Model Validation and Synthesis
Project no. 231200

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

Deliverable D.3.3.1
Model Validation and Synthesis

Submission date: 2012-02-28

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

Organisation name of lead contractor for this deliverable: CNRS

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
Chih-Chun Chen	CNRS	c.chen@abmcet.net

1.2 Other contributors

Name	Organisation	E-mail
Camille Roth	CNRS	c.roth@ehess.fr
Sergi Lozano	ETH Zurich	slozano@ethz.ch
Steve Genoud	ETH Zurich	sgenoud@ethz.ch
Nigel Gilbert	UniS	n.gilbert@surrey.ac.uk
Alistair Gill	UniS	a.gill@surrey.ac.uk
Maria Xenitidou	UniS	m.xenitidou@surrey.ac.uk
Michal Ziembowicz	UWAR	ziembowicz@gmail.com

1.3 Document history

Version#	Date	Change
V0.1	31 January 2013	Original
V0.2	20 February 2013	Revised
V0.3		
V1.0		Submitted

1.4 Document data

Keywords	science, peer review, science metrics, truth
Editor address data	c.elsenbroich@surrey.ac.uk
Delivery date	31 August, 2012

1.5 Distribution list

Date	Issue	E-mail
	Consortium members	QLECTIVES@list.surrey.ac.uk
	Project officer	Roumen.BORISSOV@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 quality

The approach of the QLectives project is unique in that it brings together a highly interdisciplinary 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

This report is concerned with the evaluation of models developed in in WP2.2 and WP2.3 and the hypotheses and assumptions on which they are based, using data collected and processed in WP3.1 and WP3.2. In line with the objectives of QLectives, the focus is on quality in online media and communities (Wikipedia and Anobii book sharing), and scientific communities (American Physical Society). Since model and algorithm development has taken place in dialogue with data analyses throughout the project, with hypotheses being validated with empirical data before contributing to the formal models, we report on both the validation of formal models (or more precisely, evaluation of model behaviour in different domain contexts) and the confirmation of the assumptions underlying these. This validation activity centres around the Quality, Trust, Reputation model developed in WP2.2.

1. Chapter 1 introduces the Quality, Trust, Reputation model (QTR) and the theories and hypotheses on which the model is based. It also identifies differences in the weight placed on different factors depending on the application context. We also give an overview of the datasets used to confirm these.

2. Chapter 2 describes the validation activities with respect to Wikipedia, the online collaboratively edited encyclopedia. We evaluate the importance of article quality, author fecundity, and external authority invocation with respect to the QTR model.

3. Chapter 3 describes the validation activities with respect to Anobii, an online community platform that allows users to share ratings of the books they are reading (or have read). We confirm the significance of homophily and identify interactions between homophily and social links. These findings are then used to evaluate behaviour of the QTR model with the data.

4. Chapter 4 describes the validation activities with respect to citations in a scientific domain using journal citation data released by the American Physical Society.

The report concludes with a discussion of the main findings and their implications for improving both social and non-social recommendation systems.

Contents

Chapter 1

Introduction

The underlying premise of the model developed in WP2.2 is that quality, trust and reputation are intimately linked. Objects produced by members of a community have certain properties that are valued (or deemed undesirable) by that community. However, individuals are not able to inspect each of the objects directly to determine the extent to which they possess these properties. Instead, individuals who are identified as being more capable of producing objects possessing the properties valued acquire a reputation, whereby objects produced by these individuals are more likely to be adopted and deemed to have higher quality. At the same time, the adoption of the objects produced by an individual serve as a signal as to how capable they are of producing objects possessing to a high degree the desirable properties.

In addition, there may be other factors external to the individual-object network that serve as indicators of individuals' capabilities, such as their invocation of other authority sources, the awards they have received, or even direct encounter. This can be formalised as an additional network of confidence or 'trust' between individuals. Furthermore, even within a particular context or domain, individuals can differ in the degrees to which they value different properties so that they agree to different degrees with the community level evaluation (i.e. they can deviate from the population average). The 'trust' network can also reflect the extents to which different individuals are similar to each other, such that individuals who are more similar to each other tend to also trust each others' evaluations more (homophily, [39]).

The work conducted in WP3.3 serves to evaluate using both qualitative and quantitative methods, the extent to which these assumptions hold within different contexts, as well as the 'Quality, Trust, Reputation' (QTR) model developed in WP2.2. An important thing to bear in mind is that the importance of the trust network can differ greatly between different contexts, and in many online contexts, it can be largely absent (e.g. few users of Wikipedia would take any notice of which editors had edited an article when evaluating its reliability).

One of the key strengths of the QTR framework is that it is extremely flexible and different paremetisations can be used to better fit a given context. We should therefore think of the QTR model as a group of models, each with a different parameter configuration and/or different weight assignments to interaction factors. In turn, the assignment of these weights and parametisations should be driven by social theory (about how individuals interact with objects and each other). The QTR model therefore allows us to confirm individual level social theory at the community level. If a particular parameter configuration is not predictive with respect to the underlying social theory assumptions, it may be that these assumptions are incorrect or insufficient to account for the resulting data. In such cases, we try to find alternative or additional theories to re-configure the model to find the most plausible set of theories to account for the data.

The QTR framework can also be seen as a generalisation of the well-known HITS algorithm (and its successor PageRank [**?**] and bipartite generalisation [**?**]), in which a node's reputation (usually called 'authority') is a function of the (scaled) number of 'hubs' (node with many incoming links) it points to. In terms of QTR, the HITS configuration is simply one in which all the parameters are set to zero.

1.1 The QTR model

The QTR model has two key features:

- Different weights can be assigned to different actions which contribute more/less to their status in the community (their reputation);
- Quality is socially constructed in that it is a function of the reputation of the users who produce the objects.

$$Q_\alpha = \frac{1}{k_\alpha^{\theta_Q}} \sum_{i=1}^{N} w_{i\alpha}[R_i - \rho_R \overline{R}] \tag{1.1}$$

$$R_i = \frac{1}{k_i^{\theta_R}} \sum_{\alpha=1}^{M} w_{i\alpha}[Q_\alpha - \rho_Q \overline{Q}] + \frac{1}{f_i^{\theta_T}} \sum_{j=1}^{M} [R_j - \rho_R \overline{R}][T_{ji} - \rho_T \overline{T}] \tag{1.2}$$

where:

- $k_i = \sum_\alpha a_{i\alpha}$ (user degree i.e. the number of objects user i has interacted with);
- $k_i^W = \sum_\alpha w_{i\alpha}$ (user weight i.e. the summed weight of all his interactions);
- $k_\alpha = \sum_i a_{i\alpha}$ (object degree i.e. the number of users who have interacted with the object);

- $k_\alpha^W = \sum_i w_{i\alpha}$ (object weight i.e. the summed weight of interactions with it);
- $f_j = \sum_i b_{ij}$ (the number of users who trust user j);
- $f_j^W = \sum_i T_{ij\alpha}$ (total amount of trust in user j);
- $\overline{Q} = \sum_\alpha \frac{Q_\alpha}{M}$;
- $\overline{R} = \sum_i \frac{R_i}{N}$;
- $\overline{T} = \sum_{ij} \frac{T_{ij}}{N(N-1)}$;
- θ_Q, θ_R, θ_T, ρ_Q, ρ_R, ρ_T are control parameters in the range $[0, 1]$.

Initially, for M objects and N users:

- $Q_\alpha^0 = \frac{1}{\sqrt{M}}$
- $R_i^0 = \frac{1}{\sqrt{N}}$

Then the mutually interconnected equations (Equation 1.1 and Equation 1.2) are updated iteratively:

$$Q'_\alpha \leftarrow \frac{1}{k_\alpha^{\theta_Q}} \sum_{i=1}^{N} w_{i\alpha}[R_i - \rho_R \overline{R}] \tag{1.3}$$

$$R'_i \leftarrow \frac{1}{k_i^{\theta_R}} \sum_{\alpha=1}^{M} w_{i\alpha}[Q_\alpha - \rho_Q \overline{Q}] + \frac{1}{f_i^{\theta_T}} \sum_{j=1}^{M} [R_j - \rho_R \overline{R}][T_{ji} - \rho_T \overline{T}] \tag{1.4}$$

To prevent divergence, Q'_α and R'_i are normalised such that:

- $\sum_{\alpha=1}^{M} (Q'_\alpha)^2 = 1$; and
- $\sum_{i=1}^{N} (R'_i)^2 = 1$.

Iteration terminates when the algorithm converges to a steady state:

$$\sum_{\alpha=1}^{M} |Q'_\alpha - Q_\alpha| + \sum_{i=1}^{N} |R - R_i| < \delta$$

1.2 Datasets

In order to explore the behaviour QTR in different contexts, we use data from both online platforms and scientific communities. Due to the low volume of active users in the QScience platform for the most part of the project's duration and the changes in functionality that occurred in QMedia, we used external datasets from more mature platforms in our validation activity.

1.2.1 Online media

In the context of online media, we analysed data from two very different platforms with respect to the QTR framework:

1. Wikipedia: We used extracts of English language Wikipedia data dumps (available at: `http://dumps.wikimedia.org/enwiki/`) containing the entire histories of articles, with all the details and content of each edit. We used different types of editing activity to assign weights to the interactions between editor and article, including the length of contribution and whether or not it contained a reference (invocation of external authority). Using the QTR model, we then evaluated the importance of each of these factors and their combined effect, by using the model to identify featured versus non-featured articles.

2. Anobii: Anobii is an online community platform which allows users to share the books and their evaluations of these books (which can be quantitative ratings or qualitative comments). It also allows users to make directed social connections between each other so that they can share and/or follow other users' book choices and evaluations. The data were collected by the authors of [4], who performed an analysis of the interaction between social network evolution and profile similarity. The authors took six snapshots of the user-book and user-user networks 15 days apart. We used only the first and fourth snapshots (hence 45 days apart) since per book, change in ratings was slow; we excluded the fifth and sixth snapshots where platform administrators had changed the rating scale from a "1-4" range to a "1-5" range. These data were used to evaluate the effects of social connections (or 'trust' in the QTR framework) and homophilly (proxied by user adoption and/or rating similarity) with respect to QTR using the book adoptions between the two snapshots.

1.2.2 Scientfic communities

Science prides itself on standards of objectivity, whereby work is judged according to individual merit. If this is the case, the association between quality and reputation should be fairly weak.

1. American Physical Society (APS): We used authorship and citation data obtained from the American Physical Society (available at: `https://publish.aps.org/datasets`) and conducted a thorough exploration of the effects of different factors by assigning interaction weights in several different ways. Specifically, we evaluated the effect of assigning different weights to first and last authors and the effect of including authors' citations in

Context	Dataset	user-object interaction weight assignment	trust network initialisation	proxy for quality used in validation
Online collaboration	Wikipedia article edit histories	Length of edit, whether or not edit contains a reference	No trust network	Presence on featured article list
Online evaluation/rating	Anobii book ratings	Score 3, Score 4	Friendship and/or neighbour links between users	Increase in mean score in later time snapshot
Scientific community	APS journal authorship and citation data	First authorship, last authorship	Co-authorship, citation rate between authors	Article number of citations

Table 1.1: Overview of datasets used in validation activities.

their interaction weights (effectively modelling the idea that authors who are well-cited can contribute more to the article's perceived reliability). Citation rates of the articles were then used as a proxy for evaluating the model's ability to obtain the correct Q values.

2. ISI: We used data obtained from ISI to explore the temporal factors that might be related to quality, trust and reputation. In Science, it is often the case that articles and/or authors are more relevant or revered at different times. 'Rebirth' periods can occur, during which work conducted in the past can again become very pertinent. We try to identify such patterns and also to determine the implications for the QTR model, resulting in an extended formulation that also allows quality, trust and reputation to be related dynamically in a time-dependent fashion.

Chapter 2

Reputation and Quality in non-social online resources: Wikipedia

Wikipedia is a collaboratively edited online resource that serves as a trusted knowledge base. An important characteristic of Wikipedia is that the majority of its users do *not* contribute to it. This contrasts with the scientific context described in Chapter 4, where consumers of the resource are also its producers. Furthermore, most users of Wikipedia are unaware of who has contributed to an article, making it difficult to justify any role of editor reputation in determining the perceived quality of an article. Another point to note is that collaboration is largely asocial and mediated by the resource itself rather than being direct. Although editors may interact with each other through discussion threads while editing an article, the basis of their interactions is for the most part with respect to the article rather than being social. This contrasts with social media platforms such as that described in Chapter 3, where interactions can take place both directly and through shared objects.

Given these observations, we would expect that if reputation (as formulated in the QTR framework) is determined purely by editor contribution to an article, editor reputation and quality should be independent. The question then arises as to how an article's quality or reliability is assessed. One possibility is that an article can invoke authority from external sources. As a corollary of this, editors contributing a reference can be seen as making a greater contribution to the article's quality.

In this chapter we describe two validation activities. The work described in Section 2.1 explores empirically the activity of referencing with respect to article maturity and other editing activity to evaluate the hypothesis that referencing might also be an indicator of article maturity and/or relevance (as indicated by increased editing activity). The second studies the behaviour of the QTR model when different interaction weights are assigned, including the condition

where a revision involving references is given extra weight.

2.1 Referencing activity in Wikipedia

The extent to which a Wikipedia article refers to external sources to substantiate its content can be seen as a measure of its externally invoked authority. We introduce a protocol for characterising the referencing process in the context of general article editing. With a sample of relatively mature articles, we show that referencing does not occur regularly through an article's lifetime but is associated with periods of more substantial editing, when the article has reached a certain level of maturity (in terms of the number of times it has been revised and its length). References also tend to be contributed by editors who have contributed more frequently and more substantially to an article, suggesting that a subset of more qualified or committed editors may exist for each article.

The reliability of Wikipedia as an information source has always been a subject of controversy, with some arguing that it is comparable to encyclopedias that centrally monitor and curate their content [23], and others casting doubt on its trustworthiness (see [19] for a more detailed discussion). At the same time, significant effort has been made to understand the processes and mechanisms that underlie the editing of Wikipedia articles, in particular the conditions under which collective editing is most productive [32], [48] and likely to lead to higher quality articles [51].

Within Wikipedia itself, different measures have been used to compare the quality, reliability, and trustworthiness of articles, editors and edits. For example, one approach is to use the proportion of the edit retained in the article and then to cast editor reputation in terms of the retention rate of his/her edits [19]. Article-centric approaches can be qualitative and based on article content, or quantitative and based on certain article features, such as length or review history [11, 28]. Editor-centric approaches define quality in terms of the composition (e.g. number, diversity) of contributors and/or their reputation [31] (where editor reputation might be defined in terms of network position e.g. [33]). There have also been studies addressing the dynamics of editing and article construction [50, 35]. These analyses have been invaluable in giving us insight into how the range of editing behaviour might give rise to articles that differ vastly in terms of quality.

While it is possible to frame quality and reliability in terms of only article content and consider articles' trustworthiness only in relation to other Wikipedia articles, this neglects the invocation of external sources by articles to substantiate content. The extent to which an article does this can be treated as a measure of its externally invoked authority (an even more intricate

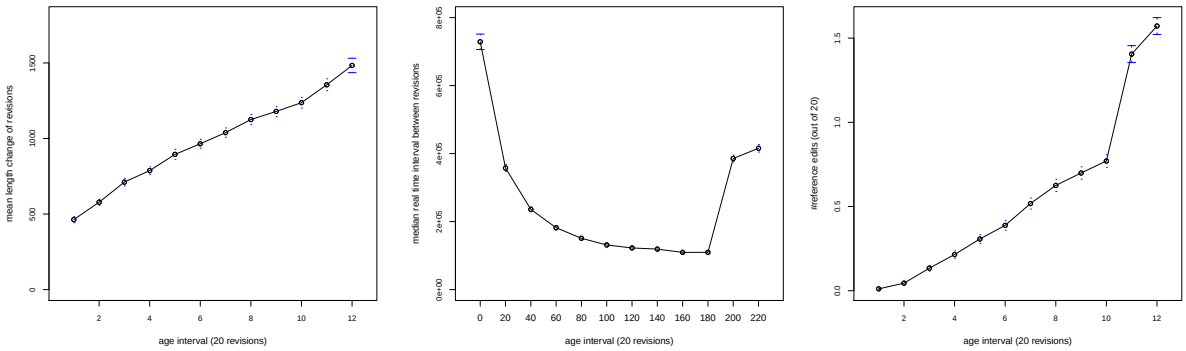


Figure 2.1: Left: Mean length change with article age; Middle: Inter-revision delay (median real time interval between edits) with article age; Right: Proportion of reference revisions with article age. Article age measured in 20-revision intervals.

measure would also take into account the reliability of the external sources referred to). The goal of our work is to better understand the dynamics of referencing with respect to an article's maturity and its editors.

In order to gain more insight into the referencing activity in Wikipedia, we analysed data from a sample of article histories. Our findings support our hypothesis that substantiation of articles by referencing external sources only occurs after articles have reached a certain level of maturity. Furthermore, referencing tends to occur during periods in which edits are more substantial. We also found that the reference density of edits is auto-correlated and that editors who contribute references are those who have contributed more frequently and more substantially to the article. In other words, reference contribution indicates editor 'maturity' as well as article maturity. These two important findings justify taking into account the invocation of external authority sources when assigning editor-article interaction weights when applying the QTR model. A more detailed description of this work can also be found in [16].

2.1.1 Dataset and methods

We extracted a sample of articles from the entire English Wikipedia as of 5th April, 2011. Because the timescales of articles can vary greatly (some may grow quickly and/or be frequently edited, while others may see very slow growth or little editing attention over the same real time period), real time would be unlikely to reflect the maturity of an article.

We therefore chose to take number of revisions rather than time stamp as the age indicator. In this initial study, we considered only a small sample of 137,104 Wikipedia articles which were randomly selected from the ≈3.6 million articles in the English Language Wikipedia and included only the most mature articles (roughly the top decile in terms of number of revisions) with 240 or more revisions with at least one reference, giving us only 5,434 articles (10,930

of the 137,104 articles had 240 or more revisions, but only 5,434 of these had one or more references at the end of the evaluated period).[1] Although we acknowledge that this is only a small fraction of the entirety of Wikipedia, the main contribution of this paper is not to provide an empirically exhaustive analysis of referencing in Wikipedia but to demonstrate a protocol for characterising the referencing process. Nevertheless, we expect the main premises of our findings to generalise.[2]

2.1.2 Referencing and article maturity

The editing dynamics of Wikipedia articles can differ enormously; some articles may have intensive periods of high activity but remain largely untouched outside of these (e.g. event-based or media-related articles) while others may grow with more regularly distributed contributions.

Considering this, we take the number of revisions as a proxy for article age, the real time between revisions as a measure of activity intensity, and the length of a revision as a measure of its substantialness. The findings reported here refer to the states of articles at the same age, 240 revisions old. We did not control for reverts since we found their frequency to be negligible in our dataset (this may have been due to the fact that our sample did not include many controversial articles).

Article growth and referencing

The growth rate of articles appears to change through time. Initially, revisions are fairly insubstantial, with small changes in article length (see Figure 2.1: Left). They are also sparse in time, with large intervals between each edit (see Figure 2.1: Middle). During this period, there appears to be very little referencing. After a certain number of revisions, the article goes through a period of higher activity when edits happen at greater frequency (more revisions within a given time) and referencing starts to happen with more regularity (see Figure 2.1: Right). After the highly active period, there is a phase of more lengthy edits and referencing.

Reference density and length

The reference density (number of references per unit length) of an article is a measure of the degree to which external sources are used to support and substantiate the article's content. Not taking into account the number of revisions that articles have gone through, longer articles tend

[1] References were extracted using the "¡ref¿" tag (depending on the article format, these might be represented to the user in different ways in different articles, e.g. as footnotes or in a References section at the end of the article).

[2] Since we only included articles that had reached a certain level of maturity, we did not consider the effect that referencing (or lack of referencing) might have on the survival rate of an article at different stages.

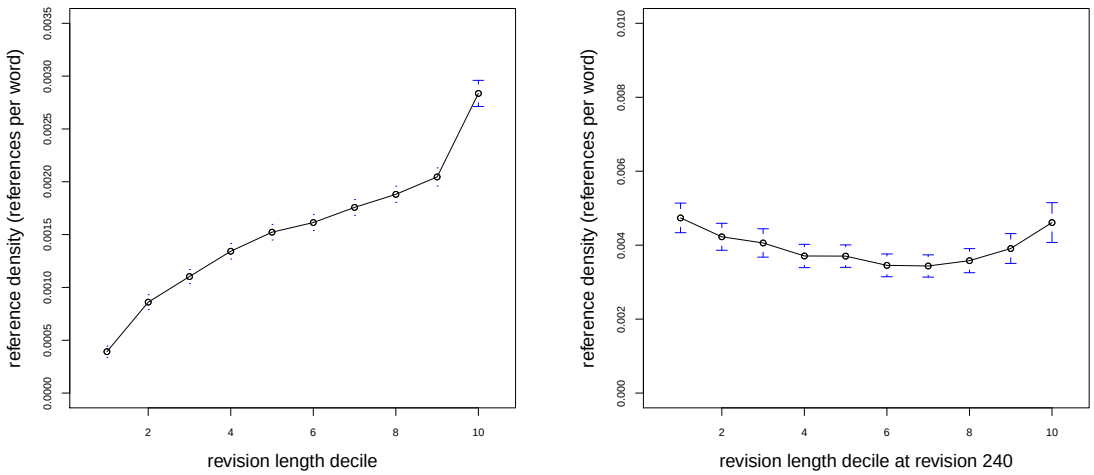


Figure 2.2: Left: Mean plot of page reference densities and article length for all revision intervals; Right: Mean plot of page reference densities and article length at 240 revisions.

to have a higher reference density (see Figure 2.2: Left). This seems to suggest a state of article maturity in which content becomes better substantiated. However, the relationship between reference density and length is not straightforward. For any given revision period, length on its own does not appear predictive of reference density: see e.g. Figure 2.2 (Right) where reference density is plotted against article length at an age of 240 revisions, without demonstrating any clear correlation. This suggests some non-trivial interaction between length, age (in terms of number of revisions) and reference density. The section that follows considers the dynamics of reference editing to try to identify some of the underlying processes.

