RANDOM WALKS ON THE REPUTATION GRAPH

SABIR RIBAS

RANDOM WALKS ON THE REPUTATION GRAPH

Thesis presented to the Graduate Program in Computer Science of the Universidade Federal de Minas Gerais — Departamento de Ciência da Computação in partial fulfillment of the requirements for the degree of Doctor in Computer Science.

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

Belo Horizonte
April 2017

Ficha catalográfica elaborada pela Biblioteca do ICEx - UFMG

Ribas, Sabir

S161r Random Walks on the Reputation Graph / Sabir Ribas. — Belo Horizonte, 2017.
xv, 111 f. : il. ; 29cm.

Tese (doutorado) — Universidade Federal de Minas Gerais — Departamento de Ciência da Computação.

Orientador: Berthier Ribeiro-Neto
Coorientador: Nivio Ziviani

1. Computação – Teses. 2. Recuperação da Informação. 3. Passeio Aleatório. 4. Sistemas de Recomendação. 5. Academic Search. 6. Reputation Flows. 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

Random walks on the reputation graph

SABIR RIBAS

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

PROF. EDMUNDO ALBUQUERQUE SOUZA E SILVA
COPPE - UFRJ

PROF. RODRYGO LUIS TEODORO SANTOS
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 06 de abril de 2017.

To Cristiene, Róvia and Sávio.

Acknowledgments

I have a dream team, and I wish I were a poet to find the best words to express how grateful I am. During my Ph.D. I was surrounded by the most reputable people in my knowledge area. I learned a lot - not only about research and academia but also about life. Berthier, the most interesting and magnetic person I have ever met, thanks for answering my first email and go all the way long till the end of my Ph.D. showing absolute confidence in this work. Thanks for all the lessons, the technical and the personal ones. Nivio, always challenging people to do their best, thanks for receiving me on your laboratory and for all support you gave me during this journey. Also, thanks for helping me build layers of resilience, which are quite important for me now. Edmundo, always positive believing in this work, thanks for your passion and unconditional dedication to this research. Also, thanks for everything you taught me about stochastic processes, Markov Chains, and Random Walks. Rodrygo, one of the most productive researchers I have ever worked with, you make publishing papers on top venues looks easy. Thanks for transforming our drafts into gold and for all soccer matches on Wednesday nights. Altigran, always asking challenging questions, thanks for offering the most clever solutions for your own challenging questions after making me think for a while. Also, thanks for your trust in this work and all your insights.

I also have a dream team in my personal life. How lucky I am! Cristiene, my wife, my other half, a person to die for, you are the most enlighted person I have ever met. Thanks for being a symbol of intelligence, productivity, strength, and love. Without you, this thesis would not happen. Thanks for everything. There are no limits on what we can do together. Rovadávia, my mother, you have the most beautiful heart I have ever seen. I know you had to pause your dreams to take care of my brother and me. I am proud of you because you had the strength to come back to your dreams and become a university professor after taking care of ours. I consider myself blessed to have you as my mother. Antônio, my father, always proud of our achievements and spreading the news about us. Thanks for my name. No matter where I go, people always find it quite interesting, and I always say that it was you. Then, people from all around the

world know how creative you are. I know you made everything you could to take care of us. Thanks for everything, may God bless you in heaven. Sávio, my brother, funny guy, clever man, thanks for showing me that there is no limit on what we can achieve if we love what we do and if we are committed to reaching the next level of excellence. Your award-winning mathematical thinking influences many people around you. I am proud to be your brother. Who else can say that received a medal from the president of the country? Also, thanks for finishing your Ph.D. precisely on the same day and time as me, and not before. Otherwise, I would have to explain to people why my younger brother started later and finished before. Karine, thanks for being a good friend and supporting Sávio in his journey. It is great having you in the family. Lili, thanks for always helping us carefully and teach us math. Reginaldo, my journey in computer science started with you saying to me when I was a child: "Hey Sabir, computers are the future. You should check it.". That triggered my curiosity about the field. Samuel and Alberto, thanks for being close friends and for all the discussions we had about life. Thanks, Cristiene's family, Eunice, Adélia, Ana, Alexandre, Ricardo, and Val, for your support and confidence. People I have worked with during my Ph.D., especially Ivan, Douglas, Leonardo, and Marlon. People from Latin, Cristiano, Aécio, Thales, Jordan, Bruno, Rafael, Felipe, Arthur, Itamar, and Anisio. Thanks to all friends I made on my soccer teams at DCC/UFMG, Meia Boca Jr, Bier Leverpilsen, and River Template. My friends from SEEK, Emanuel, Thyago, Eder, Fernando, Diego, Tiago, Andryw, Hélio, Marcus, Bruno, Luís, Sampaio, Rafael, Vinicius, Takeo. Genivaldo, thanks for introducing me to the academic world. When I was at high school, you received me as a student and became my advisor on that program for young scientists in Physics. Despite choosing another area, I am still a scientist, triggered by you. Marcone, the most dedicated person I have ever met, thanks for all your lessons. You taught me how to write academic papers. With you, I started to go to conferences around the country, which opened many doors for me. Thanks for your dedication to work, to teaching, and to help people. I know many people that were positively impacted by crossing lines with you. Satoru, a beautiful soul who has the precious ability to create a great team environment and putting people to work together towards the same goal. Thanks for receiving me at the IC/UFF and for understanding my choices. Thanks, friends from OptHouse, Igor, Pablo, Mário and Matheus. Thanks, PPGCC/UFMG staff, especially Linda, Sônia, Sheila, and Adriana. Thanks for your dedication in making sure people always get what they need with no roadblocks. Thanks, DECOM/UFOP professors Lucília, Haroldo, Eduardo, Elton, Fernanda, Tiago, David, Túlio, and José Maria, for receiving me as a lecturer. Thank you all friends I have not visited for a long time, for understanding my absence. How lucky I am by having you all around!

Abstract

The identification of reputable entities is an important task in business, education, and in many other fields. In general, the reputation of an entity reflects its public perception, which touches upon a variety of aspects that may impact the identity of the entity, such as its prowess, integrity, and trustworthiness. Indeed, more reputable entities are presumably a better fit for most purposes. Thus, while reputation is a widespread notion in society, it is albeit an arguably ill-defined one. As a consequence, quantifying reputation is challenging. Indeed, existing attempts to quantify reputation rely on either manual assessments or on a restrictive definition of reputation.

In this thesis, instead of relying on a single and precise definition of reputation, we propose to exploit the *transference* of reputation among entities in order to identify the most reputable ones. To this end, we introduce a conceptual framework of reputation flows and propose a metric based on it, which we call *P-score*. This framework consists of a random walk model that allows inferring the reputation of a target set of entities with respect to suitable sources of reputation. By using it, we can better understand how reputation flows between distinct entities in a reputation graph.

We instantiate our model in an academic search setting to address three common ranking tasks namely, research group ranking, author ranking, and publication venue ranking. By relying on publishing behavior as a reputation signal, we demonstrate the effectiveness of our model in contrast to standard citation-based approaches for identifying reputable venues, authors, and research groups in the broad area of Computer Science. In addition, we demonstrate the robustness of our model to perturbations in the selection of reputation sources. Finally, we show that effective reputation sources can be chosen via the proposed model itself in a fully automatic fashion.

Contents

Chapter 1

Introduction

Reputation is a widespread notion in society, albeit an arguably ill-defined one. In general, the reputation of an entity reflects the public perception about this entity developed over time. This public perception may be either good or bad, and touches a variety of aspects that may impact the identity of the entity before the public, such as its prowess, integrity, and trustworthiness. Moreover, the reputation of an entity can change rapidly following an event in which the entity is involved, by means of word-of-mouth dissemination—whether traditional or electronic. As a result, reputation management is actively pursued by public relations departments of corporations and institutes. Further, it is a topic of continuous interest of online communities, such as question-answering forums and online marketplaces [Hutton et al., 2001].

The identification of reputable entities is an important task in many fields. Indeed, more reputable entities are presumably a better fit for most purposes. However, the subjective nature of reputation makes its quantification—and hence the identification of reputable entities—challenging. As a result, existing attempts to quantify the reputation of an entity rely on either manual assessments or on a restrictive definition of reputation, e.g., in terms of authority [Kleinberg, 1999; Page et al., 1998], influence [Bakshy et al., 2012], or expertise [Balog, 2012]. In contrast, in this thesis, we take an agnostic view of reputation. In particular, instead of relying on a single, precise definition of reputation, we propose to exploit the *transference* of reputation among entities in order to identify the most reputable ones.

1.1 Thesis Statement

The statement of this thesis is that effective rankings of entities can be attained by explicitly representing the transference of reputation among them. Modelling the transference of reputation, instead of assigning to it a single and precise definition, may offer additional standpoints when compared to existing metrics over an application scenario. In particular, by investigating how reputation flows from one entity to another, one can grasp interesting insights regarding the relative importance of each one. To illustrate, by adopting highly reputable entities as sources of reputation, it is possible to infer or reason about the relative importance of directly or indirectly related entities.

1.2 Thesis Contributions

The key contributions of this thesis can be summarized as follows:

A Conceptual Framework of Reputation Flows. We propose a conceptual framework of reputation flows, which consists of a novel random walk model for ranking entities according to the reputation collectively transferred to them from a set of reputation sources. In this framework, reputation flows through the nodes of a special graph, so-called the reputation graph, which has three types of nodes: the reputation sources, the reputation targets, and the reputation collaterals. This special graph allows us to model flows of reputation among distinct entity types, which we use to compute the P-score metric for ranking entities in a reputation graph.

Reputation Flows in Academia. We instantiate our conceptual framework of reputation flows in the academic search setting. To evaluate our instantiation, we perform an empirical validation on the effectiveness and robustness of the model by applying it to three academic search tasks, namely: venue ranking, author ranking, and research group ranking. Our results suggest that our reputation-based metric P-score can indeed be used as an alternative method to produce academic rankings. Indeed, its experimental results led to more effective rankings than those produced using classic citation-based metrics.

Automatic Selection of Reputation Sources. We characterize the impact of choosing distinct sets of reputation sources in the academic search setting and use the acquired knowledge to investigate the suitability of automatically choosing effective sets of reputation sources. In particular, we discuss a variety of methods to perform such choices.

One of them uses the P-score itself through a randomized process, and leads to results that closely resemble those produced by human experts.

1.3 Publications

Most of the material presented in this thesis appears in the following publications:

1. Sabir Ribas, Berthier Ribeiro-Neto, Rodrygo L.T. Santos, Edmundo de Souza e Silva, Alberto Ueda, Nivio Ziviani, Marlon Dias. "Reputation flows in academia." *Journal of American Society for Information Science and Technology, submitted.* Feb, 2017.

2. Sabir Ribas, Alberto Ueda, Rodrygo L.T. Santos, Berthier Ribeiro-Neto, Nivio Ziviani. "Simplified Relative Citation Ratio for Static Paper Ranking: UFMG/LATIN at WSDM Cup 2016." *ACM International Conference on Web Search and Data Mining, WSDM Cup.* San Francisco, USA. Feb 22-25, 2016.

3. Sabir Ribas, Berthier Ribeiro-Neto, Rodrygo L.T. Santos, Edmundo de Souza e Silva, Alberto Ueda, Nivio Ziviani. "Random walks on the reputation graph." *ACM International Conference on the Theory of Information Retrieval.* Northampton, Massachusetts, USA. Sep 27-30, 2015.

4. Sabir Ribas, Berthier Ribeiro-Neto, Edmundo de Souza e Silva, Alberto Ueda, Nivio Ziviani. "Using reference groups to assess academic productivity in computer science." *Proceedings of the International World Wide Web Conference, BigScholar.* Florence, Italy. May 18-22, 2015.

5. Sabir Ribas, Berthier A. Ribeiro-Neto, Edmundo de Souza e Silva, Alberto Ueda, Nivio Ziviani. "P-score: A Publication-based Metric for Academic Productivity." *Technical Report: CoRR, Vol. abs/1503.07496.* Mar, 2015.

6. Sabir Ribas, Berthier A. Ribeiro-Neto, Edmundo de Souza e Silva, Nivio Ziviani. "R-Score: Reputation-based Scoring of Research Groups." *Technical Report: CoRR, Vol. abs/1308.5286.* Aug, 2013.

Among the aforementioned publications, the only work that does not focus on the P-score metric is the one that describes our participation in the WSDM Cup 2016. In that work, we have proposed the S-RCR metric to rank academic papers, see Appendix G. The interesting point here is that a single well-designed feature was able to produce effective results, promoting our team to the 3rd place of the competition.

1.4 Thesis Outline

The remainder of this thesis is organized as follows:

- Chapter 2 provides a background and a related work discussion. It starts by presenting a background material on random walks, from the basic concepts behind random walks to Markov Chains and PageRank. Next, it discusses core related works on random walks, on academic search, as well as on the use of random walks to address academic search tasks.

- Chapter 3 introduces our proposed random walk model for reputation-oriented ranking i.e., our conceptual framework of reputation flows. In this chapter, we model the flow of reputation through a stochastic process. We start by defining the structure of the so-called *reputation graph*, which has three types of nodes: the reputation sources, the reputation targets, and the reputation collaterals. We then formalize how we model the interaction between nodes of distinct types in a reputation graph and define the P-score metric as a means to rank these nodes. Finally, we discuss how to select the set of entities to be used as sources of reputation to rank other entities in a reputation graph.

- Chapter 4 discusses possible instantiations of our conceptual framework of *reputation flows* in academic search settings. We start by showing how it can be used to model the transference of reputation between authors and papers, as well as between research groups and publication venues. The argument here is that the relations between these scientific entities may be captured through distinct metrics and, as far as we know, the most important ones (including citation-based metrics) fit well in our conceptual framework. Next, we highlight our assumptions when instantiating the conceptual framework in an academic search setting. An example is used to facilitate the understanding. Finally, we provide suggestions on how to instantiate the reputation graph to solve three common academic search tasks, as detailed in our experiments.

- Chapter 5 describes the experimental setup that supports the empirical evaluation of our proposed model for reputation-oriented ranking. The discussion starts by specifying three research questions regarding the effectiveness and robustness of the proposed model, as well as a research question regarding the suitability of automatically choosing entities to compose the set of reputation sources. Then, we discuss our dataset and present the ground-truths, a dataset description, and the baselines. Finally, we conclude the definition of our evaluation procedure.

- Chapter 6 validates the effectiveness of the conceptual framework of reputation flows to address three academic search tasks, namely: ranking research groups, ranking individual researchers, and ranking publication venues. For each task, we compute P-scores using research groups as reputation sources, authors as reputation sources, and venues as reputation sources. That is, all configurations suggested in Chapter 4 are explored. We compare distinct instantiations of P-score against classic citation based metrics and we provide evidences regarding the effectiveness of distinct instantiations of P-score to address these three ranking tasks. These evidences allow us to verify the effectiveness of the proposed model to address distinct ranking tasks.

- Chapter 7 discusses a critical step of our method, the selection of the reputation sources. This chapter aims to address the robustness of our proposal with regard to the of reputation sources, and on the suitability of automatically choosing reputation sources. We start by discussing manual and automatic approaches to choose entities to compose the set of reputation sources, and whether the manual approach should be considered whether automatic approaches are desirable. We also explore distinct sets of reputation sources by providing a characterization of these sets, by investigating intuitive sets, and by measuring the impact of their sizes. Finally, we analyse the coverage and the robustness of P-scores when used with distinct reputation sources.

- Chapter 8 closes this thesis by providing a summary of the contributions and the conclusions obtained throughout the chapters. Finally, it presents some future directions regarding the conceptual framework of reputation flows and its instantiation in the academic search setting.

Chapter 2

Background and Related Work

In this chapter, we discuss background and related work. Our objective here is to provide a contextual overview of the theoretical basis of our method and also regarding applications of the notion of reputation and various attempts to quantify it.

2.1 Background

Link Analysis techniques explore associations between objects to model or explain a network behavior. It comprises a set of techniques of network analysis for studing graphs built to represent relations between objects. While graph theorists are interested in arbitrary questions about graphs, network theorists are more interested in questions that are relevant to situations modeled by graphs (as in this work). Some well-known link analysis algorithms include PageRank [Page et al., 1998] and HITS [Kleinberg, 1999], discussed further in this chapter.

In the immediately following, we provide a brief discussion of random walks and Markov chains, techniques that offer the theoretical basis to our proposed method.

2.1.1 Random Walks

The term random walk was introduced by Karl Pearson, an English mathematician and biostatistician. In 1905, Pearson stated the random walk problem and appealed to the readers of Nature for a solution, as follows.

> *The Random Walk Problem.* A man starts from a point O and walks l yards in a straight line; he then turns through any angle whatever and walks another l yards in a second straight line. He repeats this process n times. I require the probability that after these n stretches he is at a distance between r and $r + dr$ from his starting point, O. (Pearson [1905])

Since then, the term random walk has been used to refer, in short, to a mathematical formalization of a path that consists of a succession of random steps. Nowadays, the applications of random walks are diverse and many types of them are of interest.

While in the original statement of the random walk problem the *man* walks a fixed distance (l yards) at each step, there is a different type of random walk where the step size is variable, not fixed. In this case, the distance walked in each step can be a function or even a random step size. Another variant considers that the time between each step is not discrete, but continuous. In continuous-time random walks, the next step is performed at a random amount of time after finishing the current step, where this *random* value may follow (or be sampled from) a given distribution, e.g. the exponential distribution. Also, some random walks, like the original proposal, consider that an object moves with equal probability in any direction, but not all random walks follow this rule, some of them are *biased*. A *biased random walk* is a random walk that is biased in some direction. One way to have a biased random walk is to consider that instead of an equal probability of choosing the next direction to move, there is a higher probability of moving to a given direction. Another way is to retain equal probabilities of choosing the next direction, but whenever the object moves to the west, it moves two units, and when it moves to the east, it only moves one unit.

All those aforementioned random walk variants are useful to model specific scenarios and are not limited to only objects in a 2-dimensional space (as Pearson asked for). They can also be used to model particles in 3-dimensional space (think about gas particles bouncing around) or to model a network behaviour using graphs. Indeed, one of the most important/famous applications of random walk nowadays is the PageRank algorithm, which has a huge impact in Web search, as discussed further.

The method we propose in this thesis, the P-score metric, consists on a random walk on a graph which has a special structure, and which we call the *reputation graph*. It can also be defined as a Markov Chain, as we now discuss.

2.1.2 Markov Chains

In 1913, the Russian mathematician Andrei Markov founded a new branch of probability theory by applying mathematics to poetry [Markov, 2006] through the investigation of the patterns of consonants and vowels. While his analysis did not alter the understanding of the poem, the technique he developed extended the theory of probability in a new direction. His methodology went beyond coin-flipping and dice-rolling, where each event is independent of all others, to chains of linked events, where what happens next depends on the current state of the system [Hayes et al., 2013].

A Markov Chain is defined as a stochastic process that satisfies the Markov property, denoted as "memoryless", and a process satisfies the Markov property if one can make predictions for the future of a process based exclusively on its present state. That is, the probabilities must depend only on the present state of the system, not on its earlier history. Also, to be considered a proper Markov Chain, a system must have a set of distinct states with identifiable transitions between them.

To illustrate, a graph made up of nodes and directed edges shows the structure of a Markov Chain. Nodes represent states and edges indicate transitions. Each edge has an associated number, which gives the probability of that transition. Given that these numbers are probabilities, they must lie between 0 and 1, and all the probabilities emanating from a node must add up to exactly 1.

For instance, consider a three-state Markov Chain. Its transition probabilities can be arranged in a three-by-three matrix. Notice that we can trace a pathway that defines a sequence of states in this graph. The probability of a specific sequence can be obtained by multiplying the probabilities associated with the corresponding transition edges, which makes the process for computing multi-stage transitions equivalent to a matrix multiplication. Concretely, let's consider a matrix P representing a Markov Chain in the context of weather forecast — possible states are: sunny, rainy, and cloudy. The matrix P itself predicts tomorrow's weather, while P^2 gives weather probabilities for the day after tomorrow, P^3 defines the probabilities for three days hence, and so on. The future of the system unfolds from this one matrix. If the chain is ergodic, the successive powers of the matrix rapidly converge to a stationary configuration in which all the rows are identical and all the columns consist of a single repeated value. The interpretation of this outcome is that if you let the system evolve long enough, the probability of a given state no longer depends on the initial state.

The method we propose in this thesis, the P-score metric, can be defined as a Markov Chain where the states and transitions are represented by the reputation graph. Also, in this case our interest is on the steady state probability distribution.

2.2 Related Work

In this chapter, we review the related literature on ranking based on random walks, as well as approaches devoted to generating rankings in an academic search setting.

2.2.1 Ranking with Random Walks

Page et al. [1998] designed the PageRank algorithm to calculate the importance of pages on the Web. PageRank simulates a web surfer's behavior. In particular, with probability $p < 1$, the surfer randomly chooses one of the hyperlinks of the current page and jumps to the page it links to; otherwise, with probability $1 - p$, the user jumps to a web page chosen uniformly at random from the collection. This defines a Markov chain on the web graph, where each probability of the stationary distribution corresponds to the rank of a web page, referred to as its *pagerank*.

Kleinberg [1999] divided the notion of "importance" of a web page into two related attributes: *hub*, measured by the authority score of other pages that the page links to, and *authority*, measured by the hub score of the pages that link to the page. These attributes are calculated in his Hyperlinked-Induced Topic Search (HITS). Both algorithms, PageRank and HITS, have been successfully applied to rank the importance of different web pages through analyzing the link structure of the web graph.

Extensions of the random walk model were also studied for scoring several types of objects—e.g., products, people and organizations—in different applications. For instance, Nie et al. [2005] presented PopRank, a domain-independent object-level link analysis model to rank objects within a specific domain, by assigning a popularity propagation factor to each type of object relationship. Different popularity propagation factors for these heterogeneous relationships were assessed with respect to their impact on the global popularity ranking. Xi et al. [2004] proposed a unified link analysis framework, called Link Fusion, which considers two different categories of links: intra-type links, which represent the relationship of data objects of a homogeneous data type (e.g., web pages), and inter-type links, which represent the relationship of data objects of different data types (e.g., between users and web pages). Regarding the recommendation of generic types of object, Jamali and Ester [2009] proposed TrustWalker, a random walk method that combines trust-based and item-based recommendation, considering not only ratings of the target item, but also those of similar items.

Under the context of social networking systems, social friendship and random walks have been shown to be beneficial for collaborative filtering-based recommendation systems. These works argue that social friends—for instance, in Facebook or Twitter—

tend to share common interests and thus their relationships should be considered in the process of collaborative filtering [Ye et al., 2011]. In this context, a random walk sees a social network as a graph with probabilistically weighted links that represent social relations and thus is able to accurately predict users' preferences to items and their social influence with respect to other users. Backstrom and Leskovec [2011] proposed an algorithm based on supervised random walks that combines the information from the network structure with node and edge level attributes, using these attributes to guide the random walk on the graph. Konstas et al. [2009] showed that extra knowledge provided by the users' social activity can improve the performance of a recommendation system using random walk with restarts. Weng et al. [2010] proposed TwitterRank to measure the influence of users in Twitter, considering both the topical similarity between users and the link structure of the social network.

2.2.2 Ranking in Academic Search

Ranking has traditionally played an important role in academic search, particularly for tasks related to assessing the scientific productivity of academic entities. In particular, one of the earliest metrics proposed to quantify academic impact was Garfield's Impact Factor [Garfield, 1955]. Despite its wide usage since it was proposed in 1955, it has been largely criticized [Saha et al., 2003]. As a result, many alternatives have been proposed in the literature, such as other citation-based metrics like the H-Index [Hirsch, 2005], download-based metrics [Bollen et al., 2005], and PageRank-like metrics [Yan and Lee, 2007]. However, as argued by Leydesdorff [2009], each metric has its own bias and there are both advantages and disadvantages associated with each one.

Citation-based metrics have been applied to rank computer and information science journals [Katerattanakul et al., 2003; Nerur et al., 2005]. Also, several citation-based metrics have been proposed to measure the quality of a small set of conferences and journals in the database field [Rahm and Thor, 2005], and to rank documents retrieved from a digital library [Larsen and Ingwersen, 2006]. Mann et al. [2006] introduced topic modeling to further complement the citation-based bibliometric indicators, producing more fine-grained impact measures. Yan and Lee [2007] proposed two measures for ranking the impact of academic venues which aim at efficiency and at mimicking the results of the widely accepted Impact Factor. An alternative method was presented by Zhuang et al. [2007], who proposed a set of heuristics to automatically discover prestigious and low-quality conferences by mining the characteristics of program committee members.

Given the importance of citations in academic environments, some works try to

predict future citation counts. Castillo et al. [2007] study the problem of predicting the popularity of items in a dynamic environment in which authors post continuously new items and provide feedback on existing items. As a case study, authors show how to estimate the number of citations for an academic paper using information about past articles written by the same author(s) of the paper. Nezhadbiglari et al. [2016] tackle the problem of predicting the popularity *trend* (e.g., number of citations) of scholars as a classification problem by applying K-Spectral Clustering and regression models.

