDD	SIGIR
Journal	TODS	SIGKDD	TOIS

We thus focused on the CS subareas of DB, DM and IR. For each subarea, three experts have classified each of the 50 venues of the pre-selected list into one, two or three subareas chosen among the 37 subareas listed in Table 2. To reconciliate the multiple classifications, we used a majority criterion: if a publication venue v was associated with a subarea s at least twice, s was considered one of the subareas of v. Hereafter, we will refer to the full lists of publication venues and their subareas as our *venues ground-truth*.

Experimental Results

In this section, we discuss the results of ranking publication venues and research groups in CS on the three subareas we selected: DB, DM and IR. In particular for groups, our results are restricted to the 126 US graduate programs considered by NRC in 2011.

Subarea Venues

Figure 1 presents precision-recall curves of normalized P-scores from Equation (2) for ranking venues on a per subarea basis, as discussed in Section 3.2. For comparison, we also show results produced using H-index and standard P-scores from Equation (**??**). H-indices were obtained from Google Scholar.

As it is clear from Figure 1, normalized P-scores allow identifying the correct venues consistently better than H-indices and standard P-scores. Furthermore, for all the three subareas, the normalized P-scores yield maximum precision (100%) for the initial 30% of recall. This means that the first 15 venues in the normalized P-score ranking are strongly related to IR, according to the assessments of specialists.

To further illustrate, Table 4 shows the top 20 publication venues for the subarea of IR, produced by P-scores and normalized P-scores when we consider SIGIR and TOIS as seed venues. On the one hand, the standard P-score metric places venues such as the International World Wide Web Conferences (WWW) and the International Conference on Multimedia (MM), among the top 10 positions. These two conferences cover topics of the subarea of IR, but indeed have a larger scope than IR only.

On the other hand, such conferences do not appear in the normalized P-score ranking, even among the top 20 positions on the ranking. Besides, in the normalized P-score ranking, venues mainly focused in IR venues such as the International Conference on the Theory of Information Retrieval (ICTIR) and Transactions on Information Systems (TOIS) appear among the top 10 publication venues. In particular, ICTIR appears at the top. This power of discrimination of the normalized P-score is important to allow selecting venues that better represent the subarea of IR.

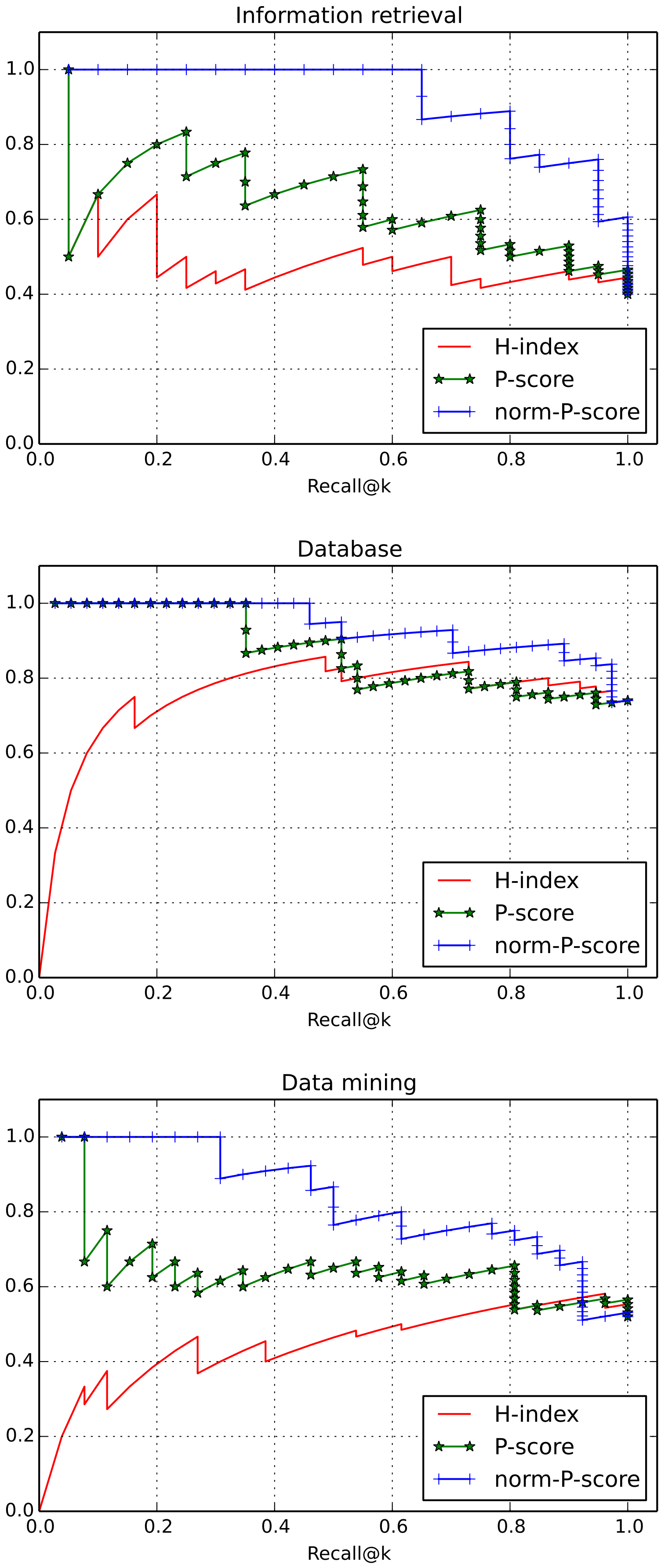
It is noteworthy to mention that the normalized P-score ranking of venues should not be interpreted as an impact or productivity ranking. We only use this output to define the most representative venues of a subarea, in a semi-automatic fashion.