2.1.3 Processes underlying referencing activity

Figure 2.1 (Right) suggests that referencing only starts after a critical number of revisions, and Figure 2.2 (Left) suggests that this might be due to the fact that referencing only occurs when articles reach a critical length. However, for any given article, it still remains an open question what initiates this process of referencing and what leads to subsequent referencing.

In the previous section, we already saw that longer articles tend to have higher reference densities and that periods during which more substantial edits are made tend to be better referenced. This may be due to the fact that as articles become longer and more substantial (containing more assertions or 'points'), they require more external substantiation.

We find support for two underlying processes (not mutually exclusive):

1. Substantiation of articles reinforces itself so that better referenced edits provoke further better referenced edits. In this case, there should be an auto-correlation for reference density of contributions throughout the lifetime of the article.

2. Referencing occurs when a set of committed and qualified editors are attracted to the article and start to make more substantial, referenced edits, in which case we would expect editors making reference edits to also make longer edits.

Referencing as substantiation

Figure 2.3 suggests that periods in which the article grows in more substantial chunks (with lengthier revisions) are also those in which referencing is more frequent, i.e. where the reference density of revisions (number of revisions per unit of length change) is higher.

We also found an auto-correlation for the reference density of revisions. To illustrate, the heatmap in Figure 2.4 (Left) shows the correlations between the reference density of revisions (number of revisions per unit of length change) of different revision intervals with the reference density of the final revision interval. The lighter regions indicate increasing Pearson correlation coefficients values towards cells corresponding to successive revision intervals $(t, t + 20)$. Figure 2.4 (Right) plots the reference density at t revisions against the reference density at $t + 20$

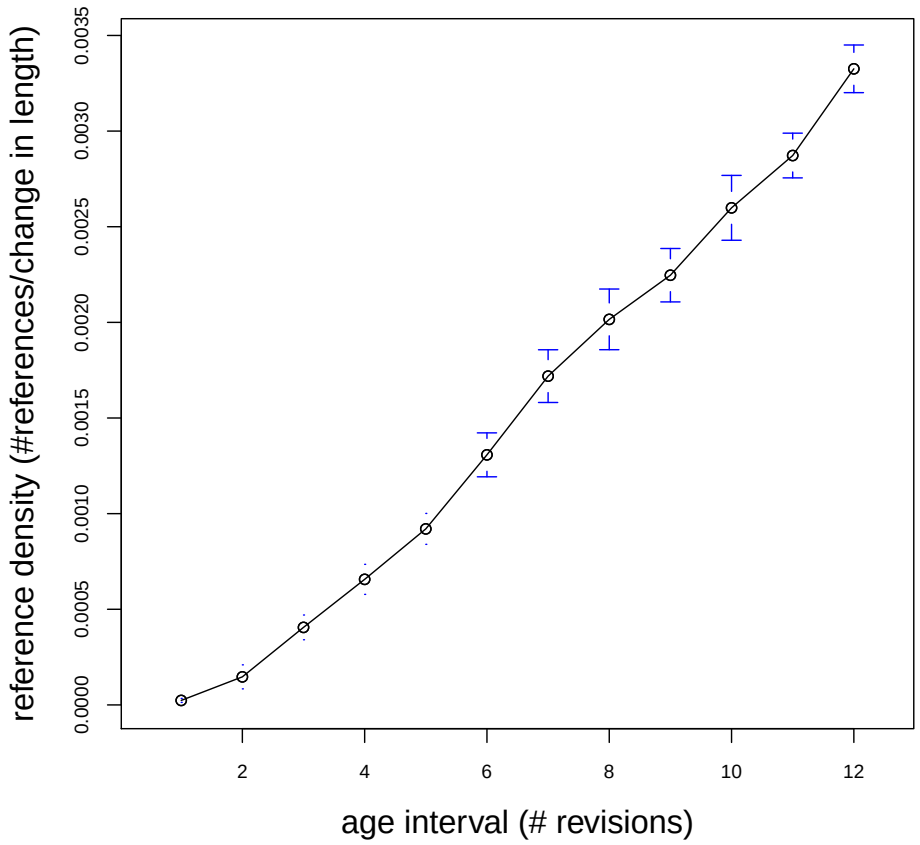


Figure 2.3: Reference density of revisions with article age (20-revision intervals)

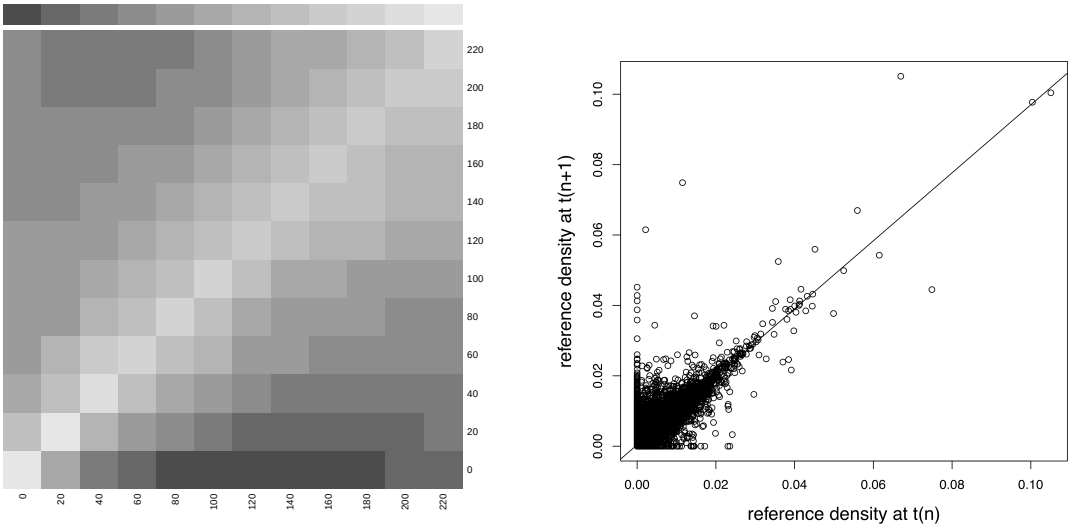


Figure 2.4: Left: Heatmap of the correlations between the reference densities of articles at different ages (20-revision intervals). Right: Scatter plot of reference densities at revision interval t and reference densities at revision interval $t + 20$

revision, with a Pearson correlation of 0.877.

Editors contributing references

One means by which reference density becomes dependent on article maturity is that once an article has reached a certain state, it begins to inspire more more conscientious editing behaviour by a set of qualified editors or attract more 'serious' editors [3]. Figure 2.5 shows that editors who add references tend to be those who edit both more substantially and more frequently and also that those who have contributed more than 2 references edit more (both in terms of contribution length and in terms of frequency) than those who only contribute one or two references. Analyses of variance confirmed these differences were significant at $p = 0.001$ between reference contributors and non-contributors, and between reference contributors contributing 2 or more references and those contributing fewer than 2 references (in terms of both length and frequency).

It should be emphasised that although those who contribute references tend to be those who contribute more often and longer revisions, this does not necessarily say anything about the quality of their contributions (e.g. [6], who found that shorter contributions tend to have higher retention). Rather, we can say that editors who bother to substantiate their contributions by invoking the authority of external sources are those whose contributions tend to be longer and more frequent with respect to that article. It is also an open question whether editors contribut-

[3] Although bots also play a role in the editing process, the proportion of edits they were responsible for in our sample was less than 1%. We therefore retained their edits in the analyses. It is also not clear that removing bot edits from the analyses would be valid since it might obscure the responses of human editors to bot edits.

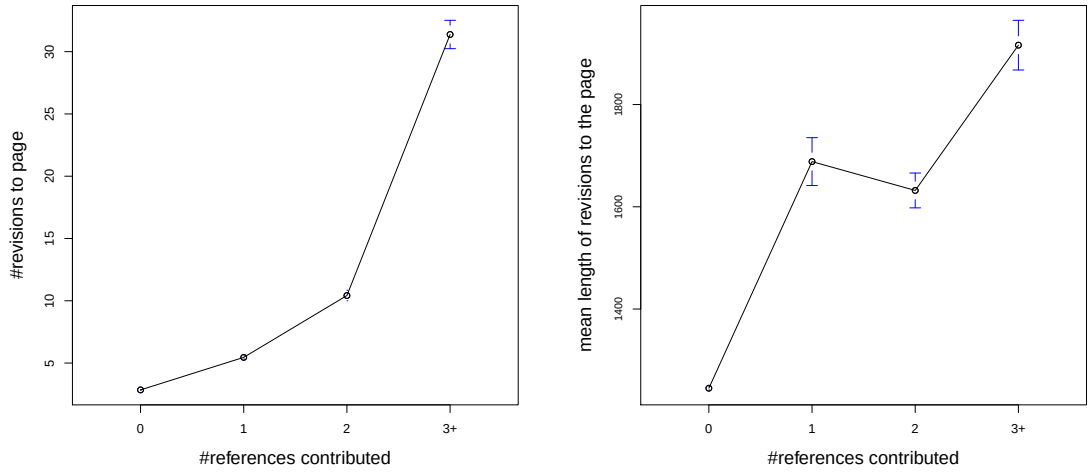


Figure 2.5: Meanplots comparing editors who have made 0, 1, 2, and > 2 reference edits in terms of (Left) median length of contributions (with respect to an article) and (Right) number of contributions to the article.

ing references tend to do so in support of their own content contributions and assertions, or whether they tend to add references in response to dissent by other editors (e.g. in response to debate to reinforce their position or to address requests for substantiation).

2.2 QTR model behaviour with the invocation of external authority sources

2.2.1 Dataset and methods

The data used for the validation activities come from Wikipedia article edit histories, as already described in Section 2.1. However, since we are evaluating performance by the correct identification of featured articles, we only consider articles from the same categories. It is also likely that there is also significant variation between categories in terms of sensitivity to different parametisations and weight assignments. This in itself sheds light on how the category 'community' is behaving with respect to quality and reputation. We illustrate this with article data from two different high level categories: Music (D1) and Mathematics (D2).

Using Wikipedia articles' editing histories, k_α is the number of edits received by an article and k_i is the number of articles within the set of articles sampled that are edited by the editor. Since we consider articles of comparable degrees of maturity and our evaluation criterion (featured article status) is not sensitive to time factors such as temporal relevance, we do not

consider time in our evaluation of the model. The following weight assignments are used for editor-article interaction to try to distinguish the effects of external authority invocation and size of contribution (as proxied by article length):

- WA1 (edit length): $w_{i\alpha} = 1$ if the length of the edit is above the mean edit length for the category; $w_{i\alpha} = 0.5$ otherwise;
- WA2 (referencing external source): $w_{i\alpha} = 1$, if the edit is a reference edit; $w_{i\alpha} = 0$ otherwise;
- WA3 (referencing ¿ length): $w_{i\alpha} = 1$ if the edit is a reference edit; $w_{i\alpha} = 0.75$ if the edit has length exceeding the mean length for the category (and does not contain a reference); $w_{i\alpha} = 0.5$ otherwise;
- WA4 (referencing ¡ length): $w_{i\alpha} = 1$ if the edit has length exceeding the mean length for the category; $w_{i\alpha} = 0.75$ if the edit is a reference edit (but does not exceed the mean length for the category); $w_{i\alpha} = 0.5$ otherwise;
- WA5: (referencing = length): $w_{i\alpha} = 1$ if the edit has length exceeding the mean length for the category or if it contains a reference; $w_{i\alpha} = 0.5$ otherwise;

Our validation measure is presence on the featured articles list (`http://en.wikipedia.org/wiki/Wikipedia:Featured\_articles`). Since the criteria for the allocation of Featured Article status are very demanding, we take an article's having once been awarded it as an indicator of its superior quality even though the label can be removed over time. For each featured article, we extract pages of comparable length (within 10%) from the same topic category. We then evaluate the accuracy with which different parametisations of the model correctly classify featured versus non-featured articles compared to the acccuracy obtained when the other two quality measures are used, namely revision frequency and number of distinct editors.

We compare performance of the QTR model to correctly identify featured articles with three empirical quality measures, the first being the page editing rate (with respect to length) and the second being the number of distinct editors (it has been found that articles with a greater frequency of edits and distinct editors were more likely to be higher in quality [51]), and the third being number of references. We also evaluate the association between these measures and the QTR model generated Q values.

Dataset	$r_{rev,ref}$	$r_{rev,ed}$	$r_{ref,ed}$
D1 (Music)	0.742	0.857	0.454
D2 (Mathematics)	0.100	0.469	-0.451

Table 2.1: Correlations between the three features: number of revisions, number of references, number of distinct editors.

Param	WA1		WA2		WA3		WA4		WA5	
	D1 (105, 8 feat)	D2 (33, 17 feat)	D1	D2	D1	D2	D1	D2	D1 WA5	D2
0000	7	13	7	13	7	13	7	13	7	13
0001	6	13	6	13	6	13	6	13	6	13
0010	8	12	8	12	8	12	8	12	8	12
0011	6	14	6	10	6	13	6	10	6	10
0100	6	13	6	13	6	13	6	13	6	13
0101	3	12	3	12	3	12	3	12	3	12
0110	7	13	7	13	7	13	7	13	7	13
0111	2	12	3	12	2	12	3	12	3	12
1000	0	11	0	11	0	11	0	11	0	11
1001	0	11	0	11	0	11	0	11	0	11
1010	3	10	3	10	3	10	3	10	3	10
1011	1	10	1	10	1	10	1	10	1	10
1100	0	10	0	10	0	10	0	10	0	10
1101	4	10	4	10	4	10	4	10	4	10
1110	0	10	0	10	0	10	0	10	0	10
1111	4	10	4	10	4	10	4	10	4	10
n_{rev}	6	11	6	11	6	11	6	11	6	11
n_{ed}	4	6	4	6	4	6	4	6	4	6
n_{ref}	6	12	6	12	6	12	6	12	6	12

Table 2.2: Number of correctly identified featured articles for different parametisations of the QTR model with different weight assignments, compared to when empirical features are used (number of revisions (n_{rev}), number of distinct editors (n_{ed}) and number of references (n_ref)) as described in Section 2.2.1. The first column denotes the parameter values of θ_Q, θ_R, ρ_Q and ρ_R (since there is no trust network, θ_T and ρ_T are irrelevant). r_{QEdits} denotes the correlation between the page editing rate and the terminating Q value.

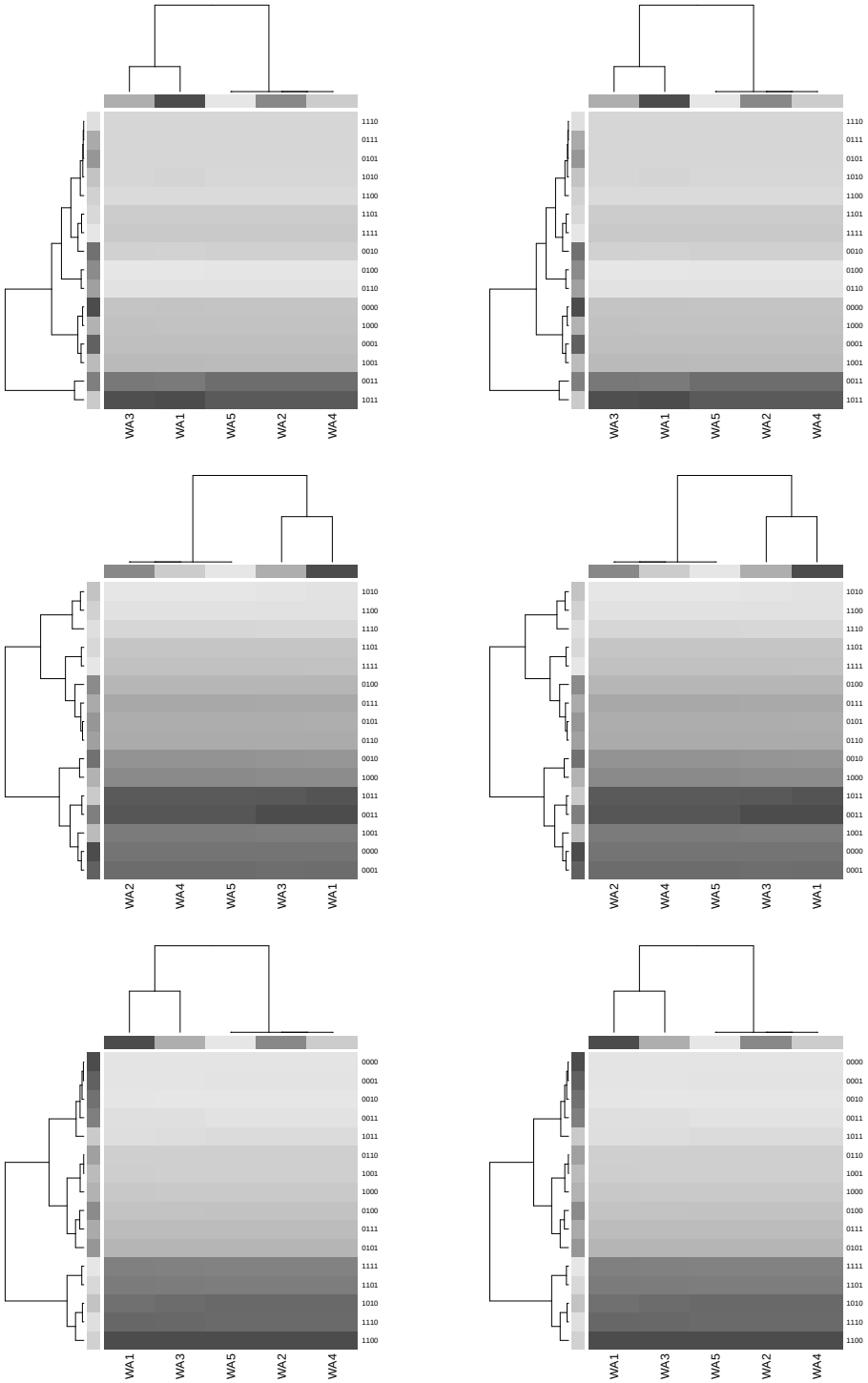


Figure 2.6: Heatmaps showing the correlations between end Q values and the different empirical measures (number of revisions n_{rev}, number of distinct editors n_{ed} and number of references n_{ref}) under the different weight assignments and parameter configurations.

2.2.2 Model behaviour

Table 2.2 shows the number of correctly identified featured articles when the QTR model is run under different parametisations, weight assignments, and datasets. These are compared to the rates obtained when empirical features are used, namely number of revisions (n_{rev}), number of references (n_{ref}) and number of distinct editors (n_{ed}; see Figure 2.7). Correlations between the Q values and these features are shown in Figure 2.6 and reported in Appendix A.1. As well as comparing with the empirical features, we also compared with random ordering of articles and found that the results obtained with the QTR model were consistently better (see Figure **??**).

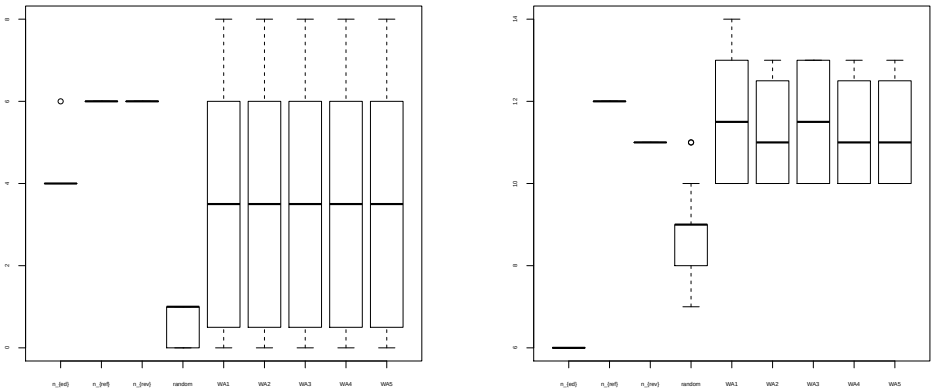


Figure 2.7: Boxplots of model performance for the two datasets. WA1, WA2, WA3, WA4, WA5 refer to the QTR model performances under the different weight assignment conditions across all parameter configurations. n_{rev}, n_{ed} and n_{ref} refer to the performance of simply using respectively number of revisions, number of distinct editors and number of references to rank articles. random refers to the random condition where items are randomly ordered.

2.2.3 Summary of QTR model evaluation

The findings reported in Section 2.2.2 suggest that while many of the parameter configurations and weight assignments of QTR correctly identify featured articles more often than the random condition, similar rates are obtained purely by using empirical features to order articles. The empirical features we considered revision rate (number of revisions per unit length, distinct editors (per unit length) and number of references (per unit length). For articles in the Music category (D1), the revision rate and referencing rate gave superior performance compared to most QTR parameterisations and compared to distinct editors per unit length. In the case of Mathematics (D2), performance of revision rate and referencing rate were comparable to most parameterisations and number of distinct editors per unit length performed very poorly.

The correlations between each of the empirical features and the end Q values differ between datasets, weighting and parametisations. For both datasets, this appears to be more sensitive to the parametisations than to the weightings. In the case of D1, this might be attributed to the fact that there is a fairly high correlation between the empirical features themselves (see Table 2.1), which would imply that the weightings are themselves associated with one another. The behaviour with respect to D2 is more surprising.

The results also show that performance of certain parametisations of QTR is superior to that which would be obtained by HITS (the 0000 configuration)

Chapter 3

Trust, homophilly and local reputation in online rating platforms: Anobii

, but this is dependent on the dataset. For example, for D1, most of the other parametisations seem to perform better than HITS, while for D2, HITS does relatively well compared to the others.

