Smart Recommendation Systems: Supporting Innovation and Media Communities
Project no. 231200

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

Deliverable D.1.4.2
Smart Recommendation Systems

Submission date: 2012-08-31

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

Organisation name of lead contractor for this deliverable: UniS

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
Corinna Elsenbroich	UniS	c.elsenbroich@surrey.ac.uk

1.2 Other contributors

Name	Organisation	E-mail
Maria Xenitidou	UniS	m.xenitidou@surrey.ac.uk
Alistair Gill	UniS	a.gill@surrey.ac.uk
Matus Medo	Zurich	matus.medo@unifr.ch
Camille Roth	CNRS	camille.roth@polytechnique.edu
Nigel Gilbert	UniS	n.gilbert@surrey.ac.uk

1.3 Document history

Version#	Date	Change
V0.1	29 October 2012	Original
V0.2	3 January 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

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

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

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

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

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

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

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

Executive summary

This report is concerned with the interplay of conceptions of quality in science, processes supporting or undermining quality and how the lives of scientists can be supported by recommendation systems for quality in science. It:

1. Explicates why the notion of quality in science is important (Section 1).

   The failure of the programme to put down formal criteria for quality in science (logical positivism) suggests, science needs to be seen as a social endeavour. This does not mean that it has to give up all special epistemic status but this status has to be justified by distinguishing science from other social processes. Different conceptions are discussed and their relation to quality explored.

2. Discusses two models of quality in science that influence the day to day work of scientists and compares them (Sections 2, 3 and 4).

   Different kinds of quality in science are distinguished. Two models of scientific quality are discussed, the Ideal Science model following a set of norms and the Metric model applying a set of metrics. Smart Recommendation Systems need to be clear what ’kind of quality’ they want to support.

3. Presents recommendation systems to support the production of quality in science (Section 5).

   Three existing recommendation systems are discussed.

The report concludes by discussing the use of quantitative measures for quality, what scientists see as important aspects of scientific work and the potential for different measures to base recommendation systems on.

Contents

Chapter 1

Philosophy of Science versus Reality of Science

In 2010 the *Lancet*, a high profile medical journal, withdrew a paper published in 1998 that linked the MMR[1] vaccine to autism and bowel disease. In the 12 years since the original paper, many studies were conducted but none replicated the findings. The paper appears to report falsehoods rather than truth. This on its own, however, is not a reason for the withdrawal of a paper from the scientific literature. In fact falsifying scientific findings is the stuff science is made of (e.g. [43]).The problem with the paper was a) Professor Wakefield, the researcher, was receiving money from solicitors of parents who were suing in the US for their children's autism being caused by the MMR vaccine, b) dishonesty in reporting findings and c) using ethically non-approved procedures on children in the study. So 'false' science can remain in the public sphere as long as it is 'good quality' false science, i.e. it conforms to scientific quality standards. But what are these standards? Are they unique and unchanging? Are they really able to separate the wheat from the chaff? And can they be measured and if they can, are current metrics appropriate?

This chapter looks at the notion of quality in science in general. We all want good quality, eat good food, have a good education and travel on a good public transport system. But what does good quality mean? For example, good quality food could be food that is very healthy, high in fibre, low in saturated fats, etc. but equally, good food could be an amazing French five-course dinner, with meat lathered in cream sauce and not a fibre in sight. What is good quality food depends on the purpose of the food, e.g. to be healthy or to be tasty (not that they are necessarily exclusive although they often

[1] Mumps, Measles and Rubella combined in one shot.

seem to be). In order to understand quality in science we might need to understand something about the function of science.

The Oxford Dictionary describes science as "the intellectual and practical activity encompassing the systematic study of the structure and behaviour of the physical and natural world through observation and experiment." This definition (and other standard definitions of science) have three components:

1. science is about an *external world*, here the 'physical and natural world'
2. science is *systematic*
3. science has certain *methods* (e.g. observation, experimentation)

So the function of science is to find out about the world. If one was that way inclined one could say that the function of science is finding *truth or truths*. Truth has become an uneasy companion in discussions of science, however. Although it has been clear for a long time that scientific truths become falsehoods and are superseded by 'new truths', e.g. Newton's theory of gravity being replaced by Einstein's theory of relativity, we are used to thinking that it is simply a matter of finding out more and findings being more precise and accurate. Science is generally seen as a cumulative endeavour.

The question of what is good science is much younger than one might suspect. Prior roughly to the 16th Century, science was simply describing/explaining nature as done for example by ancient Greek philosophers such as Ptolomy and Aristotle. There was disagreement between thinkers but little effort to resolve it. Different accounts coexisted as different schools with little attempt at unification. Then came a methodological revolution showing that description is not enough but that instead the 'phenomena behind the phenomena' need analysis.

