QLectives – Socially Intelligent Systems for Quality
Project no. 231200

Instrument: Large-scale integrating project (IP)
Programme: FP7-ICT

Deliverable D.1.3.2

Social quality selection in science and media

Submission date: 2011-03-01

Start date of project: 2009-03-01 Duration: 48 months

Organisation name of lead contractor for this deliverable:
University of Fribourg

Project co-funded by the European Commission within the Seventh Framework Programme (2007-2013)		
Dissemination level		
PU	Public	**X**
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidental, only for members of the consortium (including the Commission Services)	

Document information

1.1 Author(s)

Author	Organisation	E-mail
Matúš Medo	University of Fribourg	matus.medo@unifr.ch

1.2 Other contributors

Name	Organisation	E-mail
Nigel Gilbert	University of Surrey	n.gilbert@surrey.ac.uk
Alastair Gill	University of Surrey	a.gill@surrey.ac.uk
Stanislao Gualdi	University of Fribourg	staslanio@gmail.com
Linyuan Lü	University of Fribourg	babyann519@hotmail.com
Maria Xenitidou	University of Surrey	m.xenitidou@surrey.ac.uk
Chi Ho Yeung	University of Fribourg	chbyeung@gmail.com
Yi-Cheng Zhang	University of Fribourg	yi-cheng.zhang@unifr.ch
Tao Zhou	University of Fribourg	zhutouster@gmail.com

1.3 Document history

Version#	Date	Change
V0.1	07 December, 2010	Starting version, template
V0.2	21 December, 2010	First draft completed
V0.3	16 January, 2011	Extended and corrected version submitted for internal revision
V0.4	09 February, 2011	Corrected version
V1.0	13 February, 2011	Approved version to be submitted to EC

1.4 Document data

Keywords	complex networks, algorithms, quality, reputation, leadership
Editor address data	matus.medo@unifr.ch
Delivery date	17 February, 2011

1.5 Distribution list

Date	Issue	E-mail
	Consortium members	QLECTIVES@list.surrey.ac.uk
	Project officer	Jose.FERNANDEZ-VILLACANAS@ec.europa.eu
	EC archive	INFSO-ICT-231200@ec.europa.eu

QLectives Consortium

This document is part of a research project funded by the ICT Programme of the Commission of the European Communities as grant number ICT-2009-231200.

University of Surrey (Coordinator)
Department of Sociology/Centre
for Research in Social Simulation
Guildford GU2 7XH
Surrey
United Kingdom
Contact person: Prof. Nigel Gilbert
E-mail: n.gilbert@surrey.ac.uk

University of Fribourg
Department of Physics
Fribourg 1700
Switzerland
Contact person: Prof. Yi-Cheng Zhang
E-mail: yi-cheng.zhang@unifr.ch

Technical University of Delft
Department of Software Technology
Delft, 2628 CN
Netherlands
Contact Person: Dr Johan Pouwelse
E-mail: j.a.pouwelse@tudelft.nl

University of Warsaw
Faculty of Psychology
Warsaw 00927
Poland
Contact Person: Prof. Andrzej Nowak
E-mail: nowak@fau.edu

ETH Zurich
Chair of Sociology, in particular
Modelling and Simulation
Zurich, CH-8092
Switzerland
Contact person: Prof. Dirk Helbing
E-mail: dhelbing@ethz.ch

Centre National de la Recherche Scientifique, CNRS
Paris 75006,
France
Contact person : Dr. Camille ROTH
E-mail: camille.roth@polytechnique.edu

University of Szeged
MTA-SZTE Research Group on
Artificial Intelligence
Szeged 6720, Hungary
Contact person: Dr Mark Jelasity
E-mail: jelasity@inf.u-szeged.hu

Institut für Rundfunktechnik GmbH
Munich 80939
Germany
Contact person: Dr. Christoph Dosch
E-mail: dosch@irt.de

QLectives introduction

QLectives is a project bringing together top social modelers, peer-to-peer engineers and physicists to design and deploy next generation self-organising socially intelligent information systems. The project aims to combine three recent trends within information systems:

- **Social networks** - in which people link to others over the Internet to gain value and facilitate collaboration
- **Peer production** - in which people collectively produce informational products and experiences without traditional hierarchies or market incentives
- **Peer-to-Peer systems** - in which software clients running on user machines distribute media and other information without a central server or administrative control

QLectives aims to bring these together to form Quality Collectives, i.e. functional decentralised communities that self-organise and self-maintain for the benefit of the people who comprise them. We aim to generate theory at the social level, design algorithms and deploy prototypes targeted towards two application domains:

- **QMedia** - an interactive peer-to-peer media distribution system (including live streaming), providing fully distributed social filtering and recommendation for quality
- **QScience** - a distributed platform for scientists allowing them to locate or form new communities and quality reviewing mechanisms, which are transparent and promote

The approach of the QLectives project is unique in that it brings together a highly inter-disciplinary team applied to specific real world problems. The project applies a scientific approach to research by formulating theories, applying them to real systems and then performing detailed measurements of system and user behaviour to validate or modify our theories if necessary. The two applications will be based on two existing user communities comprising several thousand people – so-called "Living labs", media sharing community tribler.org; and the scientific collaboration forum EconoPhysics.

Executive summary

The body of scientific literature is growing at an accelerating pace and to keep an overview of relevant publications hence requires more and more effort. Automated methods for evaluating contribution, composition, and similarity of papers can be welcome tools to master this information abundance. While we mostly speak about scientific publications in this deliverable, majority of its content can be applied equally well to quality perception/detection/promotion in media and hence used directly in the media-related part of the QLectives project, QMedia.

In Chapter 2 we report present preliminary results from work which seeks to better understand quality in scientific online settings. We then make recommendations for the design and implementation of self-organizing quality systems (such as scientific communities in which scholars share, evaluate, and access relevant information). To better understand "quality", we use quantitative and qualitative methods to study results from experimental studies, data-mining of weblogs, scientific resources and discussions with scientists.

Based primarily on the experimental part of our study and secondarily on insights from our research overall, we note that:

1. Quality in scientific contexts is usually defined in distinct—but also partly in overlapping—terms, in comparison with more general concepts of quality.

2. Quality in science specifically relates to issues such as replicability, novelty, independence, methods, and clarity of contribution. In addition, quality is treated as a process of mentoring and supervision.

3. As we can see from (2) there is an overlap between how quality is conceptualized and current filters or processes used to establish quality, such as mentoring and supervision, peer-reviewing, and being published in (high quality) journals.

To summarize, we examine the fundamental concept, or concepts, relating to "quality". We have noted that aspects of quality are in some cases specific to science, and that these relate to both the core characteristics of the science, as well as to the apparent misappropriation of measurements of quality. Based on these findings, we make recommendations for the design of self-organizing quality systems, reiterating the need for: (i) science-specific measures and algorithms, and (ii) different measures, roles and levels within the community, including experts, peers and self-report/self-filtering.

In Chapter 3 we introduce a novel theoretical framework based on a simple random walk process on a directed network. The main contribution of this chapter relies in introducing the concept of influence (or, equivalently, passing probability in random walk) which is particularly motivated and suitable for directed acyclic networks (this class of networks is typically represented by citation data). We illustrate the use of this concept on three distinct examples: detecting seminal papers (using the total influence), tracing the composition of scientific papers (using the influence itself), and creating a new similarity measure (which we employ in a recommendation algorithm and evaluate the performance of resulting recommendations).

In Chapter 4 we again study a directed network. This time it is a network which connects individuals. Similar networks underlie popular online social services such as Delicious.com or Academia.edu where a user may opt to receive information approved by other users (bookmarks in the case of Delicious.com and scientific papers or research interests in the case of Academia.edu). Inspired by the classical PageRank algorithm, we present a new method for uncovering leading individuals in the network. An extensive data set obtained from Delicious.com is used to show that our new algorithm has several advantages over PageRank: it is parameter-free, it discovers the leaders more effectively, and it is more resilient to imperfections of the data.

Contents

Chapter 1

Introduction

Due to limited personal perception ability, abundance of sources (omnipresent in modern society) is an important issue: When an individual receives many pieces of information (often mutually contradictory), rational cognitive mechanisms are often insufficient to deal with them. Consequently, the possibility to recognise the quality of a source is important. We report in Chapter 2 preliminary results from a survey which seeks to better understand quality in scientific online settings. Then we turn to algorithms that can be used to assess or promote quality in different situations. In particular, we discuss in Chapter 3 algorithms for automatic evaluation of similarity and importance of scientific papers and in Chapter 4 algorithms for automatic detection of important users—leaders—in directed social networks.

We stress that while we mostly speak about scientific publications in this deliverable, majority of its content can be applied equally well to quality perception, detection and promotion in media and hence used directly in the media-related part of the QLectives project, QMedia.

Chapter 2

Scientific quality: perceptions and recommendations

Almost a decade ago, Neus (2001) noted the negative impact of the Internet upon matters of ”information quality”. With the Internet making publishing virtually cost-free, many of the processes which had evaluated the ”quality” of information in hard printed copy are now irrelevant, but what are the alternatives? The problem is that quality cannot be directly measured, it is dependent both on items being evaluated and the point in time of the evaluation, and it is often emotive and politicized. It is also personal, with similar measures proving divergent in practice (e.g., Martens and Martens (2001); Bartneck and Hu (2009)). Here we report preliminary results from work which seeks to better understand quality in scientific online settings. We then make recommendations for the design and implementation of self-organizing quality systems (such as scientific communities in which scholars share, evaluate, and access relevant information).

2.1 Data collection

Exemplar quality words were collected using an on-line survey, run between August and October 2010. These were collected as part of a study of the prototypical structure of "quality" descriptions (Rosch, 1978; Kearns and Fincham, 2004). The survey contained a number of questions relating to quality, but we only describe two relevant to the current study here: these asked participants to "write down the words and phrases that you associate with quality", specifically in relation to scientific research activity. Participants were encouraged to produce both positive and negative examples. Participants were recruited via UK universities through a personal network of contacts of the first author, with participants encouraged to distribute information to friends/contacts inside and outside universities. Students, staff, and members of the public were all encouraged to participate. Demographic information about participants is as follows: Of the 249 participants who provided valid responses, 173 provided information about their gender, 80 were male (46.2%); 175 provided age information according to 6 categories, 20 years or under (53; which is 30.2%), 21–30 years (89; 50.9%), 31–40 years (18; 10.3%), 41–50 years (12; 6.9%), 51–60 years (2; 1.1%), and 61 years and over (1; 0.6%); 121 considered English to be their first language (70.3% of 172), and the highest qualification of 173 participants was "school leaving" for 77 (44.5%), "undergraduate degree" for 44 (25.4%), "taught postgraduate degree" for 28 (16.2%), and "research post-graduate degree" for 24 (13.9%). Of 168 participants, 114 identified themselves as taught students, that is consumers of science (67.9%), whereas 26 identified themselves as involved in research, and thus science producers (15.5%).