The main function served by content-sharing social networks is to allow users to share content and content annotations. These annotations can serve as signals to help users ascertain the quality of the content items so that they can decide whether or not to adopt them. An additional factor is the direct connections between users, which might be seen as an indicator of trust between users, such that connected users are more likely to adopt content owned by each other. In other words, for a given user, trust in another user can play a role in addition to reputation in determining the user's evaluation (quality score) of an item. In addition, users may be more likely to trust another user if they know that s/he is more likely to have similar quality judgements (taste similarity). This latter effect is confounded by the fact that users who are likely to share quality judgements are also more likely to give the same evaluations anyway. The main goal of the work described in this chapter is to distinguish between the effects of homophilly and social trust and determine whether there is interaction between them, using the QTR model to predict community level behaviour.

Our validation activities relate to data obtained from Anobii, an online community platform that allows users to share ratings of the books they are reading (or have read). As described in Section 3.1 (and also in more detail in [17]), we first conduct analyses of the data to confirm the role of 'taste-based' trust (which in turn comes from perceived homophilly) and show that this type of trust enhances the effect of social trust so that users who are both socially linked and who are similar or those who are socially linked *because* they believe they are similar

(as indicated by Anobii's 'neighbor' relationship), are more likely to have the same quality judgements than either those who are simply similar to each other or those who are just socially linked (as indicated by Anobii's 'friend' link, though of course homophilly may have played a major role in forming the role in the first place [39]). Specifically, our analyses support the hypothesis that explicit social connections (both neighbour and friendship) are more strongly associated with agreement than homophilly alone and that this is a plausible mechanism for maintaining the stability of rating distributions.

After identifying these effects from the dataset, we then then compare the behaviour of the QTR model initialised with different trust networks (friendship only, neighbour only, and the union of friendship and neighbourhood), as well as with no trust network to study the model's behaviour in these differnet scenarios (see Section 3.2). We find that the best correspondence with the real data occurs when the trust network is initialised with the union of the two types of links.

3.1 Taste homophilly and the social network in product evaluation

For the items we studied, the large majority had a modal score, which can be seen to represent the community evaluation of the item. This can be attributed to the tendency of items to evoke similar degrees of satisfaction across users (even if this might be for different reasons).

In addition, the set of scores for an item can be generated by users who already have a higher likelihood of giving similar scores. Our findings suggest that social links can provide a means for keeping both the central tendency and distribution of scores stable. Firstly, we find that socially linked users are more likely to give the same score to an item (possibly due to similarities in taste). Secondly, we eliminate the possibility that distributions of scores arise through attracting users with particular ratings styles (e.g. tendency to agree). Thirdly, we find that a large mean shift is much rarer for items with a large proportion of added scores from socially linked users and that this is more likely to be due to maintaining a stable *distribution* of scores than to added scores converging to the mean.

3.1.1 Related work

Although a significant body of work has addressed semantic annotation (tags), less attention has been paid to evaluative annotations (ratings) in the context of a social network platform, especially with respect to user characteristics. While community convergence in semantic an-

notation allows us to place an item in some semantic space, for many rating systems it is not immediately clear that evaluative annotations do the same for the item in 'taste' space (unless of course the rating system explicitly defines the semantic dimensions being evaluated, e.g. [2]). In many online communities, the rating system is only a one-dimensional, undefined scale. For a given item, two users might give scores at opposite ends of the rating scale in response to the very same intrinsic property (e.g. 'overly sentimental', score 1 vs. 'touching', score 4). And different users can give the same score for very different reasons. In addition, users can differ in their rating behaviour (for example, some users may try to conform to the ratings of their friends or the ratings they see; others may only rate when they disagree with the existing mean rating).

Collaborative filtering recommender systems are often based (at least partly) on user rating systems [26], [3] which are used to build similarity profiles between users. Users who share many items and rate them similarly are believed to have similar tastes, so that a high rating of an item by one of the users can be used as a basis for recommending this item to the second user. A major issue in collaborative filtering is the sparsity of shared item ratings, which means that it usually has to be enhanced with content-based information [9], [40]. More recently, measures taking into account the non-independence of relationships between users and objects have been introduced, such as the generalised model of relational similarity [34], which is based on the fact that similar objects are also rated similarly by similar actors.

An issue that has been troubling those studying online communities and social networks is distinguishing the effects of homophily, social influence, and external common causes [5, 7, 47]. With the data available to us, it is not possible to identify causal relationships between social links and ratings, but the data are consistent with there being underlying taste commonalities between users which affect the likelihood of users adopting objects.

Apart from these issues of user and item similarities, several relatively recent studies focused on the nature of reviews themselves. As regards quantitative evaluations, [29] study ratings on a dataset provided by *amazon.com* where they conclude that item ratings are essentially bimodal; they attribute this to an underrepresentation of moderate reviews. Another stream aims at predicting future ratings from the structure of existing ones, such as using various aggregation and/or machine-learning methods as in [38], or carrying regressions on the content of reviews, using for instance review length and rating evaluations as in [42], sometimes using full text models which go beyond "bags of words" or document vectors [8]. Finally, some authors addressed the credibility of reviews by examining which factors and conditions reinforced review impact [**?**, **?**] – i.e., in short, the issue of meta-rating: how ratings are subjectively rated. In general, these works examine rating distributions independently of (platform-wide) user char-

acteristics or user-item relationships. Yet an open question remains as to whether users who are more similar in taste are also more likely to follow each other when this similarity is made known to them. Are socially linked individuals with similar tastes more likely to adopt the same objects than individuals with similar tastes who are not socially linked?

3.1.2 Dataset and methods

The dataset we use to evaluate model behaviour was obtained from the authors of [4], who performed an analysis of the interaction between social network evolution and profile similarity. The authors took six snapshots of the user-book and user-user networks 15 days apart. We used only the first and fourth snapshots (hence 45 days apart) since per book, change in ratings was slow; we excluded the fifth and sixth snapshots where platform administrators had changed the rating scale from a "1-4" range to a "1-5" range.

In Anobii, the number of ratings per item and number of items rated per user exhibit heterogeneous, power-law like distributions (Figure 3.1: left). The distribution of items rated by users has an initial slow decay followed by a sharper one (Figure 3.1: right).

So as to avoid statistical artifacts arising from over-representation of distributional differences for items with small numbers of ratings (for example, agreement is over-represented), we consider only books with 10 or more ratings (hereon we shall refer to these as 'popular books'). This represents around 12% of the books (98 995 out of 847 984), and around 74% of the ratings (4 574 764 ratings out of 6 115 183). 60 093 distinct users rated books in this set, and of these, 31 362 rated 10 or more books in the set. So as to have confidence in distributional analyses at the user level, we consider only this latter set of users and will hereon refer to them as the 'rating community'.

General trends

In order to put our analyses in context, it is worth first highlighting some key features of the rating community:

- Regarding ratings:
  - 3 and 4 scores have a tendency to dominate for raters (Figure 3.2(a)),
  - agreement is relatively common, and rating distributions tend to be unimodal, with little divergence from the single mode, as exemplified by Figure 3.3 which represents the aggregate distribution, over all books, of rating divergence with respect to the median rating of each book.

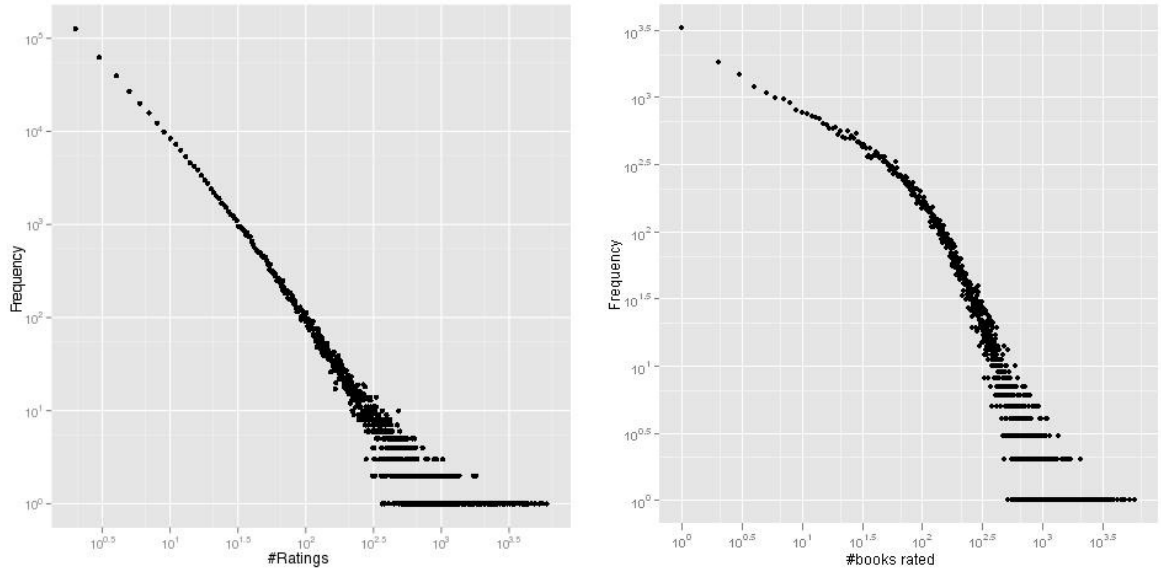


Figure 3.1: Left: Frequency plot of number of ratings per book. Right: Frequency plot of number of items rated per user.

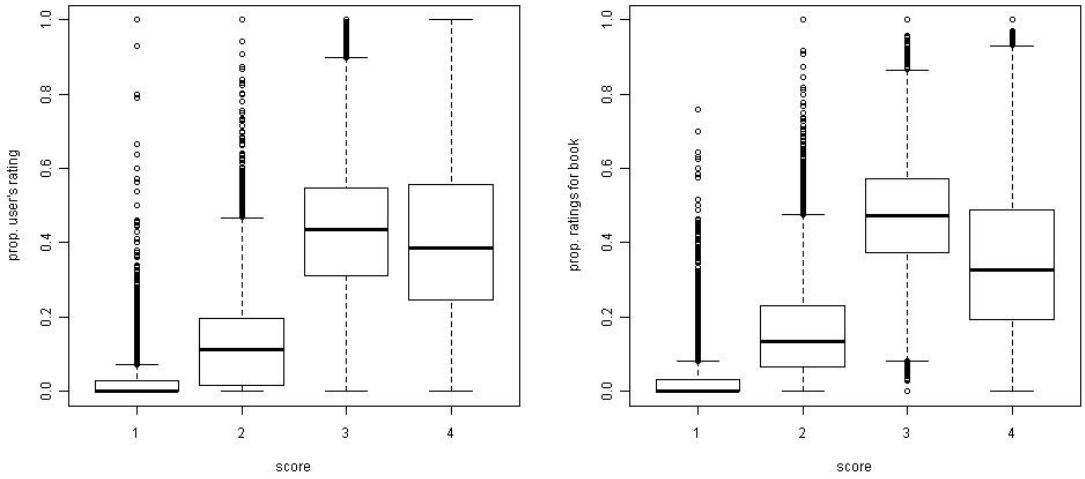
These results diverge from [29] who show that raw ratings on Amazon.com are bimodal, and who suggest that this bimodality is due to bipolar/extreme ratings. In other words, the aNobii platform seems to prevent this rating style.[1]

- Regarding sociability, our dataset displays characteristics similar to those reported earlier by [4]:
  - Social links are rare, with an overall density of $1.9 \cdot 10^{-4}$ for the union of friendship and neighbour networks; this is higher than reported for the whole user-base in [4] ($9.3 \cdot 10^{-5}$), most likely because the subset of users in our dataset are also likely to be more active and hence more social. Connectivity is also weak, with an average out-degree of 10.75 (though slightly higher than for network of the whole user-base, which has an average out-degree of ≈ 8 [4]).
  - Reciprocal links are more common than one-way links (58% for the union of friendship and neighbour networks for raters of popular book raters, only slightly higher than that for the whole network, 57%).

Since agreement between pairs of users is common (and deviations tend to be small at the item level) and links between users are rare, analyses based on linear relationships are unlikely to yield much insight. For this reason, the majority of our analyses try to identify differences between classes or communities of users, rating behaviour and items.

[1] We may perhaps speculate that this discipline is caused by some community aspect of the platform.

Items exhibit a similar pattern to that observed for users in terms of rating distribution, with the majority of items having highest proportions of 3 and 4 scores (Figure 3.2(b)).



(a) By user (b) By book

Figure 3.2: Distributions of scores.

3.1.3 Individual differences in rating styles

It is important to distinguish between ratings and rating *style*. The score that a user comes to give to an item can be seen as the combined outcome of both his/her response to the item (taste), and his/her typical rating style, which determines how this response is manifest. For example, a user with a tendency to give extreme scores may express his disappointment with an item with a score of 1 where another user with more moderate score assignment would assign a score of 2. Although we found variety among users in their distributions of scores, this can be difficult to interpret. For example, users giving high proportions of their items a score of 1 might be severe in the way they rate, but they might also have been more unlucky in their choice of books.

The rating that a user eventually gives an item can be seen as a function of both taste and the user's typical rating behaviour (rating style) (e.g. tendency to give extreme scores). As with tagging [41] and contribution in general to online communities [10], users may well differ in their motivations for rating and/or the way they use the ratings system [45]. Some users may for instance only rate an item when it has evoked a strong response, some might rate to keep a personal record of how much they have enjoyed an item and so rate nearly every item they have read, and still others may only add their rating if it differs significantly to those already existing.

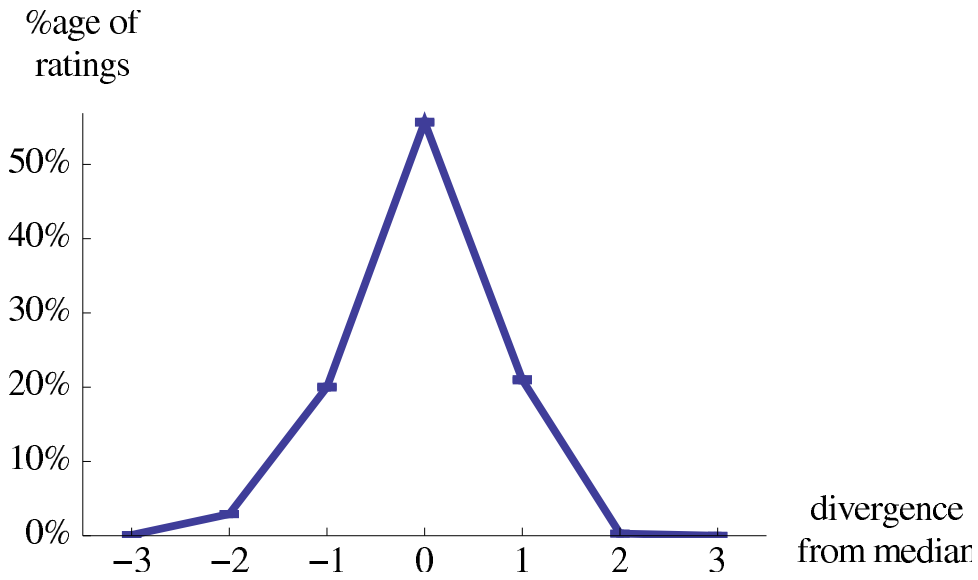


Figure 3.3: Distribution of ratings translated with respect to the median rating of each book, and averaged all books (Note: for the sake of simplicity, these distributions are computed on books whose ratings median is an integer, which represents around 95.5% of books with more than 10 ratings.)

Therefore when talking of rating behaviour, we do not give any speculative interpretation of what is giving rise to these individual tendencies. Rather, we focus on identifying associations between these tendencies and users' other activity.

Users with robust rating behaviours

As noted in Section **??**, for a given item, there tends to be consensus, with ratings clustering around a modal score, which can be seen as a community-generated evaluation of the item. Within this context however, a given user can show a high degree of divergence.

To get an idea of how members of the rating community vary in terms of their tendency to diverge, we computed the means of absolute and signed divergence of scores across all a user's n items expressed in terms of item standard deviations for all members of the rating community. More formally, for a given user with n ratings we define:

$$\overline{d_{\text{abs}}} = \frac{\sum_{i=0}^{n} |d_i|}{n} \qquad \text{and} \qquad \overline{d_{\text{sig}}} = \frac{\sum_{i=0}^{n} d_i}{n} \tag{3.1}$$

where

$$d_i = \frac{x_i - \mu_i}{\sigma_i} \tag{3.2}$$

($\overline{d_{\text{abs}}}$ and $\overline{d_{\text{sig}}}$ correlated strongly with their median equivalents; r=0.913 and hence did not appear to be sensitive to bias by extreme values.)

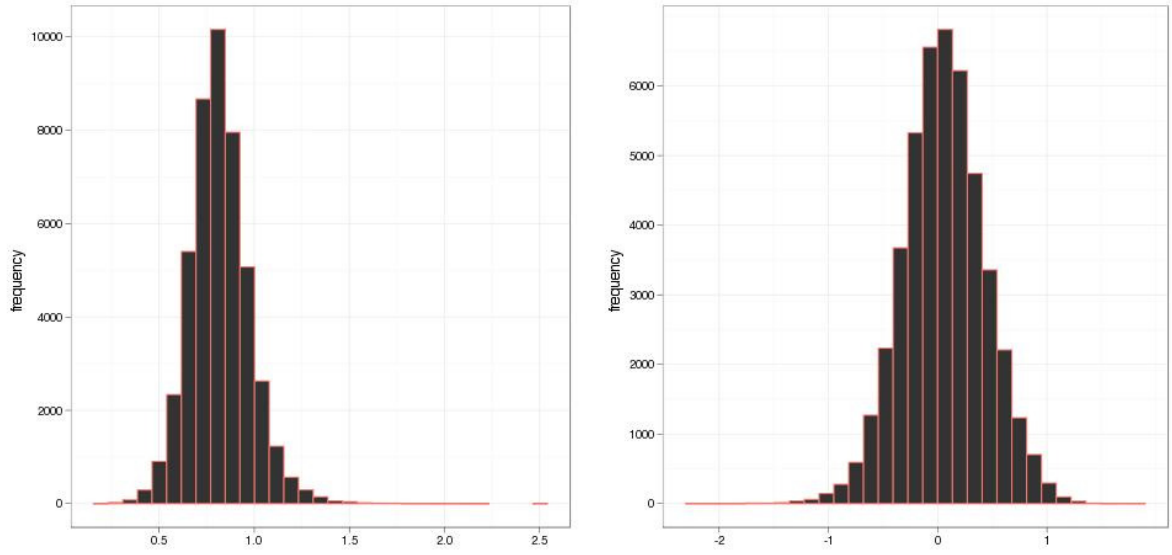
Absolute divergences indicate how much a user consistently holds an opinion different from the mean (as such a marker of 'non-conformism'), while signed divergences indicate whether this opinion consistently tilted towards a certain direction (as such a marker of negative or positive 'bias'). Figure 3.4 shows the distribution of $\overline{d_{\text{abs}}}$ and $\overline{d_{\text{sig}}}$ across users. Both the distribution of absolute and signed factors appear to be clustered around a central value (≈ 0.8 and ≈ 0.05 respectively). Consistent with the tendency to give more positive scores, the distribution of signed divergence factors is slightly negatively skewed (skewness -0.05); since raters have a tendency to rate positively, on average those who diverge more will also be those who diverge negatively.

In what way may the rating style be related to conformism or bias? In other words, do non-conformist (high $\overline{d_{\text{abs}}}$) or severe (negative $\overline{d_{\text{sig}}}$) users use a particular palette of ratings, compared with other users? To check this, we examine the association between the mean and spread of their ratings: Figure 3.5 plots the means and standard deviations of users' raw scores, against the means of their absolute and signed item divergences.

There is a positive association between the mean of users' raw ratings μ_{raw} and $\overline{d_{\text{sig}}}$, and between the standard deviation of their raw ratings σ_{raw} and $\overline{d_{\text{abs}}}$:

- The association between $\overline{d_{\text{abs}}}$ and σ_{raw} (Figure 3.4(a): left) suggests that 'non-conformism' is associated with having more different scores, although the last decile also seems to have greater variation in σ_{raw}. The last decile also has lower μ_{raw} (Figure 3.4(a): right), implying a tendency for these non-conformists to diverge negatively.

- The association between μ_{raw} and $\overline{d_{\text{sig}}}$ (Figure 3.4(b): right) suggests that users are in general equally likely to rate 'high quality' items (items with high means) as they are poorly rated ones (therefore users who have a tendency to rate more positively also have higher means, and those with a tendency to rate more negatively tend to have lower means; a lack of association would have suggested that users diverging positively were rating more 'poor quality' items and/or those diverging negatively were rating more 'high quality' items).
  The final three deciles of $\overline{d_{\text{sig}}}$ appear to exhibit a negative trend with respect to σ_{raw} (Figure 3.4(b): left), suggesting that those who tend to diverge more positively show less variation in their scores. This seems to suggest that consistency is stronger on the positive end of the spectrum.



(a) $\overline{d_{\text{abs}}}$ (absolute values) (b) $\overline{d_{\text{sig}}}$ (signed values)

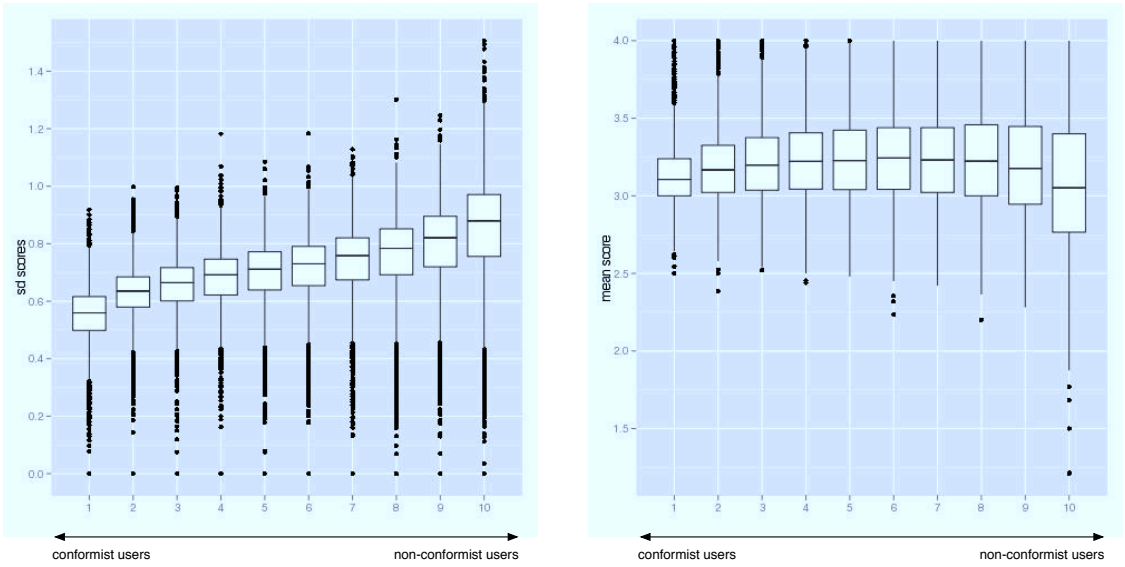
Figure 3.4: Distribution of user divergences.

