REPUTATION IN COMPUTER SCIENCE ON A PER SUBAREA BASIS

ALBERTO HIDEKI UEDA

REPUTATION IN COMPUTER SCIENCE ON A PER SUBAREA BASIS

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Berthier Ribeiro de Araújo Neto
Coorientador: Nivio Ziviani

Belo Horizonte
Junho de 2017

ALBERTO HIDEKI UEDA

REPUTATION IN COMPUTER SCIENCE ON A PER SUBAREA BASIS

Dissertation presented to the Graduate Program in Computer Science of the Universidade Federal de Minas Gerais in partial fulfillment of the requirements for the degree of Master in Computer Science.

Advisor: Berthier Ribeiro de Araújo Neto
Co-Advisor: Nivio Ziviani

Belo Horizonte
June 2017

Ueda, Alberto Hideki

U22r Reputation in Computer Science on a per Subarea Basis / Alberto Hideki Ueda. — Belo Horizonte, 2017
xxiv, 59 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais — Departamento de Ciência da Computação.

Orientador: Berthier Ribeiro de Araújo Neto
Coorientador: Nivio Ziviani

1. Computação — Teses. 2. Bibliometria. 3. Ferramentas de busca. 4. Indicadores de ciência. 5. P-score. 6. Classificação da ciência da computação. I. Orientador. II.Coorientador. III. Título.

CDU 519.6*73(043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Reputation in computer science on a per subarea basis

ALBERTO HIDEKI UEDA

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. BERTHIER RIBEIRO DE ARAÚJO NETO - Orientador
Departamento de Ciência da Computação - UFMG

PROF. NIVIO ZIVIANI - Coorientador
Departamento de Ciência da Computação - UFMG

PROF. ALBERTO HENRIQUE FRADE LAENDER
Departamento de Ciência da Computação - UFMG

PROF. ALTIGRAN SOARES DA SILVA
Departamento de Ciência da Computação - UFAM

PROF. EDMUNDO ALBUQUERQUE SOUZA E SILVA
COPPE - UFRJ

Belo Horizonte, 14 de julho de 2017.

To Camila, for staying with me in the important and also not important moments of my life over the past decade.

Acknowledgments

I would like to express my gratitude to several people for their support over the course of my master's degree. First of all, I am specially grateful to Berthier for, in a few words, giving me all support in my life-changing decision of come back to the academia. I always felt he believed in me but I never could understand why. However, his trustfulness and guidance were vital to bring me here. I am also thankful to Nivio, for giving me a key to LATIN at my first day at UFMG. His pieces of advice, self-motivation, and full commitment to work will always remain with me. I would also like to thank Sabir Ribas very warmly, for being a reference of a good graduate student for me, a focused person, and, most of all, a dear friend.

A great deal of gratitude is due to all professors who also guided or deeply inspired me somehow: Edmundo, Rodrygo, Wagner Meira, Newton, Jussara, Omar, Vinicius dos Santos, and José Coelho. I am also extremely grateful to all my peers at LATIN, who taught me what a good research lab is. In particular, Bruno Laporais, for motivating me by simply staying at the lab, Marlon Dias, for our great partnership and his inspirational quest for excellence, Jordan, Felipe, and Rafael.

To my wife Camila, my Wonder Woman, my role model, and my greatest friend, for making this dream possible. Her incommensurable support at every single moment of this journey was more than I could ever have asked for. No words of mine can describe how I am grateful to her.

I must also thank my friends Victor Melo, Rensso Mora, Carlos Caetano, Evelin Amorim, and Sonia Borges, for all their support. And heartfelt thanks to Chris Cornell, whose songs and great voice I listened so much while writing this dissertation.

The next and last paragraph is only in Portuguese because there is no sense of writing in another language.

Para meus pais, Akira e Yumiko, que deram a mim e à minha irmã uma formação muito além das expectativas de todos, apesar de tantas dificuldades. Obrigado também pelos valores transmitidos que, como tudo que vale realmente a pena, não podemos comprar ou vender.

"And again when it shall be thy wish to end this play at night,
I shall melt and vanish away in the dark,
or it may be in a smile of the white morning,
in a coolness of purity transparent."
(Rabindranath Tagore)

Resumo

Nesta dissertação, analisamos a reputação de veículos de publicação e programas de pós-graduação em Ciência da Computação (CC) com foco em suas sub-áreas. Para realizar esta tarefa, consideramos as 37 sub-áreas em CC definidas pela Microsoft Academic Research e estendemos uma métrica de reputação baseada em redes de Markov, denominada P-score (*Publication Score*). Mais especificamente, examinamos o impacto obtido na reputação de conferências, periódicos e programas de pós-graduação no Brasil e nos Estados Unidos (EUA) em CC, ao considerarmos suas sub-áreas. Nossos experimentos sugerem que a metodologia proposta produz resultados melhores que métricas basedas em citações. Também apresentamos um panorama das direções de pesquisa atuais do Brasil e dos EUA, que seja, em quais sub-áreas estes países possuem mais trabalhos de destaque no momento. Esta análise de reputação sob a perspectiva de sub-áreas fornece informações adicionais para administradores de universidades, diretores de agências de fomento a pesquisa e representantes do governo que precisam decidir como alocar recursos de pesquisa limitados. Por exemplo, em CC, sabemos que o volume de publicações científicas nos EUA é significantemente superior ao volume de publicações brasileiras. Porém, este trabalho mostra que as sub-áreas em CC em que cada país possui maior impacto científico são basicamente disjuntas.

Palavras-chave: P-score, Indicadores de Ciência, Classificação da Ciência da Computação.

Abstract

In this dissertation, we study the reputation of publication venues and graduate programs in Computer Science (CS) with focus on its subareas. For that we adopt the 37 CS subareas defined by Microsoft Academic Research and extend the usability of a reputation metric based on Markov networks, called P-score (for Publication Score). More specifically, we study the impact to the reputation of CS conferences, journals, and graduate programs in Brazil and US when subareas are taken into account. Our experiments suggest that the extended P-scores yield better results when compared with citation counts. We also present an overview of current research directions of Brazil and US, i.e. on which subareas they have the most prominent work nowadays. This analysis of reputation on a per subarea basis provides additional insights for university officials, funding agencies directors, and government officials who need to decide how to allocate limited research funds. For instance, it is known that the volume of US scientific publications in CS is significantly larger than to the volume of Brazilian CS research. However, this work shows that the CS subareas in which each country has major scientific impact are basically disjoint.

Keywords: P-score, Scientometrics, Classification of Computer Science.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Funding agencies, university officials, and department chairs constantly face the challenge of understanding how well the researchers they employ or sponsor are doing in terms of academic productivity and scientific impact. The assessment of their research quality usually includes the use of quantitative measures (e.g., citation counts, H-index) or qualitative surveys within their scientific community. For instance, in the United States (US), the National Research Council (NRC) applied in 2010 an extensive survey to gather information about more than one hundred US graduate programs.

Overall, such assessments aim to capture the *reputation* of each scholar or graduate program. That is, the perception of the researcher (or group of researchers) before the general public or before her (or its) academic peers. In principle, researchers with high reputation should receive priority treatment in the allocation of grants, research funds, scientific awards, and graduate students than less reputable ones within the same community.

However, even the aforementioned NRC attempt to present an evidence-based ranking of US graduate programs has not been accepted without controversy [Vardi, 2016]. Additionally, allocating research funds with basis on an unidimensional ranking of graduate programs in a given area ignores the complex interplay between individual traits and program's unique patterns of strengths and weaknesses. While private organizations are also able to produce such rankings, it has been argued that in this way third-party business interests may influence academic values [Brembs et al., 2013].

In this dissertation, we provide a method for assessing the reputation of publication venues and graduate programs on a per subarea basis. For that, we rely on a reputation metric previously introduced in the literature [Ribas et al., 2015b] and modify it to the context of subareas. We apply this novel metric to rank conferences, journals, and graduate programs in Computer Science (CS) in Brazil and US.

1.1 Motivation

The decision on which researchers should be hired, promoted, granted funding or awarded is typically based on criteria such as number of publications, impact of publications, number of students under supervision, number of advised MSc and PhD theses, and participation in committees (conferences, program committees, journal editorial boards, technical committees). The decision ultimately depends on how each criterion is assessed and the period of time covered in the assessment. To assist in the decision making process, several indices have become widely used to measure the productivity of researchers. Examples include the raw number of citations, h-index [Hirsch, 2005], g-index [Egghe, 2006] and citation z-score [Lundberg, 2007]. Likewise, most academic search platforms, such as Google Scholar,[1] Microsoft Academic Search,[2] and AMiner[3] provide estimates for such indices.

However, ranking researchers without regarding the specificity of their subareas or sub-fields of research is arguably unfair and potentially error-prone [Lima et al., 2013]. For instance, consider the problem of comparing the research output of 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 should take the differences among subareas into account. Otherwise, the comparison between them would be unfair.

As an illustration, consider the subarea of Human-Computer Interaction within the broad field of CS. Experimental evaluation in this area usually takes more time than in other CS subareas when arranging and assessing users' feedback is necessary [Wainer et al., 2013]. On the other hand, CS subareas such as Databases and Computer Graphics do not usually face the same problem because their experimental evaluations depend on assessing the outcome of an automatic process, such as a query evaluation or a graphics rendering engine. Likewise, researchers working on a given subarea may have fewer publications than others, but with a potentially higher impact in their community.

To have a better understanding of this current problem in academic rankings, see the scenario of CS in Table 1.1. It contains a list of 37 subareas of CS according to Microsoft Academic Search. Although many studies in the literature show that there are significant differences in CS subareas' publication patterns [Hoonlor et al., 2013; Lima et al., 2015; Wainer et al., 2013], several institutions responsible for evaluating

[1] `http://scholar.google.com`
[2] `http://academic.research.microsoft.com`
[3] `http://arnetminer.org`

Table 1.1. The Microsoft 37 Subareas of Computer Science

CS Subareas	
Algorithms	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 systems
Computer hardware	Operations research
Computer networks	Parallel computing
Computer security	Pattern recognition
Computer vision	Programming languages
Data mining	Real-time computing
Data science	Simulation
Databases	Speech recognition
Distributed computing	Telecommunications
Embedded systems	Theoretical computer science
Human-computer interaction	World Wide Web
Information retrieval	

graduate programs do not take those differences into account.

In Brazil, for instance, as in several other countries, the graduate programs are evaluated by a funding agency. This agency, called CAPES,[4] runs periodic evaluations of graduate programs and uses citation-based metrics to rank all publication conferences and journals. Specifically, it ranks the graduate programs based on their publication records and other features such as courses taught, number of students graduated, average time to graduate a student, and department infrastructure. In this evaluation model, the ranking of publication venues is decisive on identifying the most reputable graduate programs. Thus, if the ranking of venues does not take into account the differences in subareas, the evaluation of graduate programs runs the risk of being very biased towards criteria such as popularity or raw publication volume [Souza e Silva and Maldonado, 2009].

Moreover, the distribution of grants, scholarships, and awards are also based on metrics computed from the perspective of the broad area of CS. Further, funding agencies as CNPq[5] impose quantitative limits on the number of research grants awarded

[4]Foundation for the Coordination and Improvement of Higher Level or Education Personnel. `http://capes.gov.br`

[5]Brazilian National Research Council. `http:/cnpq.br`

to researchers working on any given broad area of knowledge, such as CS [Navaux et al., 2017]. Thus, one can realize that the most popular subareas in CS have large advantages in this classification process over the other subareas in CS. Given that, there is a necessity to take a closer look at these academic scenarios from a perspective of subareas.

1.2 Dissertation Statement

We claim that general purpose rankings of CS graduate programs, rankings based on weights associated with the venues in the broad area of knowledge without consideration to particularities of subareas, may not reflect important information about the academic excellence of these programs. For instance, if the government intends to stimulate research groups working on a specific subarea, such as Machine Learning, a general ranking in the broad area of CS will not facilitate the decision.

In particular, this dissertation aims to answer the following research questions:

- Q1. How to quantify the reputation of publication venues and graduate programs on a per subarea basis?

- Q2. How does the reputation of Brazilian and US graduate programs in CS vary per subarea?

- Q3. Are there differences between the current research directions in CS of the top Brazilian and US graduate programs?

To answer these questions we extend the usability of a reputation metric called P-Score (for Publication Score) [Ribas et al., 2015a] proposed in the literature. We apply this extended metric in an academic dataset to produce rankings of publication venues and graduate programs in CS on a per subarea basis. Then, we analyze the reputation of Brazilian graduate programs in CS by subarea and compare them to US graduate programs working in the same subareas. We also provide an overview of current research directions of the two countries, i.e., on which subareas they have the most prominent work nowadays.

1.3 Contributions

Our main contributions are:

- Contribution 1: A novel reputation metric for academic rankings on a per subarea basis.
- Contribution 2: An analysis of research output in Computer Science in Brazil and US on a subarea level.
- Contribution 3: A comparison between the current research directions in CS of Brazil and US and insights on the differences in research interests of each country.