Subarea Groups

In a broad area as CS, one should consider each subarea separately rather than evaluating the whole area, since they are not the same. Therefore, we propose to rank US universities considering subareas. As before, we consider three CS subareas: DB, DM and IR.

Table 5 presents a ranking of groups using standard P-scores for IR. The values presented are normalized to 1 as follows:

$$\overline{P}(g)_f = \frac{P(g)_f}{max(P(g)_f)} \tag{8}$$



Information retrieval

Database

Data mining

Figure 1. Precision-Recall curves of H-index, P-score and normalized P-score for the subareas of Information Retrieval, Database and Data Mining.

Table 4. Top 20 venues in Information Retrieval, using P-score and normalized P-score. The suffixes (c) and (j) are used to differentiate conferences and journals with the same name.

#	P-score	Norm-P-score
1	SIGIR (c)	ICTIR
2	CIKM	SIGIR (c)
3	TREC	ADCS
4	ECIR	IR
5	CLEF	ECIR
6	WWW	TREC
7	JASIS	SIGIR (j)
8	IPM	IIIX
9	SIGIR (j)	TOIS
10	MM	WSDM
11	JCDL	INEX
12	TOIS	SPIRE
13	IR	AIRS
14	WSDM	CIKM
15	NTCIR	TWEB
16	KDD	RIAO
17	TKDE	CLEF
18	ACL	NTCIR
19	ICDM	LA-WEB
20	SPIRE	JCDL

where $\overline{P}(g)_f$ is the normalized score of group g in f, $P(g)_f$ is the score and $max(P(g)_f)$ is the highest score. We apply Equation 8 to either standard and weighted P-scores.

The University of Massachusetts Amherst has the premier research group IR in the US and thus, the fact that it was not in first place in the rank was surprising to us. This led to an in-depth analysis of the ranking and the consequent understanding of the encroachment problem, as discussed in Section 3.4.

Tables 6 presents the top 10 universities on IR using our proposed approach of the weighted P-score, according to Equation (7), instead. We observe that the University of Southern California, the Georgia Institute of Technology, Stanford University and the University of California at Berkley are no longer among the top 10 IR groups. This seems appropriate given these universities are not active on research in IR.

To better understand the results in Table 6, we produced a list of the top 20 researchers on IR. Table 7 shows their affiliation. As we observe, our top 10 IR groups are those whose researchers are also among the top 20 authors on IR in the US. In particular, the top 3 groups have each one 2 or more researchers among the top 20.