Rating behaviour and book popularity

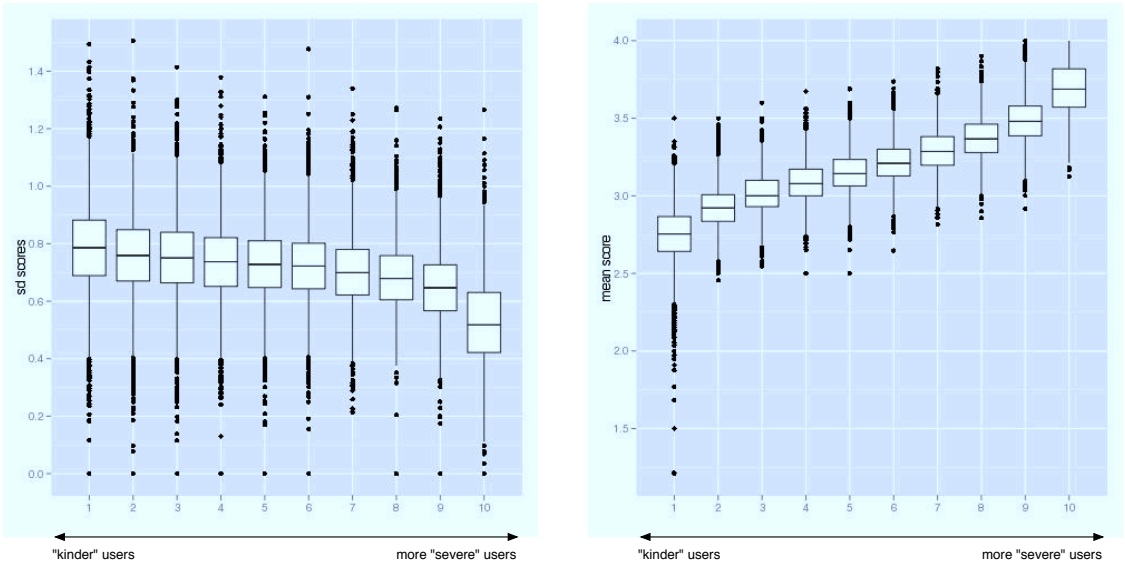
Of the 31362 users rating 10 or more popular books, 1777 rated only popular books, while the overwhelming majority of 29585 also rated less popular books. Not surprisingly though, for most users, a larger proportion of their ratings were for popular books (see Figure 3.6). More interestingly, there is a small but significant association between the proportion of popular books a user rates and his/her behaviour in terms of divergence, in the sense that more conformist users tend to have a slightly higher proportion of popular books, on average.

Rating behaviour and social activity

In terms of social activity, raters at both extremes of the divergence scale (top and bottom deciles of $\overline{d_{\text{abs}}}$) tended to be less social, with both fewer friends and fewer neighbours. They also had a slightly lower likelihood of being socially connected to each other (even taking into account the lower number of connections overall); this implies that a lower proportion of those socially linked to them had the same rating habits. The 5th decile was both the most social, with a mean of 26.2 links per user (compared to 19.6 for the first decile and 19.3 for the last decile) and had a slightly higher proportion of links coming from other users in that decile. (See Figure 3.7.)

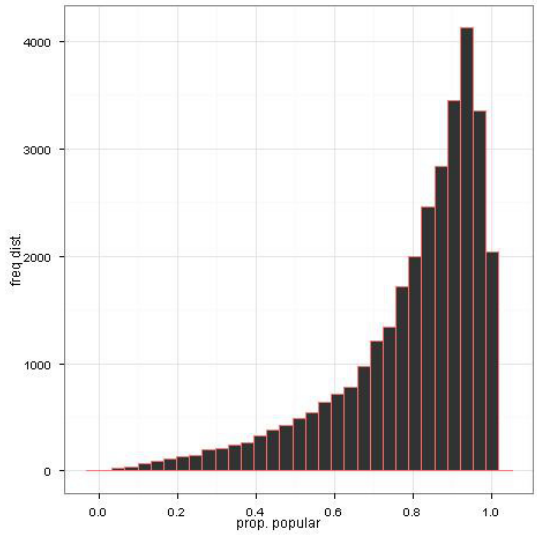


(a) $\overline{d_{\text{abs}}}$

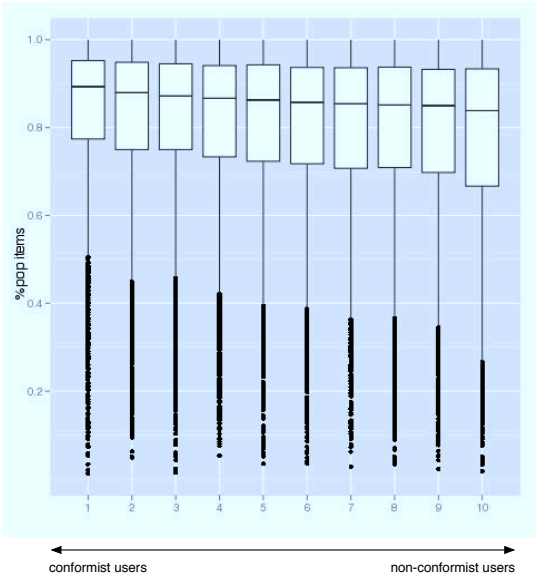


(b) $\overline{d_{\text{sig}}}$

Figure 3.5: Boxplots of $\overline{d_{\text{abs}}}$ and $\overline{d_{\text{sig}}}$ deciles against the standard deviations (left) and means (right) of raw user scores.



(a) Distribution of proportion of popular books among the books users rate



(b) Boxplot of proportions of popular book split into $\overline{d_{\text{abs}}}$ deciles (i.e. vs. 'non-conformism')

Figure 3.6: Rating behaviour and book popularity

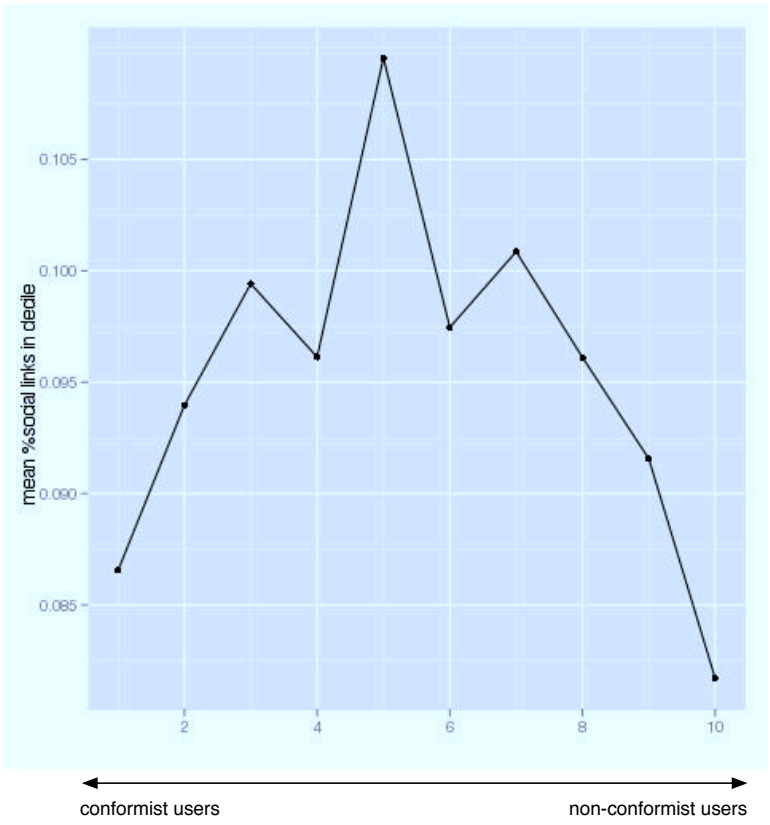


Figure 3.7: Mean proportion of users' social links that come from within the same $\overline{d_{\text{abs}}}$ decile.

3.1.4 Score distributions as the outcome of community rating

Items as communities of raters

Just as there are individual differences among raters in the way they typically rate, there may also be differences among items in the types of raters they attract. Although part of this question is answered by the rating distributions, the answer is not complete. For example, it is possible for an item to show less clustering around its central score, but the raters of this item are not necessarily those who typically diverge in their ratings (as measured by their individual-level divergences, as described in Section 3.1.3); its ratings could instead be dominated by raters who usually do not diverge much but do for this item. For each item, the mean of users' $\overline{d_{\text{sig}}}$ can be used as a measure of the polarity of users' responses; while the mean of users' $\overline{d_{\text{abs}}}$ as a measure of users' tendency to diverge. Formally, for an item with m ratings, we may define:

$$D_{\text{abs}} = \frac{1}{m}\sum_{j=0}^{m}\overline{d_{\text{abs}}(j)} \quad \text{and} \quad D_{\text{sig}} = \frac{1}{m}\sum_{j=0}^{m}\overline{d_{\text{sig}}(j)} \tag{3.3}$$

If the association between D_{sig} and the item rating mean μ were high, this would imply that positively rated items tended to be those attracting positive users and negatively rated items tended to attract negative/unlucky users. Similarly, if the association between D_{abs} and the item rating standard deviation σ were high, this would suggest that variety in scores could be attributed to more raters having more 'extreme' rating behaviours.

The data did not support either of these. On the contrary, we found the association between d_{abs} and σ to be rather weak (though still significant at $p = 0.001$), $r^2 = 0.011$. The same was true between item μ and D_{sig}, $r^2 = 0.015$ ($p < 0.001$). This seems to suggest that in general the distribution of item ratings is not the outcome of users with particular rating styles tending to rate the same items (e.g. raters with a tendency to diverge preferring more 'controversial' books). Instead, items seem equally likely to bring out users' typical and atypical rating behaviours.

These findings contrast with those for the distribution of scores for each user, where users who tend to rate more positively across their items (some of which may have low means) also tend to have higher means, and raters who tend to diverge across their items also show more deviation amongst their own scores (i.e. they are not strongly polarised in their divergence).

Agreement and social links

Although as already pointed out in Section **??** social links are in general relatively rare and agreement in ratings is relatively common, if the likelihood of agreement is higher between

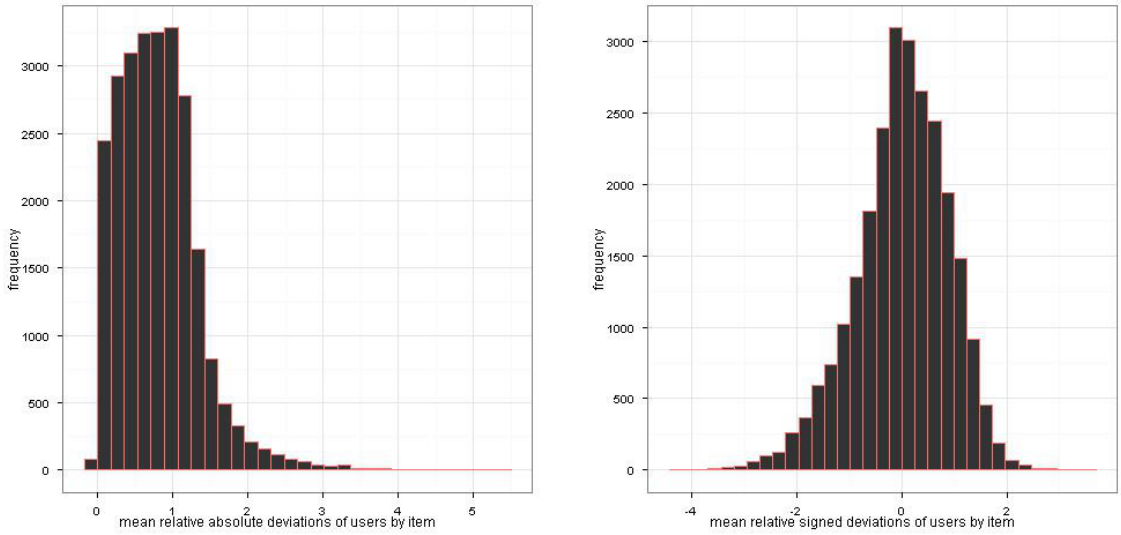
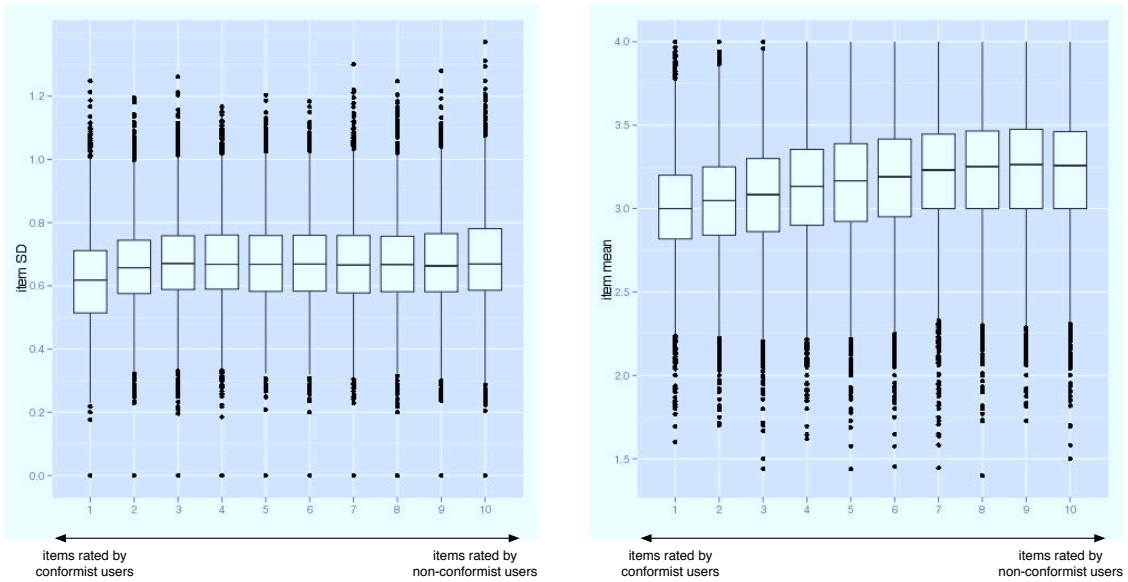


Figure 3.8: Left: Item means of the relative deviations of users' ratings with respect to their personal mean deviations. Right: Item means of the relative signed deviations of users' ratings with respect to their personal mean deviations.

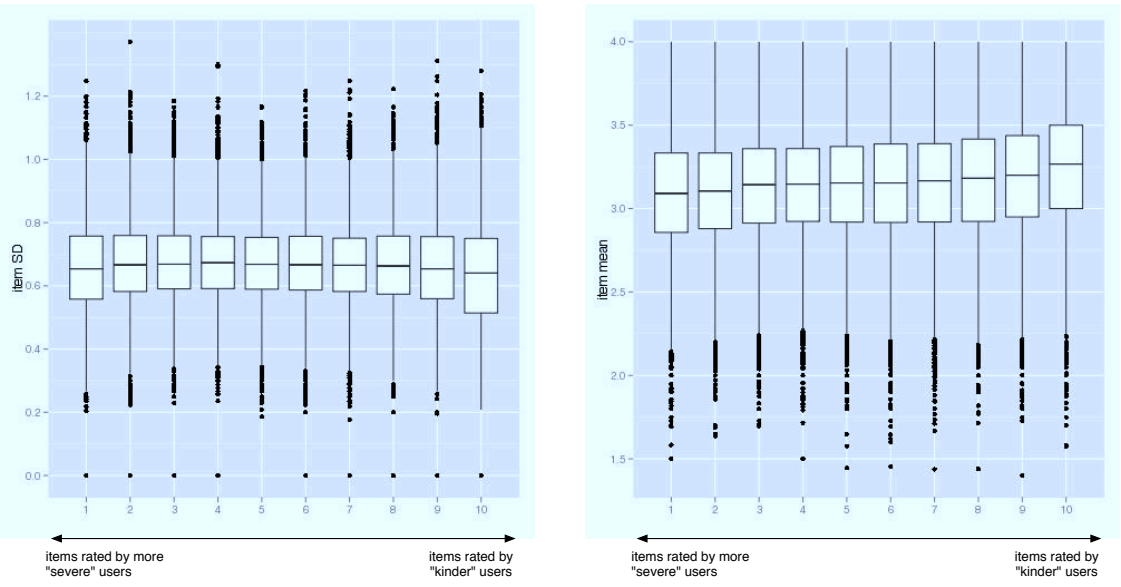
socially linked users, this could still play a significant role in shaping the rating distribution. (This need not be given a causal interpretation; indeed, social links are also likely to be associated with similarity in taste, which is the more likely reason for agreement than direct social influence.)

In terms of the proportion of pairs agreeing, the difference between socially and non-socially linked users was significant but small (solid line in Figure 3.10). This is in large part due to the high levels of agreement found for most items and the infrequency of social links.

To identify cases where social links (perhaps indicating underlying taste) would be more likely to make a difference, we took pairs of raters who agreed most with each other but disagreed most with the mode rating. For example, if the mode rating was 3, we took two of the raters giving a rating of 1. We also took into account the overall degree of consensus of an item, which we measure by the standard deviation σ_b for an item b. Given that distributions tend to be unimodal (see Figure 3.3), σ_b gives a good indication of the distribution's peakedness. The second and last deciles were then taken to represent respectively 'high consensus' and 'low consensus' items. (The second decile was used instead of the first because, although the two are qualitatively indistinguishable, a large number of books in the first had agreement for all possible pairs.) Figure 3.10 shows the proportions of social links for pairs with different agreement levels in high and low consensus items. For comparison, we also included the sixth decile as 'mid consensus' items. It is worth noting that these mid consensus items had roughly twice as many raters than low consensus or high consensus items, suggesting that a moderate



(a) $\overline{D_{abs}}$



(b) $\overline{D_{sig}}$

Figure 3.9: Boxplots of item score standard deviations (“SD”, left) and means (right) grouped by deciles of the mean absolute (top) and signed (bottom) deviations of the users rating them.

level of disagreement is associated with more rating activity.

We predicted that the difference between socially linked and non-socially linked pairs would be greater for low consensus books since for these books, the likelihood of a given pair being different from other possible pairs would be lower and hence the association between social links and agreement more easily identified.

For low consensus (high values of σ_b) items, a high agreement rate of 0.638 (3sf.) was found in pairs of raters who were neighbours, which was significantly higher (at the $p < 0.001$ level) than the agreement rate for non-neighbour pairs (0.499). For consensual items, the effect was weaker (though still significant at $p < 0.01$), with an agreement rate of 0.574 between neighbours and 0.499 between non-neighbours. We also found that agreement is more common between friends than non-friends (though the effect is weaker than for neighbours) and that disagreement is less common among friends than non-friends and that this effect is greater for low consensus items than for high consensus items.

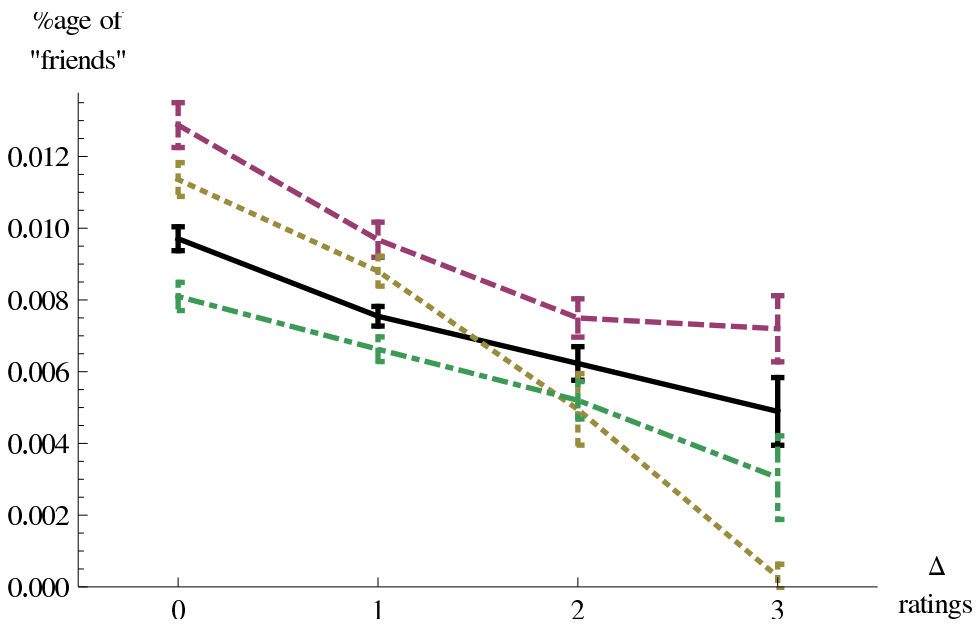


Figure 3.10: Proportion of pairs with social links for each level of agreement (Δ ratings) and for the various levels of consensus for book ratings (solid line: average over all books, dashed: low consensus, dot-dashed: mid consensus, dotted: high consensus).