Piwowar [2013] recently claimed that citation-based metrics are useful, but not sufficient to evaluate research. In particular, he observed that metrics like the H-Index are slow. Indeed, the first citation of a scientific article can take years. As a result, he argued for the development of alternative metrics to complement citation analysis. In a similar vein, Pradhan et al. [2016] argued that citation-based metrics such as H-index and its popular variants are mostly effective in ranking highly-cited authors, thus fail to resolve ties while ranking medium-cited and low-cited authors who are majority in number. Also, they discuss that these metrics are inefficient to predict the ability of promising young researchers at the beginning of their career. In the same way, Lima et al. [2013] argued that productivity indices should account for the singularities of the publication patterns of different research areas, in order to produce an unbiased assessment of the impact of academic output. Accordingly, they proposed to assess a researcher's productivity by aggregating his or her impact indicators across multiple areas. Finally, Gonçalves et al. [2014] investigated the importance of various academic features to scholar popularity and concluded that only two features are needed to explain all the variation in popularity across different scholars: (i) the number of publications and (ii) the average quality of the scholar's publication venues. In this work, we validate our proposed approach by exploiting exactly these these two features to rank different venues and different researchers.

The idea of reputation, instead of citations, was discussed by Nelakuditi et al. [2011]. In particular, they proposed a metric called peers' reputation, which measures the selectivity of a publication venue based upon the reputation of its authors' institutions. The proposed metric was shown to be a better indicator of selectivity than the acceptance ratio. In addition, the authors observed that many conferences have similar or better peers' reputation than journals. Another approach related to ours was proposed by Cormode et al. [2014], who attempted to rank authors according to their similarity with respect to a reference author. To allow the identification of comparable people in similar research areas, they first represented a researcher as a sequence of her publication records, based on research topic similarity and venue quality, and estimated the distance between any two researchers using sequence matching. As we

will discuss in Chapter 3, our approach also explores the notion of a reference source of reputation. However, in contrast to the aforementioned approaches, we allow for multiple entities—as opposed to a single one—to serve as a source of reputation. More importantly, instead of identifying similar entities from whom to propagate reputation, we explore the notion of reputation sources as part of a stochastic Markov process. As a result, our approach is able to produce a global reputation-oriented ranking of multiple interconnected entities.

2.2.2.1 Other Models

Next, we briefly describe some selected works recently published on ranking in academic search. Sastry et al. [2016] present an author ranking algorithm that interleaves the ranking of papers and authors. The algorithm is based on the idea that an author's rank is determined by the quality or rank of papers which cite him while the ranks of the papers in turn depend on the quality of its authors. Kim et al. [2016] provide a new lens to analyze the topical relationship embedded in the citation sentences in an integrated manner. To this end, authors extract citation sentences from full-text articles. By applying Author-Journal-Topic model, they identify which are the major topics shared among researchers in Oncology field and which authors and journals lead the idea exchange in its sub-disciplines. García-Romero et al. [2016] introduce an aggregation of different performance measures to build an alternative ranking of journals. Their approach is based on a pure output-oriented Free Disposal Hull. They analyze four indicators — Journal Impact Factor, Discounted Impact Factor, h-index, and Article Influence — for a set of 232 journals in Economics. Liang and Jiang [2016] propose a mutual ranking algorithm based on the multinomial heterogeneous academic hypernetwork, which serves as a generalized model of a scientific literature database. Mangaravite and Santos [2016] address expert search in academia by modeling document-person associations as non-boolean variables, reflecting the probability that a document is informative of the expertise of a candidate. Authors demonstrate the suitability of the proposed association and normalization schemes to improve the effectiveness of a state-of-the-art expert search approach. Boongoen et al. [2011] present a fuzzy qualitative classification system for academic performance evaluation using a link analysis methodology. The proposed model considers involving variables, classes and their relations as elements of a social network that can be modelled as a weighted graph. Franceschini and Maisano [2011] propose a structured method to compare academic research groups within the same discipline using some Hirsch based bibliometric indicators. Five different topologies of indicators are used so as to depict groups'

bibliometric positioning within the scientific community. SimRank [Jeh and Widom, 2002] is a well-known link-based similarity measure that can be applied on a citation graph to compute similarity of academic literature data. The intuition behind SimRank is that two objects are similar if they are referenced by similar objects. SimRank has attracted a growing interest in the areas of data mining and information retrieval recently. Despite the current success of SimRank, authors claim that it has some problems that negatively affect its effectiveness in similarity computation. Hamedani and Kim [2016] discuss existing problems of SimRank, present SimRank variants that have been proposed to solve those problems, and evaluate the effectiveness of SimRank and its variants in similarity computation for academic literature.

2.2.2.2 Academic Analysis

The analysis of academic indicators is of great importance and receives attention in literature. Some of the most recent works on it are briefly describe here.

Gonçalves [2016] correlate a popularity ranking of scholars (built by sorting scholars according to their h-index and then by their total citation count) with various academic features such as number of publications, years after doctorate, number of supervised students, as well as other popularity metrics across different areas of knowledge. Silva et al. [2016] discuss the importance of characterizing the trajectories of faculty members in their academic education and their impact on the quality of the graduate programs they are associated with. Authors analyze the mobility of faculty members from top Brazilian Computer Science graduate programs as they progress through their academic education. Classen et al. [2015] study the evolution of publication activity and citation impact in Brazil during 1991-2003. Besides the analysis of trends in publication and citation patterns and of national publication profiles, an attempt is made to find statistical evidences of the relation between international coauthorship and both research profile and citation impact in the Latin American region. Lima et al. [2015] perform a data-driven assessment of the performance of top Brazilian computer science researchers considering three central dimensions: career length, number of students mentored, and volume of publications and citations. Authors also demonstrate that it is necessary to go beyond counting publications to assess research quality and show the importance of considering the peculiarities of different areas of expertise while carrying out such an assessment. Figueira et al. [2015] use correlation analysis to assess the relative importance of academic factors (conference papers, journal articles and student supervisions) to the popularity of individual scholars and groups of scholars. Authors rely on curriculum vitae data of almost 700 scholars affil-

iated to 17 top quality graduate programs of two of the largest universities in Brazil. Delgado-Garcia et al. [2014b] present a preliminary analysis of the scientific production of Latin American Computer Science research groups. Results show a clear improvement in the publication output of these groups in the last 10 years, particularly in Argentina, Chile and Mexico. Delgado-Garcia et al. [2014a] analyze the co-authorship networks of Latin American Computer Science research groups. Results show that between 2004-2013 there has been an increase in terms of publications and collaborations in Latin America. Authors also identify the influential authors in the area according to complex network metrics and analyze the research networks originated from the co-authorships. Benevenuto et al. [2016b] investigate if computer science conferences are really able to create collaborative research communities by analyzing the structure of the communities formed by the flagship conferences of several ACM SIGs. Authors show that most of these flagship conferences are able to connect their main authors in large and well-structured communities. However, they noted that in a few ACM SIG flagship conferences authors do not collaborate over the years, creating a structure with several small disconnected components. According to Benevenuto et al. [2016a], it is likely that a researcher may use her coauthors' H-indexes as a way to infer whether her own H-index is adequate in her research area. Nevertheless, the authors show that the average H-index of a researcher's coauthors is usually higher than her own H-index. They also present empirical evidence of it and discuss some of its potential consequences.

To perform analysis similar to the aforementioned ones, researchers may create their own repository by collecting data from the web or, better, researchers may collect data from some reliable academic repository. Some well known repositories include Google Scholar, Microsoft Academic Search, DBLP, and others. Besides those well known academic repositories, there are also initiatives to build repositories for more specific tasks, like expert search. Some of them include ground-truths, which allows other researchers to validate their methods. Mangaravite et al. [2016] present the Lattes Expertise Retrieval test collection for research on academic expertise retrieval, which provides graded relevance judgements performed by expert judges and encompasses candidate experts from various areas of knowledge working in research institutions in Brazil. Dores et al. [2016] give a first step towards building a large repository that records the academic genealogy of researchers across fields and countries. Authors crawled data from the Networked Digital Library of Theses and Dissertations and develop a framework to extract academic genealogy trees from this data and provide a series of analyses that describe the main properties of the academic genealogy trees.

2.2.3 Random Walks in Academic Search

Earlier works have studied the application of random walks for ranking authors, papers and venues in an academic setting. For instance, Sun and Giles [2007] proposed a popularity weighted ranking algorithm for academic digital libraries that uses the popularity factor of a publication venue. Their approach overcomes some limitations of the Impact Factor and performs better than PageRank, citation counting and HITS. Relatedly, Zhou et al. [2007] proposed a method for co-ranking authors and their publications using several networks. Similarly, Yan et al. [2011] presented a new informetric indicator, P-Rank, for measuring prestige in heterogeneous scholarly networks containing articles, authors and journals. P-Rank differentiates the weight of each citation based on its citing papers, citing journals and citing authors.

In a narrower perspective, random walks have also been used for the task of expert finding in academic search collections. For instance, Deng et al. [2012] proposed a joint regularization framework to enhance expertise retrieval in academia by modeling heterogeneous networks as regularization constraints on top of a document-centric model [Balog, 2012]. Relatedly, Wu et al. [2009] proposed to model authors and publications as nodes of a publication network, with additional edges representing co-authorship information (author-author edges). In a similar vein, Tang et al. [2008] proposed a probabilistic topic modeling approach to enrich a heterogeneous graph comprising multiple academic entities as nodes, including authors, papers, and publication venues, with directed edges representing a variety of relationships such as "written by" and "published in". The stationary distribution computed after a random walk on this graph was then used to rank these entities with respect to an input query. A very similar approach was proposed by Gollapalli et al. [2011], by assigning topics to nodes and then computing the unique stationary distribution of the associated Markov chain.

Recent research on random walks applied to academic search include the following. Pradhan et al. [2016] propose C3-index that combines the effect of citations and collaborations of an author in a systematic way using a weighted multi-layered network to rank authors. Dhanjal and Clémençon [2014] propose to use Latent Semantic Indexing and Latent Dirichlet Allocation to find authors who have worked in a query field. Authors then construct a coauthorship graph and motivate the use of a variety of graph centrality measures to obtain a ranked list of experts. The ranked lists are further improved using a Markov Chain-based rank aggregation approach. Gkorou et al. [2013] show that the properties of a node indicate accurately its reliability, and that random walks exploiting these properties are more resilient than simple random walks. Authors model reputation systems in growing synthetic random and scale-free graphs,

and in real-world graphs derived from the Bartercast reputation system [Delaviz et al., 2010] which is used in the BitTorrent client Tribler [Pouwelse et al., 2008], from the citation network of Physical Review E journal, and from Facebook [Viswanath et al., 2009]. Yu and Chen [2012] present a PageRank-like algorithm that can be used to evaluate reputation of literatures and researchers. The basic idea is that there are some relationships among literatures, its author, periodicals and readers. The relationship can be regard as recommendation for each other that is similar to web links. Another metric based on random walks is the Eigenfactor Score [Bergstrom et al., 2008], which can be viewed as the result of a random walk through the scientific literature. This algorithm models readers following chains of citations as they move from journal to journal. The frequency with which researchers visit each journal gives a measure of that journal's importance within network of academic citations.

In contrast to the aforementioned works, we use random walks to model the *transference* of reputation from *multiple* reference sources to selected targets in a reputation graph, as discussed in Chapter 3. To validate our model, we instantiate it in an academic search setting by adopting distinct configurations of the reputation graph. To illustrate, in some configurations, we use research groups as reputation sources and publication venues as reputation targets. Moreover, while previous approaches have exploited multiple ranking signals, we demonstrate the power of the notion of reputation transfer by relying on publishing behavior as the only reputation signal.

Chapter 3

Reputation Flows

Identifying reputable entities is an important task in many domains. While quantifying the reputation of a given entity is a challenging task, we argue that the flow of reputation among entities can be accurately modeled as a stochastic process. To this end, we propose here a conceptual framework for ranking entities that interact with (and hence convey reputation to) one another in some manner.

To formalize our approach, in Section 3.1 we introduce the reputation graph, a data structure that models the flow of reputation from selected sources to multiple targets. In Section 3.2, we formalize a stochastic process to estimate the amount of reputation transferred to target entities. Lastly, in Section 3.3, we discuss a simple mechanism to rank entities according to their inferred reputation.

3.1 The Reputation Graph

We define a *reputation graph* as a graph with three node types: reputation sources, reputation targets, and reputation collaterals, as illustrated in Figure 3.1. The reputation graph models the transference of reputation from a reference set of reputation sources to reputation targets, and then to reputation collaterals. To refer to the reputation graph, we adopt the following notation: S is the set of reputation sources, T is the set of reputation targets, and C is the set of reputation collaterals.

Figure 3.1: Structure of the reputation graph.

The reputation of source nodes influences the reputation of target nodes as much as the reputation of target nodes influences the reputation of source nodes. Note that the reputation of target nodes also influences the reputation of collaterals, but the reputation of collaterals has no impact in the reputation of sources and targets. The use of collaterals allows us to isolate the impact of a set of arbitrary nodes on the reputation graph, fixing reputation sources as the only set of nodes providing reputation. While this design choice aimed primarily at effectiveness, it also contributes to the efficiency of our approach, as a random walk is only performed on the selected source and target nodes. This way, the overall cost of our approach remains the same even for large sets of collateral nodes. To illustrate, in Chapter 4, we apply these concepts to model reputation flows in academia. Specifically, we instantiate research groups as reputation sources, publication venues as reputation targets, and individual researchers as collaterals.

Given that the reputation of collaterals has no effect on the reputation of nodes of other types, we can split the model in two phases. In the first phase, we propagate the reputation of the sources to the targets. In the second phase, we propagate the reputation of the targets to the collaterals. These phases are discussed following.

3.2 Reputation Flows

The interaction between reputation sources and reputation targets is inspired by the notion of *eigenvalue centrality* in complex networks [Newman, 2010], which also provides the foundation to PageRank [LangVille and Meyer, 2006; Brin and Page, 1998].

In the reputation graph, if we consider only sources and targets, it is easy to identify reputation flows from sources to sources, from sources to targets, from targets to sources, and from targets to targets. These reputation flows can be modeled as a stochastic process as we now discuss. In particular, let $\mathbf{P}$ be a *right stochastic* matrix of size $(|S| + |T|) \times (|S| + |T|)$ with the following structure:

$$\mathbf{P} = \left[\begin{array}{r|l} (d^{\langle S \rangle}).\mathbf{P}^{\langle SS \rangle} & (1 - d^{\langle S \rangle}).\mathbf{P}^{\langle ST \rangle} \\ \hline (1 - d^{\langle T \rangle}).\mathbf{P}^{\langle TS \rangle} & (d^{\langle T \rangle}).\mathbf{P}^{\langle TT \rangle} \end{array}\right], \tag{3.1}$$

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

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

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

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

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

The parameters $d^{\langle S \rangle}$ and $d^{\langle T \rangle}$ control the relative importance of the reputation sources and targets, which are modeled in the four matrices above. Specifically, $d^{\langle S \rangle}$ is the fraction of reputation one wants to transfer between source nodes and $d^{\langle T \rangle}$ is the fraction of reputation one wants to transfer between target nodes. These are useful parameters and the ability to set them is important to calibrate the impact of different reputation flows in the final score. If we do not want to consider reputation flows between nodes of the same type, it is sufficient to set both parameters to zero. If, instead, we want to consider reputation flows between nodes of the same type, we may increase these parameters according to the desired relative importance. Note that, as (i) the sub-matrices $\mathbf{P}^{\langle SS \rangle}$ and $\mathbf{P}^{\langle TT \rangle}$ are *right stochastic*, (ii) each of the rows of matrices $\mathbf{P}^{\langle ST \rangle}$ and $\mathbf{P}^{\langle TS \rangle}$ sums to 1, and (iii) the parameters $d^{\langle S \rangle}$ and $d^{\langle T \rangle}$ are both in the range [0,1), then $\mathbf{P}$ defines a Markov chain. Assuming that the transition matrix $\mathbf{P}$ is ergodic — recall that, in an ergodic process, the state of the process after a long time is nearly independent of its initial state [Walters, 2000] —, we can compute the steady state probability of each node and use it as a reputation score. Specifically, we can obtain values for ranking the set of nodes by solving:

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

where $\boldsymbol{\gamma}$ is a row matrix with $|S| + |T|$ elements, where each one represents the probability of a node in the set $S \cup T$. This system of linear equations can be easily solved by standard Markov chain techniques. Then, from Equation (3.2), we obtain the steady state probabilities of all nodes in $S \cup T$, a.k.a. reputation sources and reputation targets.

3.2.1 Flow Equations

We recursively define the reputation of sources in terms of the reputation of targets, and the reputation of targets in terms of the reputation of sources. Specifically, the

reputation γ_s of a source s is defined as:

$$\gamma_s = \sum_{t \in T} (1 - d^{\langle T \rangle}).\mathbf{P}_{ts}{}^{\langle TS \rangle}\gamma_t + \sum_{s' \in S} (d^{\langle S \rangle}).\mathbf{P}_{s's}{}^{\langle SS \rangle}\gamma_{s'}. \tag{3.3}$$

In the summation, $\mathbf{P}_{ts}{}^{\langle TS \rangle}$ is the transition probability from t to s, given by $\mathbf{P}_{ts}{}^{\langle TS \rangle} = n_{ts}/n_t$, where n_{ts} is the number of edges running from t to s and n_t is the total number of edges running from t. Finally, γ_t is the reputation of target t, defined recursively as:

$$\gamma_t = \sum_{s \in S} (1 - d^{\langle S \rangle}).\mathbf{P}_{st}{}^{\langle ST \rangle}\gamma_s + \sum_{t' \in T} (d^{\langle T \rangle}).\mathbf{P}_{t't}{}^{\langle TT \rangle}\gamma_{t'}. \tag{3.4}$$

Similarly, in the summation, $\mathbf{P}_{st}{}^{\langle ST \rangle}$ is the transition probability from s to t, given by $\mathbf{P}_{st}{}^{\langle ST \rangle} = n_{st}/n_s$, where n_{st} is the number of edges running from s to t and n_s is the total number of edges running from s. Recall that γ_s is the reputation of source s, defined according to Equation (3.3).

3.2.2 Bipartite Reputation Graph

Some scenarios can be represented as a bipartite reputation graph. In these cases, the transition matrix $\mathbf{P}$ is reduced to a periodic Markov chain with the following structure:

$$\mathbf{P} = \left[\begin{array}{c|c} \mathbf{0} & \mathbf{P}^{\langle ST \rangle} \\ \hline \mathbf{P}^{\langle TS \rangle} & \mathbf{0} \end{array}\right]. \tag{3.5}$$

From decomposition theory [Meyer, 1989], we can obtain values for ranking the set of reputation *sources* by solving:

$$\boldsymbol{\gamma}^{\langle S \rangle} = \boldsymbol{\gamma}^{\langle S \rangle}\mathbf{P'}, \tag{3.6}$$

where $\mathbf{P'} = \mathbf{P}^{\langle ST \rangle} \times \mathbf{P}^{\langle TS \rangle}$ is a stochastic matrix and $\boldsymbol{\gamma}^{\langle S \rangle}$ is a row matrix with $|S|$ elements, where each one represents the probability of a node in the set S of reputation sources. Note that matrix $\mathbf{P'}$ has dimension $|S| \times |S|$ only and can be easily solved by standard Markov chain techniques. Then, we can obtain the reputation of all reputation *targets* linked by the reputation sources, as follows:

$$\boldsymbol{\gamma}^{\langle T \rangle} = \boldsymbol{\gamma}^{\langle S \rangle} \times \mathbf{P}^{\langle ST \rangle}. \tag{3.7}$$

By modeling a scenario as a bipartite reputation graph instead of a general rep-

utation graph, we reduce the network from a graph of size $(|S|+|T|)\times(|S|+|T|)$ to a graph of size $|S|\times|S|$, which allows us to compute the steady state probabilities much more efficiently. However, by using a bipartite graph, we are certainly losing some information, which may be critical for some applications. It is important to consider this trade-off when instantiating our framework.

3.3 Reputation-based Ranking

The steady state probability of a node can be interpreted as its relative reputation, as transferred from other nodes in the reputation graph. Thus, we can directly use the value of this probability to rank reputation sources or reputation targets. Additionally, this probability can be further propagated to nodes we want to compare, which are in the collateral set. This propagation depends on a matrix $\mathbf{P}^{\langle TC\rangle}$ of size $|T|\times|C|$ representing the transitions from reputation targets to collateral nodes. More generally, we can define the reputation score of an entity e according to:

$$\text{P-score}(e)=\begin{cases}\sum_{t\in T}\mathbf{P}_{te}{}^{\langle TC\rangle}\gamma_t & \text{if } e\in C,\\ \gamma_e & \text{otherwise}\end{cases} \tag{3.8}$$

where $\mathbf{P}_{te}{}^{\langle TC\rangle}$ is the transition weight from a target node t to a collateral node $e\in C$. The P-score of all candidate entities (targets or collaterals) can then be used to produce an overall reputation-oriented ranking of these entities.

Chapter 4

Reputation Flows in Academia

In this chapter, we discuss the instantiation of our conceptual framework of *reputation flows* in the context of academia to model the transference of reputation between authors, papers, research groups and publication venues. The relations between these academic entities may be captured through distinct metrics and, as far as we know, the most important ones (including citation-based metrics) fit well in our conceptual framework. In particular, let us start by defining the relations between authors and papers. It is easy to identify *reputation flows* from authors to authors, from authors to papers, from papers to authors, and from papers to papers. Each one of these *reputation flows* is associated with a specific quadrant of an Author-Paper × Author-Paper relation matrix, as illustrated in Figure 4.1.

$$\begin{array}{c} \\ Author \\ Paper \end{array} \begin{array}{c} \begin{array}{cc} Author & Paper \end{array} \\ \begin{bmatrix} Author \rightarrow Author & Author \rightarrow Paper \\ Paper \rightarrow Author & Paper \rightarrow Paper \end{bmatrix} \end{array}$$

Figure 4.1: Reputation flows between authors and papers.

In the first quadrant, the framework represents the reputation flow from authors to authors, which can be expressed in terms of co-authorship relations or citations from an author to another. In the second and third quadrants, the framework represents author-paper and paper-author relations, respectively. An author who publishes a paper somehow transfers its own reputation to that paper or the converse, a paper may transfer its reputation or acceptance by the community to the authors who published it. In the fourth quadrant, the framework represents the reputation flow between papers. When a paper cites another, it is somehow transferring part of its reputation to the

cited paper.

This last quadrant, $Paper \rightarrow Paper$, has received much more attention by the academic community than the other ones. The raw number of citations among papers, as well as well known citation-based metrics such as H-Index and Impact Factor can be represented in this quadrant. Additionally, there are further indicators such as the number of downloads of a paper. It is an indicator intrinsically related to the papers and has nothing to do with the reputation flow from authors to papers. In other words, the number of downloads is a reputation flow from the audience of paper readers to the papers. These external indicators can be expressed as bias variables.

The idea of reputation flows is broad and encompasses a large amount of indicators. Here, we define a more specific concept called *publication flows* to refer to the study of reputation flows where the transference of reputation is made by using only publication volume and without using citation data. In some of our experiments, we study how the reputation of a reference set of research groups is propagated to the venues they publish in and to other individual researchers by applying the concept of publication flows. In this conceptual framework, publication venues are aggregations of papers and research groups are aggregations of authors, as shown in Figure 4.2. These aggregations are sufficient to establish core relations that allow ranking these entities.

$$\begin{array}{c} \\ Group \\ Venue \end{array} \begin{array}{c} \begin{array}{cc} Group & Venue \end{array} \\ \begin{bmatrix} Group \rightarrow Group & Group \rightarrow Venue \\ Venue \rightarrow Group & Venue \rightarrow Venue \end{bmatrix} \end{array}$$

Figure 4.2: Reputation flows between groups and venues.

4.1 Overview and Assumptions

The instantiation of reputation flows in academia is based on a *role model* concept, in which one should take as reference the most reputable individuals in a community. We start by stating our main assumptions.

Assumption 1 *The reputation of a graduate program is strongly influenced by the reputation of its faculty, which is largely dependent on their publication track records.*

As a consequence of this assumption, the graduate programs in the *source set* employ the most prestigious faculty. The more prestigious is a program the better chances it has to attract the most prominent PhD graduates and senior renowned scientists.

Assumption 2 *A researcher, or member of a graduate program, conveys reputation to a venue proportionally to their own reputation.*

This assumption is a consequence of what we call the role-model effect. Reputable scientists usually choose prestigious venues to publish at and, as such, the reputation of a venue is positively correlated with the reputation of the individuals that publish in that venue. As more prestigious researchers of an area choose a venue to publish their work, the venue becomes increasingly known by peer researchers and, as a consequence, attracts even more distinguished researchers and young scientists, building up its reputation.

Assumption 3 *The reputation of a faculty member is positively correlated with the reputation of the venues in which he/she publishes.*

One of the most used metrics to promote a faculty member in any reputable department or graduate program is the number of papers in prestigious venues where the faculty under consideration publishes. Clearly, if a given scientist has a reasonable number of papers in the most prestigious venues in their field of study then it is reasonable to assume he/she is a prestigious scientist. These three assumptions form the cornerstone of our reputation-based ranking model.