1.4 Dissertation Overview

The remainder of this dissertation is structured as follows. In Chapter 2 (Related Work), we present a literature review on reputation models and some instantiations of these models in academic search tasks. Chapter 3 (Reputation Flows and P-score) describes the theoretical concepts supporting our approach, by presenting the key ideas of the reputation model we used in this work and the strategies we adopted to study the academic data on a per subarea basis. In Chapter 4 (Experimental Setup), we present the experimental methodology of our reputation assessment, including information about our academic dataset, ground-truths considered and evaluation metrics we adopted in this study. Chapter 5 (Experimental Results) discusses our findings on ranking publication venues and CS graduate programs from Brazil and US, and presents a comparison between the scientific directions of the two countries. Chapter 6 (Conclusions and Future Work) concludes the dissertation, summarizes the key contributions of this work and provides directions for further research.

Chapter 2

Related Work

In this chapter, we describe related work on academic search considering the broad areas of knowledge (Section 2.1) and on a per subarea basis (Section 2.2).

2.1 Academic Search in Broad Areas

Garfield's Impact Factor [1955] is one of the first metrics proposed to quantify research impact. It has been widely used nowadays to measure relative importance of a scientific journal within its field. In a nutshell, it indicates the average number of citations per publication of a journal, in the last two years. Despite its wide usage since it was proposed in 1955, it has been largely criticized [Saha et al., 2003]. Accordingly, several alternatives have been proposed in the literature, such as other citation-based metrics [Egghe, 2006; Waltman et al., 2011; Sun and Giles, 2007; Waltman et al., 2011; Yan et al., 2011; Lundberg, 2007], download-based metrics [Bollen et al., 2005], PageRank-like metrics [Gollapalli et al., 2011; Yan and Lee, 2007], and community-based features [Silva et al., 2014].

One of the most widespread citation-based metric, the H-index, was proposed by Hirsch [2005]. 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 number 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 [Ding and Cronin, 2011; Egghe, 2006; Sun and Giles, 2007; Yan et al., 2011] and studied their impact, advantages, and disadvantages [W.Martins et al., 2009; Leydesdorff, 2009].

The idea of reputation, without the direct use of citation data, was discussed by Nelakuditi et al. [2011]. 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. [2008], who show that conference publications are important vehicles for disseminating CS research, while in other areas such as Physical Science 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 [Balog, 2012; Cormode et al., 2014; Deng et al., 2012; Gollapalli et al., 2011; Wu et al., 2009]. Particularly, Gonçalves et al. [2014] quantified the impact of various features on a scholar popularity throughout her career. Specifically, they analyzed how features that capture the number and rate of publications, number and quality of publication venues, and the importance of the scholar in the co-authorship network relate to the scholar popularity, by applying regression analysis. 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. This finding is one of the hypotheses for the reputation model introduced by Ribas et al. [2015b].

In addition, the prediction of scientific success of a researcher is also valuable for several goals, for example, hiring faculty members, guiding funding agencies in their decision processes and improving scholar rankings in academic search engines [Nezhadbiglari et al., 2016]. As a result, previous work attempted to predict if a researcher will become a principal investigator [Dijk et al., 2014], her future H-index [Dong et al., 2015; Penner et al., 2013] and the potential number of citations to her publications [Castillo et al., 2007; Mazloumian, 2012].

Although citation-based metrics are useful, they are not enough to do a complete evaluation of research. In particular, Piwowar [2013] 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 by Ribas et al. [2015b]. 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* [Ribas et al., 2015a]. Instead of relying on standard citation-based approaches for identifying reputable venues and researchers, P-score captures publishing behavior as a reputation signal, using a few highly reputable sources (e.g., reputable publication venues). For a better understanding of both the extension we propose and the methodology we adopted in this work, we provide in Chapter 3 more detailed information about the *reputation flows* model and its instantiation in the academic domain.

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, such as 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 [Kapeller, 2010; Lee et al., 2010] and in Computer Science [Hoonlor et al., 2013; Benevenuto et al., 2015; Lima et al., 2013].

2.2 Academic Search on a per Subarea Basis

Alves et al. [2013] and Benevenuto et al. [2015] analyze the structure of the communities formed by the flagship conferences of ACM SIGs. Their findings show that most of the ACM SIGs are able to connect their main authors in large and visually well-structured communities. However, they note that a few conferences, such as the ACM Symposium on Applied Computing, flagship conference of SIGAPP, and the ACM Conference on Design of Communications, flagship conference of SIGDOC, do not form a strong research community, presenting a structure with several disconnected components. They have opened their results to the research community as an interactive visualization tool[5] that allows one to browse the scientific communities, visualizing their structures and the contribution of each specific researcher to connect its coauthorship graph. The concept of scientific community and its natural development enforced the clustering approach we initially adopted in this work.

[1] https://scholar.google.com.br
[2] http://academic.microsoft.com
[3] http://aminer.org
[4] http://citeseerx.ist.psu.edu
[5] http://acmsig-communities.dcc.ufmg.br

Several comparative studies on research productivity between nations have been presented, for improving evaluation metrics [Laender et al., 2008], characterizing research development in different regions of the world [Menezes et al., 2009], or suggesting directions to improve scientific production and impact [Wainer et al., 2009]. In constrast with these aforementioned works, we are the first to perform this comparison using the reputation model proposed by Ribas et al. [2015b] and its derived metric, called P-score.

Lima et al. [2013] showed how important the subareas of expertise are when assessing research profiles in Computer Science, proposing the *ca-index*, a cross-area metric for ranking researchers by aggregating their productivity indexes across multiple areas. Likewise, they aimed to improve the reputation assessment of authors with interdisciplinary research output. In contrast, we propose to assess reputation of researchers in each CS subarea independently, treating each subarea as a different research community.

Wainer et al. [2013] 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.

The idea of using normalized metrics to assess the reputation of academic entities on a per subarea basis was inspired by previous works in the literature [Leydesdorff et al., 2013; Glänzel et al., 2011; Opthof and Leydesdorff, 2010; Waltman and Eck, 2013]. Such metrics were proposed to consider the differences in publication and citation practices among research 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 ranking graduate programs based on this information, in a semi-automatic fashion.

Chapter 3

Reputation Flows and P-score

In this chapter, we describe the reputation model we adopted as reference and the modifications we propose to it, in order to obtain rankings of academic entities on a per subarea basis. Specifically, in Sections 3.1 and 3.2, we summarize the key points of the Reputation Flows' model and its derived metric for academic search, the P-score, proposed by Ribas et al. [2015a]. This introduction is essential to comprehend the modifications we propose for the reputation model to take subareas into account. The motivation for such adjustments is described in Section 3.3. Then, in Sections 3.4 and 3.5, we discuss in detail our ideas to consider a per subarea basis and the specific modifications aimed towards this purpose.

3.1 Reputation Flows

While quantifying the reputation of a given entity is a challenging task, Ribas et al. [2015b] argue that the flow of reputation among entities can be accurately modeled as a stochastic process. To this end, they proposed a conceptual framework for ranking entities that convey reputation to one another. They introduced a *reputation graph*, a data structure that models the flow of reputation from selected sources to multiple targets and then formalized a stochastic process to estimate the amount of reputation transferred to target entities. This model was called *reputation flows*.

Figure 3.1. Structure of the reputation graph.

The interaction between reputation sources and reputation targets is inspired by the notion of *eigenvalue centrality* in complex networks, which also provides the foundation to PageRank [Brin and Page, 1998]. In particular, let S be the set of reputation sources, T the set of reputation targets, and $\mathbf{P}$ be a *stochastic* matrix of size $(|S| + |T|) \times (|S| + |T|)$ with the following structure:

$$\mathbf{P} = \left[\begin{array}{r|r} \delta_s \cdot \underset{S\times S}{\mathbf{P}_1} & (1-\delta_s) \cdot \underset{S\times T}{\mathbf{P}_2} \\ \hline (1-\delta_t) \cdot \underset{T\times S}{\mathbf{P}_3} & \delta_t \cdot \underset{T\times T}{\mathbf{P}_4} \end{array}\right], \tag{3.1}$$

where each quadrant represents a distinct type of reputation flow. Matrix $\mathbf{P}$ depends on the following matrices:

- $\mathbf{P}_1$: stochastic matrix of size $|S| \times |S|$ representing the transition probabilities between reputation sources;
- $\mathbf{P}_2$: matrix of size $|S| \times |T|$ representing the transition probabilities from reputation sources to targets;
- $\mathbf{P}_3$: matrix of size $|T| \times |S|$ representing the transition probabilities from reputation targets to sources;
- $\mathbf{P}_4$: stochastic matrix of size $|T| \times |T|$ representing the transition probabilities between reputation targets.

The parameters δ_s and δ_t control the relative importance of the reputation sources and targets, which are modeled in the four matrices above. Specifically, $\delta_s \in [0, 1]$ is the fraction of reputation one wants to transfer between source nodes and $\delta_t \in [0, 1]$ is the fraction of reputation one wants to transfer between target nodes. These parameters are useful to calibrate the impact of different types of reputation flows in the final reputation score.

Note that, as (i) the sub-matrices $\mathbf{P}_1$ and $\mathbf{P}_4$ are *stochastic* and (ii) each of the rows of matrices $\mathbf{P}_2$ and $\mathbf{P}_3$ sums to 1, then $\mathbf{P}$ defines a Markov chain. Assuming that the transition matrix $\mathbf{P}$ is ergodic, we can compute the steady state probability of each node and use it as a reputation score.[1] Specifically, we can obtain values for ranking the set of nodes by solving:

[1] Recall that, in an ergodic process, the state of the process after a long time is nearly independent of its initial state. [Walters, 2000]

$$\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}, \tag{3.2}$$

where $\boldsymbol{\pi}$ is a row matrix with $|S| + |T|$ elements, each one of them representing the probability of a node in the set $S \cup T$. This system of linear equations can be solved by standard Markov chain techniques, as the Power Method.[2] Then, we obtain the steady state probabilities of all nodes in $S \cup T$ (reputation sources and reputation targets).

In short, the steady state probability of an entity is interpreted as its relative reputation, as transferred from other entities in a reputation graph. Subsequently, the reputation of the target nodes can be further propagated to entities we want to compare, in the *collateral set* C, as shown in Figure 3.1. This propagation depends on a matrix $\mathbf{P}_C$ 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:

$$\textit{P-score}(e) = \begin{cases} \sum\limits_{t \in T} p_{t,e} \cdot \pi_t & \text{if } e \in C \\ \pi_e & \text{otherwise} \end{cases} \tag{3.3}$$

where $p_{t,e} \in \mathbf{P}_C$ is the transition weight from a target entity $t \in T$ to an entity e, $\pi_e \in \boldsymbol{\pi}$ is the reputation of entity e, $\pi_t \in \boldsymbol{\pi}$ 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.

As shown by Ribas et al. [2015b], the conceptual framework of reputation flows could be instantiated in the academic research domain, for instance, by modeling the transference of reputation between authors, papers, graduate programs and publication venues. In their experiments, they study how the reputation of a reference set of graduate programs is propagated to the venues they publish in, i.e., graduate programs are seen as reputation sources and publication venues as reputation targets. In the next section, we discuss this instantiation of reputation flows in the academic search domain.

3.2 Academic Rankings Based on Standard P-score

Let us concentrate our attention on the problem of ranking reputation in the academic research domain. In this case, key entities of interest are researchers, graduate programs, papers, and publication venues. The hypotheses supporting the *standard P-score* metric (i.e., as it was originally proposed) are:

[2] http://en.wikipedia.org/wiki/Power_iteration

1. A graduate program conveys reputation to a publication venue proportionally to its own reputation.

2. A publication venue conveys reputation to a graduate program proportionally to its own reputation.

The basic idea of the standard P-score is to associate a reputation with publication venues based on the publication patterns of a reference set of graduate programs. Given a pre-selected set of reference graduate programs, P-score associates weights with the publication venues the researchers in the reference graduate programs publish in, by solving a system of linear equations relating those entities in the reputation graph (see Equation 3.2). Further, these weights can be used to rank publication venues (by considering these weights as venue scores) and also to rank graduate programs or authors.

In particular, using graduate programs (set G) as reputation sources, publication venues (set V) as reputation targets, and adopting $\delta_s = 0$ and $\delta_t = 0$, we have an instance of matrix P of Equation 3.1 for reputation flows in the academic search domain:

$$\mathbf{P} = \left[\begin{array}{c|c} \mathbf{0} & \underset{G\times V}{\mathbf{P}_{\mathrm{i}}} \\ \hline \underset{V\times G}{\mathbf{P}_{\mathrm{ii}}} & \mathbf{0} \end{array}\right], \tag{3.4}$$

where each quadrant represents a distinct type of reputation flow in the academic search domain. Figure 3.2 shows an example with two graduate programs (Groups 1 and 2) used as reputation sources and three venues used as reputation targets.

From Figure 3.2, Group 1 published three papers in Venue 1, two papers in Venue 2, and one paper in Venue 3. Then, the number of publications of Group 1 is six. Venue 1 receives three papers from Group 1, and two papers from Group 2. The fractions of publications from groups to venues and from venues to groups are the edge weights.