Social links as a mechanism for stabilising rating distributions

Given that social links imply a greater likelihood of agreement in scores, we hypothesised that if a large proportion of an item's raters are socially linked to each other, the item should have a more peaked (distinctly unimodal) distribution. We found no such relationship. However, when we considered ratings over time (between the first and fourth time snapshots in the data

provided by [4]), we found that when a large proportion of scores were added by users who were socially linked to those who had previously rated the item, the item's mean was less likely to change by large values.

As social links are rare, when considered globally, the proportion of additional ratings coming from users who were socially linked to users who had already rated an item made up only around 10% of the additional ratings. However, for some items, a large proportion of added scores came from socially linked users. Large mean shifts were less frequent for these items (Figure 3.11, left), implying that when a large proportion of new raters are socially linked to existing ones, the book's mean is more likely to remain stable.

One explanation for this is that raters who are driven to read and rate a book by one or more of their socially connections is more likely to rate close to the mean. An alternative explanation is that it is the *distribution* of scores that is maintained. This would imply that the rates at which each of the score occur remains relatively constant over time due to the frequency of social links being proportional to the number of users giving that score.

At the user level, no general relationship was found between deviation from the mean rating and the number of social links between added and previous raters (Figure 3.11, right). In other words, a rater contributing a score in the later time slice who was socially linked to a user who had previously rated the item was not any more likely to rate close to the mean than one who had no social link with a user previously rating the item.

Consistent with this, at the item level, for low numbers of social links, there appeared to be little association between the change in mean and the existence of social links; items for which a higher proportion of the additional raters socially linked to existing raters were not any more likely to retain the same distribution (since most books tended to retain similar distributions around the same central value).

Our findings therefore suggest that it is the distribution of ratings that is sustained through the greater tendency for agreement between socially linked users.

3.2 QTR model behaviour with different trust networks

One model of the role of homophilly is that it affects the visibility of users' contributions to each other such that the contributions of users who are more similar to each other are more visible to each other than those who are less similar. Within the QTR framework, this would imply that weight assignment of interactions is dependent on similarity between user, making it different for dfferent users. An alternative model of the impact of homophilly would be that users have greater trust in those to which they are directly connected *and* to which they are

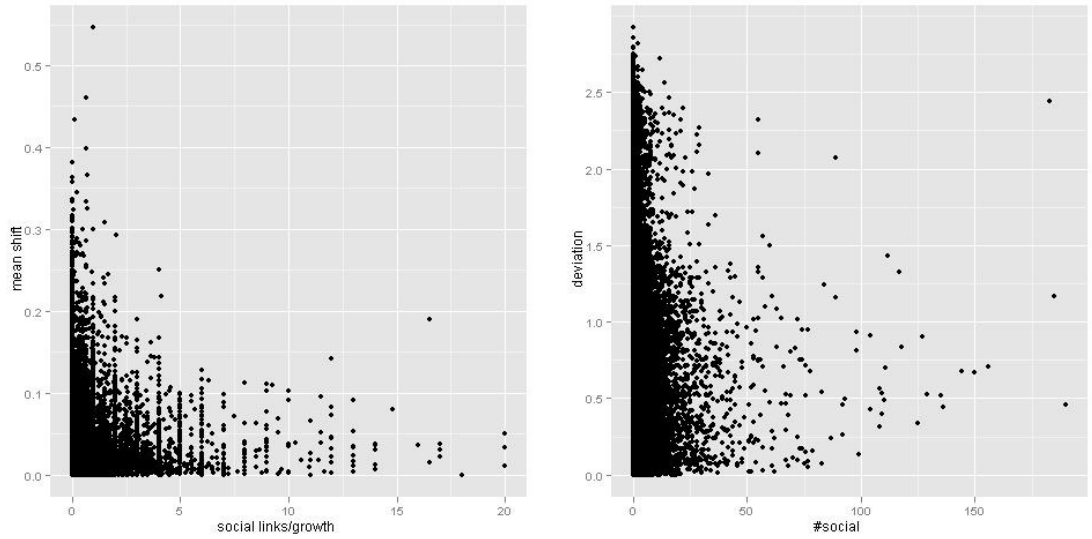


Figure 3.11: Left: Scatter plot of mean shift against social links/growth. Right: Scatter plot of user deviations against the number of social links existing in the item they have rated.

similar than in those to which they are only directly connected.

3.2.1 Social networks, homophilly and trust in Anobii

Anobii is an online community platform which allows users to share the books and their evaluations of these books (which can be quantitative ratings or qualitative comments). It also allows users to make directed social connections between each other so that they can share and/or follow other users' book choices and evaluations.

Two mutually exclusive directed social connections exist between users in Anobii: friendship and neighbour. The friendship link is derived from users' contacts in other social network platforms so for example, their facebook friends (the ones who also use Anobii) are also linked to them on Anobii. This connection is agnostic to users' taste in books in that a link may exist even when users do not share taste in books. On the other hand, homophilly implies that users who are linked socially may also be more likely to have similar tastes. The neighbour connection in Anobii is a strong indication of taste similarity since it is specific to the platform and indicates that a user has become interested in another user's book choices through using the platform. These two types of connection can be seen to indicate two distinct motivations for trust between users. In the case of neighbour connections, the user being followed (in link) is trusted on the basis of perceived homophilly (the assumption that s/he similar tastes to oneself, at least in some aspects of book choice)), while in the case of friendship connections, trust can be either socially motivated or motivated by perceived homophilly (or a combination of both).

If homophilly-based trust plays a greater role in predicting the distribution of community-

based quality evaluations, we would expect the QTR model to yield better predictions when it is initialised trust derived from the neighbour network than when it is initialised with trust derived from the friendship network. And conversely, if socially driven trust plays a greater role, we would expect better predictions with trust derived from the friendship network.

3.2.2 Dataset and methods

We evaluate the model with the same dataset as that described in 3.1.2 with only 'popular' books (books with 10 or more ratings) and the rating community (users who have rated 10 or more books). The goal was to evaluate the roles of social trust (as proxied by the friendship network) and homophilly-based trust (as proxied by the neighbour network) with the QTR model. We ran the model with different paremetisations using the following trust network initialisations:

1. Friendship network: user A 'trusts' user B if A is friends with B.

2. Neighbour network: user B 'trusts' user B if A follows B as a neighbour.

3. Union of friendship and neighbour networks: user A 'trusts' user B if either A is friends with B or if A follows B as a neighbour.

Two forms of user-object interaction weight assignment are used:

- WA1 (adoption): $w_{i\alpha} = 1$ if a rating is given to the book (irrespective of the rating value); $w_{i\alpha} = 0$ otherwise;

- WA2 (rating sensitive): $w_{i\alpha} = 1$, if the rating is 4; $w_{i\alpha} = 0.75$, if the rating is 3; $w_{i\alpha} = 0$ otherwise.

We use two different measures as a proxy for quality when validating the model:

- rate of adoption: $\frac{n_{t2}-n_{t1}}{n_{t1}}$ (n_{t1} is the number of users the book has initially while n_{t2} is the number of users the book has in the second time slice).

- mean rating increase: $\mu_{t2} - \mu_{t1}$ (μ_{t1} is the initial mean rating and μ_{t2} is the mean rating in the second time slice).

3.2.3 Model behaviour with different trust networks

We ran the QTR model under different parametisation configurations of θ_Q, θ_R, θ_T, ρ_Q, ρ_R, ρ_T for each of the trust networks (friendship only, denoted T_{fr}; neighbour only, T_{nb}; the union of friendship and neighbour, T_{frnb}, and no trust, T_0). In most cases, the parametised model did not perform well in identifying the top items (i.e. none of the top 10 or top 100 items corresponded to the top 10 or top 100 Q values) and the correlations between the Q values on termination and the validation measures (adoption rate and rating mean increase) were weak. Here, we only report the parametisations with r values greater than 0.2 for at least one of the networks. The heat map in Figure 3.12 shows the

Again, as in the wikipedia validation activities reported in Section 2.2 for the top 10 and top 100, we report precision (pr) and recall (rc) as defined in Equation **??** and **??** in Section 2.2.2, and we favour recall over precision such that if there is more than one item with the critical Q value for rank 10 or 100, we classify all items as being in the top 10 or top 100 respectively.

We only report the configurations yielding the top precision/recall values. The distribution of values across model runs is shown in Figure **??**

Adoption rate as validation measure

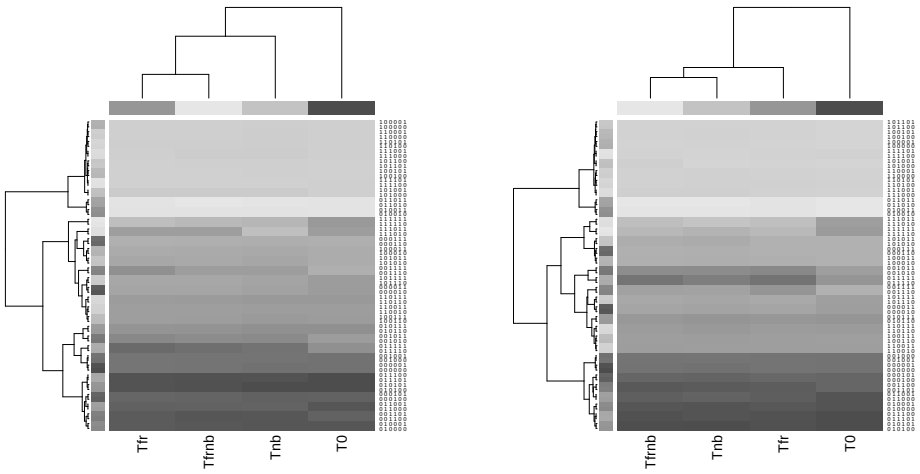


Figure 3.12: Heatmaps of correlations between the end Q values and adoption for different parametisations for different networks and weight assignments (top: WA1, bottom: WA2)

None of the configurations successfully identified the top 10 items in terms of adoption rate (both precision and recall were 0 for all configurations). However, some configurations did yield recall and precision rates over 0.1 for the top 100 (as shown in Table 3.2).

Param	W1				W2			
	T_0	T_{fr}	T_{nb}	T_{nbfr}	T_0	T_{fr}	T_{nb}	T_{nbfr}
000000	-0.394	-0.394	-0.394	-0.394	-0.394	-0.394	-0.394	-0.394
000001	-0.394	-0.394	-0.394	-0.394	-0.394	-0.394	-0.394	-0.394
000010	-0.212	-0.172	-0.174	-0.167	-0.212	-0.173	-0.175	-0.168
000011	-0.212	-0.172	-0.174	-0.167	-0.212	-0.173	-0.175	-0.168
000100	-0.446	-0.446	-0.446	-0.446	-0.437	-0.437	-0.437	-0.437
000101	-0.446	-0.446	-0.446	-0.446	-0.437	-0.437	-0.437	-0.437
001000	-0.394	-0.393	-0.380	-0.390	-0.394	-0.393	-0.380	-0.389
001001	-0.394	-0.393	-0.380	-0.390	-0.394	-0.393	-0.380	-0.390
001010	-0.212	-0.319	-0.292	-0.291	-0.212	-0.317	-0.290	-0.289
001011	-0.212	-0.319	-0.292	-0.291	-0.212	-0.317	-0.290	-0.289
001100	-0.446	-0.479	-0.474	-0.487	-0.437	-0.470	-0.460	-0.471
001101	-0.446	-0.479	-0.474	-0.487	-0.437	-0.470	-0.460	-0.471
001110	-0.104	-0.279	-0.200	-0.211	-0.106	-0.258	-0.181	-0.190
001111	-0.104	-0.279	-0.200	-0.211	-0.106	-0.258	-0.181	-0.190
010000	-0.480	-0.480	-0.478	-0.479	-0.481	-0.481	-0.480	-0.481
010001	-0.480	-0.480	-0.478	-0.479	-0.481	-0.481	-0.480	-0.481
010010	0.258	0.258	0.270	0.258	0.249	0.249	0.261	0.249
010011	0.258	0.258	0.270	0.258	0.249	0.249	0.261	0.249
010100	-0.502	-0.500	-0.500	-0.500	-0.500	-0.500	-0.500	-0.500
010101	-0.502	-0.500	-0.500	-0.500	-0.500	-0.500	-0.500	-0.500
010110	-0.273	-0.262	-0.265	-0.259	-0.252	-0.245	-0.249	-0.237
010111	-0.273	-0.262	-0.265	-0.259	-0.252	-0.245	-0.249	-0.237
011000	-0.500	-0.464	-0.440	-0.453	-0.481	-0.467	-0.443	-0.456
011001	-0.500	-0.464	-0.440	-0.453	-0.481	-0.467	-0.442	-0.456
011010	0.258	0.264	0.283	0.285	0.249	0.249	0.267	0.265
011011	0.258	0.264	0.283	0.285	0.249	0.249	0.267	0.265
011100	-0.502	-0.500	-0.485	-0.498	-0.500	-0.496	-0.478	-0.493
011101	-0.502	-0.500	-0.485	-0.498	-0.500	-0.496	-0.478	-0.493
011110	-0.273	-0.435	-0.389	-0.409	-0.252	-0.401	-0.347	-0.387
011111	-0.273	-0.435	-0.389	-0.409	-0.252	-0.401	-0.347	-0.387
100110	-0.210	-0.210	-0.213	-0.210	-0.207	-0.207	-0.207	-0.207
100111	-0.210	-0.210	-0.213	-0.210	-0.207	-0.207	-0.207	-0.207
101111	-0.210	-0.157	-0.161	-0.155	-0.207	-0.153	-0.159	-0.153
110010	-0.206	-0.206	-0.198	-0.205	-0.210	-0.211	-0.204	-0.209
110011	-0.206	-0.206	-0.198	-0.205	-0.210	-0.211	-0.204	-0.209
110110	-0.229	-0.221	-0.223	-0.220	-0.227	-0.224	-0.227	-0.221
110111	-0.229	-0.221	-0.223	-0.220	-0.227	-0.224	-0.227	-0.221
111010	-0.206	-0.205	0.005	-0.200	-0.210	-0.019	0.019	-0.022
111011	-0.206	-0.205	0.005	-0.200	-0.210	-0.019	0.019	-0.022
111110	-0.229	-0.018	-0.051	-0.045	-0.227	-0.054	-0.083	-0.060
111111	-0.229	-0.018	-0.051	-0.045	-0.227	-0.054	-0.083	-0.060

Table 3.1: Correlations between the Q values on termination and the adoption rate as defined in Section **??** for different parametisations under different trust network initialisations. The first column denotes the parameter values of θ_Q, θ_R, θ_T, ρ_Q and ρ_R, ρ_T respectively.

Param	W1				W2			
	T_0	T_{fr}	T_{nb}	T_{nbfr}	T_0	T_{fr}	T_{nb}	T_{nbfr}
010010	0.09	0.09	0.11	0.08	0.09	0.09	0.11	0.09
010011	0.09	0.09	0.11	0.08	0.09	0.09	0.11	0.09
011010	0.09	0.09	0.09	0.11	0.09	0.10	0.08	0.10
011011	0.09	0.09	0.09	0.11	0.09	0.10	0.08	0.10
100010	0.08	0.05	0.05	0.05	0.12	0.10	0.10	0.11
100011	0.08	0.05	0.05	0.05	0.12	0.10	0.10	0.11
100100	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10
100101	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10
101100	0.10	0.09	0.10	0.10	0.10	0.10	0.10	0.10
101101	0.10	0.09	0.10	0.10	0.10	0.10	0.10	0.10
110000	0.09	0.09	0.09	0.09	0.11	0.11	0.11	0.11
110001	0.09	0.09	0.09	0.09	0.11	0.11	0.11	0.11
110100	0.09	0.09	0.09	0.09	0.07	0.09	0.09	0.10
110101	0.09	0.09	0.09	0.09	0.07	0.09	0.09	0.10
111000	0.09	0.09	0.06	0.07	0.11	0.10	0.07	0.07
111001	0.09	0.09	0.06	0.07	0.11	0.10	0.07	0.07
111010	0.04	0.03	0.08	0.03	0.07	0.10	0.11	0.11
111011	0.04	0.03	0.08	0.03	0.07	0.10	0.11	0.11
111100	0.09	0.09	0.09	0.09	0.07	0.10	0.07	0.10
111101	0.09	0.09	0.09	0.09	0.07	0.10	0.07	0.10

Table 3.2: Precision and recall of top 100 Q values with respect to top 100 in terms of adoption (see Section **??**) for different parametisations under different trust network initialisations. The first column denotes the parameter values of θ_Q, θ_R, θ_T, ρ_Q and ρ_R, ρ_T respectively. In most cases precision and recall were the same (since the number of positive cases was equal), but in the cases where it was not, the unparenthesised value is the precision and the parenthesised value is the recall

Param	W1				W2			
	T_0	T_{fr}	T_{nb}	T_{nbfr}	T_0	T_{fr}	T_{nb}	T_{nbfr}
001100	-0.141	-0.179	-0.211	-0.211	-0.144	-0.184	-0.185	-0.197
001101	-0.141	-0.179	-0.211	-0.211	-0.144	-0.184	-0.185	-0.197
010000	-0.245	-0.246	-0.247	-0.247	-0.253	-0.254	-0.254	-0.255
010001	-0.245	-0.246	-0.247	-0.247	-0.253	-0.254	-0.254	-0.255
010100	-0.208	-0.204	-0.204	-0.204	-0.220	-0.217	-0.217	-0.217
010101	-0.208	-0.204	-0.204	-0.204	-0.220	-0.217	-0.217	-0.217
011000	-0.245	-0.251	-0.247	-0.250	-0.253	-0.263	-0.259	-0.262
011001	-0.245	-0.251	-0.249	-0.250	-0.253	-0.263	-0.261	-0.262
011100	-0.208	-0.221	-0.219	-0.220	-0.220	-0.230	-0.226	-0.233
011101	-0.208	-0.221	-0.220	-0.220	-0.220	-0.230	-0.226	-0.233

Table 3.3: Correlations between the Q values on termination and mean rating increase for different parametisations under different trust network initialisations. The first column denotes the parameter values of θ_Q, θ_R, θ_T, ρ_Q and ρ_R, ρ_T respectively.

Rating mean increase as validation measure

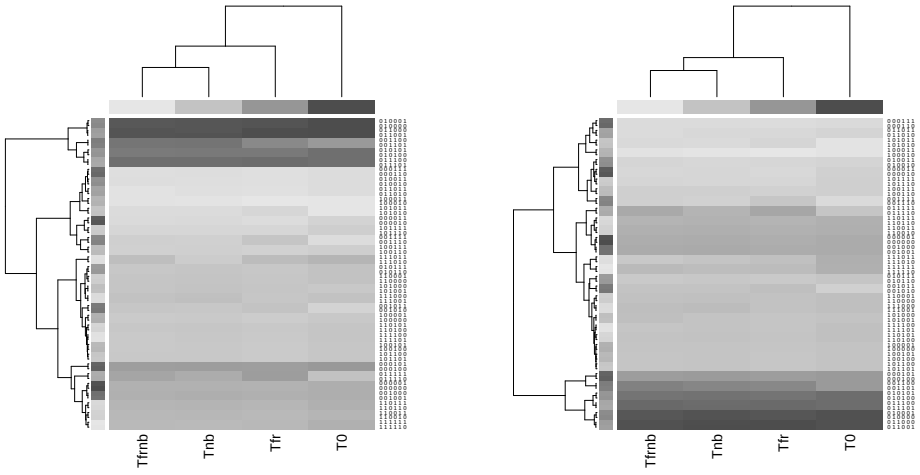


Figure 3.13: Heatmaps of correlations between the end Q values and adoption for different parametisations for different networks and weight assignments (top: WA1, bottom: WA2)

3.2.4 Summary of QTR model evaluation

The results presented in Section 3.2.3 confirm that the trust network can make a difference under some conditions, particularly for rating mean increase. However, parameterisation appears to be more important. For both weight assignments, the best performance for adoption rate was obtained with parameterisations with $\theta_R = 1$ and $\rho_T = 1$ while for rating mean increase, the best performance was obtained with parameterisations with $\theta_R = 1$ and $rho_T = 0$.

Param	W1				W2			
	T_0	T_{fr}	T_{nb}	T_{nbfr}	T_0	T_{fr}	T_{nb}	T_{nbfr}
010010	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
010011	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
011010	0.1	0	0	0	0.1	0	0	0
011011	0.1	0	0	0	0.1	0	0	0
100000	0	0	0	0	0.1	0.1	0.1	0.1
100001	0	0	0	0	0.1	0.1	0.1	0.1
100010	0.1	0	0	0	0.1	0	0	0
100011	0.1	0	0	0	0.1	0	0	0
100100	0.2	0.2	0.2	0.2	0.1	0.1	0.1	0.1
100101	0.2	0.2	0.2	0.2	0.1	0.1	0.1	0.1
101000	0	0	0	0	0.1	0.1	0	0
101001	0	0	0	0	0.1	0.1	0	0
101010	0.1	0	0	0	0.1	0	0	0
101011	0.1	0	0	0	0.1	0	0	0
101100	0.2	0.2	0.2	0.2	0.1	0.1	0.1	0.1
101101	0.2	0.2	0.2	0.2	0.1	0.1	0.1	0.1
110000	0	0	0	0	0.1	0.1	0.1	0.1
110001	0	0	0	0	0.1	0.1	0.1	0.1
110100	0	0.1	0.1	0.1	0.1	0.1	0.1	0.1
110101	0	0.1	0.1	0.1	0.1	0.1	0.1	0.1
111000	0	0	0	0.1	0.1	0.1	0	0.1
111001	0	0	0	0.1	0.1	0.1	0	0.1
111100	0	0	0.1	0.1	0.1	0.1	0.1	0.1
111101	0	0	0.1	0.1	0.1	0.1	0.1	0.1

Table 3.4: Precision and recall of top 10 Q values with respect to top 10 in terms of increase in rating mean (see Section **??**) for different parametisations under different trust network initialisations. The first column denotes the parameter values of θ_Q, θ_R, θ_T, ρ_Q and ρ_R, ρ_T respectively. In most cases precision and recall were the same (since the number of positive cases was equal), but in the cases where it was not, the unparenthesised value is the precision and the parenthesised value is the recall.