4.2 Instantiation Example

Figure 4.3 shows an example with two research groups used as reputation sources and three publication venues used as reputation targets. Notice that, in this simple

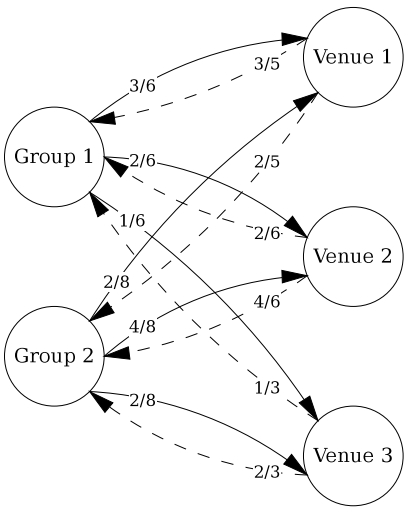

Figure 4.3: Markov chain for an example with 2 research groups and 3 venues.

example, we have not modelled co-authorships, nor have we considered paper to paper citations.

From Figure 4.3, *Group 1* published 3 papers in *Venue 1*, 2 papers in *Venue 2*, and 1 paper in *Venue 3*. The total number of publications of *Group 1* is therefore 6. *Venue 1* is composed of 3 papers from *Group 1* and 2 papers from *Group 2*. The fractions of publications from groups to venues and from venues to groups are the edge weights, as follows:

$$\mathbf{P} = \left[\begin{array}{cc|ccc} 0 & 0 & 3/6 & 2/6 & 1/6 \\ 0 & 0 & 2/8 & 4/8 & 2/8 \\ \hline 3/5 & 2/5 & 0 & 0 & 0 \\ 2/6 & 4/6 & 0 & 0 & 0 \\ 1/3 & 2/3 & 0 & 0 & 0 \end{array}\right]$$

The stochastic matrix $\mathbf{P}$ corresponds to the Markov chain displayed in Figure 4.3, which can be immediately aggregated to a two-state Markov chain, yielding:

$$\mathbf{P}' = \begin{bmatrix} 0.467 & 0.533 \\ 0.400 & 0.600 \end{bmatrix}$$

This is the stochastic matrix we use in the solution of Equation (3.6). Solving Equation (3.6) and applying Equation (3.7), we obtain the weights $\boldsymbol{\gamma}$ for our three venues:

$$\boldsymbol{\gamma}^{\langle T \rangle} = \langle 0.36, 0.43, 0.21 \rangle. \tag{4.1}$$

Venue 2 has the highest weight, followed by *Venue 1*, and then by *Venue 3*. We remark that the individual weights give the *relative reputation* of each publication venue, when *Group 1* and *Group 2* are taken as reputation sources. Then, we can apply Equation (3.8) to compute the P-score of each entity in the reputation graph.

Chapter 5

Experimental Setup

In this chapter, we describe the setup we used to evaluate our reputation flows model in the context of academia. We illustrate the effectiveness of our model by addressing three academic search tasks, namely ranking of publication venues, ranking of individual researchers, and ranking of academic departments, all in the broad area of Computer Science.

5.1 Research Questions

Let us proceed by stating the key research questions we focus on.

RQ1. *How effective is our proposed random walk model for reputation-oriented ranking of publication venues and academic groups?*

RQ2. *How robust is our model with respect to perturbations in the chosen set of reputation sources?*

RQ3. *Can we alleviate the cost of selecting reputation sources manually within our model?*

5.2 Academic Dataset

Our academic search dataset is an extension of the DBLP[1] repository. The dataset we built has all 5,000 publication venues and all 1.5 million authors which compose DBLP. We enrich the DBLP dataset by adding information regarding research groups and our ground-truths. To do so, we manually collected information about 126 graduate

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

programs from the United States and 25 graduate programs from Brazil. The author names of each North American program were collected from their Web pages (we built 126 parsers for that end) and, in the Brazilian case, we used the official lists of professors given by CAPES[2], a government agency linked to the Brazilian Ministry of Education. Salient statistics on our dataset are shown in Table 5.1.

Table 5.1: Salient statistics of the dataset used in our evaluation.

Type of Entity	Total Number of Instances
US CS Depts	126
BR Authors	383
CS Venues	1,083

To encourage the reuse of our research, we have made the full dataset available on GitHub[3] and we have developed a Web tool[4] based on our reputation flows model which allows users to quickly grasp insights about publication venues, authors, and research groups in Computer Science.

5.3 Ground-Truths

While ground-truths for the tasks of interest here may raise concern and controversy, our purpose is not to dictate a ranking of individuals and groups. Instead, we aim at showing that our reputation flows model produces results that approximate those produced by much more complex and costlier procedures. This is important because P-scores can be computed conveniently and can be updated quickly (by changing the reputation sources) to provide distinct views and insights into the reputation of the entities one intends to compare.

For the venue ranking task, we considered as ground-truth the set of venues in Computer Science classified by the Qualis system, maintained by agency CAPES, in the year of 2013. The agency assigns a committee of experts to each area of knowledge and these experts are responsible for evaluating all information acquired about the venues in that area, as well as producing a classification of venues. This classification is updated periodically and follows a set of criteria, such as: the number of publications in each venue, the number of repositories in which it is indexed, the amount of institutions publishing in it, citation information whenever available, among others. According

[2] http://www.capes.gov.br
[3] https://github.com/pscore
[4] http://pscore.dcc.ufmg.br

to Qualis, the venues in each area of knowledge are classified (in decreasing order of importance) as A1, A2, B1, B2, B3, B4, B5. In here, we adopt the classification of venues released by CAPES in 2013.

For the individual researcher ranking task, we considered as ground-truth the set of researchers with an active (as of 2014) productivity grant awarded by CNPq,[5] the Brazilian National Council for Scientific and Technological Development. That is, we restricted our evaluation to researchers in Brazil. We did so because this is one classification of individual researchers that is based on their productivity and that is repeated consistently. Indeed, to apply for a productivity grant, researchers working in Brazil must submit detailed information about their academic career to CNPq, including a research project to be conducted over the coming years. To award the grants, CNPq evaluates a set of productivity indicators including academic output, contribution to the formation of human resources, academic leadership, among others, and classifies researchers in five different levels of productivity in descending order of prestige: 1A, 1B, 1C, 1D, and 2. The starting point for any newly awarded researcher is the productivity level 2.

For the research group ranking task, we rank CS departments from the United States. For this, we considered as ground-truth the CS ranking provided by the National Research Council (NRC). The NRC issues a ranking of the graduate programs in US in many areas of knowledge, aiming at providing a general guideline for students and administrators. While these rankings raise a lot of controversy, we again emphasize that our proposal here is not to use P-scores do rank departments directly. Instead, we argue that a reputation flows model allows university administrators, government officials, and prospect students to gain alternative insights into the productivity of research groups through the proper selection of reputation sources.

In the area of Computer Science, the NRC classified the major 126 graduate programs in the US. To produce the rankings, NRC asked a sample of faculty to rate a sample of programs in their field. These rankings were then used to assign weights to a set of 20 features (such as number of publications, number of citations per paper, percentage of faculty with grants) through regression analysis. The weighted features were then used to rank all programs.

To account for uncertainty and variability in the surveys data, the regression analysis process was repeated by NRC 500 times, each time using as input a random sample of half of the surveys. As a result, 500 ranks were produced for each one of the 126 programs. Following, for each program, the top 5% and the bottom 5% ranks were

[5] `http://www.cnpq.br/`

disregarded, leaving each program with 450 ranks which were sorted. The top rank is what is referred to as the *program rank at the* 5^{th} *percentile*. The rank at the bottom is what is referred to as the *program rank at the* 95^{th} *percentile*. At the 5^{th} percentile, there are 21 programs with a rank position smaller or equal than 10 in the area of Computer Science.[6] We map the NRC ranking of all CS departments into 5 levels of relevance, as shown in Table 5.2.

5.4 Evaluation Procedure

We compare our P-score based method with a citation baseline, namely the well known H-Index. Our choice of this baseline is motivated by its wide adoption in academia. To compare the rankings of research groups, venues and individual authors produced by P-scores and H-indices, we use the discounted cumulative gain (DCG) metric introduced by Järvelin and Kekäläinen [2002].

DCG adopts a non-binary notion of relevance and allows assessing relevance on a graded scale. The metric also uses a log-based discount factor that reduces the impact of the score as we move lower in the ranking. Let l_i be the non-binary relevance level associated with the item ranked at the i-th position. The DCG at a rank position k is defined as:

$$\text{DCG@}k = \sum_{i=1}^{k} \frac{2^{l_i} - 1}{\log_2(i+1)}. \tag{5.1}$$

To bind the results within the interval [0,1], we use the normalized version of DCG, denoted nDCG, which is obtained by dividing the DCG@k value by the maximum possible value at the same ranking cutoff k.

As relevance levels, we consider a linear mapping from the classes defined by each of our ground-truths, see Table 5.2. We map the CS departments in the US into 5 relevance levels with the departments classified in positions 1-10 according to the 5^{th} percentile (there are 21 such departments) mapped into relevance level 5. Next, the departments classified in positions 11-20 (there are 11 such departments) are mapped into relevance level 4, and so on. There is nothing particular to our mapping. And again, the point we make here is not one about ranking departments using P-scores directly. Instead, we argue that reputation flows are useful to provide alternative insights into the productivity of the departments when we vary the reputation sources.

[6]Notice that more than one department might appear in a same position of the NRC ranking.

In Table 5.2, we illustrate all mappings of entity types into their ground-truths. Also, we map the classifications provided by each ground-truth with their corresponding nDCG relevance levels. The row of numbers underneath each mapping is composed of the respective count of instances.

Table 5.2: Mapping from ground-truths to the nDCG relevance levels.

		nDCG relevance level						
Type of Entity	Ground-truth	7	6	5	4	3	2	1
US CS Dept	NRC 5th perc			1-10	11-20	21-40	41-80	81-126
				21	11	24	42	28
BR Authors	CNPq class			1A	1B	1C	1D	2
				22	21	30	70	240
Venues	Qualis class	A1	A2	B1	B2	B3	B4	B5
		195	182	315	158	106	105	22

Chapter 6

Framework Validation

In this chapter, we validate the effectiveness of the conceptual framework of reputation flows by applying it to three academic search tasks, discussed in Section 6.1, and by measuring their correlations with citations, as detailed in Section 6.2.

6.1 Ranking of Venues, Authors, and CS Departments

In this section, we validate the effectiveness of the conceptual framework of reputation flows to address the following three academic search tasks: ranking publication venues, ranking individual researchers (or authors), and ranking research groups (or CS Departments). Our intent is to show that reputation flows can be used to gain insight into the productivity of individual researchers and of research groups, not that a ranking based on a single metric should be taken at face value. Instead, the decision on how to classify or rank individual researchers and research groups can only be made by committees composed of human specialists. However, even under this light, committees of specialists need data and metrics that can be used to support their decisions.

To solve the aforementioned tasks using P-score, we can use distinct types of entities as reputation sources. While using research groups as reputation sources and publication venues as reputation targets seems to be a reasonable approach to instantiate the reputation graph (as illustrated in previous sections), it is not the only one. In fact, there are many alternatives on how to instantiate the proposed metric in the academic context. In here, we experiment with the following types of reputation sources: research groups, individual researchers, and publication venues. These entities can also be used as reputation targets and collaterals, which leads us to a variety of configurations. To facilitate the discussion, we adopt the following notation:

- G^* is a subset of the set G of all research groups, which has been selected as reputation source,
- A^* is a subset of the set A of all individual paper authors, which has been selected as reputation source,
- V^* is a subset of the set V of all publication venues, which has been selected as reputation source.

In Table 6.1, we show 9 suggested configurations for the reputation graph. The sources and collaterals can be any of the three types of entities we consider i.e., venues, authors, and research groups. The targets, however, must always be entities of type *venues*. That is, we always use the reputation of venues as the key feature for classifying venues, research groups, and individual authors (or researchers).

Table 6.1: Suggested configurations for three academic search tasks: ranking publication venues, ranking individual researchers, and ranking research groups.

	Venues	Authors	Groups
Groups	$G^* \rightleftharpoons V \rightarrow V$	$G^* \rightleftharpoons V \rightarrow A$	$G^* \rightleftharpoons V \rightarrow G$
Authors	$A^* \rightleftharpoons V \rightarrow V$	$A^* \rightleftharpoons V \rightarrow A$	$A^* \rightleftharpoons V \rightarrow G$
Venues	$V^* \rightleftharpoons A \therefore A^* \rightleftharpoons V \rightarrow V$	$V^* \rightleftharpoons A \therefore A^* \rightleftharpoons V \rightarrow A$	$V^* \rightleftharpoons A \therefore A^* \rightleftharpoons V \rightarrow G$

The configuration $G^* \rightleftharpoons V \rightarrow G$ should be read as: a subset G^* of the set G of all research groups, adopted as reputation sources, transfers reputation to the publication venues in V, the reputation targets. These transfers are modeled as transition probabilities in the corresponding Markov network. The steady state probabilities of this network are taken as weights assigned to each venue in V. These weights are then used to compute P-scores for all research groups in G, the reputation collaterals.

The configuration $V^* \rightleftharpoons A \therefore A^* \rightleftharpoons V \rightarrow V$ should be read as: a subset V^* of all venues, selected as reputation sources, transfers reputation to the set A of all authors. These transfers are modeled as transition probabilities in the corresponding Markov network. The steady state probabilities of this network are taken as weights assigned to each author in A. Following, a subset $A^* \subseteq A$, composed of those authors with highest weights, is selected as reputation sources and used to transfer reputation to the set V of all venues, using a second Markov network model. In this second network, the steady state probabilities are taken as weights assigned to each venue in V. These weights are then used to compute P-scores for all venues in V, the reputation collaterals. In this

Table 6.2: Denomination of the reputation sources used in our validation.

12 Groups RandProc	The set of 12 graduate programs generated by running Algorithm 1, described in Section 7.2.2, 100 times.
Most Cited Authors	The set composed of the 1,000 most cited authors in the area of Computer Science.
Top Venues of Sub-areas	The set composed of the 4 most cited conferences and 4 most cited journals of each one of the 24 sub-areas in Computer Science, according to Microsoft Academic Search.[1]

particular case, the mapping between the collaterals, which are venues, and V is the identity matrix. Thus, the venue P-scores are simply the venue weights.

For each aforementioned task, we compute P-scores using selected research groups as reputation sources, selected paper authors as reputation sources, and selected venues as reputation sources, as detailed in Table 6.2. While somewhat simplistic, these sets of reputation sources provide enough evidence to our model and allow it to produce P-scores that lead to effective rankings.

6.1.1 Ranking Publication Venues

In this section, we discuss the effectiveness of P-score to rank publication venues. Specifically, we rank 1,083 publication venues using the three suggested configurations for this task. We display results for the first 600 positions in our ranking of venues.

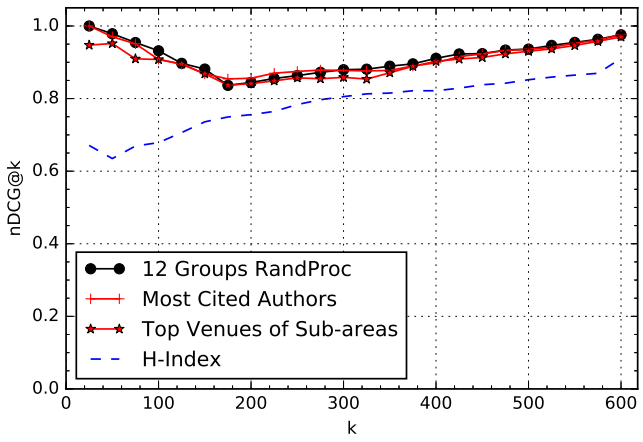

Figure 6.1: Ranking venues using as reputation sources: 12 groups identified through our randomization process ($G^* \rightleftharpoons V \rightarrow V$), most cited authors ($A^* \rightleftharpoons V \rightarrow V$), and most cited venues in sub-areas ($V^* \rightleftharpoons A \therefore A^* \rightleftharpoons V \rightarrow V$). These are compared with a ranking of venues based on H-Index.

Figure 6.1 illustrates the nDCG results for the ranking of venues when we selected as reputation sources a subset of venues, a subset of authors, and a subset of research groups (i.e., CS departments). It also displays nDCG results for a ranking of venues based on H-index. We make two key observations.

First, the three alternatives for selecting reputation sources produce almost identical results. However, while the adoption of "top venues in sub-areas" and "most cited authors" as reputation sources depend on citation information, the adoption of "12 groups randproc" generated by Algorithm 1 as reputation sources relies neither on citation nor in manual intervention. This is an important result, i.e., our reputation flows model allows ranking venues without citation information or human intervention.

Second, the ranking of venues produced by our model consistently outperforms the ranking of venues produced by H-index scores, which rely on citation data.

6.1.2 Ranking Individual Researchers

In this section, we discuss the effectiveness of P-scores to rank individual researchers. Specifically, we rank the 383 Brazilian researchers awarded with an individual research grant by CNPq.

Figure 6.2 illustrates the nDCG results of ranking individual researchers when we use as reputation sources 12 selected groups, most cited authors, and top venues in sub-areas. It also displays nDCG results when we rank the researchers by H-Index. We observe that all curves are rather similar. Thus, in this case, the adoption of selected groups, obtained in fully automatic fashion, as reputation sources produced results

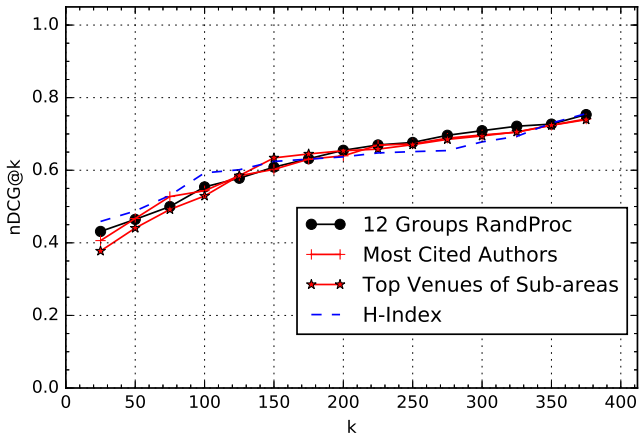

Figure 6.2: Ranking authors using as reputation sources: 12 groups selected using our randomization process, most cited authors, and top venues in sub-areas.

that are analogous to the other three alternatives, all of which depend on citation information.

We also observe that results are not as good as before with room for improvement. Fact of the matter is that the evaluation of individual researchers by CNPq is based on a conservative view of the overall production of the researcher throughout his professional life. Further, demotion of researchers to lower levels due to reduced productivity is almost never done. As a result, old timers tend to be maintained in the system at relatively high levels even when they are no longer productive in terms of research. This suggests that looking at a metric such as P-score might provide insights into how well the current evaluation system is working and how to improve the overall process.

6.1.3 Ranking Research Groups

In this section, we discuss the effectiveness of P-scores to rank research groups from the United States.

Specifically, we rank 126 CS graduate programs from the US using as reputation sources: 12 selected research groups (CS departments) produced by our randomization process, most cited authors, and top venues in each of the 24 sub-areas of CS identified by Microsoft Academic Search, as illustrated in Figure 6.3.

We observe that our reputation flows framework again yields results analogous to those produced using citation data. Further, in this case the results are quite effective and provide a good approximation of the results produced by NRC through a much more complex and costly process. We also repeated this experiment for the 25 CS

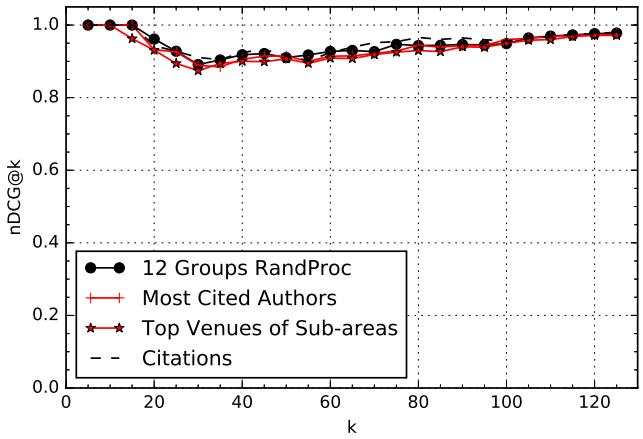

Figure 6.3: Ranking CS deparments from the US using as reputation sources: 12 selected research groups, most cited authors, and top venues in sub-areas.

departments in Brazil that are part of our dataset and obtained similar results.

The high quality of the results is a direct consequence of the fact that the highly productive departments are also those of highest reputation. That is, the departments that publish in top venues in large numbers are naturally the departments with highest reputation. Thus, given P-scores are computed as a function of venue weights and publication volume, it should not be a surprise that results are so effective. It is good news, however, in the sense that it shows that a relatively simple reputation flows model can be used to provide quick insights into the overall picture on the reputation of research groups out there.

6.2 Correlation Between P-Score and Citation Counts

Figure 6.4 presents the correlation between P-score venue weights and citation counts of all CS venues in our dataset.[2] For that we adopted as reputation sources the 12 research groups listed in Table 7.1. The correlation between venue P-score and venue citations, measured by the Kendall Tau coefficient, is 0.51 in this case, which is high given it varies between -1 and $+1$. This correlation shows that we can use P-scores to reason about the impact of publication venues, even when we do not take citations in consideration in the reputation graph.

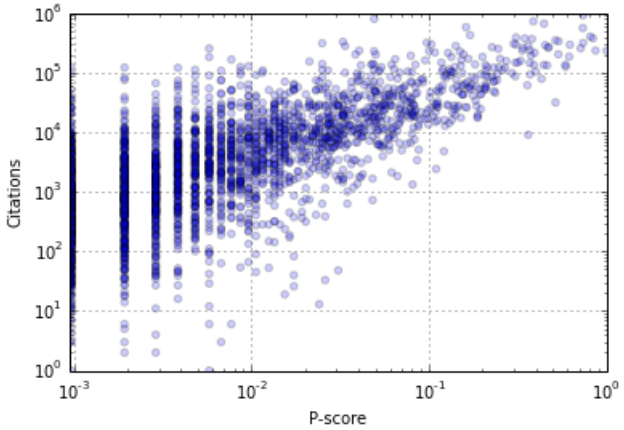

Figure 6.4: Correlation between P-score weights and citation counts

This result sheds light into the effective results presented in Section 6.1. The reason P-score weights work well for ranking publication venues, individual researchers, and whole departments in Computer Science is that they preserve a strong correlation

[2]Citation counts for whole venues were obtained from http://scholar.google.com.

with citation counts, which are a good indicator of reputation. Additionally, the framework of reputation flows provides a complementary view on the reputation of academic entities which allows implementation of new features, such as notification alerts on new venues with rising reputation, as discussed in Ribas et al. [2017].

Chapter 7

Selecting Reputation Sources

In this chapter, we discuss an important part of our model, the selection of reputation sources. As stated in Chapter 3, the proposed model scores entities based on how the reputation of a pre-selected set of reputation sources flows through a reputation graph. The reputation sources guide the network and any ranking produced by our model is *relative* to a given set of reputation sources, meaning that when we change the set of reputation sources, the results change as well. This way, the reputation sources assume a central role in our model and their selection is critical to produce effective results.

The selection of reputation sources may be performed manually, as discussed in Section 7.1, or automatically, as discussed in Section 7.2. After describing manual and automatic approaches to select reputation sources, we provide, in Section 7.3, an exploratory analysis on distinct sets of reputation sources. The main objective of this section is to better understand the impact that the choice of reputation sources with distinct properties has in the final rankings. Finally, in Section 7.4, we discuss the robustness of our proposed model.

7.1 Manual Selection of Reputation Sources

A natural approach to compose the set of reputation sources is through manual selection. By manually selecting the entities to compose the set of reputation sources, a user can quickly grasp insights about an application domain. We say that a manual selection happens when the user exerts direct influence in the composition of the reputation source set. We also say a selection process is manual when the user defines a feature to guide the process e.g., use top entities of that feature as reputation sources. Manual

approaches generally use information from outside the reputation graph, like the user's domain knowledge. An advantage of manually selecting reputation sources is that a user can take different views of an application domain by selecting thoughtful reputation sources, or reputation sources that make sense for a particular purpose. Manual selection is advised whenever the user has relevant information regarding the application domain and reliable information on the reputation of a set of entities. However, given that manual selection is often subjective, any ranking based on manual choices is prone to bias and thus might be hard to justify — in fact, our experience in this research is that people simple do not like it.

7.1.1 Subjective, Context-Dependent

Our conceptual framework of reputation flows was designed to be a tool that allows analysing the propagation of reputation from selected sources in the reputation graph. While effective rankings can be produced by using the most reputable entities, the framework is not limited to this end. In fact, the main purpose of this framework is to allow studing how reputation is propagated from certain nodes chosen by a given user with a given task or information need.

We say that a reputation source choice is subjective and context-dependent when the user explicitly chooses the reputation sources based on their own knowledge, which is often related to previous knowledge of the entities. To illustrate, lets consider the following academic example. If we ask someone to give examples of good CS departments, common answers would be Stanford and MIT. Often these answers are not based on a well defined criteria, but on word of mouth dissemination (or reputation). Despite the subjectivity associated with this process, it can provide valuable information about an area of knowledge.