Ribas et al. [2017] have also suggested different possible configurations for the reputation graph. In fact, the sources and collaterals can be any of the three types of entity we consider, i.e., publication venues, authors, and graduate programs. The targets, however, must always be entities of type venue. That is, we should always use the reputation of venues as the key feature for ranking venues, graduate programs, and individual authors.

In the next section, we present the initial approaches we adopted to rank academic entities on a per subarea basis and discuss our motivations to modify the standard P-score metric.

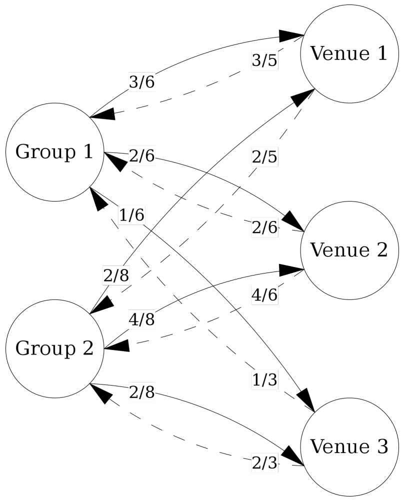

Figure 3.2. Markov chain for an example with two graduate programs and three publication venues.

3.3 The Encroachment Problem

Initially, based on the concepts and usability of the standard P-score, we consider a simple hypothesis: to rank academic entities on a per subarea basis, it would be sufficient to use the right reputation sources within a subarea and compute the steady state probabilities of all academic entities. As a result, in theory, the entities with higher P-scores would be the most reputable entities within a subarea. For instance, using as reputation sources the most important venues in Information Retrieval, we would find the most reputable authors, graduate programs, and publication venues in the subarea of IR.

Likewise, the instantiation of P-score in these first experiments, based on one of the possible configurations suggested by Ribas et al. [2015b], was as follows: given a pre-selected set of reference venues (or *seeds*) in a given subarea, e.g., conferences or journals, the metric finds the n researchers with the largest volume of publications in these reference venues (for instance, $n = 200$) and uses them to assemble the Markov network model which also includes all venues in which they published. The steady state probabilities are interpreted as venue weights that distinguish high 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 graduate programs or authors.

However, this approach proves to be inappropriate when considering subareas. The reason is that publication venues possibly cover multiple subareas and thus their reputation needs to be split among their 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 from another subarea, such as DB. This is what we call the *encroachment problem*.

We illustrate this problem 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 accepted by 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 graduate program at Purdue University on the subarea of IR.

Given that CIKM does not distinguish in its proceedings which papers are on IR, DB or 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 that 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, described in the next section.

3.4 Normalized P-score for Publication Venues

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 surely focused on that subarea). This occurs 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:

$$\textit{norm-P-score}(v) = \frac{\textit{P-score}(v)}{number\_of\_publications(v)} \tag{3.5}$$

Equation (3.5) is the starting point for obtaining the ranking of graduate programs on a per subarea basis, as presented in the following section.

3.5 Weighted P-score for Graduate Programs

In order to facilitate understanding the modifications we propose for the standard P-score to rank graduate programs by subareas, let us rewrite some equations of standard P-score in a clearer and more concise way than using the whole conceptual framework of reputation flows, focusing on ranking graduate programs itself.

Evaluating a specific graduate program requires weighting the contributions of its members who are responsible for the reputation of the program. Like Ribas et al. [2015b], we say that a group's reputation is the sum of the reputation of the venues its members have published in, taken on a per author basis. Moreover, a research publication is usually a combination of efforts by multiple researchers. In consequence, we should normalize the paper's P-score by the number of authors. Thus:

$$\textit{P-score}(g) = \sum_{p \in \theta(g)} \frac{\textit{P-score}(venue(p))}{number\_of\_authors(p)} \tag{3.6}$$

where $\theta(g)$ is the set of publications of graduate program g and *venue(p)* is the venue in which paper p was published. In other words, the reputation of a graduate program is based on the papers it published. Each paper has a reputation itself, which is given by the venue where the paper was published.

The research of a graduate program is produced by its faculty members, post-doctoral fellows, doctoral students, among others. To determine the correct affiliation of all these graduate program members is a costly task, since that information is often not available and dynamically changes (e.g., it is common for graduate students to move from a graduate program to another over the years). However, usually faculty members are responsible for research groups at which postdoctoral fellows and doctoral students work. At some point, the group publishes a paper and its lead researcher, or principal investigator, is typically involved. This allows us to use faculty members as the anchors for the transference of reputation to research groups and, subsequently, to the graduate programs these research groups belong to. Notice that, in this case, the indication of the faculty member's affiliation is key to estimate the reputation for graduate programs. Hence, Equation (3.6) may be rewritten as:

$$\textit{P-score}(g) = \sum_{a \in \phi(g)} \sum_{p \in \theta(a)} \frac{\textit{P-score}(venue(p))}{number\_of\_authors(p)} \tag{3.7}$$

where $\phi(g)$ are the researchers associated with graduate program g and $\theta(a)$ are the publications of author a.

Except for the normalization by the number of authors in a paper (which we propose), the Equation 3.7 is an alternative way of showing how the standard P-score metric ranks graduate programs, as stated by Equation 3.3, in Section 3.1. However, recap that standard P-score does not provide good ranking of graduate programs on a per subarea basis, due to the encroachment problem described in Section 3.3.

The solution we propose to 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 referred as factor γ) 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 Chapter 5.

For that, firstly we compute the normalized P-score (Section 3.4) to every publication venue based on a subarea s (e.g., Information Retrieval). We claim that the normalized P-score indicates the level of convergence between the publication records of a venue and the specific interests of the subarea. This claim is supported by our results discussed in Chapter 5. Then, the set V_s of venues more restricted to that subarea s is defined as:

$$V_s = \{v \in V \mid \textit{norm-P-score}(v) > \alpha_s\} \tag{3.8}$$

where V is the set of all publication venues, *norm-P-score(v)* is the normalized P-score for venue v and α_s is a threshold defined by manual inspection of the venues with the highest normalized P-scores for the subarea s.

Once the set V_s is defined, we can measure the reputation of authors (and consequently, graduate programs) on a per subarea basis. With V_s being the set of venues closely associated with subarea s and $\theta(a)$ the set of publications of author a, as before, we define:

$$\gamma\text{'}(a, s) = \frac{1}{|\theta(a)|} \sum_{p \in \theta(a)} \begin{cases} 1 & \text{if } p \in V_s \\ 0 & \text{otherwise} \end{cases} \tag{3.9}$$

where γ'(a, s) is the interest factor of author a in subarea s, 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:

$$\gamma(a, s) = \frac{1}{|\theta(a)|} \sum_{p \in \theta(a)} \begin{cases} \frac{1}{\log_2 (y(0) - y(p) + 2)} & \text{if } p \in V_s \\ 0 & \text{otherwise} \end{cases} \tag{3.10}$$

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 (3.7) and present the *weighted P-score* of a graduate program g in subarea s as:

$$\textit{weighted-P-score}(g, s) = \sum_{a \in \phi(g)} \gamma(a, s) \times \sum_{p \in \theta(a)} \frac{\textit{P-score}(\textit{venue}(p))}{\textit{number\_of\_authors}(p)} \tag{3.11}$$

The experimental results of these approaches are presented in Chapter 5. In the next section, we describe some considerations on using weighted P-scores to compare research output across different subareas.

3.6 Comparing Research in Different Subareas

Most assessments of this kind consist of analyzing the raw number of publications in each subarea. Likewise, they tend to select few venues that represent the area and then only consider them for the purpose of counting. In here, besides the number of publications, we also consider the reputation of each venue, according to P-score. As shown in Chapter 3, P-score provides encouraging results in ranking academic entities. We aim to show that this metric, when applied to CS graduate programs, allows us to gain valuable insights about the divergences in current publication patterns between different countries.

One final caveat. When we compute P-scores on a per subarea basis, we run a stochastic computation for each subarea. A direct side effect is that P-scores of each subarea, which represent steady state probabilities in a Markov network, are scaled up to sum up to 1. For comparisons among subareas, this is a problem. In particular, smaller venues might receive disproportionately high relative P-scores due to stochastic scaling in a given subarea.

Thus, to allow proper comparisons across subareas, we use an artifact we borrow from the computation of Pagerank [Page et al., 1998]. Inside each subarea, we consider that a fraction of the time, 85% for instance, transitions occur to nodes inside the Markov network for that subarea. The other fraction of the time (15%) transitions occur to nodes in one of the other subareas. The final result is that steady state probabilities of nodes in all subareas must now sum up to 1, which makes them directly comparable.

Chapter 4

Experimental Setup

In this chapter, we describe the setup supporting our experiments and the key assumptions we have considered on assessing reputation in academia. Specifically, we provide details of the academic dataset we used in this work, the CS subareas considered, the methods we adopted, and previous results used as starting points to our present analysis.

4.1 Dataset

We compiled a collection of academic publications records extracted from DBLP,[1] an online reference for bibliographic information on major CS publications. DBLP data has been used in related studies on CS research communities [Biryukov and Dong, 2010; Delgado-Garcia et al., 2014; Hoonlor et al., 2013; Laender et al., 2008; Wainer et al., 2013]. The dataset 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 citations.

Our collection is actually an extension of the DBLP repository. While it contains all publication venues and authors from DBLP, we have enriched it by adding data regarding graduate programs. To do so, we manually collected data about the top 126 CS graduate programs evaluated in the 2011 assessment conducted by the US National Research Council (NRC).[2] In particular, for each of these graduate programs,

[1] `http://dblp.uni-trier.de`
[2] `http://nap.edu/rdp`

we retrieved the list of its members, which was 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,[3] which ranks CS graduate programs based purely on their publications. They do so by collaboratively collecting data on authors, such as their homepage and affiliation. Therefore, we used that data to enhance our repository. Salient statistics on our dataset are shown in Table 4.1.

Table 4.1. Salient statistics of the dataset used in our evaluation.

Attribute	Value
Number of Papers	2,931,849
Number of Authors	1,595,771
Number of Venues	5,765
Number of US Graduate Programs	126
Avg. number of faculties per US Graduate Program	42.4
Number of BR Graduate Programs	25
Avg. number of faculties per BR Graduate Program	47.8

4.2 Computer Science Subareas

There are different ways of defining subareas in CS depending on the institution responsible for the classification. Two notorious classification are given by ACM[4] (through its *Special Interest Groups*) and IEEE[5] (through its *Technical Committees*). Notice that most of them divide CS into subareas rather distinct. 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 [Wainer et al., 2013] or use another source of information [Hoonlor et al., 2013].

A more recent classification is presented by Microsoft Academic Search.[6] It divides CS into 37 subareas, including relatively new ones. For the purpose of this work, we selected 20 subareas from the Microsoft Academic Search classification, as presented in Table 4.2. Along with the list, we present the abbreviation of each subarea, which we will be using from here on. We also show two venues we selected as notorious in each subarea of interest. These venues are important for identifying a group of researchers

[3] `http://csrankings.org`
[4] `http://acm.org/sigs`
[5] `http://computer.org/web/tandc/technical-committees`
[6] `http://academic.research.microsoft.com`

whose main research topics of interest are likely to be in that subarea — an essential information for the application of P-score to subareas. However, although the choice of seed venues is an important task, we have observed that any two central venues to a given subarea are sufficient to produce reasonable rankings of publication venues and graduate programs for that subarea.

Table 4.2. Subareas of Computer Science selected from Microsoft classification. The full names of the publication venues are presented in Appendix C.

Subarea	Abbreviation	Seed venues
Algorithms	Alg	SODA, Algorithmica
Artificial intelligence	AI	IJCAI, AI
Bioinformatics	Bio	BIBM, Bioinformatics
Computer graphics	CG	SIGGRAPH, TCVG
Computer networks	CN	INFOCOM, TON
Computer security	CS	CCS, TISSEC
Computer vision	CV	CVPR, IJCV
Data mining	DM	KDD, SIGKDD
Databases	DB	SIGMOD, TODS
Distributed computing	DC	ICDCS, TPDS
Human-computer interaction	HCI	CHI, TOCHI
Information Retrieval	IR	SIGIR, TOIS
Machine learning	ML	ICML, JMLR
Natural language processing	NLP	EMNLP, COLING
Operating systems	OS	SOSP, SIGOPS
Parallel computing	PC	IPPS, TPDS
Programming languages	PL	PLDI, TOPLAS
Speech Recognition	SR	INTERSPEECH, TCOM
Theoretical computer science	TCS	STOC, SIAMCOMP
World Wide Web	WWW	WWW, WS

While our subset of 20 CS subareas is not perfect or exhaustive, it is detailed enough to allow gaining insights into the scene of research in CS in Brazil, which would not be possible to obtain otherwise. In Appendix B, we present other classifications of CS subareas, according to reliable sources.

An alternative configuration of P-score consists in using a set of researchers with high reputation as seed, instead of publication venues. In particular, on a per subarea basis, one can identify the most reputable researchers in a given subarea, use them as seeds of reputation in P-score, and subsquently finds the most important venues in that subarea. To automatically identify these most reputable researchers, we made experiments using clustering techniques on a graph of coauthorships, where the nodes were individual researchers and the edges represented coautorships between them, as discussed in Appendix A. Then, for each cluster in the subarea, we selected the top

n most representative researchers as reputation seeds. This methodology produced similar results when compared to the use of venues as seeds and can be improved in the future by enriching the graph of coauthorships with more information about each researcher, such as academic productivity and centrality metrics. Moreover, clustering methods on academic networks can also help in an automatic identification of the current subareas in the broad area of CS.