Param	W1				W2			
	T_0	T_{fr}	T_{nb}	T_{nbfr}	T_0	T_{fr}	T_{nb}	T_{nbfr}
010010	0.10	0.10	0.10	0.11	0.10	0.10	0.13	0.10
010011	0.10	0.10	0.10	0.11	0.10	0.10	0.13	0.10
011010	0.10	0.10	0.16	0.11	0.10	0.11	0.15	0.12
011011	0.10	0.10	0.16	0.11	0.10	0.11	0.15	0.12
100000	0.11	0.11	0.11	0.11	0.13	0.13	0.13	0.13
100001	0.11	0.11	0.11	0.11	0.13	0.13	0.13	0.13
100010	0.06	0.06	0.06	0.03	0.06	0.08	0.07	0.11
100011	0.06	0.06	0.06	0.03	0.06	0.08	0.07	0.11
100100	0.12	0.12	0.12	0.12	0.13	0.13	0.13	0.13
100101	0.12	0.12	0.12	0.12	0.13	0.13	0.13	0.13
100110	0.01	0.01	0.01	0.01	0.03	0.03	0.02	0.03
100111	0.01	0.01	0.01	0.01	0.03	0.03	0.02	0.03
101000	0.11	0.07	0.08	0.06	0.13	0.14	0.16	0.10
101001	0.11	0.07	0.08	0.05	0.13	0.14	0.16	0.10
101100	0.12	0.12	0.12	0.12	0.13	0.13	0.14	0.13
101101	0.12	0.12	0.12	0.12	0.13	0.13	0.14	0.13
110000	0.12	0.12	0.12	0.12	0.13	0.13	0.13	0.13
110001	0.12	0.12	0.12	0.12	0.13	0.13	0.13	0.13
110100	0.11	0.12	0.12	0.12	0.11	0.11	0.11	0.12
110101	0.11	0.12	0.12	0.12	0.11	0.11	0.11	0.12
111000	0.12	0.13	0.14	0.11	0.13	0.14	0.15	0.13
111001	0.12	0.13	0.14	0.11	0.13	0.14	0.15	0.13
111010	0.07	0.07	0.13	0.06	0.08	0.12	0.15	0.11
111011	0.07	0.07	0.13	0.06	0.08	0.12	0.15	0.11
111100	0.11	0.13	0.16	0.14	0.11	0.14	0.16	0.14
111101	0.11	0.13	0.16	0.14	0.11	0.14	0.16	0.14

Table 3.5: Precision of top 100 Q values with respect to top 100 in terms of rating mean increase (see Section **??**) for different parametisations under different trust network initialisations. The first column denotes the parameter values of θ_Q, θ_R, θ_T, ρ_Q and ρ_R, ρ_T respectively.

In general there seems to be a disparity between the model's performance in terms of correctly identifying the top items and the correlations between end Q values and the quality measures. In terms of correctly identifying top items, the performance of the model is better when mean rating increase is used as the quality measure than when adoption rate is used, but the opposite holds when correlation is used to assess performance.

The results also show that performance of certain parametisations of QTR is superior to that which would be obtained by HITS (the 000000 configuration), which did not yield values exceeding out success criteria for any of the conditions.

Chapter 4

Reputation and Quality in Science: American Physical Society

4.1 Quality and Reputation in Science: Qualitative findings from a questionnaire study

In order to better understand the factors underlying individuals' quality and reputation assessments in the context of scientific communities, a questionnaire was conducted with 163 participants from a research community. This is a broader study than more recent surveys of the ways in which citations are perceived in the scientific community, for example that described in]citeaksnes09. A full summary of the findings can be found in Appendix **??**, but here we outline the findings most pertinent for QTR.

Firstly, it was found that highest reputation ('most respected') was assigned to those *commonly* regarded as experts (42%) and who have received many citations (49%). Within the QTR framework, this can be seen as introducing external information to determine the weight of an author's contribution so that author-article interaction is weighted not only by the contribution itself but also by the citation rate of the author (i.e. a well-cited author contributing to an article is given higher weight than a less well-cited author making a comparable contribution)

The remaining participants would assign highest reputation to a much cited pioneer who has only occasionally published recently. 62% assigned the lowest reputation ('least respected') to researchers 'known for many years' (assuming they do not possess the attributes of those for the more highly respected) and 44% to those who are 'well known, respected, and a friend'. These results suggest that at least in scientists' subjective experience, direct social links are not important in determining reputation.

The second major finding from the questionnaire study conducted was that at the object

evaluation level, *number* of citations is the top quality indicator (58%); this implies that the reputation of the authors plays a more minor role at this level. Within the QTR framework, we would therefore expect parametisations where $\rho_R \approx 0$ to perform better.

Another important finding is that the distribution of citations received by an author among his/her publications is important, and not just the total number. When asked to assign highest reputation to one of the following:

1. Author of one paper which received 500 citations;
2. Author of two papers which received 200 and 300 citations repsectively;
3. Author of five papers which received 100 citations each;
4. Author of ten papers which received 50 citations each;
5. Author of thirty papers which received 20 citations each;
6. I view the reputation of all of them to be at a similar level,

31% respondents chose the third option: Author of five papers which received 100 citations each. Among the choices given, this seems to be the distribution regarded as providing the optimum balance between fecundity and prestige. This also implies that simply knowing how many papers a researcher has authored may not provide the best estimation of his/her reputation in the community; one also needs to know how well cited the papers are.

4.2 QTR model behaviour with different interaction weight assignments

labelsec:qtrAPS The American Physical Society (APS) has made available a data set containing the citations and metadata of articles in its journals dating back to 1893 up till September 16, 2010 (`https://publish.aps.org/datasets`). This includes over 450 000 articles and the citations existing between them. The data included in our analyses are those for Physical Review A, Physical Review B and Physical Review C.

4.2.1 Dataset and methods

To evaluate the model's behaviour with respect to the APS dataset, we use the number of citations c_α received by an article as a proxy for quality. We then identify the parameter and weight configurations that yield significant correlations and precision/recall rates (for the top 10% articles).

Initialising the interaction network

In the study, k_α is simply the number of authors who have authored article α and k_i is the number of articles that author i has authored. Given the questionnaire responses above, we study the effects of including the citations received by an author to additionally add weight to their contributions. In this condition, the weight of an author's contribution is a function of both their contribution and their citation rate such that a given contribution weighting (as defined in Section 4.2.1 is scaled by the number of citations received by an author).

We also validate the finding reported in Section 4.1 that direct connections between authors do not play a significant role in determining reputation or quality assignment by checking whether results are significantly better when the model is run with a trust network compared to when it is run without the trust network. In the trust network condition, a directed link exists between two authors i and j if i has cited k and the weight given to the trust relation between authors is a function of the number of times articles authored by author i have cited articles authored by author j.

We do not consider the time factor, so older articles and older authors will have end up having higher Q and R. This is because the goal of the study is to determine how the model behaves with respect to the raw data rather than to study the citation dynamics between authors and articles. We also d not feel we have sufficient information to assume that citation habits have remained constant across time.

Assigning weights to author-article interactions

A key feature of the QTR model is that it allows for different weights to be assigned to different interactions. This is pertinent for cases where, as in the case of article authorship, the interactions between producers and objects (authors and articles) can differ (for example, for any given article, an author may contribute to different degrees). Correspondingly, the differences in the interaction strength should be reflected in their contribution to Q and R (intuitively, an author's reputation should be more strongly affected by papers to which he has made a large contribution than those to which he has made only a small contribution, and an article's quality should be more strongly determined by the reputation of the authors who contributed the most to it). To study such effects, we run the model (under the different network configurations) with three different weight assignments:

- WA1: $w_{i\alpha} = 1$, if i has authored α, otherwise $w_{i\alpha} = 0$.
- WA2: $w_{i\alpha} = 1$ if i is first or sole author of α, $w_{i\alpha} = 0.75$ if i is last author of α, $w_{i\alpha} = 0.5$, if i is an author (but not first or last author), otherwise $w_{i\alpha} = 0$.

- WA3: $w_{i\alpha} = 1$ if i is last or sole author of α, $w_{i\alpha} = 0.75$ if i is first author of α, $w_{i\alpha} = 0.5$, if i is an author (but not first or last author), otherwise $w_{i\alpha} = 0$.

In the condition where citation rate also contributes to weight assignment is a product of the weight and author's citation rate c_i, i.e.:

$$\frac{c_i w_{i,\alpha}}{c_i}$$

4.2.2 QTR model behaviour with different weight assignments

The role of trust

As predicted given the findings in Section 4.1, direct social links (either in terms of co-authorship or in terms of citations between authors) did not significantly improve performance of QTR under any of the parameter configurations or weight assignments. The discussion of the results below therefore relates to conditions where QTR was run without a trust network (with parameter configurations that ignore θ_T and ρ_T).

The role of externally assigned authority (author citation rate)

None of the conditions yielded significant correlations between the citation rate and the QTR model's Q values on termination. Similarly, the precision and recall rates were extremely low (< 0.1) for all conditions, including the conditions where the *author* citation rate was included to assign weights as described in Section 4.2.1.

4.2.3 Summary of QTR model evaluation

The results reported in Section 4.2.2 suggest that for the APS dataset, when author-article interactions are used as the basis for QTR, there is little connection between the model-predicted quality ratings and citation rate, even when contributions are weighted by authors' citation rates and/or the size of contribution is taken into account (by more heavily weighting first and last authors).

The results also show that performance of certain parametisations of QTR is superior to that which would be obtained by HITS (the 0000 configuration), but as for the Wikipedia findings reported in Section 2.2, this seems to also be dependent on the dataset and weight assignment. For example, in the case of WA2, the HITS configuration does not perform as poorly for the data from Physical Review B and Physical Review C as it does for the data from Physical Review A.

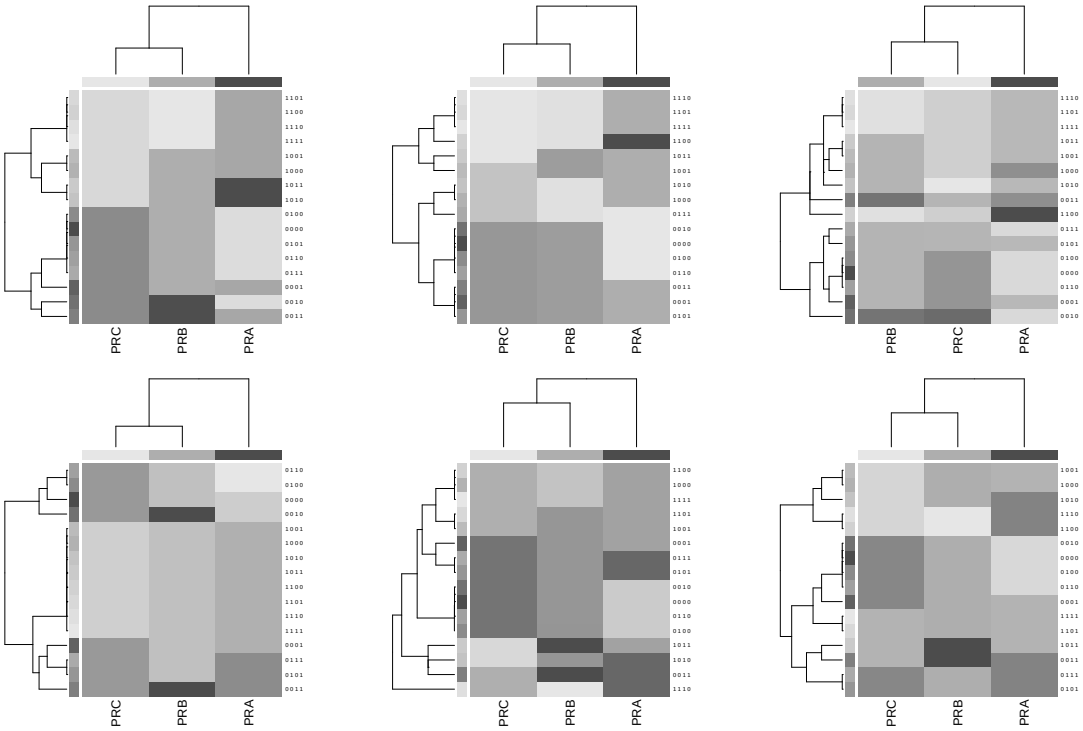


Figure 4.1: Heatmaps of the correlation between citation rate and end Q values for different weight assignments: W1, W2 and W3 (left to right) and different, and with and without the inclusion of citation rate in the weights (top: without, bottom: with)

Chapter 5

Dynamic aspects of quality and reputation assessment

In the version of QTR used in the evaluation activities described so far in Section 2.2, Section 3.2 and Section **??**, the underlying assumption was that quality and reputation are stable through time i.e. if an object has quality value x at t_i, it will also have quality value x at t_j. However, in all three contexts, this is likely to be an oversimplification. In the case of Wikipedia and scientific articles, relevance is a factor that can come into play when assessing both quality and reputation. In the case of Anobii and other rating platforms, it goes without saying that individual's tastes and interests are not static (indeed, as already discussed in Section 3.1).

Section 5.1 addresses the question of how novelty and citation rate interact in scientists' quality judgements. Section 5.2 studies the dynamics of citation through articles' lifetimes and introduces measures that capture these dynamic properties.

5.1 Further findings from questionnaire study

In the questionnaire study reported in 4.1, two questions were included that sought to probe more deeply into the role of time in assessing quality.

1. You want to learn about a new area of research and are searching for a paper to read. A bibliographic search tool returns several results; which paper would you choose to read first?

    - A paper published a month ago with 10 citations;
    - A paper published 1 year ago with 50 citations;
    - A paper published 5 years ago with 200 citations;

- A paper published ten years ago with 1000 citations;
- I view all of them as equally suitable.

2. Rank the following articles according to their quality, basing your judgement only on the information below. Start with the highest quality first.
   - Article with some recent citations;
   - Article with many citations;
   - Article with many recent Internet downloads;
   - Article with a very large number of Internet downloads;
   - Article based on work presented in many conference talks;
   - Article based on work presented in a conference key note presentation.

In response to the first question, 25% participants chose the first option, "a paper published a month ago with 10 citations" and 27% chose the second, "a paper published a year ago with 50 citations". However, 31% responded "I view all of them as equally suitable", suggesting that the interaction between citation rate and recency is far from straightforward.

In response to the second question, 58% participants ranked the second option "article with many citations" top. The majority assigned the second rank to either the first option, "article with some recent citations" (29%) and the final option, "article based on work presented in a conference key note presentation". Interestingly, the lowest ranks were assigned to the third, fourth and 5th options "article with many recent internet downloads" (28% at rank 5), "article with a very large number of internet downloads" (26% at rank 5; 28% at rank 6), and "article based on work presented in many conference talks" (31% at rank 6). These findings suggest that being cited is a stronger quality indicator than simply being downloaded, and that external signals of authority, such as being based on work presented in a conference key note presentation, are also important. The fact that an article is based on work that is presented in a conference key note presentation suggests both that the author is well-respected in the community and that the article contains work that this author considers important (important enough that s/he would present it as part of a keynote). The findings also suggest that the quantity of citations is more important than recency for more participants.

5.2 The role of local dynamics in citation networks

Rather than treating scientific articles simply as objects produced by their author(s), each paper also occupies a position in the network of all the articles to which it is related (all those it

cites and is cited by). This network changes over time. Upon publication, a paper inserts itself into the network of existing references by building on and from the state-of-the-art [43]. It contributes to creating connections between bits of existing literature of various ages, and is in turn deemed relevant by being progressively embedded in the developing network of ongoing citations [30]. As observed in [46], this continuous and cumulative process is rarely rendered by aggregate citation observations at a given date. And while a highly cited paper is certainly of relevance to the scientific community, there has also been significant debate as to how such an "impact" translates into attested quality [37, 22] when factors such as distinct [52] or biased citation practices [22, 27], excessive focus on "hubs" [18] or potentially flawed statistics derived from aggregate citation metrics [20, 21] can often lead to some articles or authors being 'over'- or 'under-represented'.

Rather than considering only (static) patterns at the global network level, we seek universal behaviours in local citation dynamics in both temporal directions, backward and forward. A more detailed account of the work described in this section can also be found in [46].

5.2.1 Dataset and Methods

Dataset

The data we use comes from Thomson Reuters Scientific "ISI Web of Science", spanning over 1960-2009 for four general *a priori* categories: computer science (denoted by "cs"), economics ("eco"), engineering ("eng") and physics ("phys"). We chose these fields so that we would have a wide range of sizes (as evidenced from Tab. 5.1), age ("cs" being younger than "phys"), as well as at least one field from social sciences ("eco") and one from applied sciences ("eng"). (We do not however rule out the possibility that each field also has its own idiosyncrocies, so we do not seek to make strong statements about the generality of the results presented).

Introducing temporal metrics requires to observe papers some time before and after publication. We therefore needed to restrict the dataset to papers having sufficient history and, in some cases, to papers having sufficient dynamic information both in terms of citations or references. We suggest that 10 years of history ahead and before provides a sufficient temporal resolution; by doing so, we still consider about 30 years of data by focusing on papers published between 1970 and 1999.

Going further, we consider several suitable constraints: "`ALL`" represents papers having at least 10 years of history before and after publication date, among which "`CMIN`" represents the subset of papers having at least 10 citations after 10 years, and "`RMIN`" the subset of papers having at least 10 references at most 10 years old. The rationale for defining `CMIN` and `RMIN`

Name	Description	Paper count			
		cs	eco	eng	phys
RAW	Papers from the raw dataset	634 592	268 785	2 056 282	2 543 769
ALL	At least 10 years of forward and backward history	258 303	142 935	1 049 626	1 365 774
CMIN	At least 10 citations after 10 years, among ALL	21 167	15 595	66 465	349 792
RMIN	At least 10 references at most 10 years old, among ALL	5 381	8 610	17 720	317 106

Table 5.1: Constraints applied to the data.

lies in the fact that average citation or reference ages cannot be defined when papers have few or no citations; most importantly, dynamics of link arrival are likely to be sparse and strongly discretized on few points. Again, we arbitrarily assume that 10 points provide a sufficient resolution on the dynamics, while it does not jeopardize statistical significance: these constraints result in diverse size reduction effects on each dataset as shown on Tab. 5.1 (see Fig. **??** too). More precisely, ALL only aims at ensuring sufficient temporal coverage before and after publication and includes 100% of papers published over 1970-1999. From this baseline, CMIN will be used for citation dynamics-related variables and does not deal with the majority of papers which have few or no citations (keeping from 6% ("eng") to 26% ("phys") of ALL papers). Similarly, RMIN will be used for reference dynamics and conserves between 2% ("cs") and 23% ("phys") of ALL papers, i.e. the minority of papers having a significant span of references from the field.[1]

History of citations and references

We use the data described above to examine the relationship between the *temporal ordering* of links arriving to and from a given paper, and its *relative citation count* with respect to its context. By *relative citation count* we mean that we consider the number of citations received by a paper published in a given year with respect to the mean number of citations that all papers published in the same field have received after an identical period of time [44, 15, 12].

By *temporal ordering of links*, we mean that we work, in the broad sense, on arrival times of links connected to a given paper, distinguishing between links originating from a paper (*references*) and those pointing to a paper (*citations*), see Fig. 5.1.

Most basic features relate to the number of references and citations over the considered time window. Various elaborate dynamics-related metrics are derivable from first neighbor-

[1]Nonetheless, we checked that our results do hold qualitatively using thresholds of 5 and 20 for CMIN and RMIN.[2] In the remainder, we then use a threshold of 10 as a decent trade-off between temporal resolution (the higher the threshold, the better) and statistical significance and representativity.

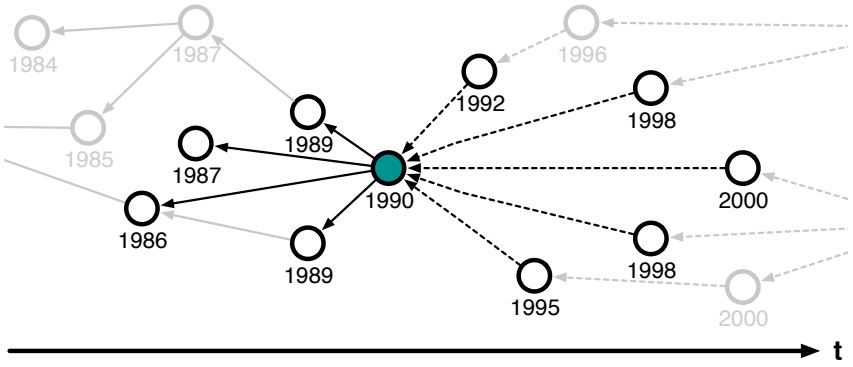


Figure 5.1: Dynamic, first neighborhood of a sample paper published in 1990. At left, references (solid links); at right, citations (dashed links).

hood linking patterns — from the mean temporal gap from publication time (i.e. average age of references or citations) to metrics indicating in a more sophisticated manner how much a citation profile experiences a late surge of interest or not (called below the "rebirth index").

For each paper i, from the dynamic citation network we define a citation vector c_i (resp. reference vector r_i) describing temporally the number of incoming links (resp. outgoing links), such that $(c_i)_j$ is the number of incoming links or citations at j years after publication (resp. $(r_i)_j$ denotes the number of outgoing links or references j years old before publication). Because of the above thresholds, j ranges from 0 to 10 for a given c_i, and from -10 to 0 for r_i. In the toy example presented in Fig. 5.1, for, say, paper "0" from "eng", we thus have: $c_0 = (0, 0, 1, 0, 0, 1, 0, 0, 2, 0, 1)$ and $r_0 = (0, 2, 0, 1, 1, 0, 0, 0, 0, 0, 0)$.

The raw *citation count* C_i and *reference count* R_i of paper i are simply the sum of its raw citation history and, respectively, reference history:

$$C_i = \sum_{j=0}^{10} (c_i)_j \qquad R_i = \sum_{j=-10}^{0} (r_i)_j \tag{5.1}$$

The relative citation count of i is thus the ratio between C_i and the mean of C'_i for all papers i' published on the same year as i. Here, $C_0 = 5$ and $R_0 = 4$.