We also manually examined our ranking of authors to observe that the top authors shown in Table 7 had more publications in venues strongly related to the subarea. Hence, the ranking of groups can be justified by the ranking of authors.

We repeated this process to the subareas of DB and DM. Tables 8 and 9 present

Table 5. Ranking of the top 10 US Universities on Information Retrieval, using standard P-scores.

#	University	P-score
1	Carnegie Mellon University	1
2	University of Massachusetts Amherst	0.8082
3	University of Illinois at Urbana-Champaign	0.6735
4	University of Southern California	0.4541
5	Georgia Institute of Technology	0.4341
6	Stanford University	0.3493
7	University of Illinois at Chicago	0.3409
8	Cornell University	0.3344
9	University of California-Berkeley	0.3337
10	Purdue University	0.3120

Table 6. Ranking of the top 10 US Universities on Information Retrieval, using the weighted P-score (Equation 7).

#	University	weighted-P-score
1	University of Massachusetts Amherst	1
2	University of Illinois at Urbana-Champaign	0.4830
3	Carnegie Mellon University	0.4625
4	University of Delaware	0.2452
5	Purdue University	0.2276
6	Northeastern University	0.1633
7	Lehigh University	0.0964
8	Cornell University	0.0552
9	University of Iowa	0.0494
10	University of Illinois at Chicago	0.0477

the top 10 universities on DB and DM, when we use the weighted P-scores produced by Equation (7).

Our analysis on a per subarea basis uncovers venues and research groups that are not always thought of as high excellence. But, once one considers their track records on a specific subarea, it is clear that they are quite productive. That is, analyzing research groups on a per subarea basis yields insights that are hidden when we apply global metrics of productivity to a broad area as CS.

Conclusions

In this paper, our first research question was: *"How do the reputation of venues and research groups vary per subarea?"* We showed that the identification of the most reputable academic entities in Computer Science depends on the subarea considered to the task. Specifically, both the most suitable venues (to use as seeds for P-score) as the ranking of US departments vary on a per subarea basis, considering the subareas of Information Retrieval, Databases and Data Mining. Besides, we described how to modify the P-score metric to find the core venues of a subarea in a semi-automatic fashion and, subsequently,

Table 7. Ranking of the top 20 US authors' universities on Information Retrieval, using Equation (7).

#	Authors' universities
1	University of Massachusetts Amherst #1
2	University of Massachusetts Amherst #2
3	Carnegie Mellon University #1
4	University of Illinois at Urbana-Champaign #1
5	Purdue University
6	University of Delaware
7	Northeastern University
8	University of Illinois at Urbana-Champaign #2
9	Lehigh University
10	Carnegie Mellon University #2
11	University of Iowa
12	University of Illinois at Chicago
13	Georgia Institute of Technology
14	University of Virginia
15	Carnegie Mellon University #3
16	Texas A&M University
17	Cornell University
18	University of Michigan
19	University of Massachusetts Amherst #3
20	New York University

how to rank research groups using this information, obtaining the top research groups of a given subarea.

Our second research question was: *"How does the reputation of venues and research groups per subarea compares to their reputation when subareas are not taken into account?"* We presented experiments suggesting that metrics that only capture broad features of a subarea such as the volume of publications, citations or the general reputation (standard P-score included) are not sufficient to produce reasonable rankings in a per subarea basis. In particular, we demonstrated that solving the venues encroachment problem allows us to improve the ranking of research groups in a given subarea. We showed a simple but effective strategy to increase or decrease the contribution of an author to the overall reputation of her research group in a given subarea, based on the author's relation to that subarea.

For future work, we intend to characterize the distribution of the most reputable research departments in CS worldwide, on a per subarea basis. We also want to do a temporal analysis of the evolution of the CS subareas communities over the last decades. Another further study is to validate the reputation model in other broad areas than CS, such as Economics, whose differences in publication patterns on a per subarea basis seem to be greater than in Computer Science.