4.3 Venues Ground-Truth

To evaluate the effectiveness of normalized P-scores as defined by Equation (3.5) 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 assess the relevance of venues (included in a pre-selected list) to their subareas. 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 4.3.

Table 4.3. Seeds of publication venues for the P-score ranking used in this work.

	Subareas		
Type of venue	Databases	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 4.2. To reconcile 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 as 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*.

Chapter 5

Experimental Results

In this chapter, we discuss the results of ranking publication venues and graduate programs in CS on the three subareas we selected: Information Retrieval, Databases, and Data Mining. In particular, for the ranking of graduate programs, our results are restricted to the 126 US graduate programs considered by NRC in 2011.

5.1 Ranking Venues

Using the normalized P-score (norm-P-score) presented in Section 3.4, we were able to better discriminate publication venues of a given CS subarea from venues of other subareas. Figure 5.1 presents the precision-recall curve obtained by norm-P-score in the task of ranking publication venues in Information Retrieval, in light of the results produced using two other methods, namely H-index and standard P-score. To produce the precision-recall curves, we use as ground-truth a venue classification done by experts in each subarea, see Section 4.3. We reproduced the same experimental methodology for ranking venues in the subareas of Databases and Data Mining. The precision-recall curves are shown in Figure 5.2.

As it is clear from Figures 5.1 and 5.2, 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 adopted in Figure 5.1 are strongly related to IR, according to the assessments of specialists.

To further illustrate, Table 5.1 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

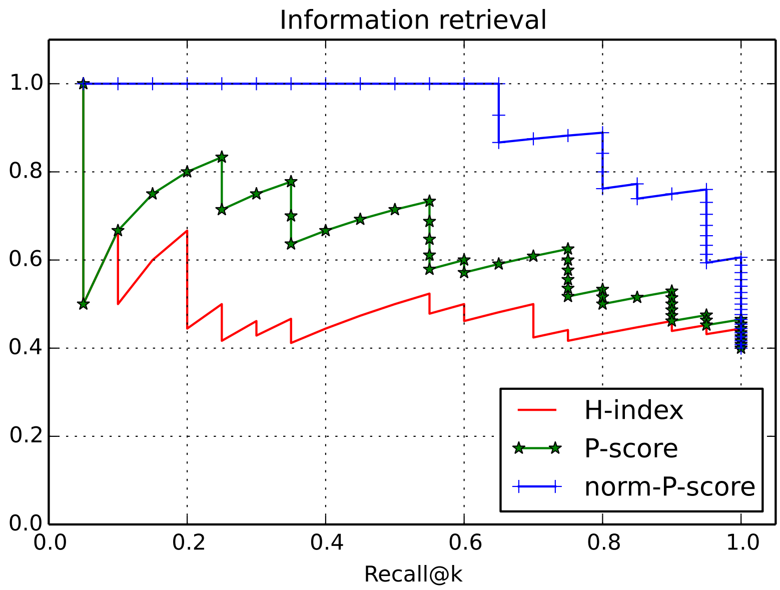

Figure 5.1. Precision-Recall curves of H-index, P-score and normalized P-score for the subarea of Information Retrieval.

as seed venues. Also, a final ranking of IR venues is computed by re-ranking the venues — previously filtered by normalized P-score — according to their standard P-scores.

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 IR subarea, 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 on IR venues such as the International Conference on the Theory of Information Retrieval (ICTIR) and Transactions on Information Systems (TOIS) appear among the top 20 publication venues. This power of discrimination of the normalized P-score is important to allow selecting venues that better represent the subarea of IR.

Similar results can be found in Tables 5.2 and 5.3, for the subareas of DB and DM, respectively. In DB, venues with a large scope such as WWW and International Conference on Data Mining (ICDM), although with high standard P-scores, are not selected among the top 20 venues according to the normalized P-score metric. In DM, venues such as the AAAI Conference on Artificial Intelligence and the Conference on Neural Information Processing Systems (NIPS), clearly not mainly focused on DM, are also filtered by the normalized P-score. In summary, for both subareas (DB and DM), the final ranking contains several publication venues highly focused on each subarea.

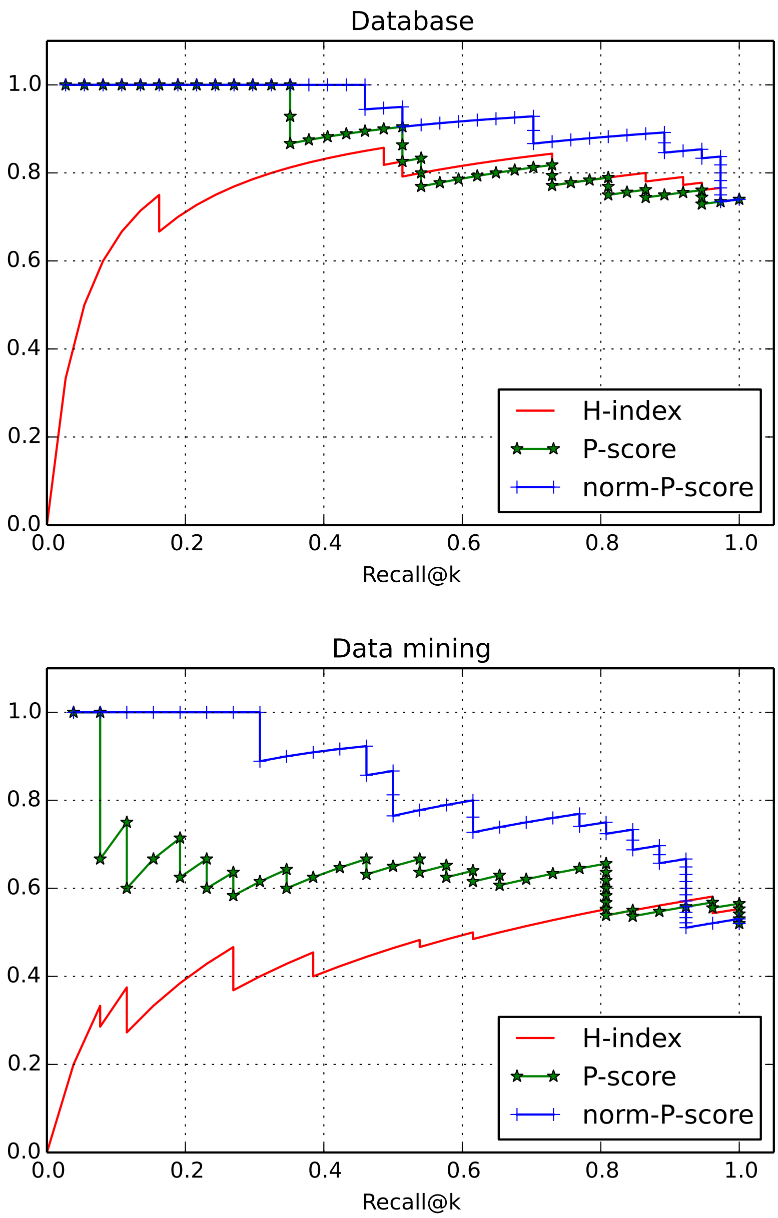

Figure 5.2. Precision-Recall curves of H-index, P-score and normalized P-score for the subareas of Databases and Data Mining.

Table 5.1. Top 20 venues in Information Retrieval using (i) standard P-score, (ii) the set of venues selected by the normalized P-score, and (iii) re-ranking the set of venues obtained in (ii) according to their P-scores. The suffixes (c) and (j) are used to differentiate conferences and journals with the same name. The full names of the publication venues are presented in Appendix C.

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

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. For this reason, in Tables 5.1, 5.2, and 5.3, we present the results obtained by normalized P-scores as sets of venues instead of usual rankings of venues.

Afterward, using this ranking of publication venues as a starting point, we can improve rankings of graduate programs in CS by considering a per subarea basis. Our results in ranking graduate programs are discussed in the next section.

5.2 Ranking US Graduate Programs by Subarea

For this section, we are interested in the distribution of publications in the CS subareas of graduate programs across US. More specifically, we want to understand which

Table 5.2. Top 20 venues in Databases, using (i) standard P-score, (ii) the set of venues selected by the normalized P-score, and (iii) re-ranking the set of venues obtained in (ii) according to their P-scores. The suffixes (c) and (j) are used to differentiate conferences and journals with the same name. The full names of the publication venues are presented in Appendix C.

#	Standard P-score	norm-P-score	#	Final Ranking
1	SIGMOD (c)	PVLDB	1	SIGMOD (c)
2	ICDE	VLDB (j)	2	ICDE
3	VLDB (c)	TODS	3	VLDB (c)
4	PVLDB	SIGMOD (c)	4	PVLDB
5	TKDE	SSD	5	TKDE
6	DEBU	CIDR	6	DEBU
7	SIGMOD (j)	DEBU	7	SIGMOD (j)
8	EDBT	TKDD	8	EDBT
9	CIKM	TKDE	9	PODS
10	PODS	DPD	10	TODS
11	TODS	ICDT	11	VLDB (j)
12	KDD	SSDBM	12	DASFAA
13	VLDB (j)	EDBT	13	SSDBM
14	WWW	ICDE	14	ICDT
15	ICDM	SIGMOD (j)	15	CIDR
16	DASFAA	WEBDB	16	WEBDB
17	SSDBM	VLDB (c)	17	SSD
18	IS	PODS	18	DPD
19	ICDT	DASFAA	19	COMAD
20	CIDR	COMAD	20	TKDD

subareas receive more attention by the top graduate programs in US.

As in Section 5.1, we consider the same three CS subareas: IR, DB, and DM.

Information Retrieval, Databases, and Data Mining

Table 5.4 presents a ranking of US universities based on standard P-scores for the IR subarea. The values presented are normalized to 1 as follows:

$$\overline{P}(g, s) = \frac{P(g, s)}{max(P(g, s))} \tag{5.1}$$

where $\overline{P}(g, s)$ is the normalized score of graduate program g in subarea s, $P(g, s)$ is the score and $max(P(g, s))$ is the highest score of the ranking. We apply Equation 5.1 to either standard and weighted P-scores.

The University of Massachusetts Amherst has the premier IR research group in

Table 5.3. Top 20 venues in Data Mining, using (i) standard P-score, (ii) the set of venues selected by the normalized P-score, and (iii) re-ranking the set of venues obtained in (ii) according to their P-scores. The suffixes (c) and (j) are used to differentiate conferences and journals with the same name. The full names of the publication venues are presented in Appendix C.

#	Standard P-score	norm-P-score	#	Final Ranking
1	KDD	TKDD	1	KDD
2	ICDM	CIKM	2	ICDM
3	CIKM	SDM	3	CIKM
4	ICDE	SIGKDD	4	ICDE
5	ICML	DATAMINE	5	ICML
6	SDM	KDD	6	SDM
7	TKDE	ICDM	7	TKDE
8	WWW	WSDM	8	PKDD
9	SIGMOD	KAIS	9	PAKDD
10	AAAI	VLDB (c)	10	SIGKDD
11	PKDD	SADM	11	DATAMINE
12	NIPS	TKDE	12	KAIS
13	PAKDD	ICML	13	PVLDB
14	VLDB (c)	PAKDD	14	WSDM
15	SIGIR	PKDD	15	TKDD
16	SIGKDD	ICDE	16	VLDB (c)
17	DATAMINE	RECSYS	17	RECSYS
18	KAIS	SSD	18	TIST
19	JMLR	TIST	19	SSD
20	PVLDB	PVLDB	20	SADM

Table 5.4. Ranking of the top 10 US Universities on Information Retrieval, based on 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

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 3.3. Table 5.5 presents the top 10 graduate programs for the IR subarea using our proposed approach of the weighted P-score, according to Equation (3.11), instead. We observe that the University of Southern California, the Georgia Institute of Technology, Stanford University and the University of California at Berkeley are no longer among the top 10 graduate programs. This seems appropriate given these universities are not active on research in IR.

Table 5.5. Ranking of the top 10 US Universities on Information Retrieval, using the weighted P-score (Equation 3.11).

#	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

To better understand the results in Table 5.5, we produced a list of the top 20 researchers on IR. Table 5.6 shows their affiliations. As we observe, our top 10 graduate programs for the IR subarea 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 showed in Table 5.6 had more publications in venues strongly related to the subarea. Hence, the ranking of graduate programs can be justified by the ranking of authors. We repeated this process to the DB and DM subareas. Tables 5.7 and 5.8 present the top 10 graduate programs on DB and DM, when we use the weighted P-scores produced by Equation (3.11).

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

Table 5.6. Ranking of the top 20 US researchers' universities on Information Retrieval, using Equation (3.11).

#	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

Table 5.7. Ranking of the top 10 US Universities on Databases, using the weighted P-score (Equation (3.11)).

#	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 5.8. Ranking of the top 10 US Universities on Data Mining, using the weighted P-score (Equation (3.11)).