Dynamics-related variables

To compute citation dynamics-related variables, we normalize each history vector c_i and r_i twice: values are first adjusted such that $(\hat{c}_i)_j$ denotes the ratio between $(c_i)_j$ and the total number of papers published on year j, and eventually $\hat{c}_i$ is normalized such that its coefficients sum to 1 — we denote twice-normalized vectors with a tilde: $\tilde{c}_i$. Adjusted proportions presented on Fig. 5.4 are averages of $\tilde{c}_i$ for all papers i in a given field. This kind of

normalisation makes it possible to compare patterns across different 'historical periods'. In our example, $\hat{c}_0 = (0, 0, \frac{1}{44\,560}, 0, 0, \frac{1}{60\,673}, 0, 0, \frac{2}{72\,680}, 0, \frac{1}{74\,693})$ as there are respectively $44\,560$, $60\,673$, $72\,680$ and $74\,693$ papers published in "eng" in 1992, 1995, 1998 and 2000. Then, $\tilde{c}_0 \simeq (0, 0, 0.28, 0, 0, 0.21, 0, 0, 0.34, 0, 0.17)$. A similar computation goes for $\tilde{r}_0$.

Citation and reference age We define a variable related to the citation delay of paper i as the expectation of the distribution across time of its normalized citation history. We denote it by $\langle c_i \rangle$. We similarly define the reference age $\langle r_i \rangle$ on the normalized reference history distribution. Formally, we have:

$$\langle c_i \rangle = \sum_{j=0}^{10} j \cdot (\tilde{c}_i)_j \qquad \langle r_i \rangle = \sum_{j=-10}^{0} j \cdot (\tilde{r}_i)_j \tag{5.2}$$

While these variables are computed on citation and reference counts normalized by the number of papers published on each corresponding year, note that the results are qualitatively stable if we consider non-normalized mean citation and reference age, i.e. actual mean delay of citation, irrespective of the volume of papers published on each respective year. In the case of paper "0", $\langle c_0 \rangle = 6.03$.

Rebirth index In addition to the basic metrics introduced above, we can also study the concept of quality of a publication by focusing on more sophisticated dynamic patterns. A significant example of this are the so-called 'Sleeping Beauties' or 'Early Birds' papers. In the literature [49, 14, 13], such publications are characterized by making a significant contribution that triggers no or very little interest upon publication, but is recognized by the community only long time afterwards. These works are based on a binary operationalization of the idea of delayed recognized articles, i.e. papers belong or not to this category.

Building upon this literature, in order to further facilitate the systematical characterization of papers having a second life after a period of sleep, here we propose a second-order dynamic pattern named "*Rebirth index*" and denoted as ρ. This index aims at describing how likely it is for a publication to have two periods of citations, that is, two distinct modes in their citation history distribution. In comparison to previous approaches, it allows for a continuous analysis, i.e. different degrees of delayed recognition. It also focuses on papers which, upon publication, exhibit at least a first surge of interest rather than no impact at all. Notice that this feature allows us to avoid relying on an a priori "rebirthing" threshold. Our Rebirth index is defined as

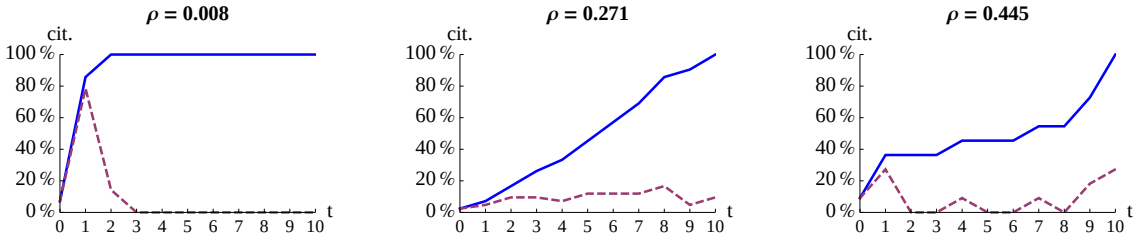


Figure 5.2: Behavior of citation profiles with respect to three typical rebirth index ρ values (from left to right: low to high). Samples were taken from the '*phys*' dataset.

follows for a given paper i of normalized citation history $\tilde{c}_i$:

$$\rho_i = \frac{\sum_{t,t' \in \{0,..,10\}^2} \tilde{c}_i{}^2(t)\, \tilde{c}_i{}^2(t')\, \frac{|t-t'|}{10}}{\sum_{t,t' \in \{0,..,10\}^2} \tilde{c}_i{}^2(t)\, \tilde{c}_i{}^2(t')} \tag{5.3}$$

in such a way that ρ_i gives more weight to plateaus (i.e. irregularities in $\tilde{c}_i(t)$). In other words, ρ is higher when a paper exhibits a first burst of citations, then a kind of stagnation period and, finally, a second burst.

Some examples representing typical possible scenarios can be found on Fig. 5.2. Considering paper "0" again, $\rho_0 = 0.31$. See also **??** for more details regarding how c_i, $\hat{c}_i$, $\tilde{c}_i$ and ρ_i vectors and indices may be computed.

5.2.2 Key findings

Universal scaling in the preference towards the recent past

We first focus on the age of forward and backward links, i.e. citations and references. Their average ages are, remarkably, similarly distributed across all datasets, as shown on Fig. 5.3. Besides, both distributions are extremely modal, and their mean does not seem to evolve in time: in other words, papers published in either 1970 or 1999 are, on average, getting cited after almost exactly the same time.

For a given paper, the probability of being cited after y years or of featuring references y' years old is represented on Fig. 5.4. Notice that probabilities are normalized in order to effectively allow comparison across historical periods: as said in the above section, we normalize quantities of references or citations on a given year by the number of papers published on that year in the given field. On a paper-by-paper basis, reference lists are more likely to include papers that are around three years old. The same goes for citations: papers are generally getting

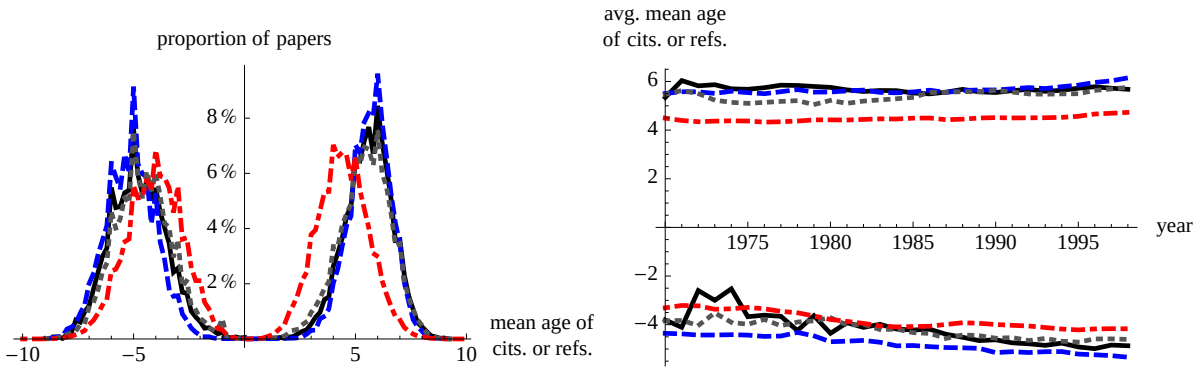


Figure 5.3: Left: Distributions of papers having given mean ages of citations or references, respectively for CMIN and RMIN papers (bins of width 0.2). Right: temporal evolution of the *average* mean ages of citations or references (put simply, it is the temporal evolution of the mean of distributions featured on the left). Legend: *cs*, black solid; *eco*, blue dashed; *eng*, gray dotted; *phys*, red dot-dashed.

the highest rate of citation within around 3-5 years of their original publication. These findings are, actually, neither surprising nor new. De Solla Price already discussed the concept of *Immediacy Factor* in Ref. [43], using a limited sample of data. Moreover these results are in line with those of Glänzel and collaborators in [25, 24]. However, as far as we know, no previous work has generalized such an observation by comparing temporal citation profiles across scientific fields, as we are doing here. The cross-field similarity shown in Fig. 5.4 can indeed be stressed by re-scaling the time variable of plots. As shown in the inset of the figure, once re-scaled the different citation patterns almost overlap. This result suggests the generality of the underlying citation dynamics, just differentiated from one field to the other by temporal factors like the duration of publishing processes (economics and, to a lesser extent, computer science being comparatively slower than engineering which itself has a slower pace than physics).

References and future impact, recent and normal science

In order to get a deeper understanding of the relationship between temporal linking aspects and paper impact, we examined how linking patterns are correlated with relative citation counts. Results are summarized on the panel on Figure 5.5 (three first columns) and suggest essentially that:

1. Papers with a larger number of references *from the field* are very significantly more likely to get cited *within the field* — up to 4 times more than the average for papers having about 10 references, and 4 times less for papers having none.

2. Papers with a low (but not too low) average reference age are more likely to be more

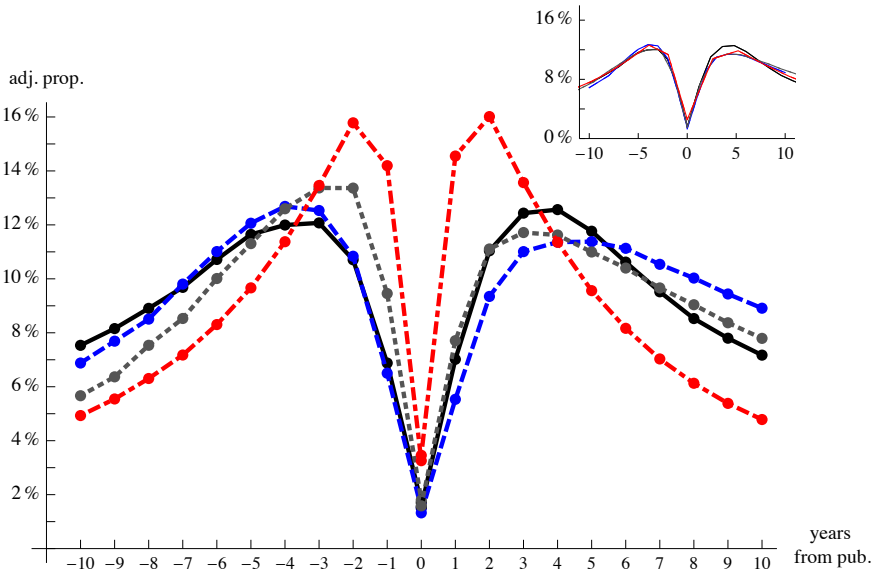


Figure 5.4: Adjusted proportions of reference links pointing back to $y < 0$ years before publication, or respectively, citation links being $y > 0$ years old. Adjusted proportions are averages of $\tilde{c}_i$ for all papers i in a given field. (Observations made on papers from `RMIN` and respectively `CMIN`). Inset: The same adjusted proportions after time re-scaling. More concretely, $t_{\text{eco}} = t_{\text{cs}} = 1.37 \cdot t_{\text{eng}} = 1.95 \cdot t_{\text{phys}}$ for references and $t_{\text{eco}} = 1.2 \cdot t_{\text{cs}} = 1.37 \cdot t_{\text{eng}} = 2.6 \cdot t_{\text{phys}}$ for citations. Profile overlapping is highly significant. Colors are set as defined in Fig. 5.3.

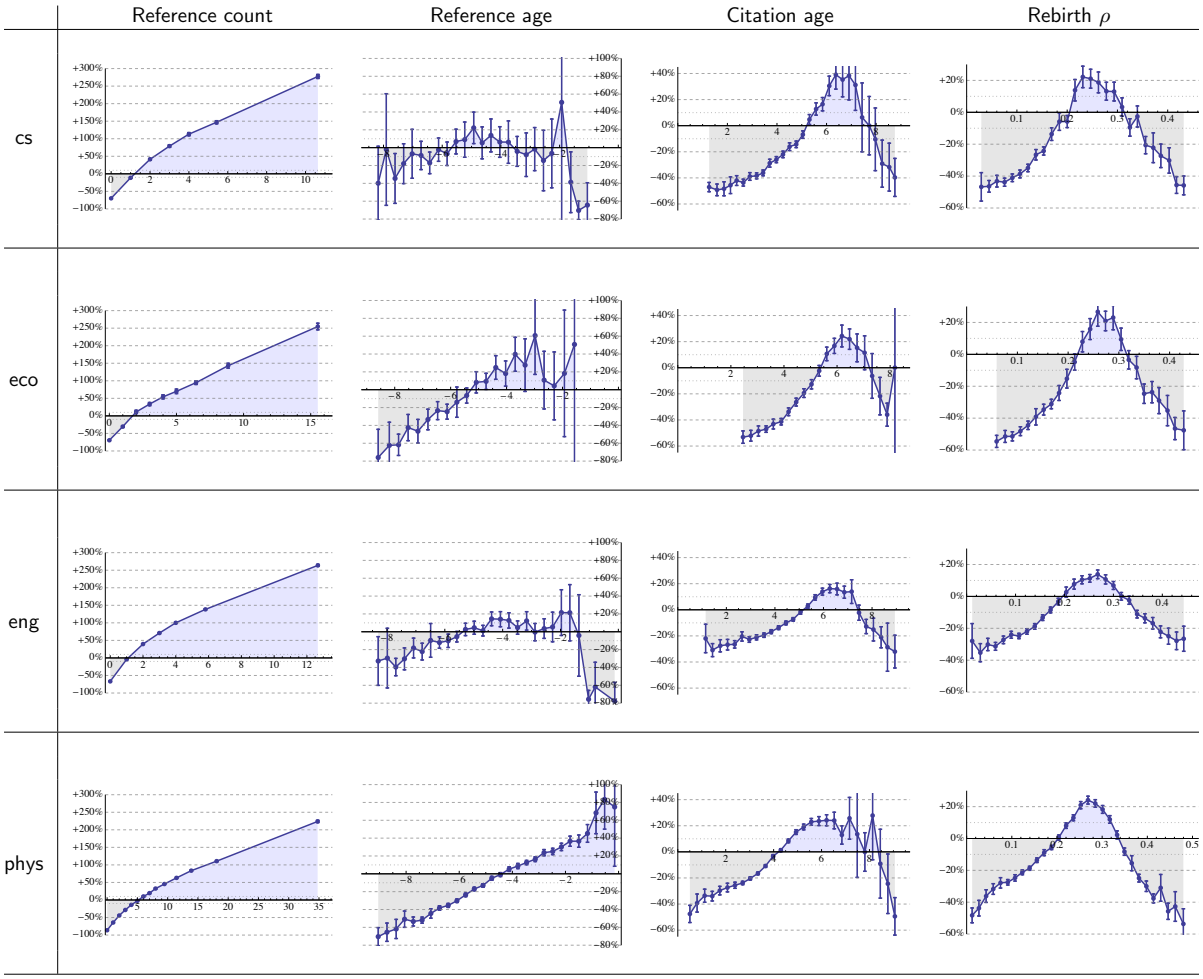


Figure 5.5: Relationship between relative citation counts (expressed in percentages above or below the average of the given field) and various citation dynamics variables. Dots indicate bin means, error bars correspond to confidence intervals ($p = .95$).

cited; in particular, papers with a high mean reference age (papers mostly citing "old publications") are significantly less cited.

This regularity is particularly striking in physics, and much less obvious in computer science; it is also the only feature which is not similar across all fields.

3. Papers with a medium citation age are those which are generally more cited. This last feature possibly indicates that papers with highest relative impact are having a longer career, yet not too long: lately cited papers might get cited too late to raise sufficient mainstream interest.

Interestingly, features 1 and 2 may globally both act as predictors, at publication time, of the likely future impact of a paper in its field. In any event, on the whole, these results emphasize that the most discipline-focused papers and those citing essentially the most recent science are getting the highest interest from the community. Notice that these are traditional markers of "normal science", but could also correspond to review articles.[3] By contrast, interdisciplinary papers and/or papers referring back to older times are raising a markedly weaker interest.

Rebirth as a complementary quality criterion

As observed above, papers being cited late (higher average citation age) are receiving less citations from the community. Such kind of papers are likely to belong to an especially interesting type (e.g. 'sleeping beauties'), yet their empirical occurrence has also been shown to be rare [24, 1]. More precisely, from these previous studies, which also address the validity of citation-based indicators and the role of time windows when analyzing rebirthing papers [**?**, in particular]]Glanzel:longterm, "non-impact" on the early years after publication appears to be a good predictor of future "non-impact" — in other words, from papers which exhibit "non-impact" initially, few eventually surge. A similar effect is observable in our study: in "phys", for instance, papers with an average citation age over 7.5 years (corresponding to less-than-average citations, from Fig. 5.5) represent only 0.37% of `CMIN`, or about a thousand papers (to be compared with the 1.37 million papers present in `ALL`).

However, these works were based on an *a priori* threshold where a "sleeping beauty" is defined to be such that after s steps we have c citations. By contrast, our re-birth index ρ aims at discriminating papers with two periods of citations, among those which have a minimal citation count (`CMIN`). Its continuous nature additionally allows for a more complete study of

[3] Unfortunately, a comprehensive examination of the dataset in order to identify separately these two classes of publication was out of the range of this work.

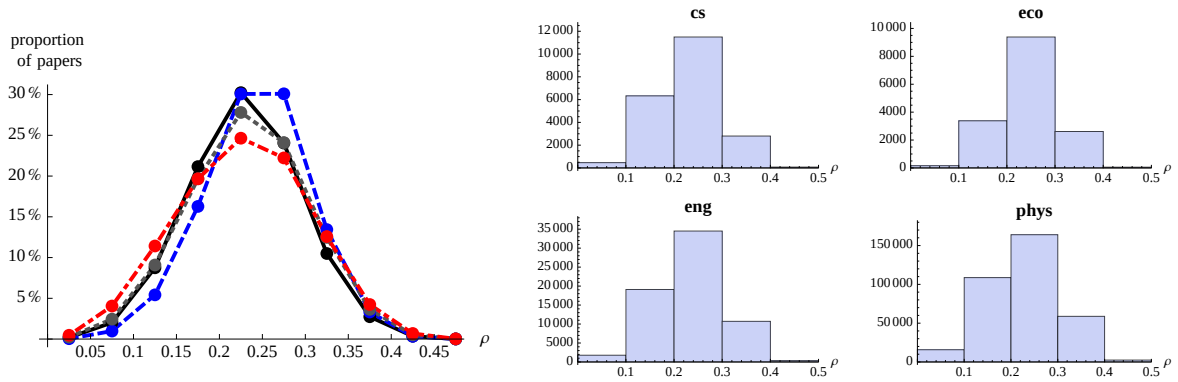


Figure 5.6: *Left:* distribution of papers according to ρ for all datasets (markers located at the middle of each bin of width 0.05). *Right:* absolute number of papers having a given rebirth index (histogram bins of width 0.1).

these cases. First we have computed the distribution of ρ values across the four discipline datasets, as shown in Fig. 5.6. In all four cases ρ appears to be modally distributed around ≈ 0.25, with a small majority of cases falling within the $[0.2, 0.3]$ range, while a still significant proportion of papers exhibits lower values ($[0, 0.2[$, around 30% of papers on average) and higher values ($\rho > 0.3$, around 15% on average).

Then, we have calculated the relative impact of articles as a function of their rebirth index value, which can be found in Fig. 5.5. These results can be interpreted in terms of low, mid and high rebirthing characteristics without pre-defining such categories and specific range values:

- *Low values of ρ:* These are short-lived papers whose impact is generally limited to its publication time. It is not surprising that they got fewer citations than average.

- *Mid values of ρ:* These are papers that have been receiving citations, more or less continuously, during 10 years. This means that either they made an interesting contribution to the community, or the authors have been continuously supporting them through self-citation. At least in the former case, it is normal (and deserved) that they have a high relative citation count.

- *High values of ρ:* These papers all had a first surge, then a plateau, then a boost again, 7-10 years after publication. These papers are not getting much success in terms of eventual total citation impact. This could sound counterintuitive: papers which are rediscovered later and, therefore, experiencing a "second life", are plausibly helpful for a future field.

Put shortly, we rather show that papers with two periods of life after a first surge of interest, which are plausibly interesting in some respect as they exhibit delayed (re-)recognition, have lower-than-average citation counts compared with other `CMIN` papers. In other words, from

papers which exhibit some impact, the ones that have two citation periods (possibly pioneers during their first period, at least with respect to the second period) are not well retributed by citation counts.

Revisiting these results in light of the debate about the appropriateness of the citation count as a impact (and frequently, by extension, quality) measure for publications, we find them contradictory. On one side the first two cases are in agreement with the role of citation count as a quality discriminator. On the other, the important role of 'rediscovered' or 'early birds' papers is not recompensed by a high citation rate. This latter point is crucial, because in these cases citation count would fail as a quality measure; the rebirth index, subsequently, would be an appropriate complementary and novel quality measure. Put differently, we show here that rebirth and impact are two distinct concepts, in that low impacts corresponds both to high and no rebirth. In this respect, the rebirth index could be used as a quality-defining criterion complementing final citation counts and able to distinguish for instance, among papers with lower-than-average final citation counts, those which are eventually rebirthing from the others.

5.3 Conclusions and proposed extensions to QTR

The results presented in Section 5.2.2 above suggest that there are dynamic temporal signatures in citation (both citing and being cited) which hold across fields and which are are in the main consistent across different time periods. This complements the findings by Radicchi and co-authors in [44], who found static universality in the citation distributions across disciplines. It was also found that articles scoring highly in ρ (the rebirthing index) and articles with a higher mean citation age tended to be (relatively) weakly cited even though they might be 'sleeping beauties'. Together with the findings in Section 5.1, this suggests that in the context of Science, there is more to consider than object degree (number of citations).