2.2 Data processing

Responses (i.e. words/phrases) from each participant were combined together. This produced a total of 1507 science exemplar items (1099 unique), an average of just over 6 items per participant, and 2816 general exemplar items (1559 unique), and an av-

erage of 11.3 per person. Lists of the individual items were then processed using the part-of-speech tagger within the Wmatrix corpus comparison tool (Rayson, 2009), identified commonly occurring multi-word units and set phrases, and provided frequency counts. Each item (word/multi-word unit) with a frequency of at least 2, was then saved (this resulted in 157 science and 320 general exemplar words or phrases). The second stage of processing combined words with a frequency of 1 and items consisting of longer phrases and descriptions into a text; n-gram software (Banerjee and Pedersen, 2003) was then used to identify 1-, 2-, 3-, and 4-grams with a frequency of 2 or more, which occurred in the data. N-grams which did not relate to content (e.g., 'of the'), or which duplicated items already identified in first stage of analysis were removed by hand. The second stage of analysis identified a further 155 science tiems and 92 general items resulting overall in 312 science items and 412 general items. Note that since we are examining how the exemplar terms are used in the blog texts, we do not stem these terms to give a root word form.

2.3 Coding of data

Coding was conducted following basic principles in the grounded theory approach. Words were initially coded "in vivo" or using synonyms. Codes were then clustered into categories taking into account the phenomenon in question—"quality"—and the contexts in which it was prompted—scientific and general respectively (see Strauss and Corbin (1998)). The generalization of codes into categories (Charmaz, 1994) consisted of a conceptual labeling guided by the questions "of which category is the item before me an instance?" and "what can we think of this as being about?" (Lofland and Lofland, 1995). Coding categories for the science context are shown in Tab. 2.1.

Coding category	Definition
Appearance	visible characteristic
Clarity	clear, explanation
Context	specific, theme-wise
Correctness	correctness and measures of it
Depth	inner understanding
Ethics	ethics and measures of it
Evaluation	evaluation and assessment
Function	function or application
Impact	impact
Information	information-oriented
Intelligence	intelligence
Novelty	novelty, time (and significance)
Process	the research process
Professional	professional
Progress	progress oriented
Proof	results, evidence, testing and verification
Quantity	quantity
Relational	in relation to the scientific community
Resilience	strength
Standards	quality standards and quality assurance
Structure	research/paper structure
Trust	trust, reputation and trust structures
Value	value and material

Table 2.1: Coding categories used for grouping exemplar words and phrases.

General Word	Freq.	Science Word	Freq.
good	80	accurate	29
expensive	59	reliable	28
well made	43	thorough	20
long lasting	40	accuracy	12
reliable	37	useful	11
excellence	26	precise	11
durable	26	good	10
value	25	valid	8
perfection	22	unbiased	7
high standard	17	safe	7
bad	15	reproducible	7
worth	14	relevant	7
value for money	14	bad	7
price	14	well written	6
characteristic	14	trustworthy	6
strong	13	research	6
reliability	13	replicable	6
useful	12	precision	6
standard	12	peer reviewed	6
excellent	12	well thought out	5

Table 2.2: Twenty most frequently reported items relating to quality in scientific and general contexts.

2.4 Results

To better understand "quality", we use quantitative and qualitative methods to study results from experimental studies, data-mining of weblogs, scientific resources and discussions with scientists. The most frequently reported words from our prototype conceptualization survey relating to quality in scientific and general contexts are presented in Table 2.2.[1]

Based primarily on the experimental part of our study and secondarily on insights from our research overall, we note the following findings:

1. Quality in scientific contexts is usually defined in distinct—but also partly in overlapping—terms, in comparison with more general concepts of quality. For example, scientific concepts of quality tend to be more detailed and specific; in

[1]These results are preliminary as complete evaluation of our survey has not been finished yet.

general contexts, quality was mainly viewed from a consumer-product perspective.

2. Quality in science specifically relates to issues such as replicability, novelty, independence, methods, and clarity of contribution. In particular, the main analytic categories are: clarity, correctness, depth, novelty, process-oriented (referring to the research process), results or proof oriented (empiricism and experimental methods), peer-reviewing, trust, appearance and value (the last two are also found in general contexts). One of the main ways of framing 'quality in science' in discussions with scientists is in terms of a distinction between global, commonly recognized, established external metrics on the one hand (such as RAE/REF in the UK), and personal understandings, on the other. In addition, quality is treated as a process of mentoring and supervision.

3. As we can see from (2) there is an overlap between how quality is conceptualized and current filters or processes used to establish quality, such as mentoring and supervision, peer-reviewing, and being published in (high quality) journals. In the case of this latter issue, such definitions lead to a circular process of defining quality by its outcomes, which themselves, do not capture quality research (for example, even journals which could be described as "low quality", may also be peer-reviewed).

The ways that these are treated by their users—scientists—should be acknowledged in self-organizing quality systems. Our recommendations for the design and implementation of self-organizing quality systems are three-fold (linking quality concepts and understandings to users and processes):

1. When evaluating quality in science, the metrics used need to be science-specific. From our study, we note that important science-specific concepts are replicability, novelty, independence, methods, clarity of contribution, correctness as well

as peer-reviewing, mentoring and supervision. We also note that value and appearance are more general descriptions of quality that also are seen as relevant to science.

2. Quality can be evaluated in a variety of different ways, for example, using a combination of expert and general peer evaluation, as well as automated and self-reported metrics.

(i) With specific reference to the concepts of scientific quality, expert evaluation seems best suited to topic and topic relevance, novelty, methods replicability and correctness. Peer or community evaluation can be applied to more general concepts, such as independence, clarity, and possibly in some general sense, value. Note that these evaluations (and contributions) are not envisioned to be performed by the traditional peer-review gatekeepers, but rather are performed by community members, with their academic reputation and credibility as an author or reviewer, and additional information (such as their network, personal preferences, or how mainstream they are). This information could then be used to weight their assessment by the recommender algorithm used to search the site.

(ii) A significant part of quality assessment in future must, if only out of necessity, be automated. Martens and Martens (2001) talk about how an artist "feels" quality, and many of our findings for general concepts of quality relate to the idea of quality resulting from "craftsmanship". However, given the amount of information available, this concept, and as some suspect, peer-review, is not sustainable. Therefore, the more that this process can be automated - as well as distributed across a community of users - the better. Although some of the evaluations noted above must relate to human judgment, this information can be fed into algorithms which use it to calculate the relevance or importance of a scientific article. One example could be 'translating' expert and community evaluations into

topic and community specific rating scales and descriptions. Another approach to streamlining the 'quality' filtering process is that adopted by the open access publisher Public Library of Science (PLoS; `http://www.plos.org`), which places burden of meeting strict guidelines relating to ethical procedure and independence of the research upon the author before submission.

In addition, increasingly sophisticated natural language processing techniques (see Paterson et al. (2010)) can be used to evaluate quality of writing (appearance), similarity to other work, and objectivity/independence - all of which can be usefully incorporated in recommender system algorithms. Finally, we think that mentoring and supervision pertain to user role and level. These could be embedded in the algorithm in terms of 'type of contribution', 'activity' (using activity metrics) and 'type of instances' installed.

To summarize, although the burgeoning amount of scientific information available puts strains on existing ways of managing scientific literature and knowledge, it does open up exciting opportunities to consider basic assumptions at the heart of science. In this work we examine the fundamental concept, or concepts, relating to "quality". We have noted that aspects of quality are in some cases specific to science, and that these relate to both the core characteristics of the science, as well as to the apparent misappropriation of measurements of quality. Based on these findings, we make recommendations for the design of self-organizing quality systems, reiterating the need for: (i) science-specific measures and algorithms, and (ii) different measures, roles and levels within the community, including experts, peers and self-report/self-filtering. Finally, we note the importance of automated measures, both in terms of recommender system algorithms, and also in automated measures derived from natural language processing techniques.

Chapter 3

Information extraction from citation data

The body of scientific literature is growing at an accelerating pace and to keep an overview of relevant publications hence requires more and more effort. Automated methods for evaluating contribution, composition, and similarity of papers can be welcome tools to master this information abundance. In this chapter we propose a simple random walk process as a basis for a theoretical framework. After showing that this framework is able to measure influence of one paper on another, we use it to answer specific questions such as how to detect seminal papers or how to measure similarity of papers using solely citation data. We test our ideas on an extensive data set of scientific citations. Due to our choice of the dataset, this work can be considered as a contribution to the active field of bibliometrics (Borgman and Furner, 2002). However, ideas and methods presented are general and can be applied to any other data which can be represented by an acyclic directed network.

3.1 Data

Our database consists of all 463 154 papers published in journals of the American Physical Society (APS) before 2010, with some papers dating as far back as to 1893. The set of journals comprises Physical Review, Physical Review A/B/C/D/E/I, Physical Review Letters, Physical Review STAB, Physical Review STPER and Review of Modern Physics.[1] For each paper we have a list of references and extensive additional metadata related to the paper (such as the title, print date, and PACS codes). Citation data largely obeys the time ordering of papers. To fully benefit from this feature, we do not consider the tiny part of citations that are between papers of the same print date. Finally, we remove all papers that do not appear in the list of references (*i.e.*, they do not refer to the others and are not referred by them). After the described cleaning of the data, we are left with 449 394 papers and 4 667 863 references (which is 97% and 99.6% of the numbers in the original data set, respectively).

In the network representation, this data corresponds to a directed network $\mathcal{G}(N, L)$ composed of N nodes (papers) and L directed edges (citations). We assume that links follow from a citing paper to cited papers and hence each node's in-degree equals to the node's citation count. Due to removing links connecting papers with identical print dates, the time ordering is strictly preserved and the network is hence acyclic.

3.2 General framework

When studying citation data, the network approach provides us with two different perspectives. The first one is based on spreading influence of a paper through the citation network from a given paper to later works. The second one is based on determining credit of a paper based on credit of papers that cite this paper (similarly as node score is evaluated in the classical PageRank algorithm). In the following paragraphs

[1] This data was obtained from the APS by a personal request and hence cannot be automatically shared within the QLectives project. Anyone interested in obtaining and studying this data can submit an access request via the APS page `https://publish.aps.org/datasets`).

we establish an analytical framework which allows us to consider both perspectives in a simple and systematic way.

Formally, for a given node x we assign an N-dimensional vector $\boldsymbol{G}_x$ whose ith component represents the probability that the random walk starting at node x passes through node i. We further define $\mathcal{A}_x$ as the set of nodes reachable from x (x's ancestors) and $\mathcal{P}_x$ as the set of nodes from which x is reachable (x's progeny); $A_x := |\mathcal{A}_x$ and $P_x := |\mathcal{P}_x$ are the sizes of respective sets. Since the network is acyclic, $\mathcal{A}_x \cap \mathcal{P}_x = \emptyset$. The master equation for $\boldsymbol{G}_x$ reads

$$\boldsymbol{G}_x = \mathsf{A}\boldsymbol{G}_x \tag{3.1}$$

where $\mathsf{A}_{ij} = 1/k_i^{\text{out}}$ if i cites j and $\mathsf{A}_{ij} = 0$ otherwise and the boundary condition is given by setting $(\boldsymbol{G}_x)_x = 1$ (because random walk certainly passes through the starting point). Components of $\boldsymbol{G}_x$ are positive for all papers in $\mathcal{A}_x$, for all other papers they are zero.

Although the equation above defines dynamics similar to that of the Google's PageRank algorithm, there are two important differences: (i) by contrast to PageRank, $\boldsymbol{G}_x$ depends on the choice of the starting node x, (ii) passings through various nodes are not mutually exclusive events and thus $\boldsymbol{G}_x$ is not normalized to one. Though $\boldsymbol{G}_x$ is not normalized, the norm $\|\boldsymbol{G}_x\|_1$ is constrained by the network itself as it is bounded by one plus the number of x's ancestors as well as by one plus the maximum number of steps that one can walk from x (the latter bound is usually much more restrictive than the former).