Table 8. Ranking of the top 10 US Universities on Databases, using the weighted P-score (Equation (7)).

#	University	weighted-P-score
1	University of Wisconsin-Madison	1
2	Stanford University	0.6570
3	University of Illinois at Urbana-Champaign	0.5687
4	Massachusetts Institute of Technology	0.4975
5	Duke University	0.4616
6	University of Massachusetts Amherst	0.4243
7	University of Michigan	0.4195
8	University of California-Irvine	0.4120
9	University of Maryland-College Park	0.4101
10	University of California-Santa Cruz	0.3982

Table 9. Ranking of the top 10 US Universities Data Mining, using the weighted P-score (Equation (7)).

#	University	weighted-P-score
1	University of Illinois at Chicago	1
2	Carnegie Mellon University	0.6857
3	University of Illinois at Urbana-Champaign	0.6344
4	University of Minnesota	0.5350
5	Arizona State University	0.4276
6	University of California-Riverside	0.4212
7	Georgia Institute of Technology	0.3955
8	University of Michigan	0.3275
9	Rensselaer Polytechnic Institute	0.2761
10	University of California-Davis	0.2593

Acknowledgements

We thank the partial support given by the Brazilian National Institute of Science and Technology for the Web (grant MCT-CNPq 573871/2008-6), Project Models, Algorithms and Systems for the Web (grant FAPEMIG/PRONEX/MASWeb APQ-01400-14), and authors' individual grants and scholarships from CAPES, CNPq and Google.

References

[1] Benevenuto, F., Laender, A., and Alves, B. How connected are the ACM SIG communities? *SIGMOD Record 44*, 4 (2015), 57–63.

[2] Biryukov, M., and Dong, C. Analysis of computer science communities based on dblp. In *Proceedings of European Conference on Research and Advanced Technology for Digital Libraries* (2010), pp. 228–235.

[3] Castillo, C., Donato, D., and Gionis, A. Estimating number of citations using author reputation. In *Proceedings of String processing and information retrieval* (2007), pp. 107–117.

[4] CORMODE, G., MUTHUKRISHNAN, S., AND YAN, J. People like us: mining scholarly data for comparable researchers. In *Proceedings of World Wide Web* (2014), pp. 1227–1232.

[5] DENG, H., HAN, J., LYU, M. R., AND KING, I. Modeling and exploiting heterogeneous bibliographic networks for expertise ranking. In *Proceedings of Joint Conference on Digital Libraries* (2012), pp. 71–80.

[6] DIJK, D., MANOR, O., AND CAREY, L. Publication metrics and success on the academic job market. *Current Biology 24*, 11 (2014), R516–R517.

[7] DING, Y., AND CRONIN, B. Popular and/or prestigious? measures of scholarly esteem. *Information Processing & Management 47*, 1 (2011), 80–96.

[8] DONG, Y., JOHNSON, R., AND CHAWLA, N. Will this paper increase your h-index? scientific impact prediction. In *Proceedings of the International Conference on Web Search and Data Mining* (2015), pp. 149–158.

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

[10] GARFIELD, E. Citation indexes for science. *Science 122*, 3159 (1955), 108–111.

[11] GOLLAPALLI, S., MITRA, P., AND GILES, C. Ranking authors in digital libraries. In *Proceedings of Joint Conference on Digital Libraries* (2011), pp. 251–254.

[12] GONÇALVES, G., FIGUEIREDO, F., ALMEIDA, J., AND GONÇALVES, M. Characterizing scholar popularity: A case study in the computer science research community. In *Proceedings of Joint Conference on Digital Libraries* (2014), pp. 57–66.

[13] HIRSCH, J. An index to quantify an individual's scientific research output. *Proceedings of National Academy of Sciences* (2005), 16569–16572.

[14] HOONLOR, A., SZYMANSKI, B., AND ZAKI, M. Trends in computer science research. *Communications of the ACM 56*, 10 (Oct. 2013), 74–83.

[15] KAPELLER, J. Citation metrics: serious drawbacks, perverse incentives, and strategic options for heterodox economics. *American Journal of Economics and Sociology 69*, 5 (2010), 1376–1408.