7.1.2 Top Entities According to a Given Feature

An alternative to select reputation sources less subjectively is to use a feature of the entities which are candidates to reputation sources to guide the selection process. To illustrate, if we ask someone to provide features related to reputation, common choices would be citation counts or number of publications. The reason is that there is a correlation between citations and reputation, that is reputation frequently overlaps with citations.

7.2 Automatic Selection of Reputation Sources

We define a selection of reputation sources as automatic when it does not take any information from outside the reputation graph into account i.e., the user is not part of the process. The main advantage of automatic methods of selecting reputation sources is subjectivity avoidance. People often do not like ranking based on subjective choices. Thus, it is important to provide automatic selection techniques alongside our model. Additionally, automatic techniques are better indicated when the user has no (or little) information regarding the application domain.

Following, we discuss some approaches to automatically choose reputation sources. The first one is based on the idea of node centrality. The second is a randomized algorithm referred to in previous sections. Finally, we discuss how clustering the candidate entities can help selecting reasonable reputation sources.

7.2.1 Top Entities According to a Centrality Measure

A natural approach to automatically choose reasonable reputation sources, given an instantiation of the reputation graph, is to compute the centrality of each reputation source candidate. To do so, we place all entity candidates in the set of reputation sources and select the top k entities, according to a given centrality measure, to compose the set of reputation sources. Notice that this is similar to the procedure discussed in Section 7.1.2, the difference is that here no information from outside the reputation graph is used. Some examples of centrality measures are: degree, betweeness, closeness, eigenvector and PageRank. To illustrate, we could compute the PageRank of the nodes in a reputation graph composed by all research groups and publication venues of a given domain and use the entities with highest PageRank as the set of reputation sources.

7.2.2 A Randomized Process

An alternative to the use of top entities according to a centrality measure is to adopt a randomization process, which we first described in Ribas et al. [2015b] and which we review in greater detail in the immediately following.

Consider that we intend to use whole university departments as reputation sources. The motivation is that good university departments have a brand and a reputation that is widely known and respected. Initially, the set S of reputation sources is populated with a small random sample of all academic departments of interest. For

instance, this could be a sample of the 126 CS departments[1] (and their graduate programs) considered in the 2011 study done by the US National Research Council (NRC). Further, the set T of targets is populated with all venues in which professors (i.e., researchers) of these departments published in. In this context, Algorithm 1 presents a randomized process to select a subset S^* of all CS departments as reputation sources.

Algorithm 1: RandProc(S, T, k)

S: sources (all university departments);
k: input parameter, an integer, $k < |S|$;
S^* = random sample of k elements of S;
T: targets (all venues);
$d^{\langle S \rangle} = 0.5$ (fraction of reputation transfer between source nodes);
$d^{\langle T \rangle} = 0$ (do not consider citation information);
$C = S$ (the set of all collaterals to rank);
repeat
  P-score$(S^*, T, d^{\langle S \rangle}, d^{\langle T \rangle}, C)$;
  S^* = top k elements of C;
until *convergence*;
return S^*

In Algorithm 1, each university department is modeled as a source node and each venue is modeled as a target node. We start by randomly sampling a small subset of all university departments in consideration. Experimentation indicated that restricting the sample size k to 10 departments suffices. We also set parameter $d^{\langle S \rangle} = 0.5$, a value we determined empirically, and $d^{\langle T \rangle} = 0$, which implies that we do not use citation information here. Next, we use this random sample of CS departments as reputation sources to compute steady state probabilities for all venues modelled as targets. These probabilities are then used to compute P-scores, as defined by Equation (3.8), for all university departments taken as collateral nodes. The P-scores define a ranking of the departments. The top departments in this ranking are then used as reputation sources in the next iteration. This process is repeated until convergence, i.e. until there is no change in the set of reputation sources. To avoid falling into a local minimum, we run this method many times (e.g., 100 times) and count the number of times each department appeared in the final reputation source set.

We applied Algorithm 1 considering a sample size $k = 10$, which we determined empirically. Once convergence was achieved, the 10 reputation sources produced were

[1]For each CS department, we retrieved the list of its members and their publications, which were then reconciled against the DBLP repository, see `http://dblp.uni-trier.de/`.

annotated on the side. Following, application of Algorithm 1 was repeated 99 times. At the end of each run, the 10 reputation sources produced were annotated on the side.

Table 7.1 lists the 12 CS departments that appeared at least once in the list of 10 reputation departments produced at the end of each run. We observe that all 12 departments are among the top 5th percentile in the ranking produced by NRC. This suggests that our recursive procedure, as detailed in Algorithm 1, was able to take advantage of patterns in the publication streams of the various CS departments to determine the most reputable ones in fully automatic fashion.

Table 7.1: CS Departments that appeared at least once among the top 10 reputation sources after 100 runs of Algorithm 1.

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

7.2.3 Clustering Reputation Source Candidates

To understand how clustering helps the automatic selection of reputation sources, consider an academic search setting in which publication venues are the reputation sources. We call V the set of reputation source candidates, i.e. the set of all venues of a given domain. Consider also the 2-dimensional space of Figure 7.1 (a) where venues are plotted according to their topics so that the distance between venues represent topic distance. Thus, venues are closer when they have similar topics.

To produce our rankings using P-score, we need a set V^* which is a subset of V. Notice that, in this interpretation, if all venues are in the area of Computer Science, clusters represent sub-areas of knowledge, such as information retrieval or databases. It is noteworthy that some sub-areas are more popular than others. If we run a centrality measure in this scenario, popular sub-areas would be in advantage, which is the case in Figure 7.1 (b). In this figure, the set of reputation sources is composed by the three black venues. Notice that one of the sub-areas is not covered in this case.

By clustering the set of reputation source candidates, we can alleviate this popularity issue. Instead of selecting the top candidates as reputation sources, we can select the top candidates in each cluster. To illustrate, in Figure 7.1 (c) we select the top venue in each sub-area. Notice that in this case, the set of reputation sources covers all sub-areas. In addition to help automatic selection of reputation sources, clustering also helps when ranking in specific sub-areas. Also, we could use clustering to enhance manual selection of reputation sources through visualization.

Example: Venue Topics Using LDA

To illustrate these ideas, let's provide a concrete example by performing the following experiment. We build an algorithm whose objective is to find effective publication venues to compose the set of reputation sources by clustering them in order to keep the coverage of sub-areas. The final set of reputation sources should thus be composed of publication venues that cover a certain number of sub-areas of Computer Science. At the end of the algorithm, we verify whether the set of reputation sources satisfy this property.

There are many ways to identify a venue's sub-areas. For example, we can apply a variety of clustering algorithms available in the literature. In this experiment, we adopt Latent Dirichlet Allocation (LDA) [Blei et al., 2003], a generative probabilistic model for collections of discrete data. LDA is a three-level (documents, words and topics) hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document.

The most natural way to represent the input data of LDA in this context is to adopt venues as documents and paper words as venue words e.g., to represent a venue as a document composed by all words in the papers published by that venue. However,

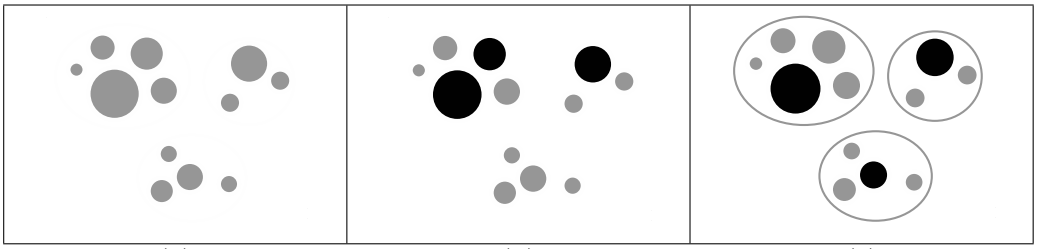

Figure 7.1: (a) Publication venues in a 2-dimentional space. (b) Top 3 venues (in black), considered as reputation sources. (c) Reputation sources in a clustered environment.

by doing so, the input data would be large. Thus, we modelled it in a different way that requires much less data and produces effective results. We represent venues as documents and authors as words. This way, each venue (or document) is represented by the set of authors that published papers in that venue.

In Table 7.2, we show the venues most likely to belong to each topic (or sub-area) according to LDA. By observing this table, we conclude that the desired property is valid, because each cluster represents a sub-area of Computer Science.

Once clustered, we can select a subset of publication venues in each cluster to compose the final set of reputation sources, ensuring that each cluster/sub-area is properly represented. As discussed previously, one way to make this choice is by adopting the probabilities returned by LDA itself or by applying a centrality measure inside each cluster. Applying the randomized method inside each cluster is also an option.

Table 7.2: Most likely venues of each sub-area according to LDA

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
TCS	AAAI	ICASSP	ICSE	GECCO
DM	ATAL	InterSpeech	ENTCS	CEC
STOC	IJCAI	TSP	TSE	IJCNN
SIAMComp	SemWeb	NIPS	TCS	ISCI
SODA	ECAI	ACL	JSS	FUZZ-IEEE

Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
IACR	CDC	IPPS	ICIP	MICCAI
TIT	Automatica	TPDS	ICPR	ISBI
IGARSS	TAC	ICPP	CVPR	CAISE
ISIT	AMC	EuroPAR	ICMCS	ICEIS
CCS	FSKD	AINA	TIP	TMI

Topic 11	Topic 12	Topic 13	Topic 14	Topic 15
ICDE	CHI	ICC	ISCAS	ICRA
CIKM	HICSS	Globecom	DATE	IROS
SIGMOD	TOG	VTC	DAC	TROB
WWW	AMIA	Infocom	TCAD	ROBIO
SIGIR	CGF	WCNC	ICCAD	AR

7.3 Exploring Reputation Sources

As we discussed, the composition of the set of reputation sources affects the final rankings. But, how exactly does the set of reputation sources affect the final results?

Let us start our discussion by analysing the results produced by intuitive sets of reputation sources. In Figure 7.2, we compare the effectiveness of venue rankings given distinct sets of reputation sources composed by individual researchers. Specifically, we adopted configuration: $A^* \rightleftharpoons V$. This figure is similar to the ones in Chapter 6, where we plot the nDCG values relative to a ranking of venues. An interesting effect occurs when we adopt Turing Awards as the set of reputation sources. While these are among the most reputable researchers in Computer Science, there are not enough of them, or their publications are not in enough numbers, to ensure proper transfer of reputation to publication venues.

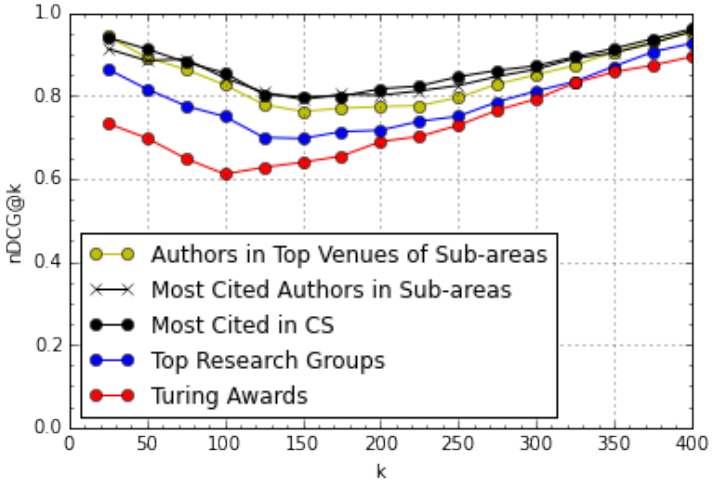

Figure 7.2: Results for distinct types of reputation sources

In Sections 7.3.1 to 7.3.3, we provide an exploratory analysis on distinct sets of reputation sources. Specifically, we seek to understand how distinct choices of reputation sources affect the final rankings and also what makes a good set of reputation sources. As the results in Section 7.3.1 suggest, one of the reasons for this strange effect regarding the Turing Awards as reputation sources is that the number of Turing Awards is quite small when compared to the other sets of reputation sources used to rank venues in Figure 7.2, which may result in a coverage issue. One alternative to address this issue is to increase the number of reputation sources.

7.3.1 Characterization

In this section, we investigate the following question: What makes a good set of reputation sources? To address this question, we start by rephrasing it as: What features F_i make an effective set of reputation sources according to a specific quality measure Q? This last question allow us to objectively formulate an experimental evaluation. To answer this question, we can analyze the plots $F_i \times Q$, for each feature F_i. The next thing we need to do is to define the set of features F and the quality measure Q.

In a general application of the framework we may be interested in evaluating the impact of features like the number of reputation sources, the number of reputation targets, and the number of source out-links. Those features will be present in any instantiation of the reputation graph. In contrast, we may also be interested in analysing features that are specific to the application context, in our case, the number of citations received by authors in a reputation source set may be of interest. Here, our set of features is $F = \{$ # sources, # targets, # source out-links, # source citations $\}$.

In Figures 7.3 to 7.6, we plot $F_i \times Q$, for each feature F_i, where the quality measure Q is nDCG@10 applied to the ranking of venues, given a set of authors as reputation sources. Each point in these plots represents a pair composed by a single set of reputation sources and the corresponding target ranking. The x-axis is the value of feature F_i of a set of reputation sources, while the y-axis is the value of the quality measure Q of the ranked venues. The number of distinct sets of reputation sources is 2^N, where N is the number of reputation source candidates (here, the number of authors). Given that it is not possible to plot the 2^N sets of reputation sources (there are more than 2 million authors in DBLP), we sample it. We observed that all analysed features have a positive correlation with ranking quality.

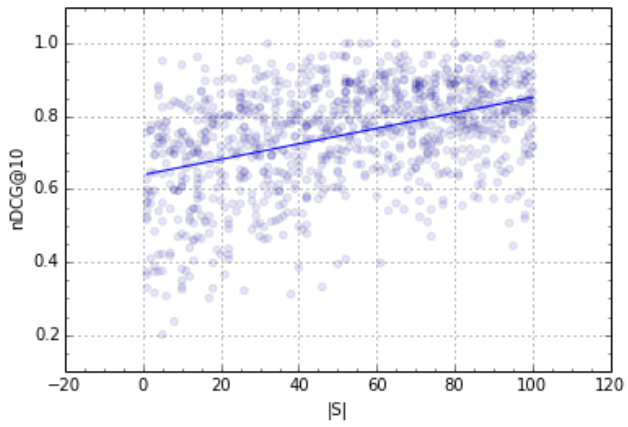

Figure 7.3: Number of reputation sources

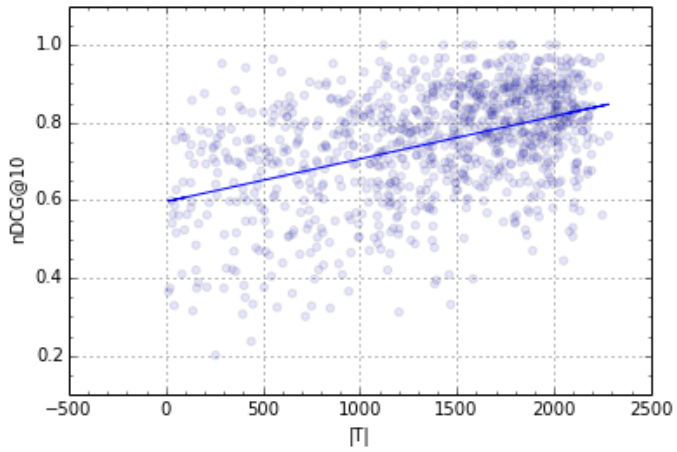

Figure 7.4: Number of reputation targets

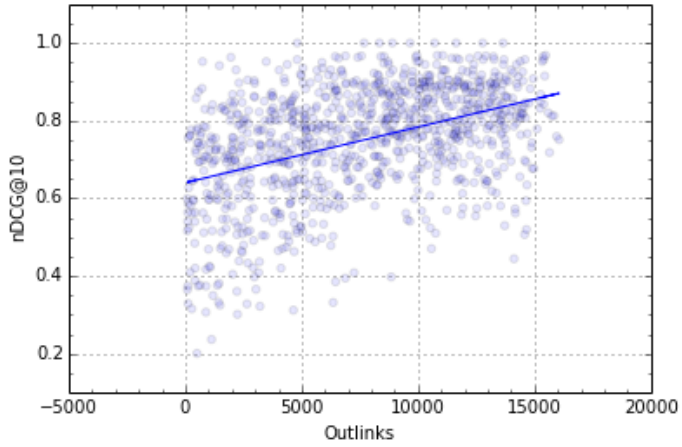

Figure 7.5: Number of source outlinks

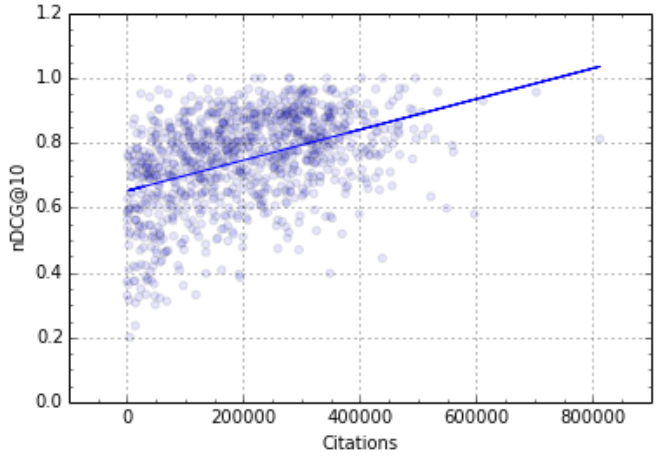

Figure 7.6: Citations received by reputation source members

7.3.2 Length Impact

To better understand the impact of the reputation sources, we analyze BPref results of venue rankings (produced by P-score), see Baeza-Yates and Ribeiro-Neto [2011], when we vary their size. To compute BPref, we consider that a venue classified by Qualis as A1 or A2 is relevant, venues with other classifications are non-relevant, and venues not classified by Qualis are unknown. In Figures 7.7, 7.8 and 7.9, the vertical axis show BPref measures relative to Qualis venue rankings and, in the horizontal axis, each integer i represents the group (or graduate program) ranked in the i-th position of the NRC ranking for Computer Science.

In Figure 7.7, we investigate the minimum size of the reference set. Using reference sets of size 1, 2, $\cdots$, N ($N = 126$), i.e., we increase the size of the reference set by 1 with each experiment, and we do so by following the NRC ranking (start with the top graduate program). By looking at the result, we observe that there is no need to choose more than 10 research groups (and we only do so to cover enough venues). After that, the ranking of venues worsens.

In Figure 7.8, we run a complementary set of experiments in which we start with a reference set of size 126 and reduce it by 1 in each new experiment, but start with the top graduate program in the NRC ranking. The reference sets are then $\{1, \cdots, N\}$, $\{2, \cdots, N\}$, $\cdots$, $\{N\}$. The results show that the quality of the rankings (in terms of BPref) decreases as we gradually remove the top groups from the reference set. Also, the quality of the ranking drops sharply at the bottom part of the graph, specifically when there are less than 10 groups in the reference set.

Since 10 seems to be a good number for the size of the reference set, in Figure 7.9 we study rankings produced using different reference sets of size 10. From this figure we observe the behavior of P-score through a sliding window of size W, here $W = 10$ and thus the reference sets are $\{1, \cdots, 10\}$, $\{2, \cdots, 11\}$, $\cdots$, $\{N - 9, \cdots, N\}$. The results show that the method becomes less stable when we choose reference sets of size 10 in arbitrary positions of the NRC Ranking. There is clearly a trend of getting worse BPref values as we slide the reference set towards the bottom of the NRC ranking.

Our overall conclusion is that, as expected, using top groups as reference set is better than using non-top groups and that we need no more than 10 graduate programs in the reference set. This fact supports our claim that the reputation of venues is inherited from the reputation of groups publishing in them. It also makes easier the use of P-scores to rank venues of other areas, since it is not necessary to get data from a large amount of groups to produce reliable rankings of venues.

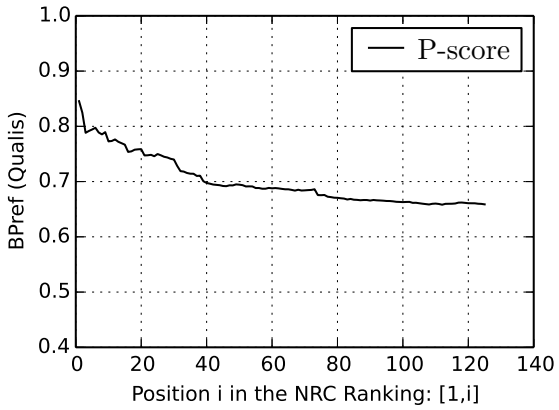

Figure 7.7: BPref of P-score venue rankings relative to Qualis produced by reference sets in the range $[1, i]$.

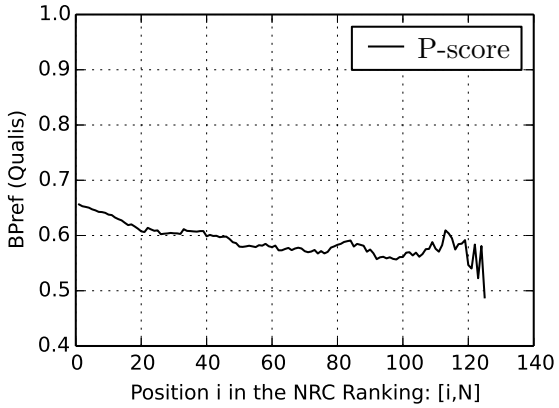

Figure 7.8: BPref of P-score venue rankings relative to Qualis produced by reference sets in the range $[i, N]$.

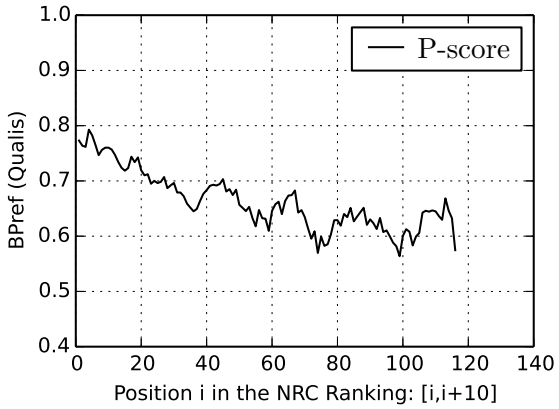

Figure 7.9: BPref of P-score venue rankings relative to Qualis produced by reference sets of size 10.

7.3.3 Coverage Analysis

In this section, we perform a coverage analysis. The objective here is to characterize the amount of reputation targets a given set of reputation sources can directly reach. As the proposed method is based on propagating reputation from a given set of reputation sources, it is important to know what is the expected size of the set of reputation sources. Such characterization is helpful to establish a better understanding of the method and of the application domain.

An intuitive approach to perform a coverage analysis would be plotting the number of reputation targets reached for each one of the compositions of the reputation sources. However, enumerating all the possible compositions of reputation sources is not viable — with N reputation source candidates, the number of distinct sets is $2^N - 1$. One alternative is to sample compositions of reputation sources. To provide a more concrete analysis, let us consider individual items such as authors and venues instead of sets of items.

In Figures 7.10 and 7.11, we consider authors as reputation sources and we inspect the number of venues (or targets) they reach. The x-axis in Figure 7.10 is the number of publications of a given author and the y-axis is the number of venues this author has published on. Notice that, as expected, the number of venues directly reached by authors with more publications is likely to be greater. Each dot in this graph represents an author and the line corresponds to the linear regression fitted using all authors. The slope of this line indicates how fast (on average) the number of venues increases with the number of publications. Specifically, the number of venues increases with a rate of 0.40 regarding the number of publications.

Considering all 5,213 venues and all 771,069 authors who have more than one publication in DBLP, Figure 7.11 displays the fraction of venues directly reached when we vary the sets of authors taken as reputation sources. The x-axis represents the fraction of authors (size of the set of reputation sources divided by the total number of authors in DBLP) and the y-axis consists of the fraction of venues directly reached when using a given set of reputation sources. To generate the values of x-axis, we sorted the set of authors in DBLP by their number of publications and a value in this axis represents a set of reputation sources composed of x% authors with most publications. Notice that it is possible to reach 80% of the venues using only 0.07% of the authors in the database — that is, 578 out of the 771,069 authors. Also, observe that this number of authors is an upper bound given that it may exist a set with a smaller number of authors with the same coverage. In fact, the minimum number of authors that reach all venues can be determined by modeling this problem as a set cover problem.

In Figure 7.12 and 7.13, we perform experiments that are similar to the aforementioned ones, considering however publication venues as reputation sources and measuring the number of authors directly reached. Specifically, the number of directly reached authors increases with a rate of 1.47 regarding the number of publication venues. Notice also that it is possible to reach 80% of the authors using just 1.75% of all venues in the dataset — that is, 91 out of 5,213 venues.