#	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

US Research Overview

For the results presented here, we used the weighted P-score (Equation 3.11) to rank graduate programs in the 20 subareas selected. To do so, we first got the P-score for each subarea, where we used as seeds the venues in Table 4.2. After that, normalized P-scores allowed us to select the top venues that more adequately represent each subarea. Following, we used them to compute the weighted P-score for each graduate program in each subarea. Finally, we selected the top 20 graduate programs in US, for each subarea. Figure 5.3 presents the results we obtained.

Figure 5.3 presents our experiments considering only CS graduate programs in the US. It shows the cumulative weighted P-score for the top 20 CS graduate programs in each of the 20 subareas. We observe that Computer Vision is the subarea with highest cumulative weighted P-score, followed by Databases, Theoretical Computer Science, and Machine Learning.

Computer Vision attracts a lot of attention and interest nowadays because it is present in some of the most cutting-edge technologies, such as autonomous vehicles which demand a massive amount of computer vision processing. On the other hand, Theoretical CS (TCS) may appear a bit surprising. It is an area where publishing is slightly more difficult, especially because it often demands time, e.g., to prove theorems, and faculty usually advise fewer students and are thus likely to produce fewer papers [Wainer et al., 2013]. Despite this, our experiments showed TCS in third place. When inspecting the number of publications, we noticed that they have fewer than others, indeed. However, P-scores for venues in TCS were rather significant, indicating that the subarea has various venues of high reputation.

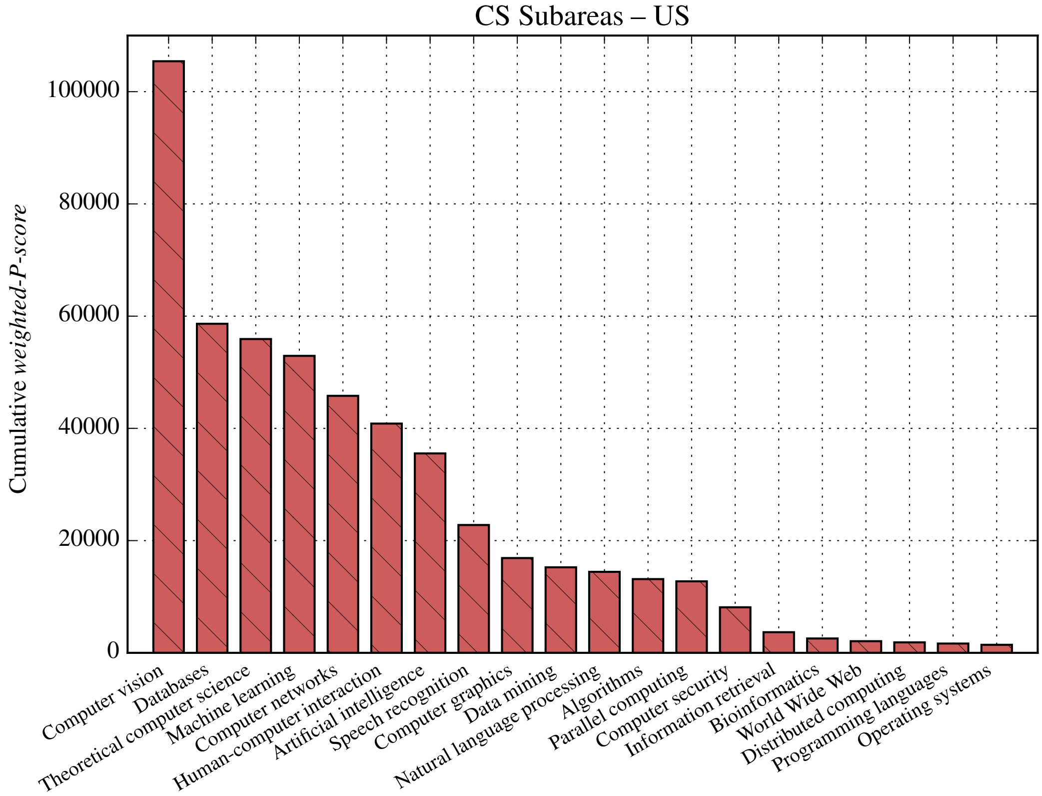

Figure 5.3. Distribution of cumulative weighted P-scores for the top 20 US graduate programs on a per subarea basis.

Another surprising subarea observed from Figure 5.3 is Machine Learning (ML). While it is a remarkable growing field of study, we observed that P-scores of ML venues were not well distributed. One problem that we noticed is that ML is extremely multidisciplinary. Therefore, several venues with notorious work for the ML subarea also contain important publications for other subareas, i.e., they are not venues dedicated exclusively to ML. Moreover, many subareas include Machine Learning related research (a good example is Computer Vision).

Databases and Computer Networks are also well positioned in our experiment. They are more traditional fields of study (TCS as well) and have a considerable amount of work developed in the past.

5.3 Ranking BR Graduate Programs by Subarea

We apply the same methodology of Section 5.2 for assessing the reputation of CS graduate programs in Brazil. Figure 5.4 presents our findings. It shows the cumulative weighted P-score for the top 20 CS graduate programs in each of the 20 subareas.

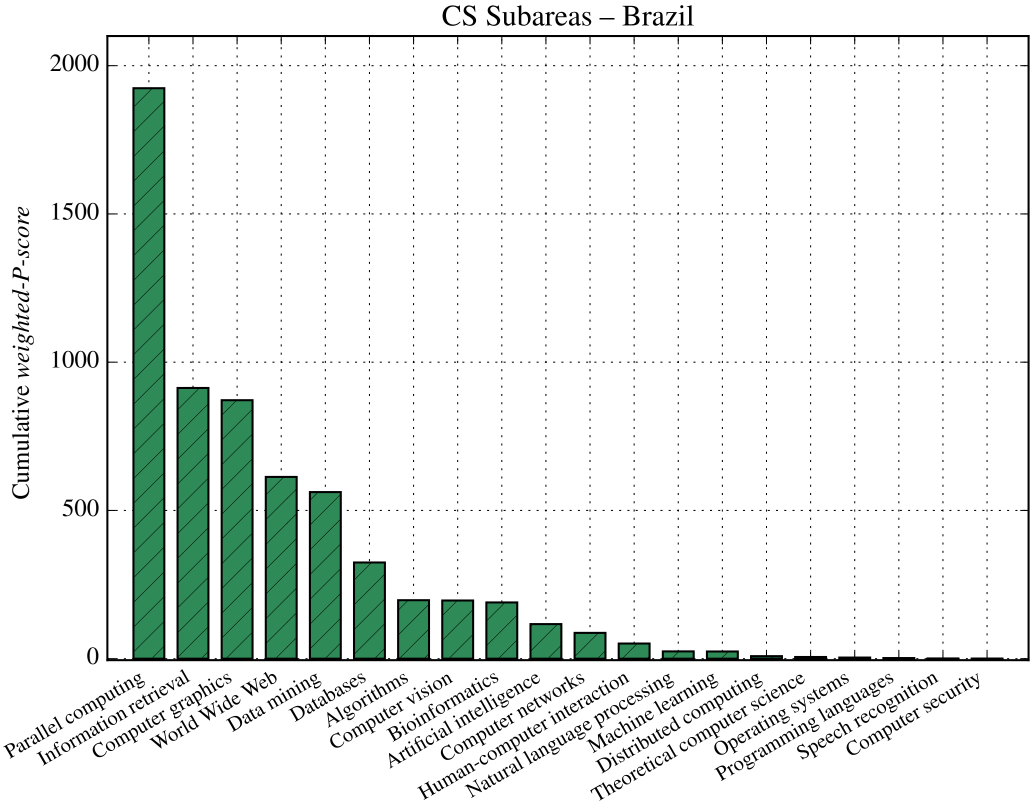

Figure 5.4. Distribution of cumulative weighted P-scores for the top 20 BR graduate programs per subarea.

According to our results, Parallel Computing is the subarea with the highest cumulative weighted P-score (considerably higher than the others), followed by Information Retrieval and Computer Graphics.

We also observe that P-scores in Brazil are almost two orders of magnitude smaller than those in the US. Thus, while the volume of publications in Brazil is relatively high [Laender et al., 2008], their impact in terms of exposure in high reputation venues is rather modest.

Computer Science production in Brazil is considerably uneven. Many subareas have extremely low scores, in some of them, the score is almost inexistent. Thus, while there are researchers in Brazil working on all subareas, the number of papers published in reputable venues in several subareas is rather small. In contrast, we noted that in some CS subareas there are Brazilian graduate programs as reputable as top US graduate programs. For instance, Figure 5.5 shows the top 20 graduate programs in Information Retrieval, according to the weighted P-score. We observe two Brazilian graduate programs among these programs. This is noteworthy, given this kind of finding would not be possible without considering a per subarea analysis.

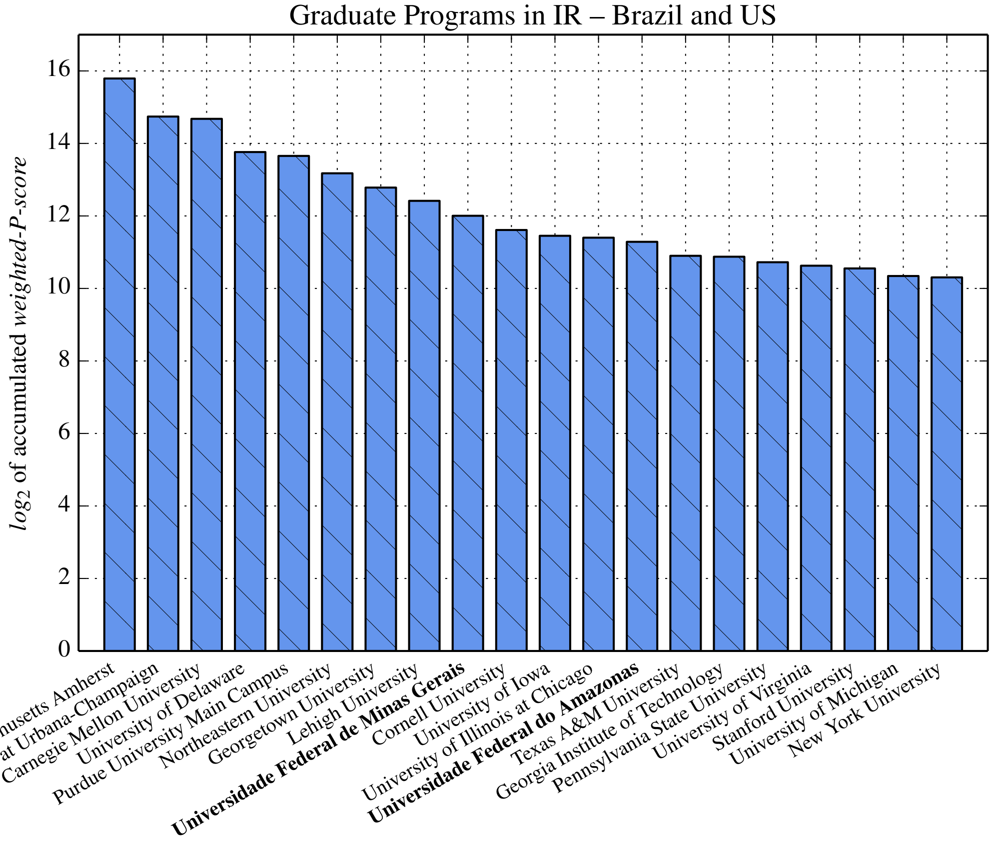

Figure 5.5. Top 20 Graduate Programs for the Information Retrieval subarea, according to weighted P-score, considering US and BR graduate programs, using logarithmic-scale.

5.4 Comparing BR and US Research in CS

In Figure 5.6 we reorder the results presented in Figure 5.4 according to the same order of subareas in the US graduate programs in Figure 5.3, with the purpose of comparing the CS research scenario in Brazil and in US.

In Brazil, Parallel Computing and Information Retrieval are the subareas with the highest scores, while in the US the two subareas of greatest interest are Computer Vision and Databases. We also observe that the five subareas of greatest interest in Brazil are very distinct from those of greatest interest in US. If we consider that the top US graduate programs are among the best programs world wide and also that their research directions are the key subareas in CS nowadays, this experiment suggests that academic research in Brazil is not shifting its focus of attention as fast as it happens in the US.