[16] KATERATTANAKUL, P., HAN, B., AND HONG, S. Objective quality rankings of computing journals. *Communications of the ACM 45* (2003).

[17] LAENDER, A., LUCENA, C., MALDONADO, J., DE SOUZA E SILVA, E., AND ZIVIANI, N. Assessing the research and education quality of the top brazilian computer science graduate programs. *SIGCSE Bulletin 40*, 2 (2008), 135–145.

[18] LARSEN, B., AND INGWERSEN, P. Using citations for ranking in digital libraries. In *Proceedings of Joint Conference on Digital Libraries* (2006), pp. 370–370.

[19] LEE, F., GRIJALVA, T., AND NOWELL, C. Ranking economics departments in a contested discipline: a bibliometric approach to quality equality between theoretically distinct subdisciplines. *American Journal of Economics and Sociology 69*, 5 (2010), 1345–1375.

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

[21] LIMA, H., SILVA, T., MORO, M., SANTOS, R., MEIRA, W., AND LAENDER, A. Aggregating productivity indices for ranking researchers across multiple areas. In *Proceedings of Joint Conference on Digital Libraries* (2013), pp. 97–106.

[22] MAZLOUMIAN, A. Predicting scholars' scientific impact. *PLOS ONE 7*, 11 (11 2012), 1–5.

[23] NELAKUDITI, S., GRAY, C., AND CHOUDHURY, R. Snap judgement of publication quality: how to convince a dean that you are a good researcher. *Mobile Computing and Communications Review 15*, 2 (2011), 20–23.

[24] NERUR, S., SIKORA, R., MANGALARAJ, G., AND BALIJEPALLY, V. Assessing the relative influence of journals in a citation network. *Communications of the ACM 48*, 11 (2005), 71–74.

[25] NEZHADBIGLARI, M., GONÇALVES, M., AND ALMEIDA, J. Early prediction of scholar popularity. In *Proceedings of Joint Conference on Digital Libraries* (2016), pp. 181–190.

[26] PENNER, O., PAN, R., PETERSEN, A., KASKI, K., AND FORTUNATO, S. On the predictability of future impact in science. *Scientific Reports 3* (2013), 3052.

[27] PIWOWAR, H. Altmetrics: value all research products. *Nature 493*, 7431 (2013), 159–159.

[28] RAHM, E., AND THOR, A. Citation analysis of database publications. *Sigmod Record 34*, 4 (2005), 48–53.

[29] RIBAS, S., RIBEIRO-NETO, B., DE SOUZA E SILVA, E., UEDA, A., AND ZIVIANI, N. Using reference groups to assess academic productivity in computer science. In *Proceedings of World Wide Web* (2015), pp. 603–608.

[30] RIBAS, S., RIBEIRO-NETO, B., SANTOS, R., DE SOUZA E SILVA, E., UEDA, A., AND ZIVIANI, N. Random walks on the reputation graph. In *Proceedings of International Conference on The Theory of Information Retrieval* (2015), ICTIR, pp. 181–190.

[31] SUN, Y., AND GILES, C. *Popularity weighted ranking for academic digital libraries*. Springer, 2007.

[32] WAINER, J., ECKMANN, M., GOLDENSTEIN, S., AND ROCHA, A. How productivity and impact differ across computer science subareas. *Communications of the ACM 56*, 8 (2013), 67–73.

[33] W.MARTINS, GONÇALVES, M., LAENDER, A., AND PAPPA, G. Learning to assess the quality of scientific conferences: a case study in computer science. In *Proceedings of Joint Conference on Digital Libraries* (2009), pp. 193–202.

[34] WU, H., PEI, Y., AND YU, J. Detecting academic experts by topic-sensitive link analysis. *Frontiers of Computer Science in China 3*, 4 (2009), 445–456.

[35] YAN, E., DING, Y., AND SUGIMOTO, C. P-rank: An indicator measuring prestige in heterogeneous scholarly networks. *Journal of the Association for Information Science and Technology 62*, 3 (2011), 467–477.