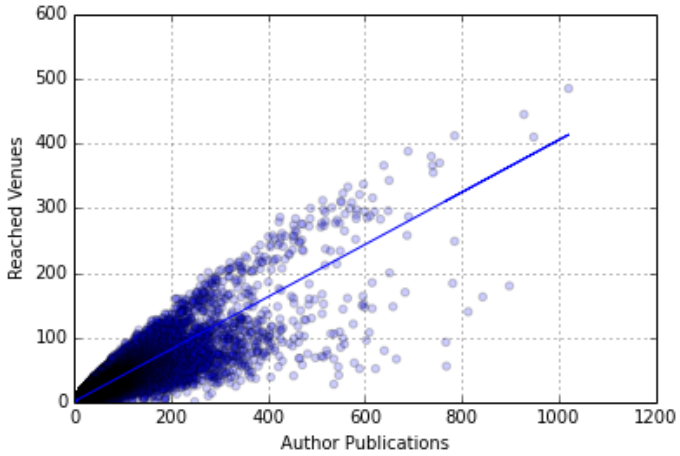

Figure 7.10: Relation between the number of publications and the number of distinct venues for all authors

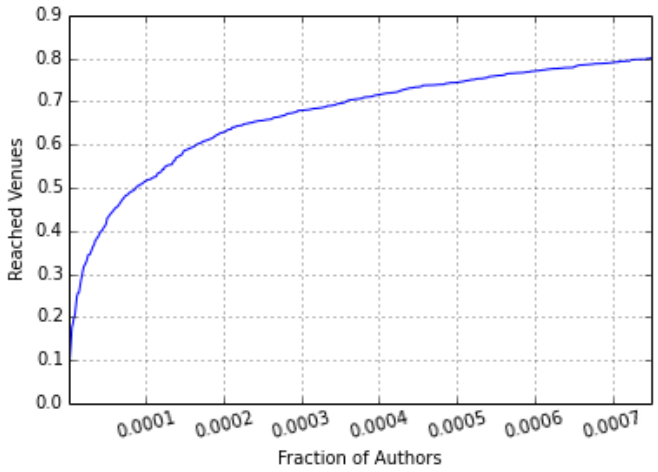

Figure 7.11: Fraction of venues reached by authors sorted by publication count

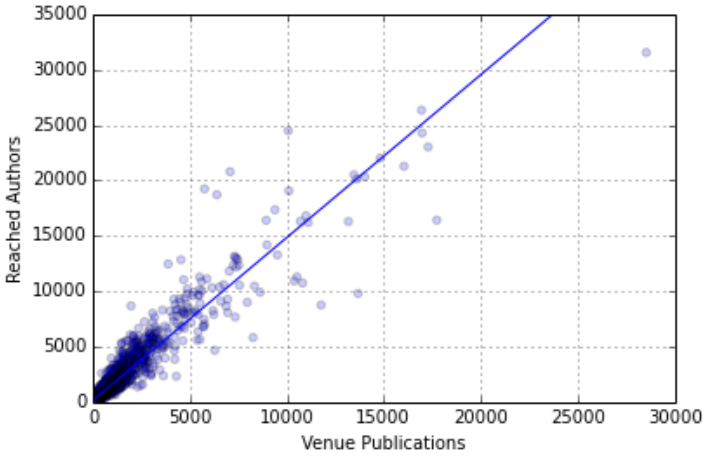

Figure 7.12: Relation between the number of publications and the number of distinct venues

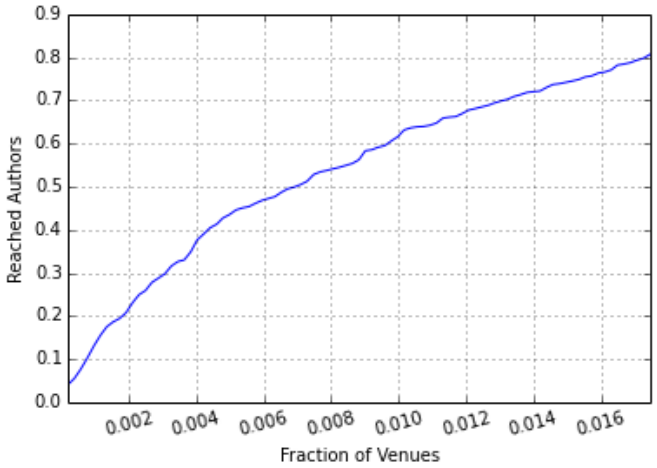

Figure 7.13: Fraction of authors reached by venues sorted by publication count

7.4 Ranking Robustness

The results in Chapter 6 attested the effectiveness of our proposed model for ranking academic entities when careful selections of reputation sources are performed. Nevertheless, any such selections may eventually include noisy reputation sources, making it sub-optimal. To address the robustness of our model, we assess the robustness of the rankings produced by P-scores with respect to random perturbations in the selected reputation sources.

Figure 7.14 shows the results of this investigation for venue rankings. In particular, the x-axis denotes the amount of noise randomly injected into a reference set of reputation sources—in our case, the top k research groups ranked by the NRC, for

$k \in \{5, 10, 20\}$. For instance, $x = 0.2$ indicates that 20% of the reputation sources are replaced by research groups randomly chosen from outside the reference set. Accordingly, $x = 0.0$ indicates no noise (i.e., the untouched top k NRC groups), whereas $x = 1.0$ indicates maximum noise (i.e., a random set of k research groups). On the y-axis, we show mean nDCG@100 figures averaged across 30 repetitions of this perturbation process, with shaded areas denoting the observed standard deviation from the mean. An additional curve including all 126 NRC groups as reputation sources is shown as a reference for comparison.

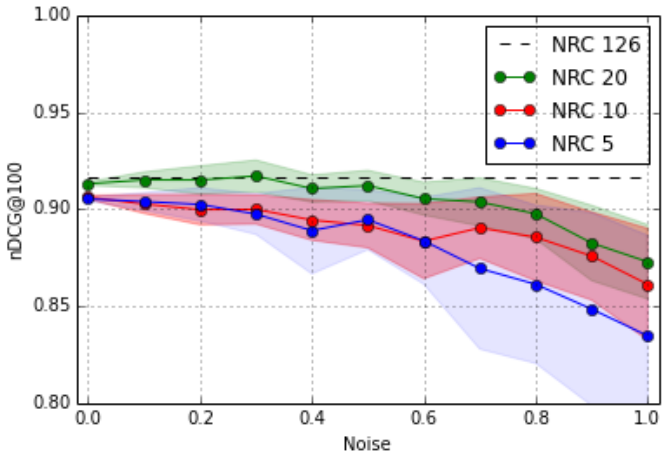

Figure 7.14: Venue ranking robustness with respect to random perturbations of the selected reputation sources.

From Figure 7.14, we observe that larger sets of reputation sources are generally more robust to noise, as demonstrated by the green curve (NRC 20). Indeed, this setting delivers nearly the same ranking effectiveness as the one achieved when using all 126 NRC groups as reputation sources. More importantly, all venue rankings produced by our model are relatively stable up to a noise level around 0.3 (i.e., when 30% of the reputation sources are randomly chosen). These results attest to the robustness of the rankings produced by our model with respect to random perturbations in the set of selected reputation sources.

7.5 Discussion on Reputation Sources

A reasonable question that may emerge is how this idea of reputation sources or reference items, impact reality. That is, are few entities, selected to compose a reputation set, sufficient to represent the overall scenario?

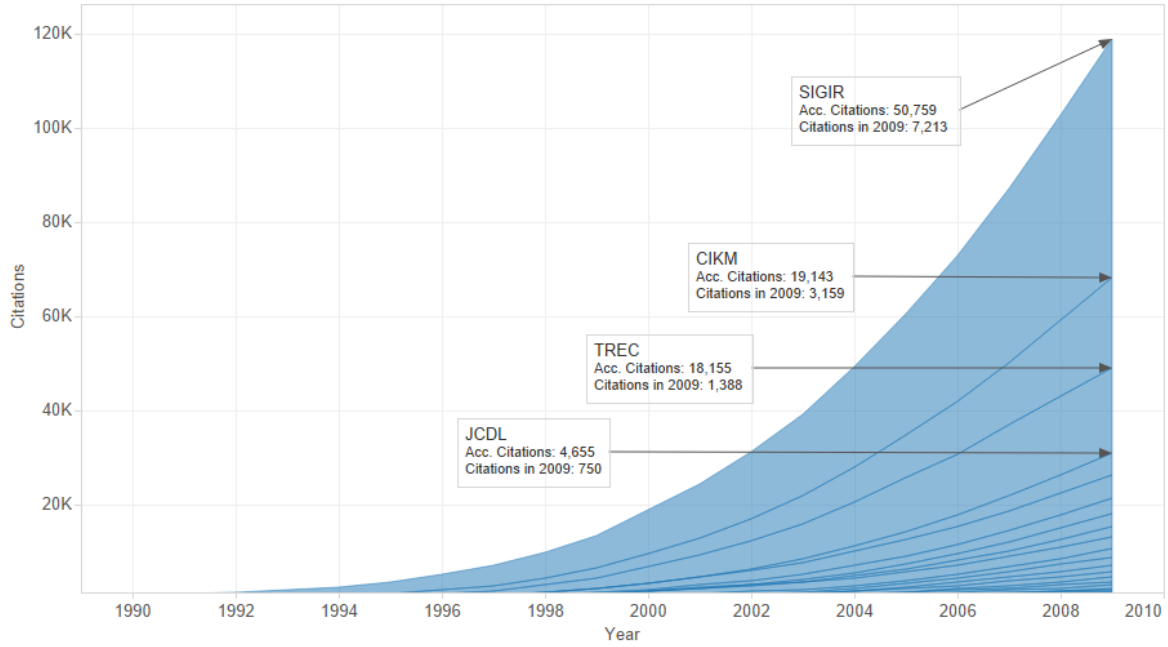

Figure 7.15: Citation count time series of IR publication venues

To better understand the implications of this question, consider the universe of publication venues in the sub-area of Information Retrieval. Specifically, let us analyze how the number of citations, one key variable of major interest in academia, varies overtime. In Figure 7.15, we show a time series with the accumulated number of citations of conferences in the sub-area of Information Retrieval. It is noteworthy how evident is the fact that just a few venues outgrow the other venues and dominate the sub-area, in terms of the accumulated number of citations received.

This fact is not an exclusivity of the Information Retrieval area, it can be observed in many other areas and sub-areas as well. Additionally, this kind of outgrowth or domination can be observed not only when looking at publication venues, but also when we analyze authors or even groups of authors. In fact, in most real life scenarios, including scenarios outside the academic ones, we can observe the fact that "*few have much and many have little*", see Pareto's Principle [Newman, 2005].

Another interesting aspect of P-score is that it is less prone to gaming. That is, while it is relatively easy for authors to hack some academic indicators, hacking P-score is not that easy (when the reputation sources are highly reputable entities). To illustrate, researchers can publish a high volume of papers in less restrictive venues to increase their total number of publications or, to boost their citation-based metrics, researchers can build citation farms with their colleagues. However, consistently publishing papers at the most reputable venues in a given area is a much harder task.

Chapter 8

Conclusions and Future Work

In this thesis, we have proposed a novel random walk model to identify the most reputable entities of a domain, based on a conceptual framework of reputation flows. Our model overcomes the challenges of quantifying reputation (arguably, a subjective and multi-faceted concept) by focusing on the transference of reputation among different entities. We instantiated our model in an academic search setting and empirically validated its effectiveness and robustness for three academic search tasks in the broad area of Computer Science, namely publication venues, individual researchers and research groups ranking. Specifically, we demonstrated the effectiveness of our model in contrast to standard citation-based approaches for identifying reputable venues, authors, and research groups, as well as its robustness to perturbations in the selection of reputation sources. Furthermore, we showed that effective reputation sources can be chosen in an automatic fashion using our proposed random walk model itself.

8.1 Summary of Contributions

In the following, we summarize the main contributions of this thesis.

A conceptual framework of reputation flows. In Chapter 3, we proposed a conceptual framework of reputation flows and a metric based on it, the P-score metric. This framework allow us to model the transference of reputation between distinct entities. To that end, the framework encodes the information of a given application scenario in what we call a reputation graph, which is composed of three types of nodes: the reputation sources, the reputation targets, and the reputation collaterals. The basic idea behind this framework, and its derived P-score metric, is to propagate the reputation from the reputation sources to the reputation targets through a random walk approach, and

then propagate the reputation from the targets to the reputation collaterals using a weighted sum.

A reputation-based method to rank academic entities. In Chapter 4, we proposed a reputation-based method to rank publication venues, individual researchers, and research groups, by instantiating our conceptual framework of reputation flows. In particular, we proposed distinct academic instantiations and we showed how to use the proposed P-score metric to produce rankings in such academic search settings. P-scores were able to produce consistently good results in the academic context. By using P-score we could produce rankings that are more effective than the ones produced by classic citation-based methods. This is of interest because citation-based metrics are used as a core feature in the assessment of academic productivity and also because retrieving the data necessary to compute P-score is easier than retrieving the data necessary to compute citation-based metrics.

An experimental evaluation on the properties of P-score. In Chapter 6 and Chapter 7, we provided an in depth evaluation of our conceptual framework by analysing its academic search instantiation. First, we evaluated the effectiveness of P-scores in three academic search tasks namely, venues ranking, authors ranking, and research groups ranking. To address these tasks, we used distinct configurations of our framework, as follows: using research groups as reputation sources, using authors as reputation sources, or using publication venues as reputation sources. Our intention of testing P-scores in many tasks and using many configurations is to better understand how our method behaves in distinct scenarios. Next, we evaluated the robustness of our method. To do so, we introduced some noise in the set of reputation sources and measured the impact of each noise level in the quality of the final ranking. Also, we explored distinct sets of reputation sources in order to characterize what defines a good set of reputation sources. Using the knowledge learned in the aforementioned characterization, we evaluated distinct approaches to select effective reputation sources in an automatic fashion.

8.2 Summary of Conclusions

In the following, we summarise the conclusions of this thesis.

On the effectiveness of P-score. In Chapter 6, we provide some evidences regarding the effectiveness of the proposed model. Specifically, we instantiated our model in an academic search setting and empirically validated its effectiveness for three academic

search tasks in the broad area of Computer Science, namely ranking of publication venues, ranking of individual researchers, and ranking of research groups. For this, we ran extensive experimentation in which we explored the adoption of selected research groups, selected individual authors, and selected venues as reputation sources for each of the three search tasks of our interest, in a total of 9 configurations. While the use of selected individual authors and selected venues as reputation sources relies on citation data, the use of selected research groups as reputation sources does not depend on citation data nor in manual intervention.

On the robustness of P-score. In Chapter 6, we attest the effectiveness of our proposed model for careful selections of reputation sources. Nevertheless, this selection may eventually include noisy reputation sources, making it sub-optimal. In Chapter 7, we provide some evidences regarding the robustness of P-score in an academic search setting. Our results suggest that the model is indeed robust to random perturbations, all rankings produced by our model were relatively stable up to a noise level around 0.3 (when 30% of the reputation sources are randomly chosen). We also observed that larger sets of reputation sources are generally more robust to noise. These results attest the robustness of the rankings produced by our model with respect to random perturbations in the selected reputation sources. Moreover, they open up an interesting direction towards automatically identifying a robust set of reputation sources.

On the selection of reputation sources. Motivated by the effectiveness and robustness of P-score, in Chapter 7, we discuss a critical step of our method, the selection of the reputation sources. While our model was primarily developed to study how the reputation from a manually pre-selected set of reputation sources is transferred to other entities in a reputation graph, in this chapter, we show how to automatically emerge with effective reputation sources from the data, without manual effort.

8.3 Directions for Future Research

Both the conceptual framework and its instantiation in an academic context open opportunities for future work.

Model Level. In this thesis, we show how distinct entities of an academic search setting can be consistently modeled as a reputation graph. However, this is not the only possible application scenario that can be modeled by our framework. Indeed, our conceptual framework allows the modeling of any context in which we could identify flows of reputation between distinct types of entities. Therefore, at the model level,

a future direction is to verify the generality of the concept of reputation flows when applied to other domains, such as enterprise search.

Another direction is to deeper explore visualization and clustering techniques to help both manual and automatic approaches to identify suitable reputation sources. By better visualizing a reputation graph, one could improve her manual selection of reputation sources. Clustering techniques may be applied to enhance automatic selections of reputation sources. In many applications it is desirable that the set of reputation sources covers as many regions as possible of the reputation graph.

Instantiation Level. At the instantiation level, a future direction is to test our model for academic search tasks in areas other than Computer Science, like Biochemistry, Economics, History, among others. Applying our model in Computer Science sub-areas is important as well, some examples are: Information Retrieval, Databases, Computer Networks, and others.

Bibliography

Backstrom, L. and Leskovec, J. (2011). Supervised random walks: Predicting and recommending links in social networks. In *Proc. of ACM International Conference on Web Search and Data Mining*, pages 635--644.

Baeza-Yates, R. and Ribeiro-Neto, B. (2011). *Modern Information Retrieval*. Pearson.

Bakshy, E., Rosenn, I., Marlow, C., and Adamic, L. (2012). The role of social networks in information diffusion. In *Proc. of WWW International World Wide Web Conference*, pages 519--528.

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

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

Benevenuto, F., Laender, A. H., and Alves, B. L. (2016b). How connected are the acm sig communities? *ACM SIGMOD Record, Special Interest Group on Management of Data*, 44(4):57--63.

Bergstrom, C. T., West, J. D., and Wiseman, M. A. (2008). The eigenfactor$^{\text{TM}}$ metrics. *Journal of Neuroscience*, 28(45):11433--11434.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993--1022. ISSN 1532-4435.

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.

Boongoen, T., Shen, Q., and Price, C. (2011). Fuzzy qualitative link analysis for academic performance evaluation. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 19(03):559--585.

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

Castillo, C., Donato, D., and Gionis, A. (2007). Estimating number of citations using author reputation. In *Proc. of String Processing and Information Retrieval*, pages 107--117.

Classen, J., Braun, J., Volk, F., Hollick, M., Buchmann, J., and Mühlhäuser, M. (2015). A distributed reputation system for certification authority trust management. In *IEEE Trustcom/BigDataSE/ISPA*, volume 1, pages 1349--1356.

Cormode, G., Muthukrishnan, S., and Yan, J. (2014). People like us: mining scholarly data for comparable researchers. In *Proc. of WWW International World Wide Web Conference*, pages 1227--1232.

Delaviz, R., Andrade, N., and Pouwelse, J. A. (2010). Improving accuracy and coverage in an internet-deployed reputation mechanism. In *International Conference on Peer-to-Peer Computing (P2P)*, pages 1--9.

Delgado-Garcia, J. F., Laender, A. H., and Meira, W. (2014a). Analyzing the coauthorship networks of latin american computer science research groups. In *Proc. of Latin American Web Congress*, pages 77--81.

Delgado-Garcia, J. F., Laender, A. H., and Meira Jr, W. (2014b). A preliminary analysis of the scientific production of latin american computer science research groups. In *Proc. of Alberto Mendelzon International Workshop on Foundations of Data Management*.

Deng, H., Han, J., Lyu, M. R., and King, I. (2012). Modeling and exploiting heterogeneous bibliographic networks for expertise ranking. In *Proc. of ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 71--80.

Dhanjal, C. and Clémençon, S. (2014). Learning reputation in an authorship network. In *Proceedings of the 29th Annual ACM Symposium on Applied Computing*, pages 1724--1726.

Dores, W., Benevenuto, F., and Laender, A. H. (2016). Extracting academic genealogy trees from the networked digital library of theses and dissertations. In *Proc. of ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 163--166.

Figueira, P., Pacheco, G., Almeida, J. M., and Gonçalves, M. A. (2015). On the impact of academic factors on scholar popularity: A cross-area study. In *International Conference on Theory and Practice of Digital Libraries*, pages 139--152. Springer.

Franceschini, F. and Maisano, D. (2011). Structured evaluation of the scientific output of academic research groups by recent h-based indicators. *Journal of Informetrics*, 5(1):64--74.

García-Romero, A., Santín, D., and Sicilia, G. (2016). Another brick in the wall: a new ranking of academic journals in economics using fdh. *Scientometrics*, 107(1):91--101.

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

Garfield, E. (1972). Citation analysis as a tool in journal evaluation: Journals can be ranked by frequency and impact of citations for science policy studies. *Science*, 178:471–479.

Gkorou, D., Vinkó, T., Pouwelse, J., and Epema, D. (2013). Leveraging node properties in random walks for robust reputations in decentralized networks. In *Proc. of International Conference on Peer-to-Peer Computing (P2P)*, pages 1--10.

Gollapalli, S. D., Mitra, P., and Giles, C. L. (2011). Ranking authors in digital libraries. In *Proc. of ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 251--254.

Gonçalves, G. D., Figueiredo, F., Almeida, J. M., and Gonçalves, M. A. (2014). Characterizing scholar popularity: A case study in the computer science research community. In *Proc. of ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 57--66.

Gonçalves, M. A. (2016). Dissecting a scholar popularity ranking into different knowledge areas. In *International Conference on Theory and Practice of Digital Libraries*, volume 9819, page 253. Springer.

Hamedani, M. R. and Kim, S.-W. (2016). Simrank and its variants in academic literature data: measures and evaluation. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing*, pages 1102--1107.

Hayes, B. et al. (2013). First links in the markov chain. *American Scientist*, 101(2):252.

Hirsch, J. (2005). An index to quantify an individual's scientific research output. *Proc. of the National Academy of Sciences*, pages 16569–16572.

Hutchins, B. I., Yuan, X., Anderson, J. M., and Santangelo, G. M. (2016). Relative citation ratio (rcr): A new metric that uses citation rates to measure influence at the article level. *PLoS Biol*, 14(9):e1002541.

Hutton, J. G., Goodman, M. B., Alexander, J. B., and Genest, C. M. (2001). Reputation management: the new face of corporate public relations? *Public Relations Review*, 27(3):247--261.

Jamali, M. and Ester, M. (2009). Trustwalker: A random walk model for combining trust-based and item-based recommendation. In *Proc. of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 397--406.

Järvelin, K. and Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422--446.

Jeh, G. and Widom, J. (2002). Simrank: a measure of structural-context similarity. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, pages 538--543.

Katerattanakul, P., Han, B., and Hong, S. (2003). Objective quality rankings of computing journals. *Communications of the ACM*, 45.

Kim, H. J., An, J., Jeong, Y. K., and Song, M. (2016). Exploring the leading authors and journals in major topics by citation sentences and topic modeling. In *Proc. of ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 42--50.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604--632.

Konstas, I., Stathopoulos, V., and Jose, J. M. (2009). On social networks and collaborative recommendation. In *Proc. of ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 195--202.

Laender, A. H. F., de Lucena, C. J. P., Maldonado, J. C., de Souza e Silva, E., and Ziviani, N. (2008). Assessing the research and education quality of the top brazilian computer science graduate programs. *ACM Special Interest Group on Computer Science Education Bulletin*, 40(2):135–145.

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

Larsen, B. and Ingwersen, P. (2006). Using citations for ranking in digital libraries. In *Proc. of ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 370--370.

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

Li, H. (2011). A short introduction to learning to rank. *IEICE Transactions, The Institute of Electronics, Information and Communication*, pages 1854–1862.

Liang, R. and Jiang, X. (2016). Scientific ranking over heterogeneous academic hypernetwork. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 20--26.

Lima, H., Silva, T. H., Moro, M. M., Santos, R. L., Meira Jr, W., and Laender, A. H. (2015). Assessing the profile of top brazilian computer science researchers. *Scientometrics*, 103(3):879--896.

Lima, H., Silva, T. H. P., Moro, M. M., Santos, R. L. T., Jr., W. M., and Laender, A. H. F. (2013). Aggregating productivity indices for ranking researchers across multiple areas. In *Proc. of ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 97--106.

Mangaravite, V. and Santos, R. L. (2016). On information-theoretic document-person associations for expert search in academia. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 925--928.

Mangaravite, V., Santos, R. L., Ribeiro, I. S., Gonçalves, M. A., and Laender, A. H. (2016). The lexr collection for expertise retrieval in academia. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 721--724.

Mann, G., Mimno, D., and McCallum, A. (2006). Bibliometric impact measures leveraging topic analysis. In *Proc. of ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 65--74.

Markov, A. A. (2006). An example of statistical investigation of the text eugene onegin concerning the connection of samples in chains. *Science in Context*, 19(04):591--600.

Meyer, C. (1989). Stochastic complementation, uncoupling Markov chains, and the theory of nearly reducible systems. *SIAM Review*, 31(2):240--272.

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

Nerur, S., Sikora, R., Mangalaraj, G., and Balijepally, V. (2005). Assessing the relative influence of journals in a citation network. *Communications of the ACM*, 48:71--74.

Newman, M. (2010). *Networks: An Introduction*. Oxford University Press.

Newman, M. E. (2005). Power laws, pareto distributions and zipf's law. *Contemporary physics*, 46(5):323--351.

Nezhadbiglari, M., Gonçalves, M. A., and Almeida, J. M. (2016). Early prediction of scholar popularity. In *Proc. of ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 181--190.

Nie, Z., Zhang, Y., Wen, J.-R., and Ma, W.-Y. (2005). Object-level ranking: Bringing order to web objects. In *Proc. of WWW International World Wide Web Conference*, pages 567--574.