A possible explanation for this result may be the resistance by Brazilian CS

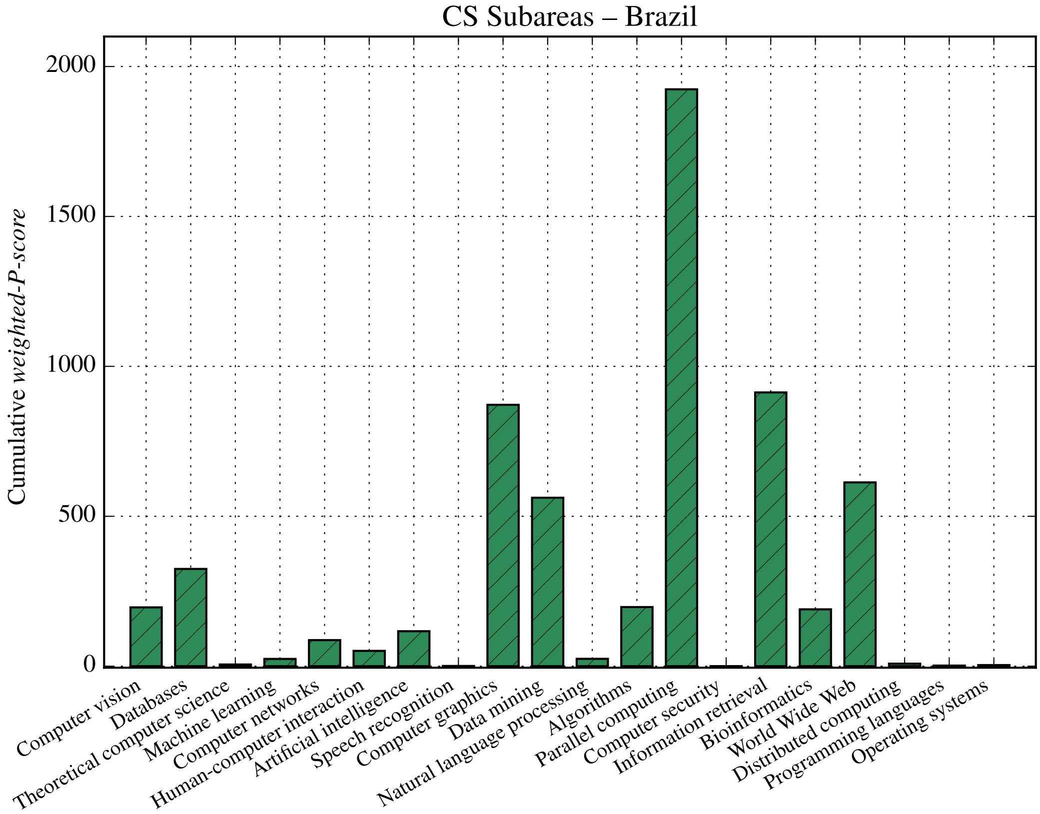

Figure 5.6. Distribution of cumulative weighted P-scores for the top 20 BR graduate programs per subarea, in the same subarea order of Figure 5.3.

researchers to move from their original subareas of study to new ones, or include newer subareas in their research interests, over the decades. This behavior can be motivated by personal reasons but also by a lack of support from government officials and funding agencies. With the increasing interest of companies in high advanced technologies and capable human resources, public-private partnerships can stimulate research of excellence in these modern subareas, which can direct benefit graduate programs in the next years.

Overall, this type of analysis can be important for those who need to decide how to allocate limited research funds. Although it is known that the volume of US scientific publications in CS is significantly larger than the volume of Brazilian CS research, this comparison shows that the CS subareas in which each country has major scientific impact are basically disjoint.

Chapter 6

Conclusions and Future Work

In this dissertation, our first research question was: *"How to quantify the reputation of publication venues and graduate programs on a per subarea basis?"* We presented experiments suggesting that metrics that only capture broad features of a subarea such as the volume of publications, citations or general reputation (standard P-score included) are not sufficient to produce reasonable rankings on a per subarea basis. In particular, we demonstrated that solving the venues encroachment problem allows us to improve the ranking of graduate programs 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 graduate program in a given subarea, based on the author's relation to that subarea.

Our second research question was: *"How does the reputation of Brazilian and US graduate programs in CS vary per subarea?"* We showed that the identification of the most reputable publication venues and graduate programs in Computer Science depends on the subarea considered to the task. Specifically, both the most suitable venues as the rankings of Brazilian and US graduate programs 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, how to rank graduate programs using this information, obtaining the top graduate programs of a given subarea.

Our third research question was: *"Are there differences between the current research directions in CS of the top Brazilian and US graduate programs?"* We showed that, in terms of reputation, the current research of the top graduate programs in Brazil and US differ considerably. It is known that the volume of US scientific publications in CS is significantly larger than the volume of Brazilian CS research. However, this work shows that the CS subareas in which each country has major scientific impact are basically disjoint. Moreover, the large distance between the reputation scores of Brazil and US programs emphasizes the focus of US researchers on publishing in the most reputable conferences and journals within their subareas.

For future work, we plan to use academic data of graduate programs from other regions of the world, such as Europe and Asia, and, thereupon, characterize the distribution of the most reputable graduate programs in CS worldwide, on a per subarea basis. Also, we expect to compare only more recent publications in the CS subareas (e.g., the last five years of scientific research), where we intend to observe the current interest of the CS community. Further, we intend to perform a more detailed temporal analysis of the evolution of the CS subareas communities over the years, looking for the shape of CS changes over time. 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.

Bibliography

Alves, B., Benevenuto, F., and Laender, A. (2013). The Role of Research Leaders on the Evolution of Scientific Communities. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 649–656.

Balog, K. (2012). Expertise Retrieval. *Foundations and Trends in Information Retrieval*, 6(2-3):127–256.

Benevenuto, F., Laender, A., and Alves, B. (2015). How Connected are the ACM SIG Communities? *SIGMOD Record*, 44(4):57–63.

Biryukov, M. and Dong, C. (2010). Analysis of Computer Science Communities Based on DBLP. In *Proceedings of the 14th European Conference on Research and Advanced Technology for Digital Libraries*, pages 228–235.

Bollen, J., van de Sompel, H., Smith, J., and Luce, R. (2005). Toward Alternative Metrics of Journal Impact: A Comparison of Download and Citation Data. *Information Processing & Management*, 41(6):1419–1440.

Brembs, B., Button, K., and Munafò, M. (2013). Deep Impact: Unintended Consequences of Journal Rank. *Frontiers in Human Neuroscience*, 7:291.

Brin, S. and Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117.

Castillo, C., Donato, D., and Gionis, A. (2007). Estimating Number of Citations Using Author Reputation. In *Proceedings of the 14th International Conference on String Processing and Information Retrieval*, pages 107–117.

Cormode, G., Muthukrishnan, S., and Yan, J. (2014). People Like Us: Mining Scholarly Data for Comparable Researchers. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 1227–1232.

Delgado-Garcia, J., Laender, A., and Meira, Jr., W. (2014). Analyzing the Coauthorship Networks of Latin American Computer Science Research Groups. In *Proceedings of the 9th Latin American Web Congress*.

Deng, H., Han, J., Lyu, M., and King, I. (2012). Modeling and Exploiting Heterogeneous Bibliographic Networks for Expertise Ranking. In *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 71–80.

Dijk, D., Manor, O., and Carey, L. (2014). Publication Metrics and Success on the Academic Job Market. *Current Biology*, 24(11):R516–R517.

Ding, Y. and Cronin, B. (2011). Popular and/or Prestigious? Measures of Scholarly Esteem. *Information Processing & Management*, 47(1):80–96.

Dong, Y., Johnson, R., and Chawla, N. (2015). Will this Paper Increase Your h-index? Scientific Impact Prediction. In *Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, pages 149–158.

Dongen, S. (2000). A Cluster Algorithm for Graphs. Technical report, Centre for Mathematics and Computer Science.

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

Garfield, E. (1955). Citation Indexes for Science. *Science*, 122(3159):108–111.

Glänzel, W., Schubert, A., Thijs, B., and Debackere, K. (2011). A Priori vs. A Posteriori Normalisation of Citation Indicators. The Case of Journal Ranking. *Scientometrics*, 87(2):415–424.

Gollapalli, S., Mitra, P., and Giles, C. (2011). Ranking Authors in Digital Libraries. In *Proceedings of the 11th International ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 251–254.

Gonçalves, G., Figueiredo, F., Almeida, J., and Gonçalves, M. (2014). Characterizing Scholar Popularity: A Case Study in the Computer Science Research Community. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 57–66.

Hirsch, J. (2005). 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.

Hoonlor, A., Szymanski, B., and Zaki, M. (2013). Trends in Computer Science Research. *Communications of the ACM*, 56(10):74–83.

Kapeller, J. (2010). Citation Metrics: Serious Drawbacks, Perverse Incentives, and Strategic Options for Heterodox Economics. *American Journal of Economics and Sociology*, 69(5):1376–1408.

Laender, A., Lucena, C., Maldonado, J., de Souza e Silva, E., and Ziviani, N. (2008). Assessing the Research and Education Quality of the Top Brazilian Computer Science Graduate Programs. *SIGCSE Bulletin*, 40(2):135–145.

Lee, F., Grijalva, T., and Nowell, C. (2010). 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):1345–1375.

Leydesdorff, L. (2009). How are New Citation-based Journal Indicators Adding to the Bibliometric Toolbox? *Journal of the Association for Information Science and Technology*, 60(7):1327–1336.

Leydesdorff, L., Zhou, P., and Bornmann, L. (2013). How Can Journal Impact Factors be Normalized Across Fields of Science? An Assessment in Terms of Percentile Ranks and Fractional Counts. *Journal of the American Society for Information Science and Technology*, 64(1):96–107.

Lima, H., Silva, T., Moro, M., Santos, R., Meira, Jr., W., and Laender, A. (2013). Aggregating Productivity Indices for Ranking Researchers Across Multiple Areas. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 97–106.

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

Lundberg, J. (2007). Lifting the Crown — Citation z-score. *Journal of Informetrics*, 1(2):145–154.

Mazloumian, A. (2012). Predicting Scholars' Scientific Impact. *PLOS ONE*, 7(11):1–5.

Menezes, G. V., Ziviani, N., Laender, A., and Almeida, V. (2009). A Geographical Analysis of Knowledge Production in Computer Science. In *Proceedings of the 18th International Conference on World Wide Web*, pages 1041–1050.

Navaux, P., Cáceres, E., and Zorzo, A. (2017). Documento de Área da Ciência da Computação. *Foundation for the Coordination and Improvement of Higher Level or Education Personnel (CAPES).*

Nelakuditi, S., Gray, C., and Choudhury, R. (2011). Snap Judgement of Publication Quality: How to Convince a Dean that You are a Good Researcher. *Mobile Computing and Communications Review*, 15(2):20–23.

Nezhadbiglari, M., Gonçalves, M., and Almeida, J. (2016). Early Prediction of Scholar Popularity. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, pages 181–190.

Opthof, T. and Leydesdorff, L. (2010). Caveats for the Journal and Field Normalizations in The CWTS ("Leiden") Evaluations of Research Performance. *Journal of Informetrics*, 4(3):423–430.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1998). The PageRank Citation Ranking: Bringing Order to the Web. In *Proceedings of the 7th International World Wide Web Conference*, pages 161–172.

Penner, O., Pan, R., Petersen, A., Kaski, K., and Fortunato, S. (2013). On the Predictability of Future Impact in Science. *Scientific Reports*, 3(3052).

Piwowar, H. (2013). Altmetrics: Value All Research Products. *Nature*, 493(7431):159–159.

Ribas, S. (2017). *Random Walks on the Reputation Graph.* Ph.D. Thesis, Federal University of Minas Gerais, Brazil.

Ribas, S., Ribeiro-Neto, B., de Souza e Silva, E., Ueda, A., and Ziviani, N. (2015a). Using Reference Groups to Assess Academic Productivity in Computer Science. In *Proceedings of the 24th International Conference on World Wide Web*, pages 603–608.

Ribas, S., Ribeiro-Neto, B., Santos, R., de Souza e Silva, E., Ueda, A., and Ziviani, N. (2015b). Random Walks on the Reputation Graph. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*, pages 181–190.

Saha, S., Saint, S., and Christakis, D. (2003). Impact Factor: A Valid Measure of Journal Quality? *Journal of the Medical Library Association*, 91(1):42–46.

Silva, T., Moro, M., Silva, A., Meira, Jr., W., and Laender, A. (2014). Community-based Endogamy As an Influence Indicator. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 67–76.

Souza e Silva, E. and Maldonado, J. (2009). Documento de Área da Ciência da Computação. *Foundation for the Coordination and Improvement of Higher Level or Education Personnel (CAPES).*

Sun, Y. and Giles, C. (2007). Popularity Weighted Ranking for Academic Digital Libraries. In *Proceedings of the 29th European Conference on IR Research*, pages 605–612.

Vardi, M. Y. (2016). Academic Rankings Considered Harmful! *Communications of the ACM*, 59(9):5.

Wainer, J., Eckmann, M., Goldenstein, S., and Rocha, A. (2013). How Productivity and Impact Differ Across Computer Science Subareas. *Communications of the ACM*, 56(8):67–73.

Wainer, J., Xavier, E. C., and Bezerra, F. (2009). Scientific Production in Computer Science: A Comparative Study of Brazil and Other Countries. *Scientometrics*, 81(2):535.

Walters, P. (2000). *An Introduction to Ergodic Theory*. Graduate Texts in Mathematics. Springer-Verlag New York.

Waltman, L. and Eck, N. (2013). Source Normalized Indicators of Citation Impact: An Overview of Different Approaches and an Empirical Comparison. *Scientometrics*, 96(3):699–716.

Waltman, L., Eck, N., Leeuwen, T., Visser, M., and Raan, A. (2011). Towards a New Crown Indicator: Some Theoretical Considerations. *Journal of Informetrics*, 5(1):37–47.