It is instructive to look on the network from a different point of view which is based on an analogy with genes spreading in population. In the context of scientific papers, we could introduce vectors of "genetic" composition of papers and assume that each paper's vector is obtained by averaging the vectors of the cited papers (inherited knowledge) and by adding a vector representing this paper's contribution

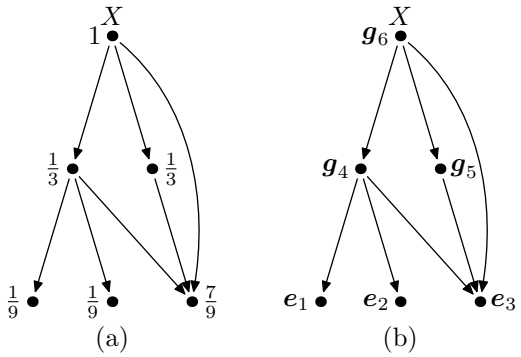

Figure 3.1: Comparison of random walk starting at X (a) with passing of "genes" (b). According to the description in the main text, $\boldsymbol{g}_4 = \frac{1}{3}(\boldsymbol{e}_1 + \boldsymbol{e}_2 + \boldsymbol{e}_3) + \boldsymbol{e}_4$, $\boldsymbol{g}_5 = \boldsymbol{e}_3 + \boldsymbol{e}_5$, $\boldsymbol{g}_6 = \frac{1}{3}(\boldsymbol{g}_4 + \boldsymbol{g}_5 + \boldsymbol{e}_3) + \boldsymbol{e}_6 = \frac{1}{9}\boldsymbol{e}_1 + \frac{1}{9}\boldsymbol{e}_2 + \frac{7}{9}\boldsymbol{e}_3 + \frac{1}{3}\boldsymbol{e}_4 + \frac{1}{3}\boldsymbol{e}_5 + \boldsymbol{e}_6$. Coefficients in $\boldsymbol{g}_6$ coincide with corresponding passing probabilities in (a).

(new knowledge). A similar model based on genetic composition of scientific papers has been shown to reproduce many quantitative features of science Gilbert (1997). Fig. 3.1b illustrates this on a simple citation network where $\boldsymbol{g}_6 = \frac{1}{2}(\boldsymbol{g}_4 + \boldsymbol{g}_5) + \boldsymbol{e}_6$ and the base vector $\boldsymbol{e}_6$ is orthogonal to vectors of all previous papers. When the genetic vector of a paper is written in terms of base vectors, coefficients of respective base vectors equal the passing probabilities obtained by the random walk approach. Since it easy to check that $\boldsymbol{G}_x = \boldsymbol{g}_x$, we can say that the previously introduced $\boldsymbol{G}_x$ represents a "genetic" composition of paper x. From this genetic perspective, the lack of normalization for $\boldsymbol{G}_x$ is due to the new contribution that each paper introduces, such that more recent papers will on average tend to accumulate more influence (in the next section we will try to clarify the consequences of introducing normalization).[2] In order to obtain a compact formalism, we finally construct an $N \times N$ matrix G where the x-th column is given by the vector $\boldsymbol{G}_x$. Components of this matrix hence have simple interpretation: G_{yx} represents what we consider to be the genetic influence of paper y on paper x.

In this work we make use of this formalism in various ways: by measuring the

[2]We briefly note that by changing the gene-spreading mechanism we could also assume that each paper takes a portion $1 - \delta$ of its genetic code from the papers it cites and the remaining δ portion of its composition consists of its new code (which represents the paper's contribution). This would result in $\|\vec{G}_x\|_1 = 1$ and, in turn, to the total influence of papers given by the usual PageRank vector.

importance of a paper by its impact on future generations, by studying "influence" composition of selected papers, and by using the information contained in the relation vectors $\boldsymbol{G}_x$ to infer paper similarity. They are discussed in detail in the following sections.

3.3 Evaluating the impact of papers

We now try to detect seminal/fundamental works in different fields. Although one may say that those works are best to be discovered based on peers' opinions, we believe that it is interesting to attempt to answer such a peculiar question using only structural properties of the citation network. Furthermore, the notion of a fundamental paper can be translated to other directed networks and hence the method that we provide could be helpful also for other purposes. In our interpretation, seminal/fundamental papers are those creating novel branches of research, so that all papers in a certain field should directly or indirectly refer to them and to just a few papers out from the given field. We believe that existing importance measures developed for directed networks (such as popularity, Page Rank, and modularity) do not solve this problem because they do not address the question of originality and focus at best on a local neighborhood of a paper.

We first quantify the overall influence of a paper in the population which is in our case represented by the paper's total "genetic" impact on subsequent papers. Following the above-introduced formalism, the aggregate impact of paper x can be evaluated as

$$I_x = \sum_i G_{xi} \tag{3.2}$$

where the number of non zero terms in the summation is equal to the size of x's progeny, P_x. The quantity I_x can be considered as an analog of Page Rank. We can benefit from the similarity with PageRank by computing $\boldsymbol{I}$ in an analogous recursive

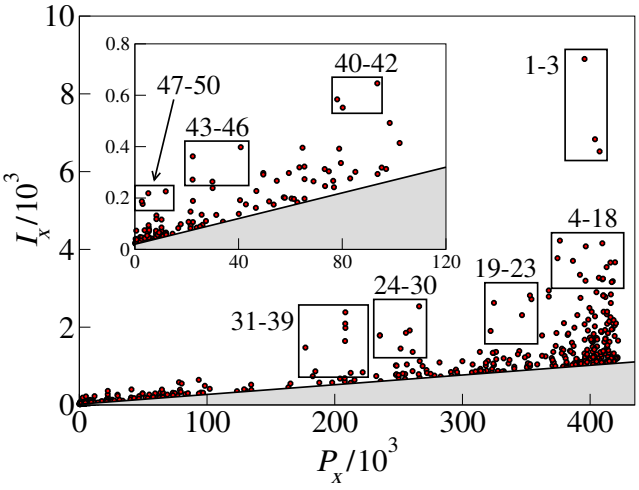

Figure 3.2: Total paper influence I_x versus its progeny size P_x. Only papers with $I_x > 20 + P_x/400$ are shown. Details about the outliers (marked with boxes) are listed in Tab. 3.1. Numbers correspond with papers listed in Tab. 3.1.

way

$$\boldsymbol{I}^{t+1} = \mathbf{1} + \mathsf{A}\boldsymbol{I}^t \tag{3.3}$$

where A is the same as in Eq. (3.1) and $\mathbf{1}$ is a vector composed of ones. The fact that the network is acyclic implies the convergence of Eq. (3.3) and $\|\boldsymbol{I}\|_1 \leq N(N+1)/2$ (the equality holds in the extreme case of a chain of citations). However, the aggregate influence contained in components of $\boldsymbol{I}$ is not informative by itself since old papers tend to have higher impact (simply because their progeny is greater than that of recent papers and their impact I_x is hence obtained by summing over many contributions). Mathematically, I_x can be high because of x has large progeny or because many contributing terms G_{ix} are high. High values of G_{ix} (*i.e*, those that are close to the upper bound one) correspond to papers i that can be almost entirely traced back to x.

To summarize the discussion above, we would like to highlight papers whose values of I_x are high in comparison with their progeny size P_x. In Fig. 3.2 we plot I_x versus P_x for all papers from 1940 to 2009 (we do not consider older papers because they tend to cite just a few papers and together with the limited scope of our data covering only APS-to-APS citations only makes results obtained for them noisy and unreliable). The vast majority of papers lie in the bottom of the graph. For the sake

of clarity we only plot points above the straight line $I_x = 20 + P_x/400$ and the rest of the graph is shaded (only 429 papers out of the total number of 450 000 pass the threshold). Some of the papers above the line are exceptional and can be considered as outliers—they are delimited in the figure with boxes and listed in Tab. 3.1 together with their basic characteristics (title, list of authors, and year of publication). It is of course difficult to judge the quality of this selection. We can however follow the line of other papers facing a similar problem (Chen et al., 2007; Radicchi et al., 2009) and use external quality measures—scientific prizes in this case—to distinguish quality of publications. In Tab. 3.1, for each paper we put a star in the prize column if any of the authors won one of the following international prizes: Nobel Prize, Dirac Medal, Wolf Medal, Boltzmann Medal, Planck Medal, Lorentz Medal, and Isaac Newton Medal.

Since most prizes need long time to be received (our table is almost time-ordered and this bias of scientific prizes towards old papers is well visible by the abundance of stars in the top half of the table), we add an additional distinguishing criterion for prize-free papers: if they are marked as pioneering works in a certain domain on Wikipedia, we mark them with $+$. As can be seen, almost all papers in top 25 are marked with stars and many of the remaining papers are marked with plus symbols—this suggests that our I_x vs P_x criterion has some merit. Yet, there are also a few papers that appear not to be seminal (namely 27,31,32) because they both have little citation count and even after closer investigation they do not seem to be particularly ground-breaking works.[3] Such problems seem to be quite rare (3 papers out of 50) and can be partially due to our limited citation information: our data consists only of citations to and from other APS publications. Some important papers can appear to have a few references (usually because they do not cite other papers of APS—take "Diffusion-Limited Aggregation" in the table below as an example: in the APS data it appears to cite only one paper, the problematic #27 by Rosenstock and Marquardt which hence receives all its credit) and hence transfer all their credit to the minor works that they

[3]Note that paper 27 appeared at top places also in (Chen et al., 2007) where the authors tried to explain this as a consequence of this paper being incidentally cited by a few influential papers.

cite. While in our context they seem to be inspired by only a few works, this may be false impression and the list of exceptional works obtained with full citation data may be biased by this effect.

It is instructive to compare our results with those obtained by evaluating how closed a community $\mathcal{P}_x$ is. For this purpose we introduce "exclusivity" q_x of node x which is obtained by dividing the number of citations by members of $\mathcal{P}_x$ to another member of $\mathcal{P}_x$, by the total number of citations by members of $\mathcal{P}_x$. As a result, $q_x = 1$ corresponds to a completely independent line of research where all papers in $\mathcal{P}_x$ only cite other papers from $\mathcal{P}_x$. In some sense, exclusivity is a generalization of the clustering coefficient which is a standard quantity in the field of complex networks (Newman, 2003). It is also similar to modularity which is frequently used to identify the community structure of networks (Newman and Girvan, 2004).

One could try to detect seminal papers based on their exclusivity but this approach has two main shortcomings. The first (and the biggest) one is that exclusivity itself does not pay any attention to the size of $\mathcal{P}_x$ and the largest possible value $m_x = 1$ is likely to be achieved by marginal papers with small progeny. The second shortcoming is that exclusivity judges all links as equal without taking into account whether citations pointing out of the community are close to the starting paper x or several generations away. By contrast, our quantity I_x naturally gives more weight to "close losses" (papers from the first or second generation that cite many papers not belonging to $\mathcal{P}_x$) and hence one can expect it to be more sensitive to the detailed structure of the citation network.