NRC (2010). United States National Research Council, National Academy of Sciences.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1998). The pagerank citation ranking: Bringing order to the web. In *Proc. of WWW International World Wide Web Conference*, pages 161--172.

Pearson, K. (1905). The problem of the random walk. *Nature*, 72:342--342.

Piwowar, H. (2013). Altmetrics: Value all research products. *Nature*, 493(7431):159--159.

Pouwelse, J. A., Garbacki, P., Wang, J., Bakker, A., Yang, J., Iosup, A., Epema, D. H., Reinders, M., Van Steen, M. R., Sips, H. J., et al. (2008). Tribler: A social-based peer-to-peer system. *Concurrency and computation: Practice and experience*, 20(2):127.

Pradhan, D., Paul, P. S., Maheswari, U., Nandi, S., and Chakraborty, T. (2016). C3-index: A pagerank based multi-faceted metric for authors' performance measurement. *arXiv preprint arXiv:1610.07061*.

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

Ribas, S., Ribeiro-Neto, B., de Souza e Silva, E., Ueda, A. H., and Ziviani, N. (2015a). Using reference groups to assess academic productivity in computer science. In *Proc. of WWW International World Wide Web Conference*, pages 603--608.

Ribas, S., Ribeiro-Neto, B., Santos, R. L., de Souza e Silva, E., Ueda, A., and Ziviani, N. (2015b). Random walks on the reputation graph. In *Proc. of ACM International Conference on the Theory of Information Retrieval*, pages 181--190.

Ribas, S., Ribeiro-Neto, B., Santos, R. L., de Souza e Silva, E., Ueda, A., Ziviani, N., and Dias, M. (2017). Reputation flows in academia. *Journal of American Society for Information Science and Technology*, (submitted).

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.

Sastry, C. S., Jagaluru, D. S., and Mahesh, K. (2016). Author ranking in multi-author collaborative networks. *Collnet Journal of Scientometrics and Information Management*, 10(1):21--40.

Silva, T. H., Laender, A. H., Davis Jr, C. A., da Silva, A. P. C., and Moro, M. M. (2016). The impact of academic mobility on the quality of graduate programs. *D-Lib Magazine*, 22(9/10).

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B., and Wang, K. (2015). An overview of microsoft academic service (mas) and applications. In *Proc. of WWW International World Wide Web Conference*.

Sun, Y. and Giles, C. L. (2007). *Popularity weighted ranking for academic digital libraries*. Springer.

Tang, J., Jin, R., and Zhang, J. (2008). A topic modeling approach and its integration into the random walk framework for academic search. In *Proc. of IEEE International Conference on Data Mining*, pages 1055--1060.

Viswanath, B., Mislove, A., Cha, M., and Gummadi, K. P. (2009). On the evolution of user interaction in facebook. In *Proceedings of the ACM workshop on Online social networks*, pages 37--42.

Walters, P. (2000). *An introduction to ergodic theory*, volume 79. Springer Science & Business Media.

Weng, J., Lim, E.-P., Jiang, J., and He, Q. (2010). Twitterrank: Finding topic-sensitive influential twitterers. In *Proc. of ACM International Conference on Web Search and Data Mining*, pages 261--270.

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.

Xi, W., Zhang, B., Chen, Z., Lu, Y., Yan, S., Ma, W.-Y., and Fox, E. A. (2004). Link fusion: A unified link analysis framework for multi-type interrelated data objects. In *Proc. of WWW International World Wide Web Conference*, pages 319–327.

Yan, E., Ding, Y., and Sugimoto, C. R. (2011). P-rank: An indicator measuring prestige in heterogeneous scholarly networks. *Journal of the American Society 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 *Proc. of ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 235--244.

Ye, M., Yin, P., Lee, W.-C., and Lee, D.-L. (2011). Exploiting geographical influence for collaborative point-of-interest recommendation. In *Proc. of ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 325--334.

Yu, K. and Chen, J. (2012). Reputation based academic evaluation in a research platform. *Journal of Software*, 7(12):2749--2754.

Zhou, D., Orshanskiy, S. A., Zha, H., and Giles, C. L. (2007). Co-ranking authors and documents in a heterogeneous network. In *Proc. of IEEE International Conference on Data Mining*, pages 739--744.

Zhuang, Z., Elmacioglu, E., Lee, D., and Giles, C. (2007). Measuring conference quality by mining program committee characteristics. In *Proc. of ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 225--234.

Appendix A

Web Tool

We developed a web tool that allows users to quickly grasp insights about publication venues, academics and research groups by using P-score. To use the tool, access:

`http://pscore.dcc.ufmg.br`

The URL above gives access to the web page shown in Figure A.1.

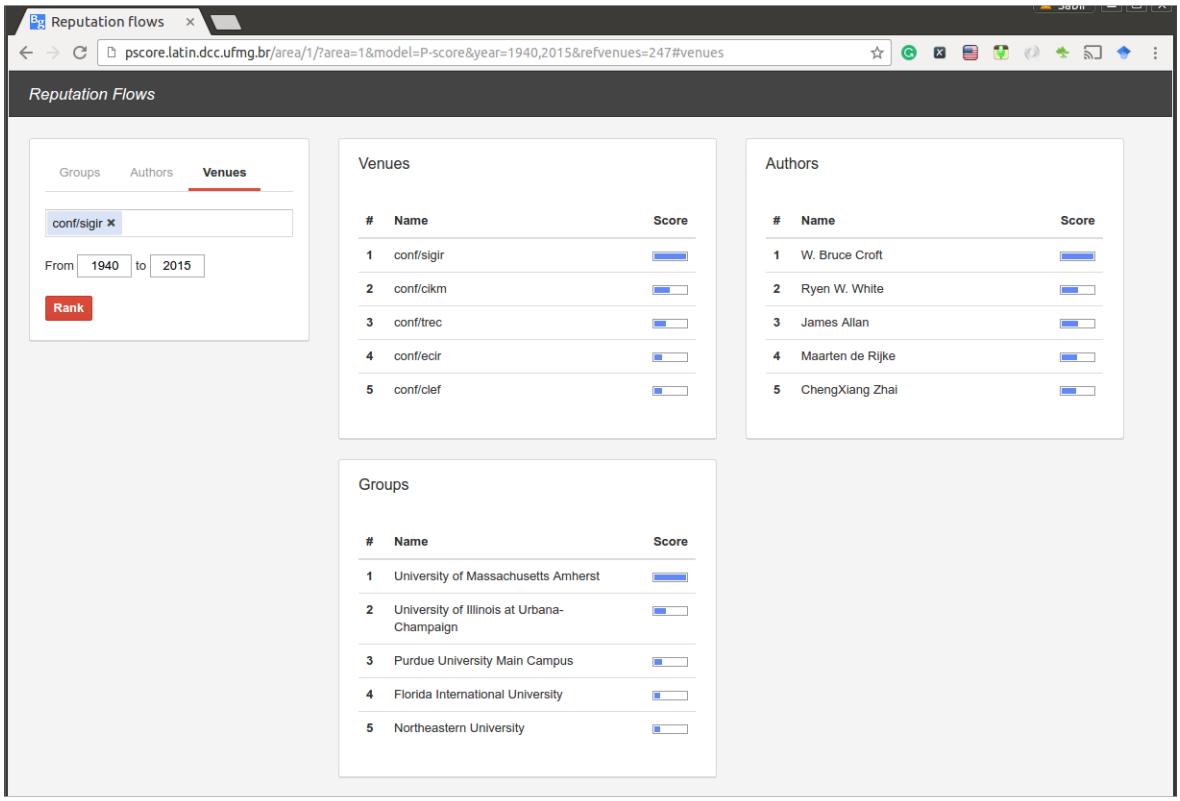

Figure A.1: P-score web tool

Next we provide a brief tutorial one may follow to use our P-score tool to analyse how reputation flows through the network:

1. Select whether the set of reputation sources will be composed by research groups, authors, or publication venues.
2. Choose the set of reputation sources using the search box at the left hand side.
3. Select the time range to be considered in the analysis.
4. Click the *Rank* button.

Once the *Rank* button is clicked, the corresponding reputation graph is instantiated and the rankings of venues, authors and groups are produced using P-score. The interface shows the top 5 positions of each ranking, to see more positions just click in the ranking you want to take a closer look. To illustrate, in Figure A.1, the rankings were produced by choosing the SIGIR conference as the only venue to compose the set of reputation sources. Notice that other conferences could be easily added to the set.

Appendix B

Academic Repositories

There are publication repositories in the Web from which we can collect academic data. In some cases we can retrieve the values of certain indicators directly from a publication repository and, in others, we may need to derive the values of an indicator from existing data in a repository (e.g., there is no authors' H-Index in Microsoft Academic Search, but we can compute this value by collecting the list of publications of an author with their corresponding citation counts). In this section, we point out some facts in regard to citation data in existing repositories.

In here, we focus our analysis in Google Scholar[1] (GS) and Microsoft Academic Search[2] (MAS) because they are largely adopted as source of academic data. One of the main advantages of MAS, in comparison with GS, is that it provides a dump of the entire database[3]. However, MAS has not being properly updated since 2010. That is, we can download an outdated MAS repository. It may be a good dataset for studying behavioural aspects of a certain method (e.g. for citation prediction), but it is not a good data source if our intent is to evaluate current scientific production of academics or to predict the number of citations an author, paper or venue will receive in the next year. While there is no GS dump available for download, the repository has being properly updated over the years. This fact makes GS a better choice in some scenarios.

In Table B.1, we compare some features of the two major repositories in regard to the availability of venue data. We check-marked a cell when the feature in the corresponding column is present in the repository. We notice that we can directly retrieve the number of citations of publication venues from MAS. From GS, we can directly retrieve the H-Index of publication venues. Neither MAS nor GS provides

[1] `http://scholar.google.com`
[2] `http://academic.research.microsoft.com`
[3] `http://datamarket.azure.com/dataset/mrc/microsoftacademic`

the Impact Factor of publication venues. Both repositories present a list with the top ranked venues.

Table B.1: Venue data availability in academic repositories

Repository	Cits	H-Index	IF	List	Dump	Updated
GS		✓		✓		✓
MAS	✓			✓	✓	

In Table B.2, we illustrate the metadata of the two most cited journals and conferences in Computer Science according to Microsoft Academic Search. The number of publications and the number of citations were collected from MAS, while the H-Index was collected from GS. The year of the first publication were collected from DBLP and the Impact Factor of the journals was collected from Cite Factor[4] and Impact Factor Search[5], which are mirrors of Thomson Reuters[6]. We notice, currently, that Thomson Reuters do not compute the Impact Factor for conferences and, as far as we know, there is no repository with conferences' Impact Factor.

Table B.2: Metadata collected from various data sources of the two most cited journals and conferences in Computer Science according to Microsoft Academic Search

Venue	Pubs	Cits	H5-Index[7]	1*st* year	IF
TIT	14742	387985	93	1953	2.650
CACM	13102	361257	77	1958	2.863
INFOCOM	6556	190568	76	1982	
ICRA	17210	181834	61	1984	

In Table B.3, we compare some features of the two major publication repositories in regard to the availability of author data. We check-marked a cell when the feature in the corresponding column is present in the repository. We notice that we can directly retrieve the number of citations of publication authors from both repositories. The H-Index of authors can be directly retrieved from GS. Only MAS presents a list with the top ranked authors.

While both the GS and MAS repositories provide the number of citations of individual researchers, we observe relatively large differences between the values presented by each. While citation counts are underestimated in MAS, they are inflated in GS.

[4] `http://www.citefactor.org`

[5] `http://www.impactfactorsearch.com`

[6] `http://thomsonreuters.com/journal-citation-reports`

[7] H5-Index is the specification of H-Index provided by Google Scholar in which only publications of the last 5 years are considered: `http://scholar.google.com.br/citations?view_op=top_venues`

Table B.3: Author data availability in academic repositories

Repository	Cits	H-Index	List	Dump	Updated
GS	✓	✓			✓
MAS	✓		✓	✓	

For example, while W. Bruce Croft has 13,733 citations in MAS, in GS he has 34,354. Another example is Andrei Shleifer who has 44,758 in MAS and 182,944 in GS. MAS considers citations from a set of selected venues only, while GS counts all types of citations — for example, citations from a technical report or from a web page. Despite that, there is an argument that while GS citations are more inflated, they are more consistent if we restrict the counting to GS.

When ranking venues, Impact Factor (IF) is an important metric to be considered. The reason we do not use it here is that we could not get the official IF values reported by Thomson Reuters for most publication venues in our analysis. Although in theory, nothing prevents us from calculating this metric, the input data needed to compute it is hard to obtain. Even Thomson Reuters uses only a small fraction of the needed data. This fact leads to rough approximations and many venues with missing IF.

There is no standard repository for research groups, and probably it will never happen. The reason is that the definition of what configure a research group may vary according to the context. Despite that, for some definitions of research groups, it is possible to retrieve from existing repositories the list of their publications, or even the values of some metrics (including citation-based ones), by directly using the group names. Repositories like Microsoft Academic Search may lead us from the group name to distinct bibliometric measures if we define that research groups are represented by all people related to an university, for instance, that published in a given area or sub-area of knowledge. To illustrate, in MAS repository, if we search for "information retrieval" we may find institutions like Stanford, MIT, CMU, and others, and we can treat them as research groups. But, publication repositories may not work if we increase the granularity and ask for more precise definitions of research groups, such as the professors of graduate programs that work in an arbitrary area of knowledge. In this case, data like publication lists or citation counts needs to be derived from the names of the researchers that compose the groups of interest.

As we discuss in this chapter, there are many citation-based metrics proposed in the literature. Each one of them tries to solve (or to soften) the problems of previously proposed ones, or to capture some unexplored semantics. However, these indicators can not be used when the existing repositories fail to provide citation data.

Appendix C

Governmental Evaluations

This chapter is composed of four sections, each of which discusses a governmental effort of evaluating research in the context of academia. In the first three sections, we detail governmental evaluations of publication venues, researchers and graduate programs run by funding agencies in Brazil. In the last section, we detail how the National Research Council evaluates graduate programs in the United States. This chapter is important because we compare the results of the proposed metric with these governmental evaluations, which are produced through much more complex processes.

C.1 Qualis Classification of Publication Venues

The Brazilian funding agency CAPES[1] maintains a method of systematically evaluating publication venues and provides a classification called Qualis. The venue evaluations provide estimates of the relative importance of publication venues in a given area of knowledge. CAPES assigns a committee of experts to each area of knowledge and these experts are responsible for evaluating all information acquired about the venues and produce a classification. According to the Qualis method, for each area of knowledge, the venues are classified (in order of importance) as A1, A2, B1, B2, B3, B4, B5 or C.

The classifications are updated annually and follow a set of criteria, such as: the number of publications in the venue, the number of databases in which it is indexed, the amount of institutions publishing in the venue, citation information whenever available, among others. Each committee decides which criteria to be considered when evaluating their respective knowledge area. For the area of Computer Science, the classification of journals is heavily influenced by citation data (whenever available) and the classification of venues is heavily influenced by H-Index information (whenever available).

[1] `http://www.capes.gov.br`

Table C.1: Classification of venues in Computer Science (CS) according to Qualis

Qualis	Conferences	Journals
A1	115	38
A2	140	40
B1	294	72
B2	203	77
B3	246	37
B4	464	24
B5	192	29
Total	1654	317

In Table C.1, we show the distribution of the publication venues in Qualis[2] classification for the area of Computer Science in the year of 2013.

C.2 CNPq Productivity Levels for Researchers

The CNPq is a well established agency dedicated to the promotion of scientific and technological research and to the formation of human resources for research in Brazil. One of the programs of CNPq is to distribute research productivity grants to individual researchers, with the main goal of stimulating research in the country. To distribute the grants, CNPq classifies researchers in 5 different levels of productivity: 1A, 1B, 1C, 1D, and 2, in descending order of prestige.

In Table C.2, we show the distribution of researchers according to CNPq productivity levels[3] for the area of Computer Science and for all areas evaluated for CNPq. The data in this table is based on all researchers with active grants in the year of 2014.

To receive a productivity grant, a researcher needs to submit to CNPq information about him/her academic career, including a research project to be developed during the next years. The starting point for any new researcher is the productivity level 2. To assign a research grant to a researcher, CNPq evaluates a set criteria and productivity indicators such academic output, contribution to the formation of human resources, academic leadership, among others.

[2] The current Qualis classification of distinct areas is available at the official web page of CAPES: `http://qualis.capes.gov.br/webqualis/publico/documentosDeArea.seam?conversationPropagation=begin`

[3] The current CNPq productivity levels of researchers is available at: `http://plsql1.cnpq.br/divulg/RESULTADO_PQ_102003.prc_comp_cmt_links?V_COD_DEMANDA=200310&V_TPO_RESULT=CURSO&V_COD_AREA_CONHEC=10300007&V_COD_CMT_ASSESSOR=CC`

Table C.2: Distribution of researchers according to CNPq productivity levels for the area of Computer Science (CS)

Level	Researchers	
	CS	All Areas
1A	23	1320
1B	22	1308
1C	31	1376
1D	70	2386
2	244	7933
Total	390	14323

C.3 CAPES Classification of Graduate Programs

A well structured effort to evaluate graduate programs is the CAPES ranking in Brazil, which has been evaluating and comparing graduate programs since 1977, on a triennial basis [Laender et al., 2008]. The process conducted by CAPES takes into account various quantitative and qualitative parameters such as coverage of courses' contents, curriculum vitae of professors, international reputation, and publication records. One of the key parameters is the publication records both in volume and in quality.

CAPES committees run a thorough comparative analysis of the publication records (based on the Qualis classification of publication venues) of the professors in each major department in Brazil to establish a classification of the graduate programs. Each graduate program receives a grading in a scale of 1-7, where 7 is the highest classification. CAPES considers that programs with a rank of 5 or higher are elite, i.e., the ones that are awarded with funds from special programs.

In Table C.3, we show the CAPES classification[4] of Brazilian graduate programs in the area of Computer Science and in all areas for the year of 2013. In that year, CAPES classified 3,337 graduate programs in Brazil, with 68 of them belonging to the area of Computer Science. That is, the graduate programs in Computer Science correspond to approximately 2% of the total number of graduate programs evaluated. Of these graduate programs, only 13 programs have a CAPES classification of 5 or higher, as shown in Table C.4.

[4]The CAPES classification is available at: `http://www.capes.gov.br/cursos-recomendados`

Table C.3: CAPES classification of graduate programs in Computer Science (CS)

Class	Programs	
	CS	All Areas
7	5	145
6	3	270
5	5	610
4	21	1216
3	33	1047
2	1	43
1	0	6
Total	68	3,337

Table C.4: Top ranked graduate programs in Brazil for the area of Computer Science, according to CAPES

Class	Programs
7	PUC-Rio, UFMG, UFRGS, UFRJ, UNICAMP
6	UFPE, USP, USP-SC
5	PUC-RS, UFAM, UFC, UFF, UFRN

C.4 Rankings of the US National Research Council

The National Research Council (NRC) issues a ranking of the graduate programs in US in many areas of knowledge, aiming at providing a guideline for students and administrators.[5] In the area of Computer Science, the NRC classified the major 126 graduate programs in the US using two approaches: S-rankings and R-rankings [NRC, 2010]. We will focus on the R-rankings here, it suffices for our purposes.

For R-rankings, NRC asked a sample of faculty to rate a sample of programs in their field. These rankings were then used to assign weights to a set of 20 features (such as number of publications, number of citations per paper, percentage of faculty with grants) through regression analysis. The weighted features were then used to rank all programs. To account for uncertainty and variability in the surveys data, the regression analysis process was repeated 500 times, each time using as input a random sample of half of the surveys. As a result, 500 ranks were produced for each one of the 126 programs. Following, for each program, the top 5% and the bottom 5% ranks were disregarded, leaving each program with 450 ranks which were sorted. The top rank is

[5] According to the NRC report, *"These illustrative rankings should not be interpreted as definitive conclusions about the relative quality of doctoral programs, nor are they endorsed as such by the National Research Council. Rather, they demonstrate how the data can be used to rank programs based on the importance of particular characteristics to various users."*

Table C.5: CS programs likely to be at top 10 in the R-ranking of NRC.

Rank		Program
5^{th}perc.	95^{th}perc.	
1	2	Stanford University
1	2	University of California-Berkeley
3	5	Massachusetts Institute of Technology
3	10	Carnegie Mellon University
3	18	University of Illinois at Urbana-Champaign
4	14	Princeton University
5	17	Cornell University
5	17	University of California-Santa Barbara
5	17	University of North Carolina at Chapel Hill
5	29	University of California-Los Angeles
6	22	University of Texas at Austin
6	28	Georgia Institute of Technology
6	28	Michigan State University
7	25	University of Maryland College Park
8	27	University of California-San Diego
8	27	University of Michigan-Ann Arbor
8	30	University of Southern California
8	26	Penn State University
9	36	University of Massachusetts, Amherst
10	27	University of Wisconsin-Madison
10	35	Harvard University

what is referred as the *program rank at the* 5^{th} *percentile*. The rank at the bottom is what is referred to as the *program rank at the* 95^{th} *percentile*.

To illustrate, in Table C.5 we list the programs with ranks smaller than 10 (i.e., the programs likely to be at top 10) at the 5^{th} percentile in the year of 2010. That is, at the 5^{th} percentile, there are 21 programs with a rank smaller or equal than 10 in the area of Computer Science. In addition, either Stanford or Berkeley may occupy the top position, while MIT, CMU or Illinois may occupy position 3 in the ranking.

Appendix D

P-score in Computer Science

In this appendix we take a closer look in the application of P-score in the computer science domain.

Given that our method is based on reference groups, it is appropriate at this point to recall and keep in mind the answer to the following question: *Why to use reference groups and how to choose them?* We use reference groups because they provide a natural way to produce relative comparisons. By computing the similarity of the research output of a group of authors with reference groups, we can gain insights on the productivity of these authors in a certain area or sub-area of knowledge.

The choice of a reference group depends on what we want to measure. In the following sections of this appendix, we discuss how the reputation of reliable research groups is transferred to publication venues, other authors and research groups in the broad area of Computer Science. In Sections D.1 and D.2, we discuss how the reputation of top research groups and highly cited authors in many sub-areas of Computer Science is transferred to publication venues, other authors and research groups. Next, in Appendix E, we discuss how the reputation of highly cited authors in the sub-area of Information Retrieval is transferred to publication venues, other authors and research groups. Thus, we choose reference groups appropriately for these purposes.

D.1 Top Research Groups as Reputation Sources

In this section we adopt the top research groups according to the randomized process proposed in Section 7.2.2 as shown in Table 7.1. We use the set of 12 research groups in Table 7.1 as the set of reputation sources to rank publication venues, individual researchers and other research groups. Recall that those groups are the ones that appeared among the top 10 at least once in a given run, after the process stabilized.

D.1.1 Publication Venues

In Table D.1, we present the top 25 venues in the ranking produced by P-score using, as reference, the research groups discussed in the previous section. We notice that only one of these top 25 venues according to P-score did not receive a Qualis classification — a precision of 96% at position 25 (P@25), if we consider A1 as relevant and other classifications as non-relevant. Also, we observed P@50, P@100, and P@200 values of 88%, 73%, and 59%, respectively. In the first 50, 100 and 200 positions of the P-score ranking, we notice that 4, 9 and 25 publication venues have no Qualis evaluation, respectively. Notice that if we consider only publication venues that received a Qualis classification, then P@25 goes up to 100%. These results show that P-score is able to consistently place high impact venues (as classified by experts and highly influenced by citation data) in the top positions of its ranking.

Table D.1: Top 25 CS venues according to P-score

Rank	Venue	P-score	Qualis	Rank	Venue	P-score	Qualis
1	AAAI (C)	1.000	A1	14	SIAMCOMP (J)	0.632	A1
2	CACM (J)	0.949	A1	15	IPPS (C)	0.608	A1
3	FOCS (C)	0.949	A1	16	PAMI (J)	0.581	A1
4	STOC (C)	0.934	A1	17	ICDE (C)	0.578	A1
5	NIPS (C)	0.895	A1	18	TOG (J)	0.575	A1
6	CVPR (C)	0.891	A1	19	TC (J)	0.539	A1
7	ICRA (C)	0.886	A1	20	JACM (J)	0.528	A1
8	IJCAI (C)	0.758	A1	21	TCAD (J)	0.527	A1
9	CHI (C)	0.738	A1	22	COMPUTER (J)	0.520	
10	DAC (C)	0.683	A1	23	ICCV (C)	0.504	A1
11	INFOCOM (C)	0.661	A1	24	ICML (C)	0.489	A1
12	SODA (C)	0.652	A1	25	ICSE (C)	0.483	A1
13	SIGMOD (C)	0.642	A1				

We also ran analogous experiments considering the top 10 Computer Science graduate programs in Brazil, according to CAPES, as reference groups. In this case, for the ranking of venues we observed P@25, P@50, P@100, and P@200 values of 12%, 24%, 29%, and 29.5%, respectively. As we see, P-scores are heavily influenced by the selection of the reference groups, but work well when the right groups are selected. Most important, despite a great increase in volume, these results show that, the Brazilian output of research papers has lots of room to improve in terms of quality and impact.