W.Martins, Gonçalves, M., Laender, A., and Pappa, G. (2009). Learning to Assess the Quality of Scientific Conferences: A Case Study in Computer Science. In *Proceedings of the 9th International ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 193–202.

Wu, H., Pei, Y., and Yu, J. (2009). Detecting Academic Experts by Topic-sensitive Link Analysis. *Frontiers of Computer Science in China*, 3(4):445–456.

Yan, E., Ding, Y., and Sugimoto, C. (2011). P-Rank: An Indicator Measuring Prestige in Heterogeneous Scholarly Networks. *Journal of the Association for Information Science and Technology*, 62(3):467–477.

Yan, S. and Lee, D. (2007). Toward Alternative Measures for Ranking Venues: A Case of Database Research Community. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 235–244.

Appendix A

Clustering Computer Science

Initially, we explored some clustering methods in order to help us characterize research communities. To do that, we built a graph of coauthorships, using the data from the DBLP dataset. In this graph, each node is a distinct researcher and each edge represents a co-work of the authors — an article written by the both authors. The higher the number of collaborations between the authors, the greater the weight of the edge connecting them. We use this weight as a measure of collaboration among the authors. This allows us to run clustering algorithms and analyze the resulting communities, as we now discuss.

A.1 Markov Cluster Algorithm

One of the main algorithms used in our initial studies, the Markov Cluster Algorithm (MCL) discussed in [Dongen, 2000] is a fast and scalable unsupervised cluster algorithm for graphs based on simulation of stochastic flow in graphs. It finds cluster structures in graphs by a mathematical bootstrapping procedure. The process deterministically computes (the probabilities of) random walks through the graph, and uses two operators transforming one set of probabilities into another. It does so using the language of stochastic matrices (also called Markov matrices) which capture the mathematical concept of random walks on a graph. The MCL algorithm simulates random walks within a graph by an alternation of two operators called expansion and inflation. The expansion coincides with taking the power of a stochastic matrix using the normal matrix product (i.e., matrix squaring). Inflation corresponds with taking the Hadamard power of a matrix (taking powers entrywise), followed by a scaling step, such that the resulting matrix is stochastic again, i.e., the matrix elements (on each column) correspond to probability values.

A.2 Preliminary Results

The first of our preliminary experiments has been the application of the MCL algorithm to the subarea of Computer Networks. We did so by taking Infocom as a single source of reputation. We then selected the most productive authors (in number of papers published) of Infocom. Our goal in this step was to take a small set of researchers that represents the academic production of the venue Infocom, which we will call *reference set* . But, as we were interested in studying the whole subarea of Computer Networks instead of a single venue, we added to the reference set all the coauthors of the researchers from the original set. A coauthor of author a is any researcher who has at least one published work together with a.

Using the researchers as nodes and the coauthorships as edges, we then obtained a graph of coauthorships. The weights of the edges in this graph are the number of papers two researchers have published together. In order to generate clusters in the graph of coauthorships, we used the number of papers which two researchers published together as a metric of similarity (or proximity) between these two researchers. We then ran clustering methods over this graph of coauthorships (e.g., the MCL algorithm, discussed in next session) and could also analyze the resulting academic communities with graph visualization tools. In our preliminary experiments we used the software Gephi[1] to visualize the graphs of coauthorships.

The result of this initial experiment is illustrated in Figure A.1. It shows that the clusters produced are good indicators of productive authors in Computer Networks.

Additionally, in Figures A.2, A.3, A.4, and A.5 we show other experiments generated with the same approach described above but changing the initial single venue as input — respectively, Transaction on Networks (TON), Computer Networks (CN), SIGIR, and Web Search and Data Mining (WSDM). The venues TON and CN are considered as publication venues of the subarea of Computer Networks while SIGIR and WSDM are considered venues of the subarea of Information Retrieval.

A.3 Next Steps

The use of clustering methods was an initial step to identify research communities and also reputable sources (main authors) in a subarea. Further directions for research include using clustering results to improve rankings of publication venues, authors, and graduate programs in a given subarea.

[1] `http://gephi.org/`

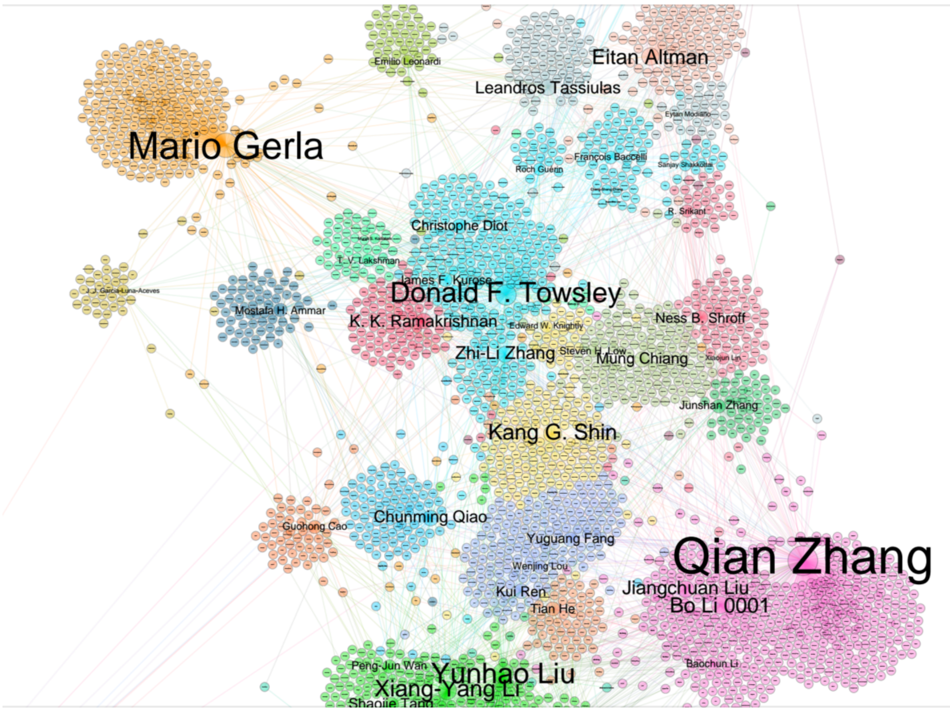

Figure A.1. Graph of couauthorships with researchers from the subarea of Computer Networks, using the venue Infocom as source of reputation. This visualization was generated by Gephi, an open-source framework for manipulating graphs.

Examples of research questions which could guide future work are (i) how to select the most representative authors of a subarea based only on the relationships between academic entities and (ii) which features provided by clustering are the most discriminative in reputation assessments in academic domain (e.g., most centered clusters, largest clusters).

These studies may lead to better characterizations of subareas of CS in Brazil and further comparisons with other countries. Also, one can be interested in applying these ideas to the problem of finding good references of reputation in any topic of study (e.g., “land policies”, “infectious diseases”, “deep learning”). In other words, to use clustering-based methods for general expert search.

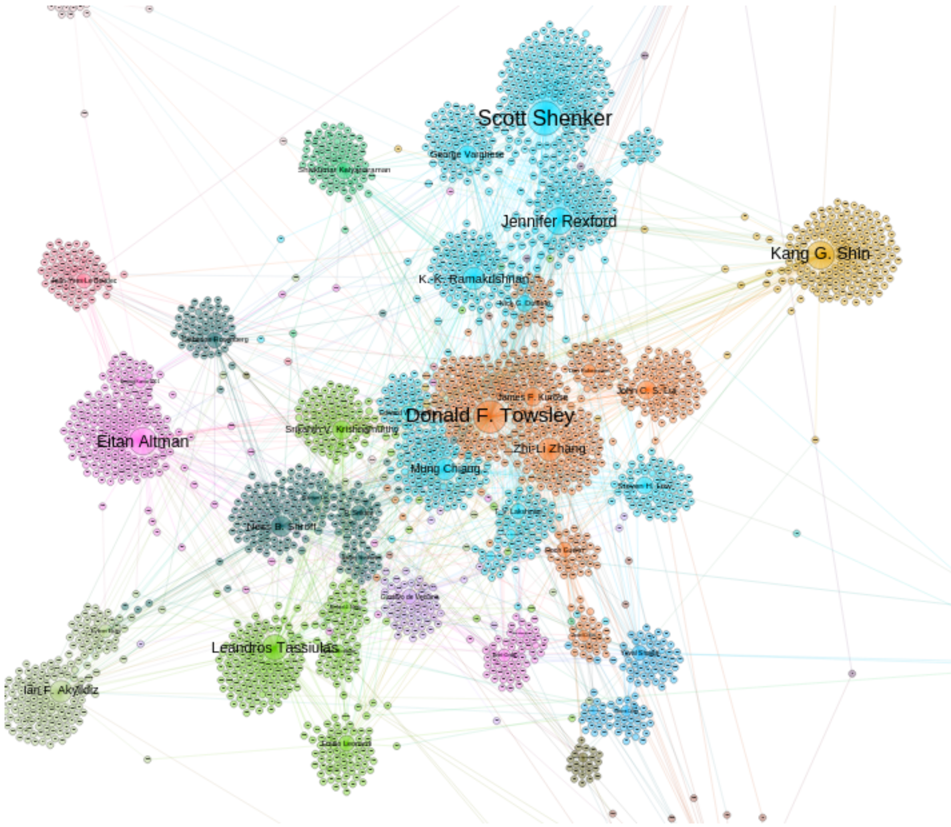

Figure A.2. Graph of couauthorships with researchers from the subarea of Computer Networks, using the venue TON as source of reputation.

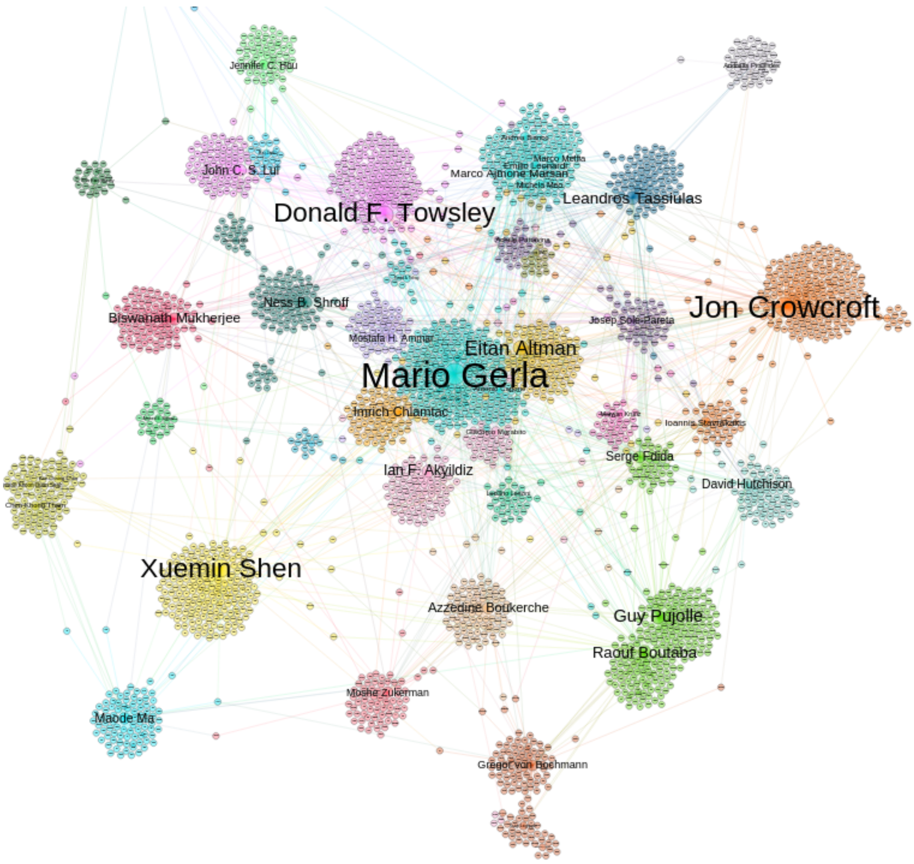

Figure A.3. Graph of couauthorships with researchers from the subarea of Computer Networks, using the venue Computer Networks as source of reputation.