As proposed in [36], the temporal aspects of quality assessment can be easily integrated in the QTR framework by introducing a time-driven function, $D(\tau_{i\alpha})$ to extend Equation 1.1 and 1.2 (see Section 1.1):

$$Q_\alpha(t) = \frac{1}{k_\alpha^{\theta_Q}} \sum_{i=1}^{N} w_{i\alpha}[R_i - \rho_R \overline{R}] D(\tau_{i\alpha}) \tag{5.4}$$

$$R_i(t) = \frac{1}{k_i^{\theta_R}} \sum_{\alpha=1}^{M} w_{i\alpha}[Q_\alpha - \rho_Q \overline{Q}] D(\tau_{i\alpha}) + \frac{1}{f_i^{\theta_T}} \sum_{j=1}^{M} [R_j - \rho_R \overline{R}][T_{ji} - \rho_T \overline{T}] D(\tau_{ij}) \tag{5.5}$$

where

- t is the current time;
- $\tau i\alpha$ is the age of the interaction of user i and object α
- τ_{ij} is the age of the trust relationship (if a trust network is relevant) between users i and j;
- $D(t)$ is a time-dependent function, e.g. decay function $D(t) = [1 + (\frac{t}{\tau_0})^{\beta}]^{-}1$

Since $D(t)$ can be any function, the findings presented in Section 5.2.2 could easily be integrated by making D_t also dependent on the rebirth rate so that while for most papers the function would decrease with time, a 'sleeping beauty' paper would have a function that increases with time, e.g.: $D(t) = [1 + (\frac{\rho^{-}1t}{\tau_0})^{\beta}]^{-}1$ (ρ being the rebirth index).

Chapter 6

Summary and Conclusions

Section 2.2, Section 3.2 and Section **??** have sought to study the effects of integrating the empirical findings of Section 2.1, Section 3.1 and Section 4.1 into the QTR framework by using the data from the respective platforms to initialise the model. We have shown that across all platforms, certain parameterisations of QTR yield better performance than would be obtained from the unparameterised HITS model. However, the performance of each parametisation is also sensitive to the data in that a parametisation that works better for one dataset can perform very poorly with respect to another. Furthermore, there are instances in which empirical features to rank items can yield comparable or better results than the model (see Chapter 2).

In platforms with explicit social (trust) networks such as Anobii in Chapter 3, it can also be important to take into account these networks, but for this particular platform, the role of parameterisation appeared to be more significant.

While the performance of QTR was poor with the American Physical Society dataset (see 4), we believe this to be largely due to the mis-identification of features (number and contribution of authors) with respect to the quality measure (number of citations), even when the interactions were weighted by the a priori 'reputation' of the authors, or even misidentification of the quality measure itself. From the empirical findings obtained from a questionnaire survey study and analyses of data from citation networks, we suggest also considering dynamic features when assigning quality to scientific articles (see Chapter 5).

Appendix A

Appendices

A.1 Wikipedia Data Analyses

Param	WA1		WA2		WA3		WA4		WA5	
	D1	D2	D1	D2	D1	D2	D1	D2	D1	D2
0000	0.825	0.058	0.823	0.051	0.824	0.054	0.823	0.051	0.823	0.051
0001	0.770	0.028	0.766	0.022	0.768	0.025	0.766	0.022	0.766	0.022
0010	0.859	0.158	0.858	0.136	0.858	0.146	0.858	0.136	0.858	0.136
0011	0.690	-0.304	0.678	-0.379	0.684	-0.318	0.678	-0.379	0.678	-0.379
0100	0.867	0.297	0.866	0.294	0.866	0.295	0.866	0.294	0.866	0.294
0101	-0.270	0.185	-0.266	0.179	-0.268	0.182	-0.266	0.179	-0.266	0.179
0110	0.883	0.274	0.882	0.268	0.882	0.270	0.882	0.268	0.882	0.268
0111	-0.275	0.179	-0.271	0.173	-0.273	0.176	-0.271	0.173	-0.271	0.173
1000	-0.112	0.047	-0.109	0.039	-0.110	0.043	-0.109	0.039	-0.109	0.039
1001	-0.363	0.003	-0.361	-0.005	-0.362	-0.001	-0.361	-0.005	-0.361	-0.005
1010	0.165	0.170	0.178	0.184	0.172	0.179	0.178	0.184	0.178	0.184
1011	-0.379	-0.433	-0.361	-0.430	-0.370	-0.432	-0.361	-0.430	-0.361	-0.430
1100	-0.265	0.211	-0.263	0.207	-0.264	0.209	-0.263	0.207	-0.263	0.207
1101	0.272	0.115	0.272	0.108	0.272	0.111	0.272	0.108	0.272	0.108
1110	-0.257	0.182	-0.257	0.174	-0.257	0.178	-0.257	0.174	-0.257	0.174
1111	0.271	0.093	0.271	0.086	0.271	0.089	0.271	0.086	0.271	0.086

Table A.1: Correlations between end Q values and article number of revisions for the two datasets and different weight assignments

Param	WA1		WA2		WA3		WA4		WA5	
	D1	D2	D1	D2	D1	D2	D1	D2	D1 WA5	D2
0000	0.857	0.483	0.862	0.490	0.860	0.487	0.862	0.490	0.862	0.490
0001	0.872	0.478	0.877	0.486	0.875	0.483	0.877	0.486	0.877	0.486
0010	0.865	0.487	0.868	0.503	0.866	0.496	0.868	0.503	0.868	0.503
0011	0.910	0.452	0.910	0.477	0.911	0.458	0.910	0.477	0.910	0.477
0100	0.703	0.285	0.707	0.290	0.705	0.288	0.707	0.290	0.707	0.290
0101	-0.020	0.214	-0.012	0.226	-0.016	0.220	-0.012	0.226	-0.012	0.226
0110	0.739	0.350	0.744	0.358	0.742	0.354	0.744	0.358	0.744	0.358
0111	-0.026	0.248	-0.019	0.259	-0.023	0.254	-0.019	0.259	-0.019	0.259
1000	-0.095	0.315	-0.086	0.331	-0.090	0.324	-0.086	0.331	-0.086	0.331
1001	-0.232	0.347	-0.227	0.359	-0.229	0.353	-0.227	0.359	-0.227	0.359
1010	-0.036	-0.049	-0.021	-0.064	-0.028	-0.058	-0.021	-0.064	-0.021	-0.064
1011	-0.026	0.445	0.001	0.435	-0.011	0.439	0.001	0.435	0.001	0.435
1100	-0.339	-0.130	-0.334	-0.123	-0.336	-0.127	-0.334	-0.123	-0.334	-0.123
1101	0.208	-0.017	0.208	-0.005	0.208	-0.012	0.208	-0.005	0.208	-0.005
1110	-0.367	-0.073	-0.367	-0.060	-0.367	-0.067	-0.367	-0.060	-0.367	-0.060
1111	0.207	0.000	0.207	0.012	0.207	0.006	0.207	0.012	0.207	0.012

Table A.2: Correlations between end Q values and number of references in the article for the two datasets and different weight assignments

Param	WA1		WA2		WA3		WA4		WA5	
	D1	D2	D1	D2	D1	D2	D1	D2	D1 WA5	D2
0000	0.673	-0.452	0.665	-0.459	0.668	-0.455	0.665	-0.459	0.665	-0.459
0001	0.583	-0.477	0.572	-0.484	0.577	-0.480	0.572	-0.484	0.572	-0.484
0010	0.714	-0.304	0.709	-0.326	0.712	-0.316	0.709	-0.326	0.709	-0.326
0011	0.442	-0.567	0.426	-0.550	0.434	-0.573	0.426	-0.550	0.426	-0.550
0100	0.817	-0.124	0.814	-0.128	0.816	-0.126	0.814	-0.128	0.814	-0.128
0101	-0.372	-0.177	-0.371	-0.186	-0.371	-0.182	-0.371	-0.186	-0.371	-0.186
0110	0.826	-0.192	0.823	-0.199	0.825	-0.196	0.823	-0.199	0.823	-0.199
0111	-0.373	-0.208	-0.372	-0.216	-0.373	-0.212	-0.372	-0.216	-0.372	-0.216
1000	-0.261	-0.351	-0.262	-0.367	-0.261	-0.359	-0.262	-0.367	-0.262	-0.367
1001	-0.469	-0.416	-0.468	-0.428	-0.468	-0.423	-0.468	-0.428	-0.468	-0.428
1010	0.117	0.193	0.127	0.211	0.122	0.204	0.127	0.211	0.127	0.211
1011	-0.479	-0.547	-0.465	-0.535	-0.472	-0.541	-0.465	-0.535	-0.465	-0.535
1100	-0.331	0.178	-0.331	0.172	-0.331	0.175	-0.331	0.172	-0.331	0.172
1101	0.357	-0.027	0.357	-0.038	0.357	-0.033	0.357	-0.038	0.357	-0.038
1110	-0.315	0.103	-0.315	0.091	-0.316	0.097	-0.315	0.091	-0.315	0.091
1111	0.356	-0.052	0.356	-0.063	0.356	-0.058	0.356	-0.063	0.356	-0.063

Table A.3: Correlations between end Q values and number of distinct editors for the article for the two datasets and different weight assignments

A.2 Summary of reponses from quality and science questionnaire for scientists

The questionnaire contained seven questions relating to scientists' views of scientific activity and opinion. Below is a summary of the responses to each question (Rx refers to the rank of the response, e.g. R1 for usefulness indicates that the respondent thought a particular choice was the most useful):

1. Imagine an online community relevant to your research, please rank the following possible actions according to how useful they would be to other community members, starting with the most useful first:

    - write a blog post (48% R6);
    - author a review of an article/book;
    - comment on a paper (31% R2);
    - rate a paper;
    - uploading a new paper (44% R1);
    - upload news or event information (26% R5);

2. You are comparing papers on the same topic published at a similar time. The first appeared in a premier journal (Nature, Science, etc...) and received 20 citations. The second was published in a less prestigious journal. How many citations would the second paper need for it to appear to be better quality than the first?

    - 20+ citations;
    - 50+ citations;
    - 100+ citations;
    - 1000+ citations;
    - Not possible for the second to be better;
    - Number of citations is not relevant to this decision (71%).

3. From the following researchers, choose the one whom you think deserves the highest reputation for their research in a field:

    - Author of one paper which received 500 citations;
    - Author of two papers which received 200 and 300 citations, respectively;
    - Author of five papers which received 100 citations each (31%);
    - Author of ten papers which received 50 citations each;
    - Author of thirty papers which received 20 citations each;
    - I view the reputation of all of them to be at a similar level (29%).

4. You need to cite a paper in an article you are writing, but only have time to read the titles and briefly scan abstracts. Which paper would be best to cite? Please rank the following options, starting with the most preferred first:

    - By one of the most highly cited experts (29% R1, 34% R2);
    - Published recently;
    - By one of the field's pioneers (46% R1; 30% R2);
    - By a well known and respected researcher who is a friend;
    - By a researcher you met briefly at a conference (71% R6);
    - By a researcher you have know for many years (31% R5).

5. You want to learn about a new area of research and are searching for a paper to read. A bibliographic search tool returns several results; which paper would you choose to read first?

- A paper published a month ago with 10 citations (25%);
- A paper published 1 year ago with 50 citations (27%);
- A paper published 5 years ago with 200 citations;
- A paper published ten years ago with 1000 citations;
- I view all of them as equally suitable (31%).

6. Imagining researchers in your field, who would you consider the most highly respected? Rank the following characteristics, with the most respected first. A scientist in the field who...

   - Is commonly regarded as an expert (42% R1; 26% R2);
   - Is a much cited pioneer, but has only occasionally published recently (18% R1; 27% R2);
   - Has published many papers;
   - Has received many citations for publications in the field (30% R1; 23% R2);
   - Is well known, respected, and a friend (44% R5);
   - You have known for many years (62% R6).

7. Rank the following articles according to their quality, basing your judgement only on the information below. Start with the highest quality first.

   - Article with some recent citations (29% R2);
   - Article with many citations (58% R1);
   - Article with many recent Internet downloads (28% R5);
   - Article with a very large number of Internet downloads (28% R6; 26% R5);
   - Article based on work presented in many conference talks (31% R6);
   - Article based on work presented in a conference key note presentation (21% R2).

Demographic information:

1. Total number of participants:

   - Completing survey at final screen: 165
   - Quitting before final screen: 113 (valid responses still included)

2. Please select your highest level of qualification:

- Post-graduate research degree (e.g. PhD) (69%);
- Post-graduate taught degree (e.g. MSc) (23%)

3. Please select the appropriate range for your age:
   - 31-40 years (30%)
   - 21-30 years (24%)
   - 41-50 years (23%)
4. Please select your gender:
   - Male (60%)
   - Female (40%)

Bibliography

[1] The myth of delayed recognition. *Scientist*, 18(11), 2004.

[2] Silvana Aciar, Debbie Zhang, Simeon Simoff, and John Debenham. Recommender System Based on Consumer Product Reviews. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, WI '06, pages 719–723, Washington, DC, USA, 2006. IEEE Computer Society.

[3] Gediminas Adomavicius and Alexander Tuzhilin. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEE Trans. on Knowl. and Data Eng.*, 17(6):734–749, June 2005.

[4] L. M. Aiello, A. Barrat, C. Cattuto, G. Ruffo, and R. Schifanella. Link creation and profile alignment in the aNobii social network. In *SocialCom '10: Proceedings of the Second IEEE International Conference on Social Computing*, pages 249–256, Minneapolis, Minnesota, USA, August 2010.

[5] Aris Anagnostopoulos, Ravi Kumar, and Mohammad Mahdian. Influence and correlation in social networks. In *In Proc. of the 14th ACM Int. Conf. on Knowledge Discovery and Data Mining (KDD'08*, 2008.

[6] Denise Anthony, Sean Smith, and Tim Williamson. Explaining Quality in Internet Collective Goods: Zealots and Good Samaritans in the Case of Wikipedia. Electronically.

[7] Sinan Aral, Lev Muchnik, and Arun Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51):21544–21549, December 2009.

[8] Nikolay Archak, Anindya Ghose, and Panagiotis G. Ipeirotis. Show me the money!: deriving the pricing power of product features by mining consumer reviews. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '07, pages 56–65, New York, NY, USA, 2007. ACM.

[9] Justin Basilico and Thomas Hofmann. Unifying collaborative and content-based filtering. In *Proceedings of the twenty-first international conference on Machine learning*, ICML '04, pages 9+, New York, NY, USA, 2004. ACM.

[10] Gerard Beenen, Kimberly Ling, Xiaoqing Wang, Klarissa Chang, Dan Frankowski, Paul Resnick, and Robert E. Kraut. Using social psychology to motivate contributions to online communities. In *Proceedings of the 2004 ACM conference on Computer supported cooperative work*, CSCW '04, pages 212–221, New York, NY, USA, 2004. ACM.

[11] Joshua E. Blumenstock. Size matters: word count as a measure of quality on wikipedia. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 1095–1096, New York, NY, USA, 2008. ACM.

[12] Lutz Bornmann and Hans-Dieter Daniel. Universality of citation distributionsA validation of Radicchi et al.'s relative indicator cf = c/c0 at the micro level using data from chemistry. *J. Am. Soc. Inf. Sci.*, 60(8):1664–1670, August 2009.

[13] Tibor Braun, Wolfgang Glänzel, and András Schubert. On Sleeping Beauties, Princes and other tales of citation distributions . *Research Evaluation*, 19(3):195–202, September 2010.

[14] Quentin L. Burrell. Are Sleeping Beauties to be expected? *Scientometrics*, 65(3):381–389, December 2005.

[15] Claudio Castellano and Filippo Radicchi. On the fairness of using relative indicators for comparing citation performance in different disciplines. *Archivum immunologiae et therapiae experimentalis*, 57(2):85–90, 2009.

[16] C. C. Chen and C. Roth. Citation needed: The dynamics of referencing in Wikipedia. In *WikiSym '12, Linz, Austria*. ACM Digital Library, 2012.

[17] Chih-Chun Chen and C. Roth. The Role of (Non-)Conformism in Rating Platforms. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 347–353. IEEE, October 2011.

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

[19] T. Chesney. An empirical examination of Wikipedia's credibility. *First Monday*, 11(11), 2006.

[20] The PLoS Medicine Editors. The Impact Factor Game. *PLoS Med*, 3(6):e291+, June 2006.

[21] Leo Egghe. The Hirsch index and related impact measures. *Ann. Rev. Info. Sci. Tech.*, 44(1):65–114, January 2010.

[22] E. Garfield. Is citation analysis a legitimate evaluation tool? *Scientometrics*, 1(4):359–375, 1979.

[23] Jim Giles. Internet encyclopaedias go head to head. *Nature*, 438(7070):900–901, December 2005.

[24] W. Glanzel, B. Schlemmer, and B. Thijs. Better late than never? On the chance to become highly cited only beyond the standard bibliometric time horizon. 58:571–586, 2003.

[25] W. Glänzel and A. Schubert. Predictive aspects of a stochastic model for citation processes. *Information Processing & Management*, 31(1):69–80, January 1995.

[26] David Goldberg, David Nichols, Brian M. Oki, and Douglas Terry. Using collaborative filtering to weave an information tapestry. *Commun. ACM*, 35(12):61–70, December 1992.

[27] Steven A. Greenberg. How citation distortions create unfounded authority: analysis of a citation network. *BMJ*, 339(jul20_3):b2680+, January 2009.

[28] Meiqun Hu, Ee P. Lim, Aixin Sun, Hady W. Lauw, and Ba Q. Vuong. Measuring article quality in wikipedia: models and evaluation. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pages 243–252, New York, NY, USA, 2007. ACM.

[29] Nan Hu, Paul A. Pavlou, and Jennifer Zhang. Can online reviews reveal a product's true quality?: empirical findings and analytical modeling of Online word-of-mouth communication. In *Proceedings of the 7th ACM conference on Electronic commerce*, EC '06, pages 324–330, New York, NY, USA, 2006. ACM.

[30] Xiaojun Hu, Ronald Rousseau, and Jin Chen. On the definition of forward and backward citation generations. *Journal of Informetrics*, 5(1):27–36, January 2011.

[31] Sara Javanmardi and Cristina Lopes. Statistical Measure of Quality in Wikipedia. In *1st Workshop on Social Media Analytics (SOMA '10)*, July 2010.

[32] Aniket Kittur and Robert E. Kraut. Beyond Wikipedia: coordination and conflict in online production groups. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, CSCW '10, pages 215–224, New York, NY, USA, 2010. ACM.

[33] N. Korfiatis, M. Poulos, and G. Bokos. Evaluating authoritative sources using social networks: An insight from wikipedia. *Online Information Review*, 30(3):252–262, 2006.

[34] Balázs Kovács. A generalized model of relational similarity. *Social Networks*, 32(3):197–211, July 2010.

[35] David Laniado, Riccardo Tasso, Yana Volkovich, and Andreas Kaltenbrunner. When the Wikipedians Talk: Network and Tree Structure of Wikipedia Discussion Pages. 2011.

[36] H. Liao, G. Gimini, and M. Medo. Measuring quality, reputation and trust in online communities. *Physics and Society*, 2012.

[37] Michael H. MacRoberts and Barbara R. MacRoberts. Problems of citation analysis: A critical review. *J. Am. Soc. Inf. Sci.*, 40(5):342–349, September 1989.

[38] Mary McGlohon, Zach Reiter, and Natalie Glance. Star Quality: Aggregating Reviews to Rank Products and Merchants. In *International Conference on Weblogs and Social Media*, May 2010.

[39] M. McPherson and L. Smith-Lovin. *Cohesion and membership duration: Linking groups, relations and individuals in an ecology of affiliation*, volume 19. Emerald Group Publishing Limited, 2002.

[40] Prem Melville, Raymond J. Mooney, and Ramadass Nagarajan. Content-Boosted Collaborative Filtering for Improved Recommendations, 2002.

[41] Vanesa Mirzaee and Lee Iverson. Tagging: Behaviour and motivations. *Proc. Am. Soc. Info. Sci. Tech.*, 46(1):1–5, 2009.

[42] Susan M. Mudambi and David Schuff. What makes a helpful online review? a study of customer reviews on amazon.com. *MIS Q.*, 34(1):185–200, March 2010.

[43] Derek D. Price. A general theory of bibliometric and other cumulative advantage processes. *J. Am. Soc. Inf. Sci.*, 27(5):292–306, September 1976.

[44] Filippo Radicchi, Santo Fortunato, and Claudio Castellano. Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences*, 105(45):17268–17272, October 2008.

[45] Al M. Rashid, Kimberly Ling, Regina D. Tassone, Paul Resnick, Robert Kraut, and John Riedl. Motivating participation by displaying the value of contribution. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 955–958, New York, NY, USA, 2006. ACM.

[46] Camille Roth, Jiang Wu, and Sergi Lozano. Assessing impact and quality from local dynamics of citation networks. *Journal of Informetrics*, 6(1):111–120, January 2012.

[47] Cosma R. Shalizi and Andrew C. Thomas. Homophily and Contagion Are Generically Confounded in Observational Social Network Studies. *Sociological Methods and Research*, 40:211–239, November 2010.

[48] Besiki Stvilia, Michael B. Twidale, Linda C. Smith, and Les Gasser. Information quality work organization in Wikipedia. In *Journal of the American Society for Information Science and Technology*, 2008.

[49] A. F. van Raan. Untitled. *Scientometrics*, pages 467–472, 2004.

[50] Fernanda B. Viégas, Martin Wattenberg, and Kushal Dave. Studying cooperation and conflict between authors with ¡i¿history flow¡/i¿ visualizations. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '04, pages 575–582, New York, NY, USA, 2004. ACM.

[51] Dennis M. Wilkinson and Bernardo A. Huberman. Cooperation and quality in wikipedia. In *Proceedings of the 2007 international symposium on Wikis*, WikiSym '07, pages 157–164, New York, NY, USA, 2007. ACM.

[52] Michel Zitt, Suzy Ramanana-Rahary, and Elise Bassecoulard. Relativity of citation performance and excellence measures: From cross-field to cross-scale effects of field-normalisation. *Scientometrics*, pages 373–401, April 2005.