D.1.2 Individual Researchers

In Table D.2, we present the top 25 academics in the ranking produced by P-score using, as reference, the research groups discussed in Section D.1. Given that the method is based on reference groups, which are composed by academics, using academics affiliated with reference groups as target may guide the analysis. Thus, we highlight the academics affiliated with the reference groups. An author that is not in a reference group and that appears before a reference author is a great signal of high productivity. The authors in positions one to four of our ranking are good reference candidates, after all they are ranked higher than any author of the reference groups.

Table D.2: Top 25 CS academics according to P-score

Rank	Expert	Pubs	Cits	P-score	Rank	Expert	Pubs	Cits	P-score
1	Toshio Fukuda	501	3861	1.000	14	Sebastian Thrun	445	21664	0.662
2	Thomas S. Huang	1172	19988	0.913	15	Kang G. Shin	731	14293	0.661
3	Philip S. Yu	788	18005	0.874	16	Michael I. Jordan	501	30735	0.657
4	A. L. S.-Vincentelli	1006	23094	0.809	17	Sudhakar Reddy	659	8119	0.650
5	Jiawei Han	655	28942	0.726	18	Luc J. Van Gool	548	12539	0.633
6	Vijay Kumar	413	6480	0.715	19	Moshe Y. Vardi	556	17698	0.611
7	Gerd Hirzinger	478	4745	0.712	20	Rama Chellappa	602	13608	0.604
8	Avi Wigderson	308	11337	0.699	21	Wolfram Burgard	395	13896	0.598
9	Oded Goldreich	379	18551	0.693	22	Robert E. Tarjan	405	29613	0.597
10	Takeo Kanade	743	32788	0.692	23	Irith Pomeranz	552	5563	0.597
11	C. Papadimitriou	506	28181	0.672	24	Xiaoou Tang	283	3275	0.582
12	Noga Alon	708	15106	0.666	25	Massoud Pedram	606	8532	0.581
13	Micha Sharir	584	14685	0.665					

In Table D.3, we filter the aforementioned ranking and present the top 25 Brazilian researchers that receive a productivity grant from CNPq. The last column presents the CNPq Level of each researcher. From this ranking we can extract some insights about the relative productivity of these authors in the national context. Many well known authors in Brazil appear at the top of the ranking. The correlation between P-score and the CNPq Levels are not so straightforward as the first 25 positions of the ranking of academics. Among the top 25, 8 researchers are 1A, which seems right. The fact that there are 1D researchers way up there, such as Marcos Gonçalves, is due to the fact that CNPq conditions the evolution in level to a minimum time period (which, for some people, makes little sense). When we compare the scores of Brazilian researchers with the scores of researchers from US, we notice that the score of Brazilian researchers is much lower. In fact, the score of the most productive Brazilian researcher, according to P-score, is approximately 10% of the score of the most productive researcher from

Table D.3: Top 25 CS academics from Brazil according to P-score

Rank	Expert	P-score	Level
1	Carlos J. P. de Lucena	1.000	1A
2	Luigi Carro	0.884	1B
3	Mário F. M. Campos	0.838	1B
4	Antonio A. F. Loureiro	0.807	1B
5	Wagner Meira Jr	0.756	1C
6	Jayme L. Szwarcfiter	0.751	1A
7	Marcos A. Gonçalves	0.721	1D
8	Nelson L. S. da Fonseca	0.718	1B
9	Paulo S. L. M. Barreto	0.677	1D
10	Alessandro F. Garcia	0.658	1D
11	Luís da C. Lamb	0.655	1C
12	Fabio G. Cozman	0.655	1C
13	Marco A. Casanova	0.636	1B
14	Agma J. M. Traina	0.627	1B
15	Eduardo S. Laber	0.626	1C
16	Caetano Traina Jr	0.619	1B
17	Virgilio A. F. Almeida	0.619	1A
18	Celina de Figueiredo	0.592	1A
19	Ana L. C. Bazzan	0.583	1C
20	Alberto H. F. Laender	0.575	1A
21	Edmundo Souza e Silva	0.564	1A
22	Jussara M. de Almeida	0.526	1D
23	Jorge Stolfi	0.525	1A
24	Yoshiharu Kohayakawa	0.508	1A
25	Alba C. M. A. de Melo	0.501	1D

US. The fact that P-scores for Brazilian researchers is so low is due to the fact that researchers in Brazil continue to publish a lot in local venues.

Additionally, we notice that 8 of these top 25 academics according to P-score receive the maximum CAPES Level — a precision of 32% at position 25 (P@25), if we consider 1A as relevant and other levels as non-relevant. Also, we observed P@50, P@100, and P@200 values of 24%, 15%, and 10%, respectively.

D.1.3 Research Groups

In Table D.4, we present the top 25 research groups according to P-score by using, as reference, the research groups discussed in Section D.1. The NRC Rank 5^{th} percentile is shown in the last column and does not include the graduate programs used as reference. From Table D.4, we notice that 6 out of 10 top groups according to P-score are considered as belonging to the top 10 graduate programs in USA according to NRC; 20 out of 25 top groups according to P-score are considered as belonging to the top 25 graduate programs in USA according to NRC; 5 European groups where considered in our analysis, two of them are at the top positions. The correlation between the rankings of P-score and NRC is of 0.58, with a p-value near to zero; and many well known research groups appear at the top of the ranking.

In Table D.5, we filter the ranking and present the top 25 Brazilian research groups in CS. In the last column, we show the CAPES Classification of each group. From Table D.5, we notice that all the groups classified by CAPES as 7 appear before the ones that are classified as 6, which appear before the other groups in the P-score

Table D.4: Top 25 CS groups according to P-score

Rank	Group	P-score	NRC
1	Un. of Michigan (AA)	1.000	8
2	Cornell Un.	0.997	5
3	UMass (Amherst)	0.888	9
4	Un. of Pennsylvania	0.883	9
5	Un. of Texas at Austin	0.882	6
6	Princeton Un.	0.778	4
7	Purdue Un. (MC)	0.751	15
8	UC-Irvine	0.743	22
9	Columbia Un. (NY)	0.729	26
10	Duke Un.	0.623	16
11	ETH Zurich	0.619	
12	Rutgers	0.596	52
13	Northwestern Un.	0.595	49
14	Un. of Wisconsin (M)	0.574	10
15	Un. of Minnesota (TC)	0.541	19
16	UC-Santa Barbara	0.536	5
17	Ohio State Un. (MC)	0.536	22
18	Arizona State Un.	0.498	25
19	Brown Un.	0.494	21
20	New York Un.	0.493	16
21	UNC at Chapel Hill	0.473	5
22	Un. of Rochester	0.413	11
23	Imperial College London	0.396	
24	Un. of Illinois (Chicago)	0.391	67
25	State Un. of NY	0.388	17

Table D.5: Top 25 CS groups from Brazil according to P-score

Rank	Group	P-score	Class
1	UFRGS	0.119	7
2	UFMG	0.098	7
3	PUC-RIO	0.090	7
4	Unicamp	0.086	7
5	UFRJ	0.082	7
6	UFPE	0.065	6
7	USP/SC	0.063	6
8	USP	0.056	6
9	UFF	0.039	5
10	PUC/RS	0.028	5
11	UFPR	0.025	4
12	UFCG	0.024	4
13	UFSCar	0.022	4
14	UFRN	0.019	5
15	UFAM	0.019	5
16	UFSC	0.019	4
17	UnB	0.017	4
18	PUC/PR	0.012	4
19	UFC	0.009	5
20	UFES	0.009	4
21	UFBA	0.008	4
22	UFMS	0.008	4
23	UNIFOR	0.008	4
24	UFU	0.007	4
25	UNISINOS	0.005	4

ranking. There are some small swaps between groups classified as 4 and 5. The Kendall Tau correlation between the P-score and CAPES is of 0.76, with a p-value near to zero. The difference of P-score value between the last 7-classified group and the first 6-classified group (UFRJ and UFPE) is relatively large, this is also true between the last group classified as 6 and the first group classified as 5.

D.2 Inherited Reputation from Top Authors of Distinct Sub-areas

The original proposal of P-score is to rely on top research groups as source of reputation to rank publication venues, academics and other research groups one want to get insights about. There are many ways to define a research group, and the proposed method is based on this choice. Despite defining graduate programs as seed research groups and trusting on their faculty members is a reasonable approach, a coverage issue may arise. Specifically, the problem is we have no guarantee about the coverage of the sub-areas that compose the major knowledge area under analysis.

In this section, we discuss the application of P-score to rank publication venues, academics and research groups in the knowledge area of Computer Science. Here, instead of relying on the top graduate programs according to the randomized procedure as we made in Section D.1, we compose a research group with the top 10 researchers for a series of Computer Science sub-areas. This alternative smooths this coverage issue by relying on small groups of top researchers in each sub-area we want to cover.

In Table D.6, we list the 24 sub-areas of Computer Science we consider in this analysis. These sub-areas are based on the list of Computer Science sub-areas extracted from Microsoft Academic Search (October 2014). To include another sub-area, one just need to get data of a small set of top researchers in this sub-area. It is sufficient to identify top 10 researchers and their corresponding entities in DBLP (see Section A).

Table D.6: Main sub-areas of Computer Science

#	Subarea	#	Subarea
1	Algorithms & Theory	13	Multimedia
2	Artificial Intelligence	14	Natural Language & Speech
3	Bioinformatics & Computational Biology	15	Networks & Communications
4	Computer Vision	17	Operating Systems
5	Data Mining	18	Programming Languages
6	Databases	19	Real-Time & Embedded Systems
7	Distributed & Parallel Computing	20	Scientific Computing
8	Graphics	21	Security & Privacy
9	Hardware & Architecture	22	Simulation
10	Human-Computer Interaction	23	Software Engineering
11	Information Retrieval	24	World Wide Web
12	Machine Learning & Pattern Recognition		

D.2.1 Publication Venues

In Table D.7, we present the top 25 venues in the ranking produced by P-score using, as reference, the 24 groups composed by the top 10 researchers according to Microsoft Academic Search in each sub-area of Computer Science.

We notice that, many important venues in the sub-area appears at the top positions of this ranking. In terms of workload, by collecting data from only 10 researchers for some sub-areas one can produce a reasonable ranking of venues in Computer Science though our metric. If we consider as relevant publication venues classified by Qualis as A1, we observe P@25, P@50, P@100, and P@200 values of 92%, 84%, 75%, and 56%, respectively. We also notice that some publication venues of Computer Science are not evaluated in Qualis. Some of them are important for the area: out of the top 25 publication venues according to P-score in this sub-area, one is not evaluated in Qualis, 3 among the top 50, 10 among the top 100, and 31 among the top 200.

Table D.7: Top 25 CS venues according to P-score, top researchers as reference

Rank	Venue	P-score	Qualis
1	ICIP (C)	1.000	A1
2	ICASSP (C)	0.927	A1
3	CACM (J)	0.926	A1
4	SIGMOD (C)	0.912	A1
5	STOC (C)	0.887	A1
6	CVPR (C)	0.868	A1
7	SIAMCOMP (J)	0.842	A1
8	NIPS (C)	0.808	A1
9	FOCS (C)	0.793	A1
10	CHI (C)	0.769	A1
11	ICDE (C)	0.767	A1
12	INFOCOM (C)	0.757	A1
13	VLDB (C)	0.751	A1
14	DAC (C)	0.734	A1
15	PAMI (J)	0.719	A1
16	WSC (C)	0.688	A1
17	JACM (J)	0.660	A1
18	KDD (C)	0.628	A1
19	COMPGEOM (C)	0.616	A2
20	TCAD (J)	0.606	A1
21	TC (J)	0.589	A1
22	IACR (J)	0.582	
23	TSE (J)	0.572	A1
24	AAAI (C)	0.565	A1
25	ICCV (C)	0.536	A1

D.2.2 Individual Researchers

In Table D.8, we present the top 25 academics in the ranking produced by P-score using, as reference, the 24 groups composed by the top 10 researchers according to Microsoft Academic Search in each sub-area of Computer Science. Reference authors are highlighted in this table. A non-reference author appearing before a reference is a great signal about this non-reference. The authors in positions three and four of this ranking are good reference candidates, after all they are ranked higher than any author of the reference groups.

Table D.8: Top 25 CS academics according to P-score

Rank	Expert	Pubs	Cits	P-score
1	Thomas S. Huang	1172	19988	1.000
2	Philip S. Yu	788	18005	0.809
3	Rama Chellappa	602	13608	0.679
4	Wen Gao	907	4412	0.663
5	Jiawei Han	655	28942	0.663
6	Aggelos Katsaggelos	503	5184	0.569
7	Micha Sharir	584	14685	0.542
8	A. L. S.-Vincentelli	1006	23094	0.532
9	Anil K. Jain	588	29175	0.529
10	Georgios Giannakis	932	21128	0.525
11	Kang G. Shin	731	14293	0.492
12	Truong Q. Nguyen	496	2986	0.470
13	K. J. Ray Liu	672	7397	0.465
14	Christos Faloutsos	484	19778	0.461
15	Edwin R. Hancock	636	3878	0.443
16	Hermann Ney	581	11206	0.433
17	Xiaoou Tang	283	3275	0.432
18	Tsuhan Chen	341	3178	0.431
19	Martin Vetterli	547	19351	0.430
20	H. Garcia-Molina	605	29773	0.429
21	Moshe Y. Vardi	556	17698	0.418
22	C. Papadimitriou	506	28181	0.413
23	Luc J. Van Gool	548	12539	0.407
24	Xin Li	550	4815	0.407
25	Oded Goldreich	379	18551	0.403

Table D.9: Top 25 CS academics from Brazil according to P-score

Rank	Expert	P-score	Level
1	Marcos A. Goncalves	0.161	1D
2	Carlos J. P. de Lucena	0.156	1A
3	Wagner Meira Jr	0.128	1C
4	Luigi Carro	0.124	1B
5	Berthier Ribeiro-Neto	0.115	
6	Agma J. M. Traina	0.111	1B
7	Caetano Traina Jr	0.108	1B
8	Marco A. Casanova	0.103	1B
9	Alberto H. F. Laender	0.103	1A
10	Altigran S. da Silva	0.102	1D
11	Jorge Stolfi	0.099	1A
12	Jayme L. Szwarcfiter	0.097	1A
13	Antonio A. F. Loureiro	0.096	1B
14	Nelson L. S. da Fonseca	0.096	1B
15	Nivio Ziviani	0.093	1A
16	Zhao Liang	0.093	1C
17	Alexandre X. Falcao	0.092	1C
18	Virgilio A. F. Almeida	0.092	1A
19	Edmundo Souza e Silva	0.090	1A
20	Edleno S. de Moura	0.090	1D
21	Antonio L. Furtado	0.089	
22	Celina de Figueiredo	0.086	1A
23	Eduardo S. Laber	0.085	1C
24	Olga R. P. Bellon	0.081	1D
25	Alessandro F. Garcia	0.081	1D

In Table D.9, we present the top 25 academics from Brazil in the ranking produced by P-score using, as reference, the 24 groups with the top 10 researchers according to Microsoft Academic Search in each sub-area of Computer Science. We notice that 8 of these top 25 academics according to P-score receive the maximum CNPq Level — a precision of 32% at position 25 (P@25), if we consider 1A as relevant and other levels as non-relevant. Also, we observed P@50, P@100, and P@200 values of 22%, 15%, and 10%, respectively In the first 25, 50, 100 and 200 positions of the P-score ranking, we notice that 2, 6, 12 and 42 academics have no CNPq Level, respectively.

D.2.3 Research Groups

In Table D.10, we present the top 25 research groups in the ranking produced by P-score using, as reference, the 24 groups composed by the top 10 researchers according to Microsoft Academic Search in each sub-area of Computer Science. We notice that 7 out of top 10 groups according P-score are considered as belonging to the top 10 graduate programs in USA according to NRC; 19 out of top 25 groups according P-score are considered as belonging to the top 25 graduate programs in USA according to NRC. The Kendall Tau correlation between the rankings of P-score and NRC is of 0.624, with a p-value near to zero. These numbers are smaller than the ones in Section D.1.3, which is expected because research groups of a major area of knowledge may be stronger (or weaker) in distinct sub-areas.

In Table D.11, we present the top 25 research groups from Brazil in the ranking produced by P-score using, as reference, the 24 artificial groups composed by the top 10 researchers according to Microsoft Academic Search in each sub-area of Computer Science. We notice that all the groups classified by CAPES as 7 appear before the ones that are classified as 6, which appear before the other groups in the P-score ranking. There are some small swaps between groups classified as 4 and 5. The Kendall Tau correlation between the rankings of P-score and CAPES is of 0.839, with a p-value near to zero. The nDCG of this ranking is of 0.998, while it is of 0.997 when using the randomized process to produce reference groups (in Appendix D.1). Despite the groups classified by CAPES as 6 and 7 appear at the top 10 positions of the ranking, here there is not a clear separation between them as in general Computer Science (see Table D.5).

Table D.10: Top 25 CS groups according to P-score, top researchers as reference

Rank	Group	P-score	NRC	Rank	Group	P-score	NRC
1	Georgia Tech	1.000	12	14	Un. of Pennsylvania	0.518	18
2	MIT	0.949	3	15	Purdue Un. (MC)	0.499	24
3	UC-Berkeley	0.918	1	16	Un. of Texas at Austin	0.490	11
4	Un. of Illinois (UC)	0.884	5	17	UC-Irvine	0.480	35
5	Stanford Un.	0.877	1	18	Princeton Un.	0.465	6
6	Carnegie Mellon Un.	0.841	4	19	Columbia Un. (NY)	0.443	40
7	UC-San Diego	0.711	15	20	Northwestern Un.	0.435	70
8	UC-Los Angeles	0.700	10	21	ETH Zurich	0.424	
9	Un. of Maryland (CP)	0.658	14	22	UC-Santa Barbara	0.401	7
10	Cornell Un.	0.615	7	23	Duke Un.	0.388	25
11	Un. of Michigan (AA)	0.612	15	24	Rutgers	0.357	74
12	Un. of Southern California	0.557	17	25	Ohio State Un. (MC)	0.353	36
13	UMass (Amherst)	0.536	19				

Table D.11: Top 25 CS groups from Brazil according to P-score, top researchers as reference

Rank	Group	P-score	Class	Rank	Group	P-score	Class
1	UFRGS	0.086	7	14	UFCG	0.017	4
2	UFMG	0.072	7	15	UFSC	0.015	4
3	Unicamp	0.071	7	16	UFRN	0.013	5
4	PUC-RIO	0.064	7	17	PUC/PR	0.012	4
5	UFRJ	0.058	7	18	UnB	0.010	4
6	USP/SC	0.048	6	19	UFC	0.007	5
7	UFPE	0.047	6	20	UFES	0.006	4
8	USP	0.039	6	21	UNISINOS	0.006	4
9	UFF	0.028	5	22	UFU	0.006	4
10	UFPR	0.023	4	23	UFBA	0.006	4
11	PUC/RS	0.020	5	24	UNIFOR	0.005	4
12	UFSCar	0.018	4	25	UFMS	0.005	4
13	UFAM	0.018	5				

Appendix E

P-score in Information Retrieval

In this appendix, we discuss the application of P-score in the sub-area of Information Retrieval to rank publication venues, academics and research groups.

To produce these rankings, we build an artificial group composed of the top 10 researchers in the sub-area of Information Retrieval according to Microsoft Academic Search (MAS), as presented in Table E.1. The results we present in this appendix were produced using this single group as reference in P-score.

Table E.1: Reference group for the sub-area of Information Retrieval

Top 10 researchers of IR according to Microsoft Academic Search
W. Bruce Croft, Gerard Salton, Ellen Voorhees, Chris Buckley, Stephen E. Robertson, Jamie Callan, Susan Dumais, James Allan, Hsinchun Chen, Justin Zobel

E.1 Publication Venues in Information Retrieval

In Table E.2, we present the top 25 venues in the ranking produced by P-score using, as reference, a group composed by the top 10 researchers in the sub-area of Information Retrieval according to Microsoft Academic Search (see Table E.1). We notice that, the most important conference in the area of Information Retrieval appears at the first position of the ranking produced by P-score. Indeed, many important venues in the sub-area appears at the top positions of this ranking. In terms of workload, by collecting data from only 10 researchers one can produce a reasonable ranking of venues in Information Retrieval though our metric.

If we consider as relevant publication venues classified by Qualis as A1, we observe P@25, P@50, P@100, and P@200 values of 48%, 56%, 42%, and 24%, respectively. We also notice that many publication venues of the sub-area of Information Retrieval (and

Table E.2: Top 25 IR venues according to P-score

Rank	Venue	P-score	Qualis
1	SIGIR (C)	1.000	A1
2	TREC (C)	0.573	A2
3	CIKM (C)	0.449	A1
4	SIGIR (J)	0.310	
5	JASIS (J)	0.284	A1
6	IPM (J)	0.266	A2
7	ISI (C)	0.257	B1
8	CACM (J)	0.223	A1
9	DSS (J)	0.173	A1
10	DGO (C)	0.172	B1
11	IR (J)	0.134	B1
12	TOIS (J)	0.133	A2
13	ECIR (C)	0.109	A2
14	COMPUTER (J)	0.108	
15	EXPERT (J)	0.104	A1
16	JCDL (C)	0.102	A2
17	HICSS (C)	0.098	A1
18	JACM (J)	0.084	A1
19	CHI (C)	0.081	A1
20	AAAI (C)	0.071	A1
21	NAACL (C)	0.070	A1
22	WSDM (C)	0.067	B1
23	TKDE (J)	0.065	A1
24	RIAO (C)	0.065	
25	IPL (J)	0.062	A2

others sub-areas, as we will see later) are not evaluated in Qualis. Some of them are quite important for the sub-area: out of the top 25 publication venues according to P-score in this sub-area, 4 are not evaluated in Qualis, 8 among the top 50, 21 among the top 100, and 75 among the top 200.

E.2 Academics in Information Retrieval

In Table E.3, we present the top 25 academics in the ranking produced by P-score using the reference group presented in Table E.1, reference authors are highlighted.

Table E.3: Top 25 IR academics according to P-score

Rank	Expert	Pubs	Cits	P-score
1	W. Bruce Croft	389	13733	1.000
2	James Allan	343	6664	0.588
3	ChengXiang Zhai	243	4645	0.584
4	Maarten de Rijke	410	3890	0.576
5	Hsinchun Chen	555	6351	0.508
6	Iadh Ounis	164	1251	0.473
7	Ryen W. White	141	1080	0.457
8	Mark Sanderson	212	2502	0.420
9	Charles Clarke	193	2025	0.396
10	Justin Zobel	255	5127	0.378
11	Ellen M. Voorhees	169	7401	0.375
12	Susan T. Dumais	229	15645	0.372
13	Craig Macdonald	75	709	0.366
14	Gerard Salton	290	22981	0.365
15	Joemon M. Jose	221	1037	0.361
16	Edward A. Fox	494	6288	0.356
17	Leif Azzopardi	111	496	0.353
18	Douglas W. Oard	208	2245	0.349
19	Nicholas J. Belkin	193	3467	0.348
20	Ophir Frieder	338	4835	0.347
21	Luo Si	106	944	0.339
22	Jaap Kamps	224	1029	0.321
23	Stephen Robertson	320	8767	0.316
24	Clement T. Yu	357	5300	0.308
25	Mounia Lalmas	250	1997	0.307

When we rank reference authors together with the authors we want to compare, we can uncover new reliable authors. A non-reference author appearing before a reference author is a great signal about this non-reference. The authors in third and fourth positions of this ranking are good reference candidates, after all only two out of ten references appear before them. Another notable event is the appearence of an author that have only 75 publications and 709 citations according to MAS before 4 out of 10 reference authors in position 13 (which is very high). The number of publications of Craig Macdonald is higher in DBLP repository, 128 publications, which is lower than any author in the top 25 positions of this ranking. According to Google Scholar, he had 2,284 citations until 2012. This value seems to be more coherent with the citation counts of the top 25 authors in the P-score ranking.

While citation counts are underestimated in MAS, in Scholar they are inflated. For example, while W. Bruce Croft has 13,733 in MAS, in Google Scholar he has 34,354. Another example is Andrei Shleifer that has 44758 in MAS and 182,944 in Scholar.

Microsoft Academic Search considers citations from a set of selected venues only, while Google Scholar counts all types of citations (for example, citations from a technical report). There is an argument that the Scholar citations are more inflated, but that they are consistent if we restrict the counting to Scholar. Another argument is that the impact of venues that are not strictly academic matters. For example, a researcher who has thousands of tweets about a result of his research, or has thousands of downloads of a technical report or a software available on the Web that produced a result of impact and should be accounted for this. If we consider this interpretation, the citation counts from Google Scholar is most appropriate.

E.3 Research Groups in Information Retrieval

In Table E.4, we present the top 25 research groups in the ranking produced by P-score using, as reference, an artificial group composed by the top 10 researchers in the sub-area of Information Retrieval according to Microsoft Academic Search. The NRC Rank is shown in the last column. We observe that the first position seems to be coherent since UMass (Amherst) is well known in the sub-area of Information Retrieval. Many well known research groups in this sub-area appear at the top of the ranking. 15 out of top 25 groups according P-score are considered as belonging to the top 25 graduate programs in USA according to NRC. The Kendall Tau correlation between the rankings of P-score and NRC is of 0.47, with a p-value near to zero. These numbers are smaller than the ones in Section D.1.3, which is expected because research groups of a major area of knowledge may be stronger (or weaker) in distinct sub-areas.