Fig. 3.3 shows m_x versus t_x where t_x represents the fraction of papers published after x (hence the larger the value, the older the paper). As can be seen, the outliers identified in Fig. 3.2 tend to have higher exclusivity than other papers which agrees with our idea of seminal papers founding their new branches of research. At the same time, papers with high exclusivity as compared to other papers from the same time are numerous which makes exceptional papers less distinguishable than in our influence-

#	title	authors	year	prize	PR	CC
1	Theory of Superconductivity	J. Bardeen, L. Cooper	1957	*	2	10
2	Crystal Statistics in a Two-Dimensional Model...	L. Onsager	1944	*	8	87
3	Statistics of the Two-Dimensional Ferromagnet...	H. Kramers	1941	*	54	1645
4	Population Inversion and Continuous Optical Maser	A. Javan, W. Bennett	1961	+	169	14517
5	Theory of the Superconducting State...	H. Fröhlich	1950	*	298	3120
6	The Maser–New Type of Microwave Amplifier,...	J. Gordon, H. Zeiger	1955	+	369	14517
7	Dynamical Model of Elementary Particles Based on...	Y. Nambu, G. Jona-Lasinio	1961	*	24	50
8	Infrared and Optical Masers	A. Schawlow, C. Townes	1958	*	171	2108
9	Resonance Absorption by Nuclear Magnetic Moments	E. Purcell, H. Torrey	1946	*	314	8345
10	The Radiation Theories of Tomonaga, Schwinger, and...	F. J. Dyson	1949	*	96	1435
11	The S Matrix in Quantum Electrodynamics	F. Dyson	1949	*	108	1248
12	Theory of the Fermi Interaction	R. Feynman, M. Gell-Mann	1958	*	28	148
13	A Simplification of the Hartree-Fock Method	J. Slater	1951		22	111
14	A Collective Description of Electron Interaction	D. Pines	1952	*	120	2108
15	Interaction Between the D Shells in the Transitition...	C. Zener	1951	+	42	1101
16	Correlation Energy of an Electron Gas...	M. Gell-Mann	1957	*	23	358
17	Nuclear Induction	F. Bloch	1946	*	43	447
18	Stochastic Problems in Physics and Astronomy	S. Chandrasekhar	1943	*	4	90
19	Self-Consistent Equations Including Exchange And...	W. Kohn, L. Sham	1965	*	1	1
20	Inhomogeneous Electron Gas	P. Hohenberg	1964	*	3	2
21	A Model of Leptons	S. Weinberg	1967	*	6	18
22	Static Phenomena Near Critical Points:...	L. Kadanoff, W. Götze	1967	*	58	355
23	Radiative Corrections as the Origin of Spontaneous...	S. Coleman, E. Weinberg	1973	*	31	75
24	Scaling Theory of Localization:...	E. Abrahams, P.W. Anderson	1979	*	11	24
25	New Measurement of the Proton Gyromagnetic Ratio	E. Williams, P. Olsen	1979	*	150	26327
26	New Method for High-Accuracy Determination of...	K. Klitzing	1980	*	32	134
27	Cluster Formation in Two-Dimensional Random Walk	H. Rosenstock, C. Marquardt	1980		109	217150
28	Diffusion-Limited Aggregation	T. Witten	1981	+	17	64
29	Instabilities and Pattern Formation in Crystal...	J. Langer	1980	+	37	187
30	Maximum Metallic Resistance in Thin Wires	D. Thouless	1977	*	144	490
31	Electronic Structure of $BaPb_{1-X}Bi_XO_3$	L. Mattheiss, D. Hamann	1983		106	4224
32	Bulk Superconductivity at 36 K in $La_{1.8}Sr_{0.2}CuO_4$	R. Cava, R. Van Dover	1987		37	1086
33	Evidence for Superconductivity above 40 K In...	C. Chu, P. Hor	1987		40	606
34	Superconductivity at 93 K in a New Mixed-Phase...	M. Wu, J. Ashburn	1987	+	19	102
35	Self-Organized Criticality: An Explanation of...	P. Bak, C. Tang	1987	+	16	47
36	Phase Organization	C. Tang, K. Wiesenfeld	1987		296	18045
37	Inflationary Universe: A Possible Solution to...	A. Guth	1981	*	46	40
38	Trapping of Atoms by Resonance Radiation Pressure	A. Ashkin	1978	+	1007	9611
39	Ergodic Theory of Chaos and Strange Attractors	J. Eckmann, D. Ruelle	1985	*	56	239
40	Teleporting an Unknown Quantum State via...	C. Bennett, G. Brassard	1993	+	53	26
41	Bose-Einstein Condensation in a Gas of Sodium Atoms	K. Davis, M. Mewes	1995	+	63	27
42	Evidence of Bose-Einstein Condensation in...	C. Bradley, C. Sackett	1995	+	99	51
43	TeV Scale Superstring and Extra Dimensions	G. Shiu, S.-H. Tye	1998		216	3991
44	Small-World Networks: Evidence for a Crossover Picture	M. Barthélémy, L. Amaral	1999	+	658	9872
45	Scaling and Percolation in the Small-World Networks	M. Newman, D. Watts	1999	+	778	3378
46	Large Mass Hierarchy from a Small Extra Dimension	L. Randall	1999	+	115	28
47	Statistical Mechanics of Complex Networks	R. Albert, A.-L. Barabási	2002		112	59
48	Negative Refraction Makes a Perfect Lens	J. Pendry	2000	*	279	192
49	Composite Medium with Simultaneously Negative...	D. Smith, W. Padilla	2000	+	433	459
50	Extremely Low Frequency Plasmons in ...	J. Pendry	2001	*	456	1058

Table 3.1: Detailed list of the outstanding papers marked in Fig. 3.2. PR and CC stand for ranking according to PageRank and citation count, respectively.

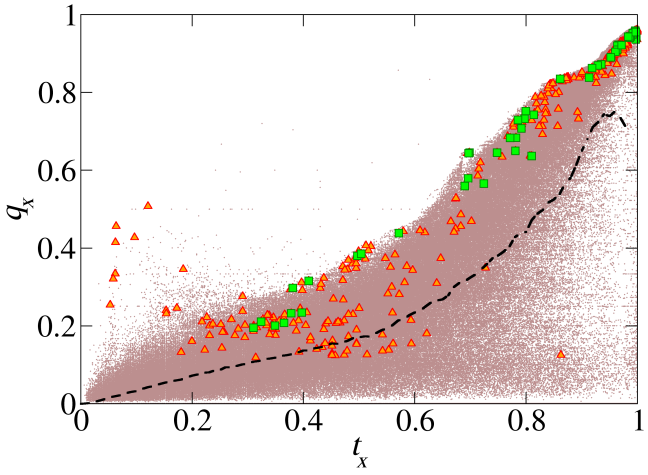

Figure 3.3: Modularity of papers m_x versus their relative age t_x. Papers shown in Fig. 3.2 are marked with triangles and the outliers reported in Tab. 3.1 are marked with squares; the dashed line shows the average exclusivity as a function of t_x.

based approach. Moreover, almost all "outliers" in Fig. 3.3 that do not correspond to the outliers in Fig. 3.2 are papers with very low citation counts which implies that exclusivity indeed performs poorly in identifying seminal papers.

3.4 Studying the composition of scientific publications

Now we use the above-established framework of influence to study which works influenced most a specific selected paper. We chose two highly cited papers from statistical physics: *Dynamic scaling of growing interfaces* by M. Kardar, G. Parisi and Y.-C. Zhang (which we refer to as KPZ) and *Self-organized criticality: An explanation of the* $1/f$ *noise* by P. Bak, C. Tang and K. Wiesenfeld (which we refer to as SOC). According to the Web of Science, these two papers received 2 556 and 3 155 citations, respectively. Tables 3.3 and 3.2 show papers with the highest influence (measured by $G_{KPZ,x}$ and $G_{SOC,x}$, respectively) on our two probe papers; $l_{\min}$ denotes the length of the shortest paths between x and SOC. For SOC, results are rather non-surprising: the most influential papers are those two that are directly cited and in general, the higher the value of $l_{\min}$, the lower the influence of x on SOC.

#	$G_{SOC,x}$	Title	Author(s)	$l_{\min}$
1	0.50	Flicker $1/f$ Noise: Equilibrium Temperature And Resistance Fluctuations	R. Voss, J. Clarke	1
2	0.50	Phase Organization	C. Tang, K. Wiesenfeld, P. Bak	1
3	0.17	Wave-Vector Field Of Convective Flow Patterns	M. Heutmaker, J. Gollub	2
4	0.17	Observation Of A Pulse-Duration Memory Effect In $K_{0.3}MoO_3$	R. Fleming, L. Schneemeyer	2
5	0.17	Discrete Model Of Chemical Turbulence	Y. Oono	2
6	0.13	Thermoluminescence And Changes Of Color Centers In...	J. Sharma	2
7	0.11	$1/f$ Noise From Thermal Fluctuations In Metal Films	J. Clarke, R. Voss	2
8	0.11	Turbulence Theory For The Current Carriers In Solids And A Theory Of $1/f$ Noise	P. Handel	2
9	0.09	Reciprocal Relations In Irreversible Processes 1	L. Onsager	2
10	0.08	Stochastic Problems In Physics And Astronomy	S. Chandrasekhar	2

Table 3.2: Ten most influential works (as measured by $G_{SOC,x}$) for the SOC paper.

The situation is quite different for KPZ where we meet our old friend: the paper on cluster formation by Rosenstock and Marquardt (we already encountered this paper in Tab. 3.1 and as an outlier when plotting PageRank vs number of citations in Chen et al. (2007)). This paper is the most influential for KPZ despite its large shortest distance $l_{\min} = 3$. In this case, the reason lies in the closeness of the part of the network between the Rosenstock-Marquardt paper and KPZ (by papers "between" we mean those that are reachable from RM and at the same time one can reach KPZ from them). This subnetwork contains 53 papers in total and citations by these papers mostly stay in the subset (in 723 cases) and rarely point out from it (in 79 cases)—this allows the influence of RM to propagate efficiently to KPZ. Finally, Fig. 3.4 shows how influence values decay (notably, this decay appears to be of a power-law kind). We can conclude that the proposed framework allows us to quantify mutual influence of papers and that in some situations it can be used to obtain non-trivial results.

#	$G_{KPZ,x}$	Title	Author(s)	$l_{\min}$
1	0.23	Cluster Formation In Two-Dimensional Random Walks: Application To Photolysis Of Silver Halides	H. Rosenstock, C. Marquardt	3
2	0.23	Diffusion-Limited Aggregation, A Kinetic Critical Phenomenon	T. Witten	2
3	0.14	Active Zone Of Growing Clusters: Diffusion-Limited Aggregation And The Eden Model	M. Plischke, Z. Rácz	1
4	0.14	Instabilities And Pattern Formation In Crystal Growth	J. Langer	1
5	0.11	Active Zone Of Growing Clusters: Diffusion-Limited Aggregation And The Eden Model In Two And Three Dimensions	Z. Rácz, M. Plischke	1
6	0.10	Statistical Dynamics Of Classical Systems	P. Martin, E. Siggia	2
7	0.09	Large-Distance And Long-Time Properties Of A Randomly Stirred Fluid	D. Forster	1
8	0.08	Fluctuation-Dissipation Theorems For Classical Processes	U. Deker, F. Haake	1
9	0.08	Calculation Of Dynamic Critical Properties Using Wilson'S Expansion Methods	B. Halperin, P. Hohenberg	1
10	0.08	Interface Motion And Nonequilibrium Properties Of The Random-Field Ising Model	R. Bruinsma	1

Table 3.3: Ten most influential works (as measured by $G_{KPZ,x}$) for the KPZ paper.