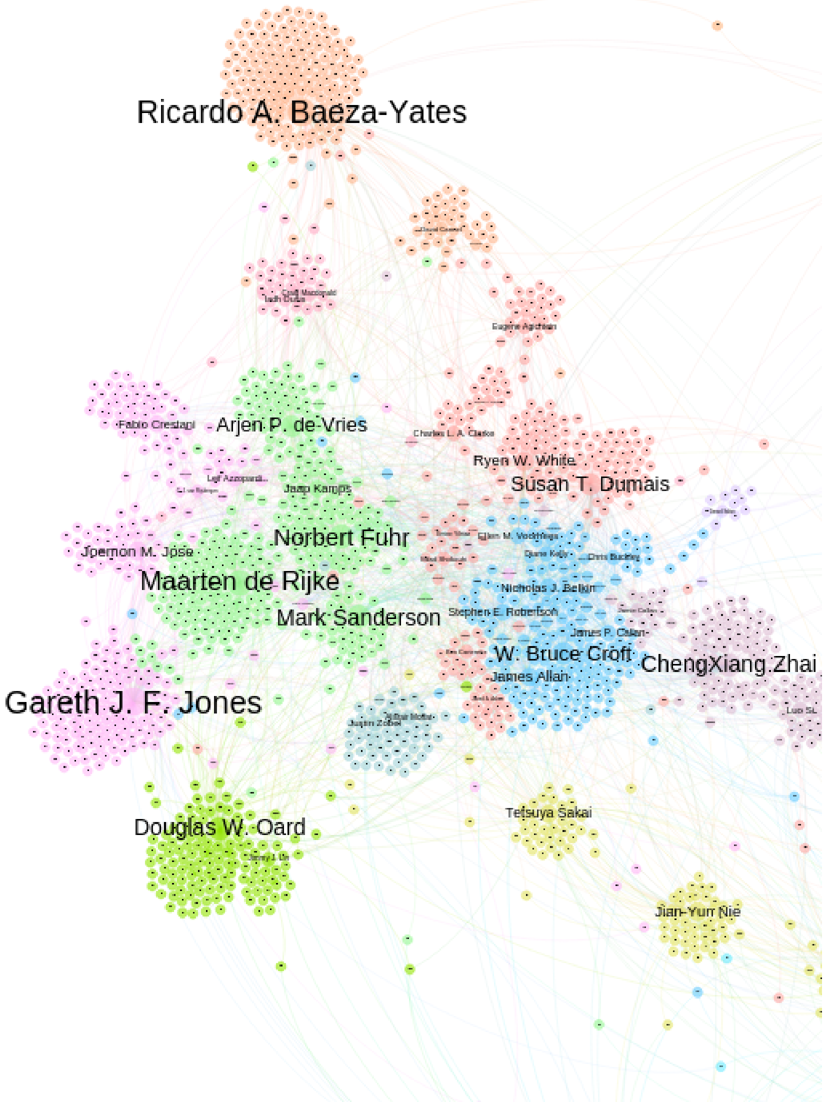

Figure A.4. Graph of couauthorships with researchers from the subarea of Information Retrieval, using the venue SIGIR as source of reputation.

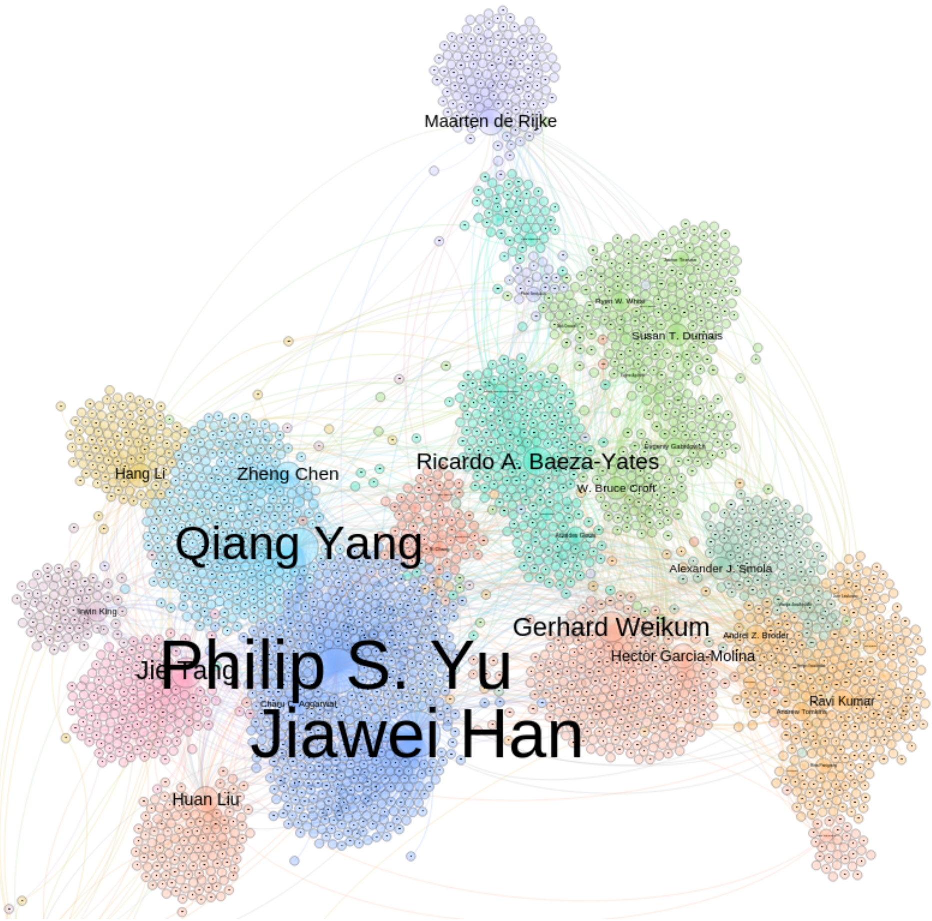

Figure A.5. Graph of couauthorships with researchers from the subarea of Information Retrieval, using the venue WSDM as source of reputation.

Appendix B

CS Subareas According to Different Sources

ACM SIG's (2016)	IEEE (2016)
SIGACCESS - Accessible Computing	Business Informatics and Systems (TCBIS)
SIGACT - Algorithms and Computation Theory	Computational LIfe Science (TCCLS)
SIGAda - Ada Programming Language	Computer Architecture (TCCA)
SIGAI - Artificial Intelligence	Computer Communications (TCCC)
SIGAPP - Applied Computing	Data Engineering (TCDE)
SIGARCH - Computer Architecture	Dependable Comp. and Fault Tolerance (TCFT)
SIGBED - Embedded Systems	Digital Libraries (TCDL)
SIGBio - Bioinformatics	Distributed Processing (TCDP)
SIGCAS - Computers and Society	Intelligent Informatics (TCII)
SIGCHI - Computer-Human Interaction	Internet (TCI)
SIGCOMM - Data Communication	Learning Technology (TCLT)
SIGCSE - Computer Science Education	Mathematical Foundations of Computing (TCMF)
SIGDA - Design Automation	Microprocessors and Microcomputers (TCMM)
SIGDOC - Design of Communication	Microprogramming and Microarch. (TCuARCH)
SIGecom - Electronic Commerce	Multimedia Computing (TCMC)
SIGEVO - Genetic and Evolutionary Comp.	Multiple-Valued Logic (TCMVL)
SIGGRAPH - Comp. Graph. and Interact. Tech.	Parallel Processing (TCPP)
SIGHPC - High Performance Computing	Pattern Analysis and Mach. Intellig.(TCPAMI)
SIGIR - Information Retrieval	Real-Time Systems (TCRTS)
SIGITE - Information Technology Education	Scalable Computing (TCSC)
SIGKDD - Knowledge Discovery in Data	Security and Privacy (TCSP)
SIGLOG - Logic and Computation	Semantic Computing (TCSEM)
SIGMETRICS - Measurement and Evaluation	Services Computing (TCSVC)
SIGMICRO - Microarchitecture	Simulation (TCSIM)
SIGMIS - Management Information Systems	Software Engineering (TCSE)
SIGMM - Multimedia	Test Technology (TTTC)
SIGMOBILE - Mobility of Syst., Data and Comp.	Visualization and Graphics (VGTC)
SIGMOD - Management of Data	VLSI (TCVLSI)
SIGOPS - Operating Systems	
SIGPLAN - Programming Languages	
SIGSAC - Security, Audit and Control	
SIGSAM - Symbolic and Algebraic Manipulation	
SIGSIM - Simulation and Modeling	
SIGSOFT - Software Engineering	
SIGSPATIAL - SIGSPATIAL	
SIGUCCS - Univ. and College Comp. Services	
SIGWEB - Hypertext and the Web	

SBC (2016)	Wainer (2013)
Arquiteura de Comp. e Proces. de Alto Desempenho	Artificial Intelligence
Banco de Dados	Bioinformatics
Biologia Computacional	Communications and Networking
Informática na Educação	Compilers and Programming Languages
Métodos Formais	Computer Architecture
Algoritmos, Combinatória e Otimização	Computer Graphics
Computação Aplicada à Saúde	Databases
Computação Gráfica e Processamento de Imagens	Distributed Computing
Computação Musical	Human-Computer Interaction
Concepção de Circuitos e Sistemas Integrados	Image Processing and Computer Vision
Engenharia de Sistemas Computacionais	Machine Learning
Engenharia de Software	Management Information Systems
Geoinformática	Multimedia
Inteligência Artificial	Operational Research and Optimization
Inteligência Computacional	Security
Interação Humano-Computador	Software Engineering
Jogos e Entretenimento Digital	Theory
Linguagens de Programação	
Processamento de Linguagem Natural	
Realidade Virtual	
Redes de Computadores e Sistemas Distribuídos	
Robótica	
Segurança da Informação e de Sistemas Computacionais	
Sistemas Colaborativos	
Sistemas de Informação	
Sistemas Multimedia e Web	
Sistemas Tolerantes a Falhas	

Appendix C

List of Abbreviations for Publication Venues

Abbreviation	Name
ADCS	Australasian Document Computing Symposium
AI	Artificial Intelligence
AIRS	Asia Information Retrieval Symposium
ALGORITHMICA	Algorithmica
BIBM	International Conference on Bioinformatics and Biomedicine
BIOINFORMATICS	Bioinformatics
CCS	Conference on Computer and Communications Security
CHI	Conference on Human Factors in Computing Systems
CIDR	Biennial Conference on Innovative Data Systems Research
CIKM	Conference on Information and Knowledge Management
CLEF	Cross-Language Evaluation Forum Workshop
COLING	Conference on Computational Linguistics
COMAD	Conference on Management of Data
CVPR	Computer Vision and Pattern Recognition
DASFAA	Conference on Database Systems for Advanced Applications
DATAMINE	Data Mining and Knowledge Discovery
DEBU	Data Engineering Bulletin
DPD	Distributed and Parallel Databases
ECIR	European Conference on Information Retrieval
EDBT	Conference on Extending Database Technology
EMNLP	Conference on Empirical Methods in Natural Language Processing
ICDCS	International Conference on Distributed Computing Systems
ICDE	International Conference on Data Engineering
ICDM	International Conference on Data Mining
ICDT	Conference on Database Theory
ICML	Conference on Machine Learning
ICTIR	Conference on the Theory of Information Retrieval
IIIX	Conference on Information Interaction in Context
IJCAI	Joint Conference on Artificial Intelligence
IJCV	Journal of Computer Vision
INEX	Initiative for the Evaluation of XML Retrieval
INFOCOM	INFOCOM
INTERSPEECH	Conference of the International Speech Communication Association
IPPS	International Parallel and Distributed Processing Symposium
IR	Information Retrieval Journal

Abbreviation	Name
JMLR	Journal of Machine Learning Research
KAIS	Knowledge and Information Systems
KDD	Knowledge Discovery and Data Mining
LA-WEB	Latin American Web Congress
NTCIR	NII Testbeds and Community for Information Access Research
PAKDD	Pacific-Asia Conference on Knowledge Discovery and Data Mining
PKDD	European Conference on Principles of Data Mining and Knowledge Discovery
PLDI	Symposium on Programming Language Design and Implementation
PODS	Symposium on Principles of Database Systems
PVLDB	Proceedings of the VLDB Endowment
RECSYS	Conference on Recommender Systems
RIAO	Open research Areas in Information Retrieval
SADM	Statistical Analysis and Data Mining
SDM	SIAM International Conference on Data Mining
SIAMCOMP	SIAM Journal on Computing
SIGGRAPH	Conference on Computer Graphics and Interactive Techniques
SIGIR (c)	International Conference on Research and Development in Information Retrieval
SIGIR (j)	SIGIR Forum
SIGKDD	SIGKDD Explorations
SIGMOD (c)	International Conference on Management of Data
SIGMOD (j)	SIGMOD Record
SIGOPS	Operating Systems Review
SODA	Symposium on Discrete Algorithms
SOSP	Symposium on Operating Systems Principles
SPIRE	Symposium on String Processing and Information Retrieval
SSD	Symposium on Spatial and Temporal Databases
SSDBM	Conference on Statistical and Scientific Database Management
STOC	Symposium on the Theory of Computing
TCOM	Transactions on Communications
TISSEC	Transactions on Information and System Security
TIST	Transactions on Intelligent Systems and Technology
TKDD	Transactions on Knowledge Discovery from Data
TKDE	Transactions on Knowledge and Data Engineering
TOCHI	Transactions on Computer-Human Interaction
TODS	Transactions on Database Systems
TOIS	Transactions on Information Systems
TON	Transactions on Networking
TOPLAS	Transactions on Programming Languages and Systems
TPDS	Transactions on Parallel and Distributed Systems
TREC	Text Retrieval Conference
TWEB	Transactions on the Web
VLDB (c)	International Conference on Very Large Data Bases
VLDB (j)	International Journal on Very Large Data Bases
WEBDB	Workshop on the Web and Databases
WS	Journal of Web Semantics
WSDM	Web Search and Data Mining
WWW	International World Wide Web Conference