Table E.4: Top 25 IR groups according to P-score

Rank	Group	P-score	NRC	Rank	Group	P-score	NRC
1	UMass (Amherst)	1.000	19	14	Cornell Un.	0.323	7
2	Un. of Illinois (UC)	0.663	5	15	Columbia Un. (NY)	0.236	40
3	Georgia Tech	0.579	12	16	Arizona State Un.	0.219	38
4	Carnegie Mellon Un.	0.517	4	17	Florida International Un.	0.219	104
5	UC-Berkeley	0.461	2	18	UC-Los Angeles	0.214	10
6	Stanford Un.	0.445	1	19	UC-Irvine	0.210	35
7	Un. of Maryland (CP)	0.402	14	20	Un. of Rochester	0.206	22
8	Purdue Un. (MC)	0.394	24	21	Un. of Texas at Austin	0.205	11
9	Un. of Southern California	0.382	17	22	UFMG	0.204	
10	MIT	0.361	3	23	Un. of Pennsylvania	0.204	18
11	Un. of Illinois (Chicago)	0.347	91	24	Un. of Minnesota (TC)	0.200	32
12	Un. of Michigan (AA)	0.342	15	25	Northeastern Un.	0.192	85
13	Virginia Poly.	0.325	53				

We need to be cautious when evaluating productivity indicators of research groups in major areas, when that is the case and a group A is better evaluated than group B, there is no guarantee that A is better than B in all sub-areas, after all B may be much more productive than A in specific sub-areas and this effect may not be captured when looking at indicators of the broad area. Despite the Brazilian graduate programs are not the top ones when compared with programs from USA and Europe in the broad area of Computer Science (see Section D.1.3), these groups may appear among the top ones when the comparison is made in sub-areas like in Information Retrieval. Indeed, there is a group from Brazil in position 22.

In Table E.5, we filter the aforementioned ranking and present the top 25 Brazilian graduate programs in the sub-area of Information Retrieval. In the last column, we show the CAPES Classification of each group. We notice that the Kendall Tau correlation between the rankings of P-score and CAPES classification is of 0.763, with a p-value near to zero. The nDCG of this ranking is of 0.940, while in general CS is of 0.997. Despite the groups classified by CAPES as 6 and 7 appear at the top 10 positions of the ranking, here there is not a clear separation between them as in general Computer Science (see Table D.5). UFAM is classified as 5, but has respectable production in Information Retrieval. We finish this section with the following question: *Should a student choose his future program based on a broad classification of the area, or based on insights about the productivity in the sub-area he is interested in?*

Table E.5: Top 25 IR groups from Brazil according to P-score

Rank	Group	P-score	Class	Rank	Group	P-score	Class
1	UFMG	0.204	7	14	UFSCar	0.006	4
2	UFAM	0.091	5	15	UNIFOR	0.005	4
3	Unicamp	0.037	7	16	UFPR	0.005	4
4	PUC-RIO	0.036	7	17	UFSC	0.005	4
5	UFRGS	0.024	7	18	UFCG	0.005	4
6	USP/SC	0.023	6	19	UFC	0.002	5
7	UFRJ	0.018	7	20	UnB	0.002	4
8	UFPE	0.016	6	21	UFMS	0.001	4
9	UFF	0.009	5	22	PUC/PR	0.001	4
10	USP	0.009	6	23	UFRN	0.001	5
11	UFES	0.009	4	24	UFBA	0.001	4
12	PUC/RS	0.008	5	25	UNISINOS	0.001	4
13	UFU	0.007	4				

Appendix F

Counting Proof

In a bipartite reputation graph $\mathbf{P}$ (see Section 3.2.2) composed by sources and targets, if the number of edges n_{st} (see Section 3.2.1) from source s to target t equals the number of edges n_{ts} from target t to source s for all sources and targets, then the steady state probability of each node is proportional to its total number of edges in the reputation network. Formally, in a bipartite reputation graph $\mathbf{P}$, for all $s \in S$ and $t \in T$:

$$n_{st} = n_{ts} \implies \boldsymbol{\gamma}_s \propto n_s \ \wedge \ \boldsymbol{\gamma}_t \propto n_t. \tag{F.1}$$

A practical implication of it is that, in scenarios like the aforementioned one, it is not necessary to compute the steady state probabilities $\boldsymbol{\gamma}$ of the network using techniques like the Power Method, we can simply count the number of edges between sources and targets. Thus, by making such kind of model, one can reason about how the reputation flows through the network with much less computation effort.

Proof

$$\begin{aligned}
\boldsymbol{\gamma} &= \boldsymbol{\gamma}\mathbf{P} && \text{(F.2)}\\
&= \left[\begin{array}{c|c} \mathbf{P}^{\langle S\rangle} & \mathbf{P}^{\langle T\rangle} \end{array}\right] \left[\begin{array}{c|c} \mathbf{0} & \mathbf{P}^{\langle ST\rangle} \\ \hline \mathbf{P}^{\langle TS\rangle} & \mathbf{0} \end{array}\right] && \text{(F.3)}\\
&= \left[\begin{array}{c|c} \mathbf{P}^{\langle S\rangle}\mathbf{0} + \mathbf{P}^{\langle T\rangle}\mathbf{P}^{\langle TS\rangle} & \mathbf{P}^{\langle S\rangle}\mathbf{P}^{\langle ST\rangle} + \mathbf{P}^{\langle T\rangle}\mathbf{0} \end{array}\right] && \text{(F.4)}\\
&= \left[\begin{array}{c|c} \mathbf{P}^{\langle T\rangle}\mathbf{P}^{\langle TS\rangle} & \mathbf{P}^{\langle S\rangle}\mathbf{P}^{\langle ST\rangle} \end{array}\right] && \text{(F.5)}\\
&= \left[\begin{array}{c|c} {n_1}^{\langle S\rangle} \cdots {n_{|S|}}^{\langle S\rangle} & {n_1}^{\langle T\rangle} \cdots {n_{|T|}}^{\langle T\rangle} \end{array}\right] \times C && \text{(F.6)}
\end{aligned}$$

where C is a constant value defined as $C = 1/(\sum_{s\in S} n_s + \sum_{t\in T} n_t)$.

A careful reader would notice that, the proof is not complete and that the next step is to prove that $\mathbf{P}^{\langle T\rangle}\mathbf{P}^{\langle TS\rangle}$ equals $[{n_1}^{\langle S\rangle}\cdots{n_{|S|}}^{\langle S\rangle}]$, necessary to go from equation D.5 to equation D.6:

$$\mathbf{P}^{\langle T\rangle}\mathbf{P}^{\langle TS\rangle} = \begin{bmatrix} {n_1}^{\langle T\rangle} & \cdots & {n_{|T|}}^{\langle T\rangle} \end{bmatrix} \begin{bmatrix} \frac{{n_{1,1}}^{\langle TS\rangle}}{{n_1}^{\langle T\rangle}} & \cdots & \frac{{n_{1,|S|}}^{\langle TS\rangle}}{{n_1}^{\langle T\rangle}} \\ \vdots & \ddots & \vdots \\ \frac{{n_{|T|,1}}^{\langle TS\rangle}}{{n_{|T|}}^{\langle T\rangle}} & \cdots & \frac{{n_{|T|,|S|}}^{\langle TS\rangle}}{{n_{|T|}}^{\langle T\rangle}} \end{bmatrix} \tag{F.7}$$

$$= \begin{bmatrix} \sum_{t\in T} {n_t}^{\langle T\rangle}\frac{{n_{t,1}}^{\langle TS\rangle}}{{n_t}^{\langle T\rangle}} & \cdots & \sum_{t\in T} {n_t}^{\langle T\rangle}\frac{{n_{t,|S|}}^{\langle TS\rangle}}{{n_t}^{\langle T\rangle}} \end{bmatrix} \tag{F.8}$$

$$= \begin{bmatrix} \sum_{t\in T} {n_{t,1}}^{\langle TS\rangle} & \cdots & \sum_{t\in T} {n_{t,|S|}}^{\langle TS\rangle} \end{bmatrix} \tag{F.9}$$

$$\mathbf{P}^{\langle T\rangle}\mathbf{P}^{\langle TS\rangle} = \begin{bmatrix} {n_1}^{\langle S\rangle} & \cdots & {n_{|S|}}^{\langle S\rangle} \end{bmatrix} \tag{F.10}$$

Finally, to conclude the proof, it is necessary to prove that $\mathbf{P}^{\langle S\rangle}\mathbf{P}^{\langle ST\rangle}$ equals $[{n_1}^{\langle T\rangle}\cdots{n_{|T|}}^{\langle T\rangle}]$. It can be done by a similar process from equation D.7 to D.10.

Appendix G

WSDM Cup 2016

Static rankings of papers play a key role in the academic search setting. Many features are commonly used in the literature to produce such rankings, some examples are citation-based metrics, distinct applications of PageRank, among others. More recently, learning to rank techniques have been successfully applied to combine sets of features producing effective results. In this work, we propose the metric S-RCR, which is a simplified version of a metric called Relative Citation Ratio — both based on the idea of a co-citation network. When compared to the classical version, our simplification S-RCR leads to improved efficiency with a reasonable effectiveness. We use S-RCR to rank over 120 million papers in the Microsoft Academic Graph dataset. By using this single feature, which has no parameters and does not need to be tuned, our team was able to reach the 3rd position in the first phase of the WSDM Cup 2016.

G.1 Ranking Papers

Finding the most relevant papers of a field of knowledge is a task with many motivations. From the researcher's perspective, it is important for instance to quickly discern the papers with major impact in his/her study area from those with less relevance. On the other hand, from an academic search engine perspective, a common task is to present the papers by using rankings, which demands a sort criterion as relevance. Also, establishing a relative order of importance of papers could help in other tasks such as providing grants or research awards for individual researchers and graduate programs. The problem of ranking papers was addressed in the WSDM Cup 2016, a competition that brought together 32 research teams from all over the world.

G.1.1 The Competition — WSDM Cup 2016

The WSDM Cup Ranker Challenge[1] was created by the WSDM[2] organizers and supported by Microsoft Research.

G.1.1.1 The Task

The task for each competitor in WSDM Cup 2016 was to provide the best static rank values for publication entities in the Microsoft Academic Graph[3] (MAG) [Sinha et al., 2015]. The goal behind it was to assess the query-independent importance of academic papers.

G.1.1.2 Evaluation

The evaluation in WSDM Cup 2016 was conducted in two phases, as we now describe. During Phase 1, submissions were scored based on the agreements with human judgement data. A group of Computer Science researchers were invited by the organizers to conduct a pairwise ranking of papers in the fields they actively conduct research. The pairwise judgement data were then randomly segregated into an Evaluation and a Test set. Submissions during Phase 1 were automatically scored against the Evaluation set and added to a public leaderboard that was sorted based on the percentage of agreements with the judgement data. At the end of Phase 1, the most recent submission from each team was evaluated against the Test set and the scores (ranked by the percentage agreements with the Test set) were announced to the leaderboard.

The top eight teams on the leaderboard at the end of Phase 1 were invited to participate in Phase 2. Each participant of Phase 2 was asked to re-run the algorithms over an updated graph and to submit the final rank values. Phase 2 of the Challenge was conducted by Microsoft Research in cooperation with Bing. Each of the finalist results was applied to Bing search results and powered the ranker used by Bing for academic queries.

G.1.2 This Report

In this work, we report the participation of our team, named UFMG/LATIN, in the WSDM Cup 2016. Before getting into the final model, we performed a set of tests considering distinct approaches, which include distinct citation-based metrics, PageRank,

[1] http://wsdmcupchallenge.azurewebsites.net
[2] http://www.wsdm-conference.org/2016/
[3] http://research.microsoft.com/en-us/projects/mag/

among others. Our final approach was based on a simplified version of a metric called Relative Citation Ratio [Hutchins et al., 2016]. In here, we describe this metric, how and why we choose to run a simplified version.

The remainder of this work is organized as follows. In Section G.2, we present the related works on paper rankings. In Section G.3, we describe the RCR metric as well as our simplified version, the S-RCR. In Section G.4, we discuss some additional techniques we applied to rank papers. In Section G.5, we report our experiments. Finally, in Section G.6, we provide a final discussion and concluding remarks.

G.2 Literature on Paper Rankings

The most common approach to produce paper rankings in the literature is by using citation-based metrics. These metrics provide a natural way to reason about the relative quality of academic entities, such as scientific papers, individual researchers and publication venues. One of the earliest metrics proposed to quantify academic impact was the Impact Factor [Garfield, 1955]. Since them, many alternatives have been proposed, including other citation-based metrics like the H-Index [Hirsch, 2005], random walks and machine learning techniques.

Another common approach to rank academic entities is by considering the structural information. The structure of the citation network can be used to produce academic rankings by applying random walk techniques, such as the PageRank [Page et al., 1998]. A natural approach is to apply random walks in the paper-paper citation network. However, some authors also apply random walks in heterogeneous graphs. In [Ribas et al., 2015b], for example, the authors propose a novel random walk model to identify the most reputable entities of a domain based on a conceptual framework of reputation flows.

Another concept that is worth mentioning is the Altmetrics movement, which points out the need for novel evaluation metrics as alternative to classic citation-based metrics. According to Piwowar [2013], citation-based metrics are useful, but not sufficient to evaluate research. In particular, they observe that citations are slow — their main argument is the fact that a paper's first citation can take years.

Learning to rank techniques [Li, 2011] have been used over the last few years to improve the quality of rankings by effectively combining multiple sources of evidence. The large amount of available features related to some ranking tasks motivates the adoption of learning to rank methods in distinct contexts, including in academic search.

Our approach is inspired by the work of Hutchins et al. [2016], which proposes a

metric called Relative Citation Ratio. It is a paper-level and field-independent score that provides an alternative to classic citation-based metrics to identify influential papers. They show that the rankings produced by their metric strongly correlate with the opinions of experts in biomedical research and suggest that the same approach should be generally applicable in all areas of science.

While complex models, such as heterogeneous random walks or learning to rank methods, are able to produce effective results, in this work, we investigate a single well-designed feature to rank papers given its citation network.

G.3 Relative Citation Ratio

In this section, we describe the metric we propose, the S-RCR, which is a simplification of a metric called Relative Citation Ratio (RCR) [Hutchins et al., 2016]. Before describing S-RCR it is worth reasoning about the basic concepts of the original RCR metric. The RCR metric is based upon the idea of using the co-citation network of each paper to normalize it in terms of time and area of study, by calculating an expected citation rate of the target paper from the aggregated citation behavior of its neighborhood. Basically, this strategy consists of computing the average citation rate of this neighborhood which is used as the RCR denominator; the numerator is the citation rate of the target paper.

G.3.1 Co-citation Network

The basis of the RCR metric and our proposed simplification is the notion of co-citation network. Hutchins et al. [2016] define this co-citation network as the *papers' area of influence*. As described in Hutchins et al. [2016], when a paper is first cited, the other papers appearing in the reference list along with this work comprise its co-citation network, see Figure G.1. As the paper continues to be cited, the papers appearing in the new reference lists alongside it are added to its co-citation network. This network provides a dynamic view of the paper's field of research, taking advantage of information provided by the experts who have found the study useful enough to cite. The co-citation network of a paper can be viewed as a representative sample of its area of research allowing us to perform a reasonable cross-field evaluation of papers.

In Figure G.1, we present the schematic view of a co-citation network for the RCR computation. The Reference Article (RA, in red) cites previous papers from the literature (in orange) and other papers (in blue) cite the RA. The co-citation network

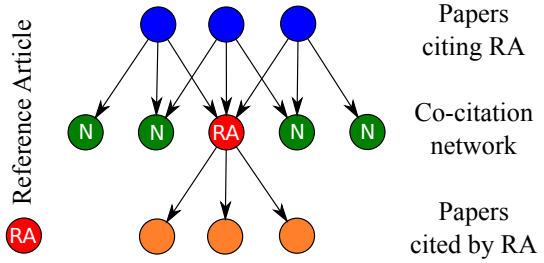

Figure G.1: Schematic view of a co-citation network 2015relative

(or neighborhood) of the Reference Article is the set of papers (in green) that appear alongside the RA.

G.3.2 Article Citation Ratio

The Article Citation Ratio (ACR) of a given paper p is defined as:

$$\text{ACR}(p) = \frac{\text{Citations}(p)}{\text{Age}(p) + 1}, \tag{G.1}$$

where Citations(p) is the total number of citations paper p received since its publication and Age(p) is the time in years since the publication date of paper p.

By keeping a counter of citations and paper ages, we can store the data in a hash table and compute the ACR of a paper p in $\Theta(1)$ time complexity.

G.3.3 The Simplified RCR

To produce field independent rankings, RCR metrics normalize the ACR of a paper p based on the information of its co-citation network. In the original version of RCR [Hutchins et al., 2016], this information is used through a complex normalization process. Computing the original RCR of a single paper p depends on performing a linear regression on its co-citation network using the journal citation ratio [Garfield, 1972] of the venues p's neighbors were published. For a full description of the RCR metric, we refer the reader to the work of [Hutchins et al., 2016].

In particular, we define the Simplified Relative Citation Ratio (S-RCR) of a given paper p as follows:

$$\text{S-RCR}(p) = \frac{\text{ACR}(p)}{(1/|N_p|)\sum_{p' \in N_p} ACR(p')}, \tag{G.2}$$

where ACR(p) is the Article Citation Ratio (see Section G.3.2) of paper p and N_p is the set of neighbors of paper p (see Section G.3.1). Similarly to the classic RCR, the

numerator is the ACR of paper p and the denominator acts as a normalizer, forcing the ACR of paper p to be relative to its neighbors.

The main difference between our proposal and the original RCR is its normalization step, which, in our metric, is much simpler. Specifically, we normalize the ACR of a paper p by the average of the ACR values of p's neighbors. While this simplification is briefly mentioned in the original paper, the authors discard it for its numerical limitations, e.g., for papers with no neighbors. In contrast, we overcome these limitations by smoothing Eq. (G.2) via additive smoothing.

If, for example, a target paper has the same ACR than the average of its neighborhood, its S-RCR value is equal to 1. An S-RCR higher than 1 indicates that the paper has a relevance signal stronger than its co-citation network. Similarly, an S-RCR value lower than 1 indicates low relevance of the paper within its neighborhood.

Since the ACR function can be computed in $\Theta(1)$, the time complexity to compute the S-RCR of a given paper p is $\Theta(|N_p|)$, where $|N_p|$ is the neighborhood size of paper p. This time complexity is lower than the time complexity of the classic RCR since, to compute the classic version of RCR, we need to perform a linear regression on p's neighborhood.

G.4 Other Features

Besides the S-RCR metric, we performed a set of tests using features that ended up not being used in our final ranking. These features include paper raw citation counts and normalizations, distinct aggregations of citation-based metrics of authors and publication venues, among others. We also tried some Random Walk techniques, like an application of PageRank on the paper citation graph and a Random Walk on a heterogeneous Pseudo-Tripartite graph composed by paper, author and venue nodes. In this last approach, there is an edge between an author a and a paper p if author a is one of the authors of paper p. Also, there is an edge between paper p and venue v if paper p was published in venue v. The interaction between papers was given by the paper citation graph. While we believe that a proper investigation of this last approach would lead to effective results, it depends on the calibration of the parameters needed to control the amount of probability mass between nodes of distinct types. A plus of our final approach based on the proposed S-RCR metric is that it produces effective results without the need of any parameter tuning.

G.5 Experiments

In this section, we report some experiments performed by our team during the 1st phase of the WSDM Cup 2016.

G.5.1 Dataset Description

The Microsoft Academic Graph (MAG) [Sinha et al., 2015] is a heterogeneous graph containing scientific publication records, citation relationships between publications – as well as authors, institutions, journals, conferences, and fields of study. The file size (zipped) is 37GB and it contains individual information about more than 120 million papers. During the competition, three versions of this dataset were released. Here, we describe only the last one (version of Nov. 6, 2015).

Table G.1: Relevant statistics

Papers with citation information	49,870,036
Papers without citation information	71,017,797
Total number of papers	120,887,833
Average neighborhood size	891

In Table G.1, we present some statistics of the MAG dataset that are relevant to this work. Notice that a large fraction (59%) of the papers in this dataset have no information on citations, that is, the paper can be represented as a node in the citation graph that has neither inlinks nor outlinks. There are two possible reasons for this to happen. The first alternative is the zero degree (i.e., both in- and out-degree) is a true representation of reality, it is a paper that in fact does not receive any citation yet and does not cite any other paper. The second alternative is due to the fact that any big repository offers an approximation of the reality, which also happens in the MAG dataset. Collecting and organizing a real-world dataset of such size is not a trivial task. In fact, it is a process that involves the treatment of huge amounts of semi-structured data, which usually causes some inconsistencies.

In Figure G.2, we plot the distribution of neighborhood sizes. To compute this distribution, we consider only the approximately 50 million papers that have information on citations. This figure helps us to characterize the neighborhood sizes of the graph. We notice that the neighborhood sizes follow a long tail distribution, there are many papers with just a few neighbors and few papers with large neighborhoods.

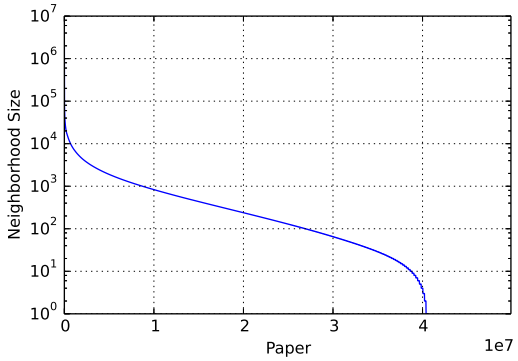

Figure G.2: Distribution of neighborhood sizes

G.5.2 Submissions

Our team made many submissions in the first phase of the competition. In this section, we discuss some of them. As mentioned previously, each submission is a ranking where each item is a paper ID and a probability of that paper being important. The only evaluation sign available was the score reported by the competition's page for each submission, a value between zero and one representing the quality of the submitted ranking — the higher the better. While we know that the evaluation score is based on previously computed pair-wise expert judgments, the exact evaluation metric was not disclosed by the organizers.

In Table G.2, we present the evaluation scores our rankings received in the first phase of the competition and also the time elapsed to produce each submission file.

Table G.2: Our ranking scores in the 1st phase of WSDM Cup

	Leaderboard		
Feature	Public	Private	Time
Citations	0.675	-	0h08m
PageRank	0.687	-	1h29m
ACR	0.685	-	0h16m
S-RCR	0.697	0.671	1h50m

A first guess for one who is somehow familiar with the problem is to count paper citations in order to rank them. It was our first submission and received the score of 0.675 in the public leaderboard. Using PageRank is also a natural approach to rank papers in a citation graph. Our PageRank-based submission scored 0.687, which represents an improvement of 1.8% over citation counts. Before submitting the S-RCR, we submitted the ranking produced by its component ACR (see Section G.3.2). The ACR submission was scored higher than citation counts but lower than PageRank.

Finally, our highest scored submission was based on the S-RCR metric. Scoring 0.697 in the public leaderboard, it corresponds to an improvement of 3.3% over citations.

Since the ranking produced by S-RCR received the highest public score among our submissions, we used it as the final ranking of the first phase. As part of the competition, it was evaluated using a private test set and received the score of 0.671, which was sufficient to bring our team UFMG/LATIN to the 3rd position at the end of the first phase. Eight teams were promoted to the second phase of the competition: the score of the 1st-placed team was 0.684, while the score of the 8th-placed team was 0.642. It is noteworthy that our score dropped a little in the final evaluation of the 1st phase, from 0.697 to 0.671. Some teams experienced higher drops, apparently due to overfitting.

We ran our experiments on a machine with 64 GB RAM, 24 processors of 3.33GHz — Intel(R) Xeon(R) CPU X5680 — under Ubuntu 14.04.2 LTS. While we did not take advantage of the full computational power available (by not using parallelism in a 24-cored machine), the processing times were pretty low for a graph of such size. The time elapsed to produce our final submission file was of only 1h50m.

Critical parts of our approach, like graph processing, were implemented in C++, while other parts, like intermediate analysis or file treating/formatting, were implemented in Python — Pandas and Jupyter were useful tools.

G.6 Discussion

In this work, we have proposed the S-RCR metric and applied it to produce static rankings of academic papers in the Microsoft Academic Search dataset. The interesting point here is that a single well-designed feature (which is a simplification of a more complex one) was able to produce effective results, promoting our team to the 3rd place in the first phase of the WSDM Cup 2016 competition. This fact reinforces the argument that feature engineering is at least as important as complex models, since we apply a single well-designed feature that leads to better results than complex models with the advantage of less tuning and less computational effort. Also, single features tend to be more interpretable. A future direction that is worth investigating is the impact of using the S-RCR metric together with other features in learning to rank techniques. Another direction is to study approaches to address ranking ties, specially how to break ties between papers with no information on citations. Using reputation-based metrics [Ribas et al., 2015a,b] seems to be a reasonable approach to address these issues.