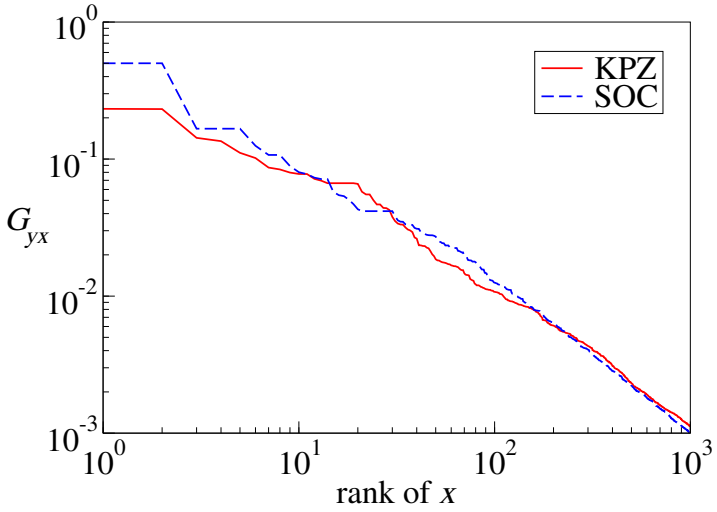

Figure 3.4: Influence of paper x on KPZ/SOC versus the rank of x.

3.5 Similarity definition

As the last example, we use the G matrix to introduce a parameter-free similarity measure for nodes in a citation network. Consider two nodes x and y and a chosen third node z. As described in Sec. 3.2, G's elements G_{xz} and G_{yz} represent the probabilities of passing through nodes x and y when starting in node z. Since the citation network is directed and acyclic, it holds that:

1. probabilities G_{xz} and G_{yz} can be non-zero only when node z is older than nodes x and y,

2. if $G_{xz} > 0$, then $G_{zx} = 0$.

Now we define a new similarity similarity measure S^* as

$$S^*(x,y) = \sum_i \sqrt{G_{ix} G_{iy}}. \tag{3.4}$$

It is also possible to base the similarity on $\min\{G_{ix}, G_{iy}\}$ or $G_{ix}G_{iy}$, for example—we present here the one performing best in our numerical tests. While the lower bound of this similarity is zero, it's upper bound is only given by $\mathcal{A}_x \cap \mathcal{A}_y$ (it's not normalized). We stress that S^* is a parameter-free similarity metric which is important for its implementation in practice.

3.6 Similarity tests

The usual approach to tests of similarity measures is based on judging how well they are able to reproduce missing links in a network (Lü and Zhou, 2010). In practice this means that small part of links (usually 10%) are chosen at randomly and removed from the network and one tries to guess the removed links by seeing which similar nodes are not connected. A similarity measure which is able to fill the network in this

way captures well the network's structure and one may use it for other purposes than link prediction.

In the case of our newly proposed similarity measure S^*, we adopt a slightly different approach and test similarity measures by how good recommendations one may obtain based on them. This change of the testing approach is motivated by potential practical use of recommendation for scientists who often face the problem of searching for relevant literature in their research field. An algorithm able to highlight possible relevant works could therefore be a useful tool for them. Common search engines, like Google Scholar, provide for example a list of relevant papers based on a given search query in combination with a global index (quality measure of individual papers). Personalized recommendation algorithms represent a more refined tool for information filtering as they are usually based on the global data and on user information such as personal view history (Adomavicius and Tuzhilin, 2005). We use various similarity measures to obtain simple recommendation algorithms and compare their resulting performance levels.

Our tests are done as follows. We first divide the data in two parts: all papers published until year 2003 (the sample set—it contains approximately 75% of papers) and all papers published after 2003 (the probe set). Then we find 20 most-cited articles published in each core APS journal in 2003 (we consider seven journals: Phys. Rev. Lett., Rev. Mod. Phys. and Phys. Rev. A–E) and take their last authors if they published at least one paper with APS also after 2003.[4] Recommendations are made for each test author separately on the basis of papers published by this author in 2003 (we could also base our recommendation on all author's papers published until 2003 but that would mix the author's all past interests together, some of which were probably long forgotten in 2003). Denoting the set of papers published by author α in 2003 as

[4]Since we are taking the last authors of important papers, these authors are presumably senior scientists with extensive history and high activity in the field.

$\mathcal{U}_\alpha$, the recommendation score of paper x is given by its similarity with all y in this set

$$r_x = \sum_{y \in \mathcal{U}_\alpha} S^*(x, y). \tag{3.5}$$

Papers that haven't been cited by α until 2003 are then sorted according to their score in a descending order and those at the top represent *personalized recommendation* for author α.

Resulting recommendations are evaluated using the probe set which allows us to label as "relevant" those papers that were eventually cited by a given author after 2003. To curb the level of noise in the results, we discard authors with less than 10 relevant papers to be guessed; then we are left with the final set of 99 test authors for whom we have on average 116 relevant items to guess out of almost 340 000 papers published until 2003. To assess the recommendations, we use metrics often used in the field of recommender systems Adomavicius and Tuzhilin (2005): (i) precision P_{100} (the fraction of the top 100 places of the recommendation list occupied by the relevant papers), (ii) recall R_{100} (the fraction of the relevant papers appearing at the top 100 places of the recommendation list), (iii) the average ranking of the relevant papers q_R (expressed as a fraction of all potentially relevant papers), and (iv) the fraction of the relevant papers with non-zero score f_R. A good recommendation list should have relevant papers at the top, i.e. high P_{100} and R_{100} and low q_R, and it should assign non-zero scores to most relevant papers, e.g. high f_R (all these quantities lie in the range $[0, 1]$).

We compare our method with some other well known similarity measures. Following the findings of Lü and Zhou (2010), we chose two highly performing local similarity measures and one highly performing global similarity measure. The global measure that we use as a benchmark is the Katz similarity, defined as

$$S^{KA}_{xy} = \sum_{l=1}^{\infty} \beta^l (\mathsf{A}^l)_{xy} \tag{3.6}$$

	P_{100}	R_{100}	q_R	f_R
S^*	0.098	0.115	0.069	0.973
S^{KA}	0.081	0.099	0.048	0.995
S^{CN}	0.104	0.124	0.304	0.396
S^{RA}	0.111	0.132	0.304	0.396

Table 3.4: Recommendation performance obtained with different similarity metrics: precision and recall for the top 100 recommendations, average ranking of the relevant items, and the fraction of ranked relevant items.

where A is the standard adjacency matrix of the network and $\beta \in (0, 1)$ is a parameter to be optimized. This measure computes the similarity between two nodes by counting paths of any length connecting these nodes and exponentially discounting paths of length l by the factor β^l. Since we don't observe any remarkable improvement in changing the value of β, we set it to 0.75 (see Fig. 3.7). We also discard paths of length greater than five which keeps the required computational time at a manageable level (including paths of grater length doesn't change the relative values of similarity).

Our first local similarity is the simple 'Common Neighbors' similarity defined as

$$S_{xy}^{CN} = \Gamma(x) \cap \Gamma(y) \tag{3.7}$$

where $\Gamma(x)$ is the set of first-order neighbors for node x. The second local similarity is the 'Resource Allocation Index' defined as

$$S_{xy}^{RA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k_z}. \tag{3.8}$$

All three similarities are defined on an undirected network and we therefore consider as neighbors both citing both cited papers. By using the undirected version of the network it is possible to evaluate the similarity between nodes sharing a common progeny in addition to the ones sharing a common set of ancestors. Since S^* can only evaluate the latter type of information, S^{RA}, S^{CN} and especially S^{KA} have access to the data which S^* does not evaluate. We will discuss the impact of this advantage later.

Similarities described above can be substituted for $S^*(x, y)$ in Eq. (3.5), leading to recommendations which can be in turn compared with those obtained with S^*. The test results are summarized in Tab. 3.4. They show good performance of local metrics S^{KA} and S^{CN} with respect to precision and recall. This is because these metrics rank only a small set of papers (local neighborhoods) where there is high probability of finding relevant papers. The drawback is that only a minor part of relevant papers can be found ($f_R \approx 0.4$) and poor overall performance results ($q_R \approx 0.3$). Global metrics S^* and S^{KA} are able to rank almost all relevant objects and achieve much lower average ranking, but they pay for this enhanced 'variety' with worse performance on top places of their recommendation lists. Our new similarity metric significantly outperforms the other global metric and, from the point of view of recommendation, provides a good compromise in terms of its performance between global and local metrics. Note that this is despite the fact that S^{KA}, S^{CN}, and S^{RA} are computed on an undirected form of the data which gives them more information: they assign similarity also to nodes with overlapping progeny, not only to those with overlapping ancestors as S^* does. The advantage hence given to local similarities can be evaluated by preventing them from accessing this information. The performance of both is then decreased by 0.008/0.010 (CN) and 0.007/0.007 (RA) for precision/recall, respectively. On equal ground, S^* would match S^{CN} in terms of precision and recall. We may conclude that S^* is a reliable similarity metric which is able to compete with other known metrics.

To conclude our tests, we investigated how recommendation results obtained with our new similarity depend on the depth into which the network is probed (see Fig. 3.5, a similar dependency for the Katz similarity (see Fig. 3.6) and the dependency on the damping parameter β for the Katz similarity (see Fig. 3.7). These figures show that the maximal path lengths of 2–3 (for S^*) and 3–5 (for S^{KA}) provide a good compromise between individual recommendation performance metrics.

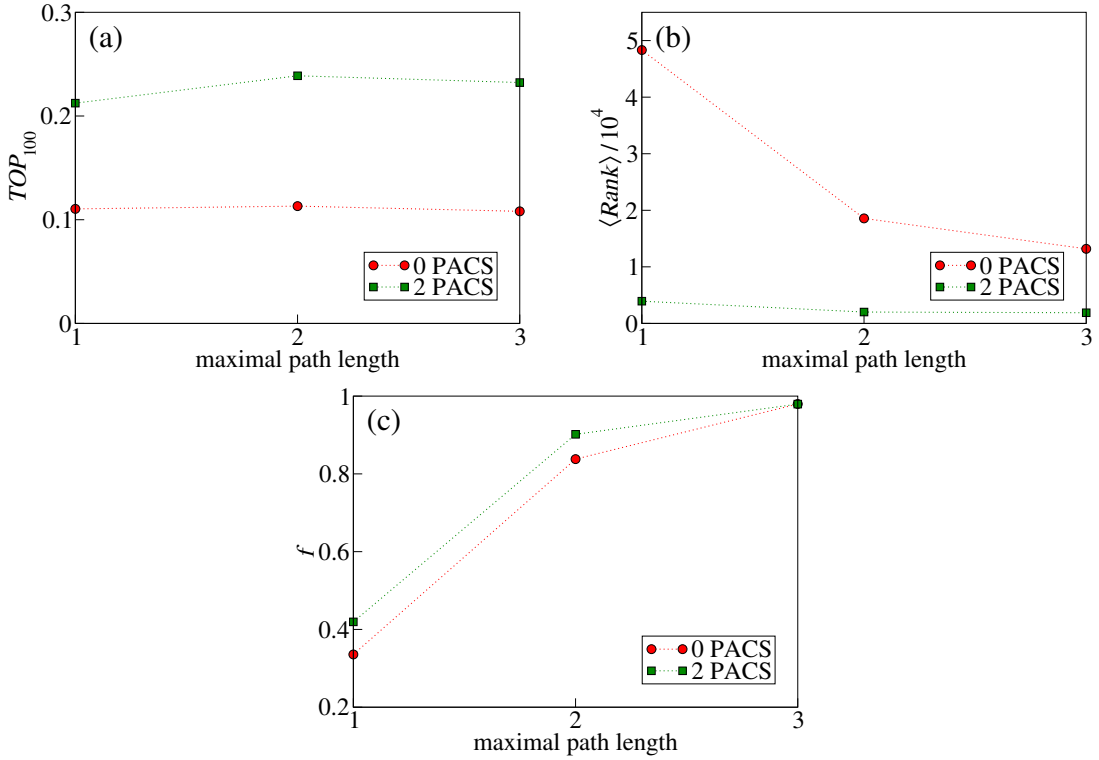

Figure 3.5: Recommendation performance as a function of the maximal path length for S^*.

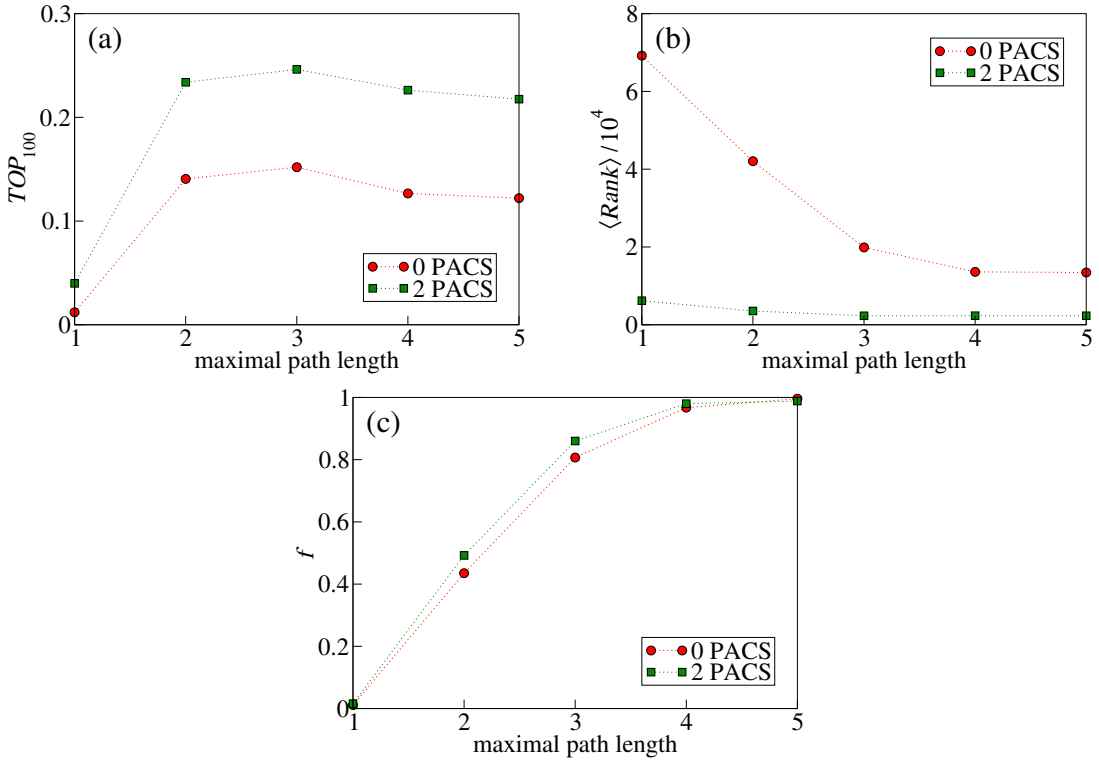

Figure 3.6: Recommendation performance as a function of the maximal path length for S^{KA}.

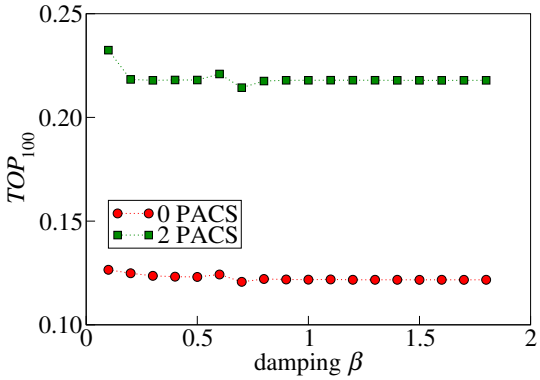

Figure 3.7: Recommendation performance as a function of the damping parameter β for S^{KA}.

3.7 Summary

The main contribution of this chapter relies in introducing the concept of influence (or, equivalently, passing probability in a random walk process) which is particularly motivated and suitable for directed acyclic networks (this class of networks is typically represented by citation data). We illustrated the use of this concept on three distinct examples: detection of seminal papers (using the total influence), tracing the composition of scientific papers (using the influence itself), and on introducing a new similarity measure (which we employed in a recommendation algorithm and evaluated performance of the resulting recommendations).

Chapter 4

Detection of leaders in social networks

Social networks such as *Twitter.com* and *Delicious.com* allow millions of users to interact. Clearly, some of these users have much greater influence than the others and contribute to the system's behavior. Identifying these influential users is not easy, yet knowing them can be help us to better understand the underlying social network. While we focus on Delicious.com in this chapter, our ideas can be applied to any other social directed network where a user may decide to receive information approved/favored by another user. Several professional social networks relevant to scientists (Academia.edu or Linkedin.com, for example) produce this kind of data and the algorithm elaborated below could be applied on them in the future.

The primary function of Delicious.com for individuals is to allow them to collect useful bookmarks with tags. But for many users its new function of creating links to other people is more interesting: it allows users to select other users to be their web "guides", in the sense that the bookmarks of the guides are often useful and relevant and subscriptions to their bookmarks will be automatic. Of course users who select their guides can in turn be the guides of others. Here we call the guides the *leaders* of their subscribers, and the subscribers the *fans* of their guides. Out of the 7 million users of Delicious.com, about half a million users are linked in a big cluster by these leader and fan relations. We call this big cluster the *leadership network*. This seemingly

minor group in fact contains the most active users of the service.

Though leadership networks are highly informative for identification of leaders, to utilize them well is still challenging (Jing and Baluja, 2008; Radicchi et al., 2009). First of all, the leadership structure is complex and simply following a particular upstream by indefinitely climbing up the ladder of leaders is not illuminating. In addition, considering just a leader alone provides no absolute measure of influence because as we will see, removing the leaders of an influential user may have a negative effect on his/her social influence. Similarly, merely counting the number of fans is not a good way to quantify a leader's significance. A sophisticated model however could reveal the intrinsic structure and identify the worthy leaders.

To better utilize the leadership network, we shall devise a method akin to *PageRank* (Page et al., 1999; Brin and Page, 1998), but with some crucial differences as relations among individual users are different from those among web pages. This new ranking algorithm—we call it *FreeRank* (or, one could also say, *LeaderRank*)—will effectively identify influential users in social networks. We will show that FreeRank is more effective than PageRank in identifying users who lead to quick and wide spreading of fresh items. Moreover, FreeRank outperforms PageRank in terms of robustness against incomplete information, manipulation and gaming activities. All these advantages are of particular importance for social networks. Unlike PageRank, FreeRank is a parameter-free algorithm and is applicable to any type of graph.

In addition to ranking, our results may shed light on the future design of community rules and advance the development of online social networks. A robust ranking algorithm also discourages people from gaming the reputation system (Masum and Zhang, 2004). All this suggests that FreeRank may serve as a prototype ranking algorithm which is widely applicable to social networks, and other general ranking tasks. Interested readers may visit the web page `http://www.rank.sesamr.com`, on which FreeRank is implemented to rank users of Delicious.com.

4.1 Algorithm

Users of many online applications are able to select other users to be their information sources. We represent these user-user relations by a network with directed links pointing from fans to their selected leaders. Popular leaders have a large number of incoming links. We use this convention as it matches the direction of random walk in our algorithm, though the direction of information flows is *opposite*, i.e. from leaders to fans. Our aim now is to rank all the users based on this network topology.

Specifically, we consider networks of N nodes and M directed links. Nodes correspond to users and links are established according to the relations between leaders and their fans. We introduce a *ground node* which connects to every user through bidirectional links (see Fig. 4.1 for an illustration). The network thus becomes strongly connected and consists of $N+1$ nodes and $M+2N$ links. To start the ranking process, we assign to each node, except for the ground node, one unit of resource which is then evenly distributed to the node's neighbors through the directed links. The process continues until a steady state is reached. This is mathematically equivalent to random walk on directed network which can be described by a stochastic matrix P with elements $p_{ij} = a_{ij}/k_i^{\text{out}}$ representing the probability that a random walker at i goes to j in the next step; here k_i^{out} denotes the out-degree of node i. $a_{ij} = 1$ if node i points to j and $a_{ij} = 0$ otherwise. Denote by $s_i(t)$ the score of node i at time t, we have

$$s_i(t+1) = \sum_{j=1}^{N+1} \frac{a_{ji}}{k_j^{\text{out}}} s_j(t). \tag{4.1}$$

The initial scores are $s_i(0) = 1$ for all nodes except for the ground node for which $s_g(0) = 0$.

The presence of the ground node makes P irreducible as the network is strongly connected. The ground node also ensures the co-existence of loops of size 2 and 3 from any node, which implies that P^6 is positive and P is primitive. By the Perron-Frobenius theorem, P has a maximum eigenvalue 1 with a unique eigenvector. The

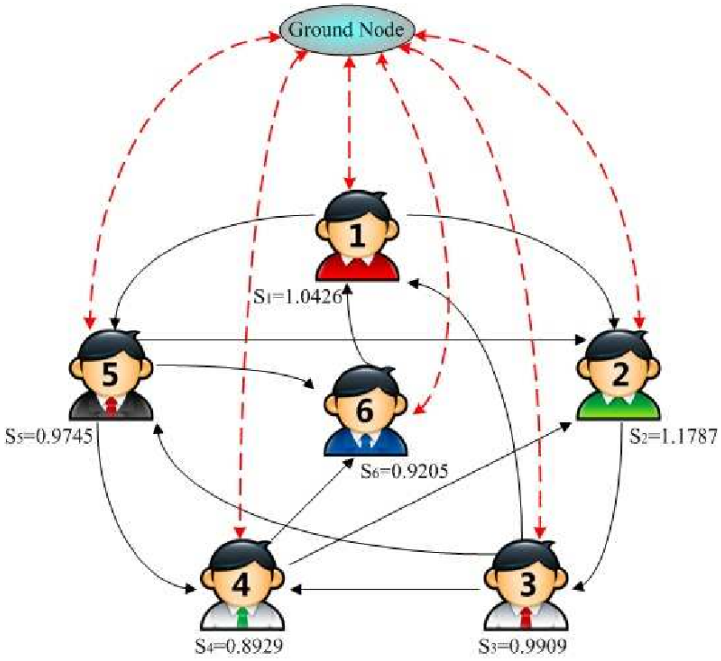

Figure 4.1: An illustration of the ground node and the FreeRank algorithm. The social network consists of six users and 12 directed links. The final ranking scores are shown next to the users.

score $s_i(t)$ thus converges to a unique steady state denoted as s_i^{EQ} for all i. At the steady state, we evenly distribute the score of the ground node to all other nodes to conserve scores on the nodes of interest. Thus we define the final score of a user to be the leadership score S, namely

$$S_i = s_i^{EQ} + \frac{s_g^{EQ}}{N}, \tag{4.2}$$

where s_g is the score of the ground node. Based on the above properties, there are several advantages of applying FreeRank: (i) it is parameter free, (ii) it is applicable to any directed graph, (iii) it convergences to a unique ranking, and (iv) it is independent of the initial conditions. To illustrate the ranking process, we provide a simple ranking example in Fig. 4.1. After 12 steps of iterations, scores converge. The final scores of the six users are $S_1 = 1.04$, $S_2 = 1.18$, $S_3 = 0.99$, $S_4 = 0.89$, $S_5 = 0.97$ and $S_6 = 0.92$, respectively. User 2 is hence ranked at the top of the FreeRank ranking.

Now we briefly recall the PageRank algorithm (Brin and Page, 1998; Page et al.,

1999) which we use as a benchmark for our ranking results. PageRank forms a basis of the Google search engine and corresponds to random walk on a hyperlink network. A parameter c (a so-called return probability) is introduced as the probability for a web surfer to jump to a random website; $1-c$ is consequently the probability for the web surfer to continue browsing through hyperlinks. In this case, $s_i(t)$ of web page i at time t is given by

$$s_i(t+1) = c + (1-c)\sum_{j=1}^{N}\left[\frac{a_{ji}}{k_j^{\text{out}}}(1-\delta_{k_j^{\text{out}},0}) + \frac{1}{N}\delta_{k_j^{\text{out}},0}\right] s_j(t). \tag{4.3}$$

where $\delta_{a,b} = 1$ when $a = b$ and 0 otherwise. The first and second term correspond to the contribution from random surfers and to surfers arriving through hyperlinks, respectively. The parameter c is essential to PageRank for when $c = 0$, its convergence is only guaranteed on strongly connected networks (i.e., every node must be reachable from any other node which implies that, for example, every node must have out-degree at least one). Not only is the algorithm not parameter free (which results in the need of extensive tests to establish an appropriate value of the parameter), the return probability c is identical for all nodes irrespective of their significance which is a strong assumption by itself.

4.2 Results

We apply the FreeRank algorithm to the leadership network of Delicious.com to rank users according to their importance. The data was collected in May 2008 and consists of 582 377 users and 1 686 131 directed links.[1] Most users (571 686) belong to the giant component which we consider in our further investigation. The number of directed links in the largest component is 1 675 008, of which 338 756 (169 378 pairs) are reciprocal links. Below we first show the difference among the rankings obtained from FreeR-

[1]This data was obtained by automated crawling of the web page. To get more detailed information or to obtain the data, contact Zi-Ke Zhang (zhangzike@gmail.com).

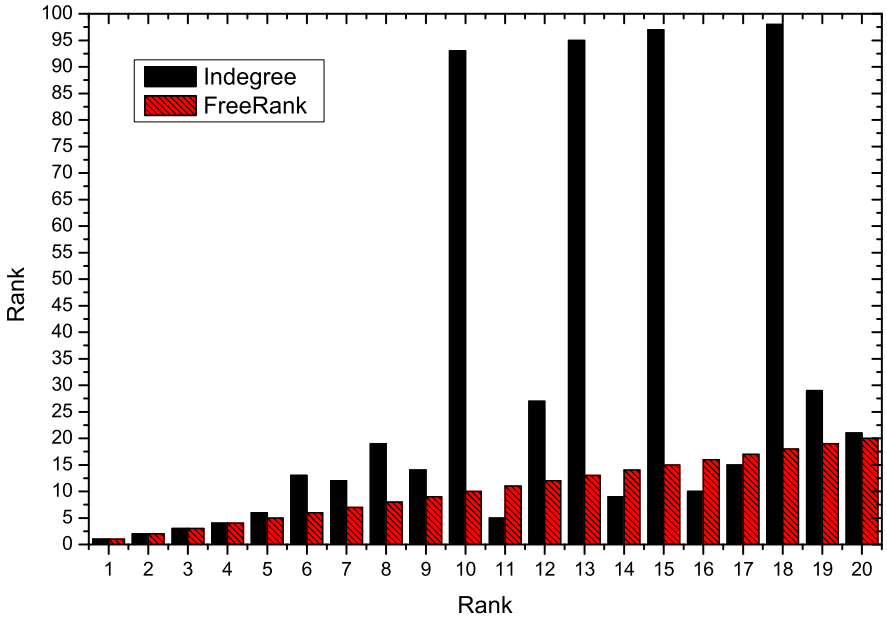

Figure 4.2: Ranking by the number of fans (horizontal) as compared to ranking by FreeRank (vertical).

ank, PageRank and by the number of fans. Tab. 4.1 shows the top 20 users ranked by the three approaches. We plot also the ranking by the number of fans as compared to the ranking by FreeRank in Fig. 4.2 which shows that there are users who have a rather small number of fans but still rank at the top. This means that our new ranking algorithm is not purely popularity-driven.

4.2.1 Comparison with ranking by the number of fans

Ranking algorithms based on analysis of the whole leadership network outperform ranking by simply the number of fans. We again compare the ranks with intrinsic qualities of the users which are independent of the ranking algorithm. One quantity which well characterizes a user's influence is how many times this user's bookmarks were saved by other users. Influential users should be able to better promote/propagate their collected bookmarks. We denote the number of bookmarks collected by user i to be B_i and the number of times these bookmarks are saved by the others to be U_i; we

	Ranking		
User ID	FreeRank	PageRank	Number of fans
adobe	1	1	1
twit	2	2	2
wfryer	3	6	3
willrich	4	7	4
joshua	5	8	6
cshirky	6	12	13
hrheingold	7	15	12
ewan.mcintosh	8	14	19
dwarlick	9	19	14
twitarmy	10	3	
merlinmann	11	16	5
blackbeltjones	12		
jdehaan	13	9	
regine	14		9
lseymour	15	10	
jonhicks	16	17	10
zephoria	17		15
isola	18	11	
djakes	19		
secondlife	20	13	
thetechguy		4	
cffcoach		5	
samoore		18	
kevinrose		20	11
steverubel			7
jgwalls			8
ambermac			16
jgates513			17
ramitsethi			18
cory_arcangel			20

Table 4.1: Top 20 users ranked by the three approaches.

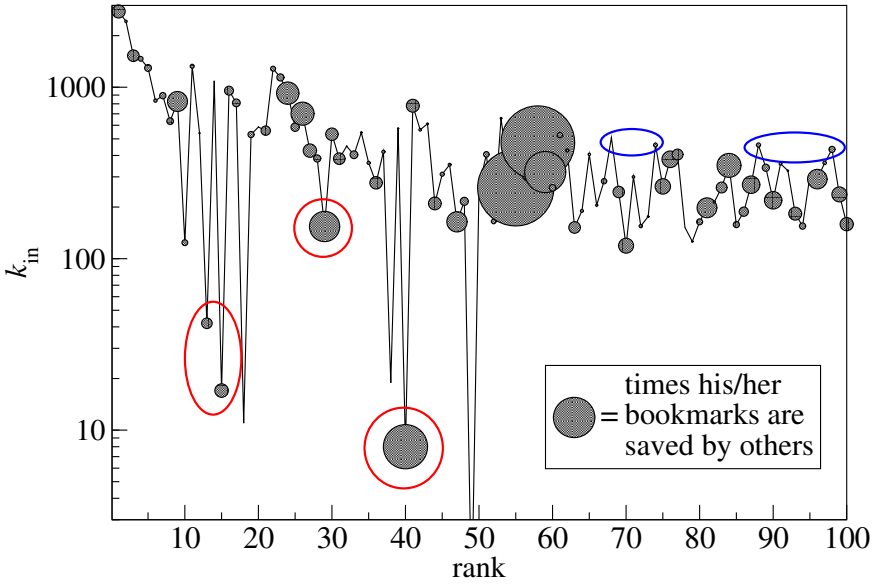

Figure 4.3: The number of fans of a user in descending order of the user rank by FreeRank. Sizes of the circles correspond to how many times the bookmarks collected by a given user are saved by other users.

shall evaluate the users by their fraction U_i/B_i: the higher, the better.

We show in Fig. 4.3 the number of fans of a user in descending order of his/her rank in FreeRank. The size of the circle is proportional to the value of U_i/B_i. As we can see, there are users who are ranked high by FreeRank but have only a small number of fans k_{in}. Their ranks would greatly deteriorate if the number of fans is used as the ranking criteria. However, the users who are highlighted by the red circles have relatively large U_i/B_i which shows that they are indeed high quality users. These users are identified by FreeRank which is based on the topology of the leadership network, but not by the number of fans. By contrast, there are users who have low FreeRank despite a large number of fans. The users who are highlighted by the blue circles corresponds to low quality users in terms of U_i/B_i but with a large number of fans. They are correctly ranked lower by FreeRank but not by the number of fans.

4.2.2 Comparison with PageRank

The better ranking by FreeRank as compared to that by the number of fans shows that network topology is crucial in ranking. Here we compare FreeRank with PageRank, which also utilizes network topology in ranking. In the following we compare the two algorithms in terms of effectiveness in identifying influential users and robustness with respect to noisy data.

How opinion spreads and forms in a community is an interesting question (Galam, 2002). To effectively spread opinions, we have to identify the influential users and create initial social inertia. For instance, companies may choose to start their adverts on influential leaders, who are capable of initiating extensive spreading through the Internet or SMS networks. Thus a smart algorithm which ranks influential users accurately is of a great commercial value. On the other hand, an effective ranking algorithm may serve its role to identify influential users for immunization and stop an epidemic outbreak (Ebel et al., 2002). Here we show that FreeRank is more capable than PageRank to identify influential users who are able to initiate *quick* and *wide* spreading.

Specifically, we employ a variant of the SIR model to examine the spreading influence of the top-ranked users (Yang et al., 2007). At each step, from every infected individual, one randomly selected fan gets infected with probability λ . Infected individuals recover with probability $1/\langle k_{\text{in}}\rangle$ at each step, where $\langle k_{\text{in}}\rangle$ is the average in-degree of the users.[2] To compare the ranking effectiveness, we set the initial infected individuals to be the users who appear in top-20 by either FreeRank or PageRank but not both (in this way we better distinguish the spreading performance of users identified by each respective algorithm).[3] Then we compare the cumulative number of infected users (which includes infected and recovered users), denoted by N_{I}, as a function of time. This experiment resembles an opinion spreading initiated from the top users and observe how the opinion propagates. Figure 4.4 shows that infecting

[2]The exact value of this parameter is not important as it only sets the time scale of the simulation.

[3]As we see from Tab. 4.1 in the top-20 case, the initial infected users by FreeRank are blackbeltjones, regina, zephoria and djakes, while that by PageRank are thetechguy, cffcoach, samoore and kevinrose.

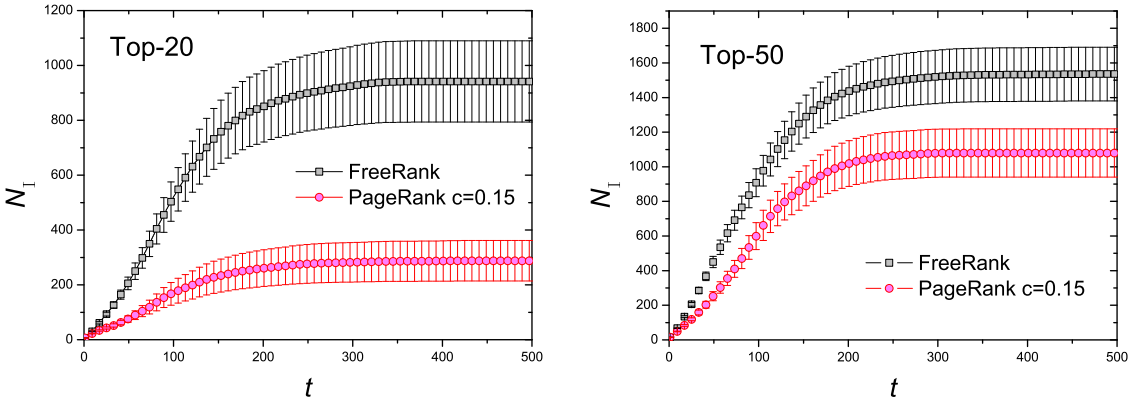

Figure 4.4: The cumulative number of infected users (including recovered user), N_I, as a function of time, with initial infected to be the users who appear in top-20 (a) and top-50 (b) when ranked by FreeRank or PageRank. Infection probability is $\lambda = 0.5$ and return probability is set to $c = 0.15$ in PageRank.

the top users from FreeRank results in faster growth and a higher saturated number of infected user, indicating a *quick* and *wide* spreading.

4.2.3 Robustness against noisy data

Robustness of ranking against spurious and missing links, i.e. false positive and false negative connections, is crucial when network structure is subject to noisy observations (Guimerá and Sales-Pardo, 2009). Social network data may be unreliable, especially when users are required to explicitly indicate relationship with others (Marsden, 1990). The same happens for networks other than social networks but with a rather different cause. For example, protein connections obtained from biological experiments often include numerous false positives and negatives (Legrain et al., 2001). Apart from experimental accuracy, it is also costly and technically demanding to explore large social networks. Efforts have thus been made to predict the missing connections (Lü and Zhou, 2010).

To examine the ranking robustness of FreeRank and PageRank, we measure the change in scores and rankings when links are added or removed randomly (these modified links correspond to the above-mentioned false positives and negatives). The

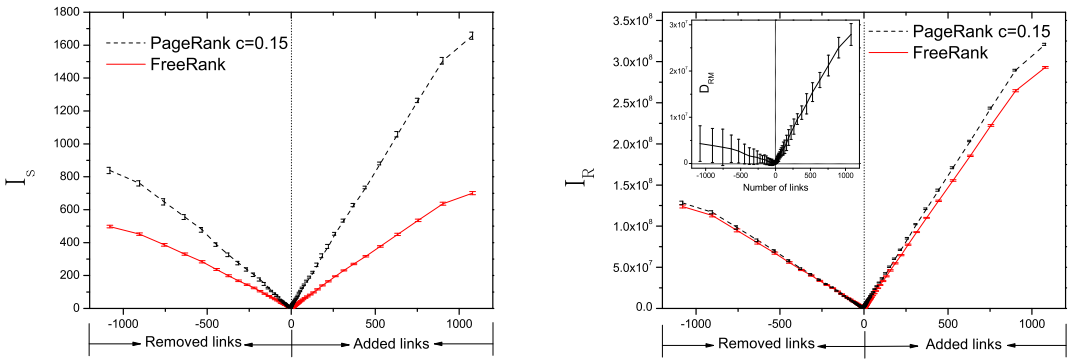

Figure 4.5: The impact on (a) scores and (b) ranking as a function of the number of links added and removed. Inset in (b): the difference in the ranking change between FreeRank and PageRank.

scores obtained from the modified graph are compared to those from the original graph, by measuring the impact I_S on score, as given by

$$I_S = \sum_{i=1}^{N} |S_i' - S_i|, \tag{4.4}$$

where S_i and S_i' correspond to the scores obtained respectively from the original and modified graph. We measure I_S for both FreeRank and PageRank subject to the same modifications. As shown in Fig. 4.5a, I_S increases with the number of links added or removed. Remarkably, much smaller values of I_S are obtained from FreeRank when compared to PageRank regardless of addition or removal of links.

Since a small change in scores does not need to correspond to a small change in the resulting ranking, we define a second measure to examine the impact I_R on ranking, given by

$$I_R = \sum_{i=1}^{N} |R_i' - R_i|. \tag{4.5}$$

As shown in Fig. 4.5b, a smaller difference between I_R of FreeRank and PageRank is observed when compared to I_S. Nevertheless, I_R of FreeRank is smaller, as shown by $D = I_R^{\text{Free}} - I_R^{\text{Page}} > 0$ in the inset. These results suggest that FreeRank is more

robust against topology randomness and hence a better candidate for ranking in noisy networks.

4.3 Discussion

After going through the above details, we may conclude that identifying influential users is not a simple task. In particular, a reliable ranking is hard to achieve with noisy data or smart manipulators, not to mention that the algorithms have to be generic to graph and problem types. This leads us to answer a much broader question than a merely identifying the leaders by devising a robust and generic algorithm. We suggest that FreeRank may serve as a prototype of ranking algorithm applicable to rank users in social networks. In addition to ranking the users, FreeRank outperforms PageRank in another important aspect—it identifies the users who are able to spread their opinions quickly and extensively. FreeRank is robust with respect to spurious and missing links. Our numerical tests show that its speed of converence is comparable with that of PageRank which makes it applicable even to very large datasets. These results already make FreeRank a good candidate for user rankings as well as other ranking tasks.

Chapter 5

Conclusions

This deliverable presents our progress in investigating the perception of quality in the scientific community and in devising and evaluating algorithms for detection of content and users of high quality. We expect that many of these theoretical achievements will be in the future implemented in respective components of QMedia and QScience.

Bibliography

G. Adomavicius and A. Tuzhilin, Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions, *IEEE Transactions on Knowledge and Data Engineering* **17**, 734–749 (2005).

S. Banerjee and T. Pedersen, The Design, Implementation, and Use of the Ngram Statistics Package, *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics* (2003).

C. Bartneck and J. Hu, Scientometric analysis of the CHI proceedings, In *Conference on Human Factors in Computing Systems (CHI2009)*, 699–708 (2009).

C. L. Borgman and J. Furner, Scholarly Communication and Bibliometrics. In B. Cronin (Ed.), Annual Review of Information Science and Technology, Vol 36. Medford, NJ: Information Today, 3–72 (2002).

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

K. Charmaz, Grounded Theory: objectivist and constructivist methods, In N. Denzin and Y. Lincoln (Eds.), *Handbook of Qualitative Research* (2nd Ed.), Thousand Oaks, CA: SAGE (1994).

P. Chen, H. Xie, S. Maslov, and S. Redner, Finding scientific gems with Google's PageRank algorithm, *Journal of Informetrics* **1**, 8–15 (2007).

H. Ebel, L. Mielsch, S. Bornholdt, Scale-free topology of e-mail networks, *Phys Rev E* **66**, 035103 (2002).

S. Galam, Minority opinion spreading in random geometry, *Eur Phys J B* **25**, 403–406 (2002).

N. Gilbert, A Simulation of the Structure of Academic Science, *Sociological Research Online* **2**, no. 2, http://www.socresonline.org.uk/2/2/3.html (1997).

R. Guimerá and M. Sales-Pardo, Missing and spurious interactions and the reconstruction of complex networks, *Proc Natl Acad Sci USA* **106**, 22073–22078 (2009).

Y. Jing, S. Baluja, PageRank for product image search, *Proceedings of the 17th international conference on World Wide Web*, 307–316 (2008).

J. N. Kearns and F. D. Fincham, A prototype analysis of forgiveness, *Personality and Social Psychology Bulletin* **30**(7), 838–855 (2004).

C. Lampe, N. Ellison, C. Steinfield, A Face(book) in the crowd: social searching vs. social browsing. *Proceedings of the 20th anniversary conference on computer supported cooperative work*, 167–170 (2006).

P. Legrain, J. Wojcik, J. M. Gauthier, Protein-protein interaction maps: a lead towards cellular functions, *Trends in Genetics* **17**, 346–352 (2001).

J. Lofland and L. H. Lofland, *Analysing Social Settings: A Guide to Qualitative Observation and Analysis* (3rd Ed.), London: Wadworth (1995).

L. Lü, T. Zhou, Link Prediction in Complex Networks: A Survey, arXiv:1010.0725 (2010).

P. V. Marsden, Network data and measurement, *Annual Review of Sociology* **16**, 435–463 (1990).

H. Martens and M. Martens, *Multivariate analysis of quality: an introduction*, Chichester: Wiley (2001).

H. Masum H, Y.-C. Zhang, Manifesto for the reputation society, *First Monday* **9**, 7–5 (2004).

A. Neus, Managing Information Quality in Virtual Communities of Practice. In: Pierce, E. & Katz-Haas, R. (Eds.) *Proceedings of the 6th International Conference on Information Quality at MIT*, Boston, MA: Sloan School of Management (2001).

M. E. J. Newman, The structure and function of complex networks, *SIAM Review* **45**, 167–256 (2003).

M. E. J. Newman and M. Girvan, Finding and evaluating community structure in networks, *Physical Review E* **69**, 026113 (2004).

L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank Citation Ranking: Bringing Order to the Web, *Technical Report Stanford InfoLab* 1999-66 (1999).

J. Paterson, C. Lange, I. Akhtar, F. Iacobelli, P. Anderson, and A. Leonhard, Exploring the use of computational linguistics for automated formative feedback in the humanities, In *Proceedings of ICERI2010 Conference*, 5303–12 (2010).

F. Radicchi, S. Fortunato, B. Markines, and A. Vespignani, Diffusion of scientific credits and the ranking of scientists, *Physical Review E* **80**, 056103 (2009).

P. Rayson, Wmatrix: a web-based corpus processing environment, Computing Department, Lancaster University, http://ucrel.lancs.ac.uk/wmatrix/ (2009).

E. Rosch, Principles of categorization, In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization*, 27–48, Hillsdale, NJ: Erlbaum (1978).

A. Strauss and J. Corbin, *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory* (2nd Ed.), London: SAGE (1998).

M. V. Vieira, B. M. Fonseca, R. Damazio, P. B. Golgher, D. de Castro Reis, B. Ribeiro-Neto, Efficient Search Ranking in Social Networks, *Proceedings of the 16th ACM conference on information and knowledge management*, 537–572 (2007).

R. Yang, B.-H. Wang, J. Ren, W. J. Bai, Z. W. Shi, W. X. Wang, T. Zhou, Epidemic spreading on heterogenous networks with identical infectivity, *Phys Lett A* **364**, 189–193 (2007).