Bacon advocated that 'real science' needs an empirical basis, advocating observation and experimentation rather than 'fanciful rationalisation' in the *Novum Organon*. For him, science was an empirical endeavour, practical in purpose and had to follow stringent methods.

At about the same time, Copernicus' *On the Revolutions of the Celestial Spheres* describing a solar system, in which the sun replaces the earth as the central body in the solar system; Galileo's *Discourses and Mathematical Demonstrations Relating to Two New Sciences* analysing the motion of bodies and Newton's *Mathematical Principles of Natural Philosophy* are cases where simple intuition and description no longer suffice, using mathematics and thought experimentation (in addition to observation) to explain the phenomena.

Whereas before there was no difference between philosophy and science, as we can also see from Newton's *Principia* still referring to natural philosophy, these changes led to a split of philosophy ('fanciful rationalisation') and science. Science was now a methodologically well defined, separate activity, using mathematisation and empiricism. Following a certain methodology was seen as the path to truth. This dream of a methodology for truth was destroyed by major shifts in science and mathematics in the beginning of the 20th century leading to questions about the method of science and its access to truth. Philosophy of science was born.

1.1 Verification and the Loss of Certainty

Philosophy of science has been a serious endeavour since roughly the beginning of the 20th century. Through the discovery of paradoxes in mathematics (e.g. set theory) and radical theory change in physics (e.g. relativity, quantum mechanics) philosophers became concerned with the certainty and truth of science. The Vienna Circle [28] (a group of mathematicians, physicists and philosophers) took up the programme to provide a formal system for the scientific method to guarantee certainty and truth, a movement often called "logical positivism". The hallmarks of the movement were still in line with science as a) empirical and b) mathematical or logical and the purpose of the movement was to provide formalisations to stringently define the scientific method. Their endeavour of finding certainty and truth was undermined almost immediately by competing positions. Among them was for example Popper's [43] criticism that statements can never be verified and that what we should do instead is to try and falsify scientific hypotheses. However, falsification is not a simple method either, with similar issues of under-determination of theory by data as besets verification (Duhem-Quine-Thesis) ([14]). If an experiment fails it does not necessarily mean that the theory is false. It might be that something in the experiment was wrong instead. Many other problems were found with the formalisation programme of science such as Putnam's attack on observation/theory distinction [45] or Quine's attack on the analytic/synthetic distinction [46]. As Ayer put it in an interview in the 1970s: "I suppose the most important [defect of logical positivism]. . . was that nearly all of it was false." [28].

Then came a publication that changed not only the game but the whole ballpark.

Thomas S. Kuhn's *Structure of Scientific Revolution* [32] did not only throw doubt on the endeavour of finding criteria for certainty and truth but on the whole endeavour of science making any progress towards certainty and truth. By analysing scientific theory changes, Kuhn showed that, rather than moving further and further towards one eternal truth, theory changes in science were so radical that theories could not be seen as cumulative. Scientific theories do not just become more precise and accurate and contain more knowledge. There are conceptual shifts so severe that the theories bear no relation to each other. This sits uneasy with the notion that science is about truth. There is only one truth so there should not be many radically different conceptualisations of it. This means the endeavour of approaching an eternal truth using the scientific method becomes dubious.

Three general developments ensued from this publication.

1. SAVING CERTAINTY: Finding ways of saving the scientific endeavour by providing criteria by which knowledge can still be rendered cumulative. This group contains varying forms of realism trying to save science's continuity by finding criteria that allow even radical theory changes to retain cumulative knowledge. For example 'structural realism' argues that scientific theories represent structures and these structures remain continuous even if the representations do not ([33],[59]). Some philosophers of science claim that science finds 'approximations to the truth' rather than truth, has 'truthlikeness' rather than 'truth content' or, as Psillos puts it, how 'science tracks truth' [44]. Good science is still science that gets to the truth.

2. ANTI-REALISM: Science has no justified claim to truth and certainty and essentially cannot be distinguished from any other human activity. This group contains a range of positions of anti-realism, stripping science of its special epistemic status to arrive at truth. There is no such thing as good or bad science; in the most extreme form, there is no such thing as science at all as it cannot be distinguished from other human endeavours such as story telling [16], leading to post-modernism, post-structuralism and social contructivism [8]. More moderate versions of antirealism claim that science is not about truth at all but simply a pragmatic way of dealing with the world (e.g. [14], [55]).

3. SCIENCE AS A SOCIAL PROCESS (SOCIOLOGY OF SCIENCE): Science is a social process with human scientists that make mistakes, are not objective, defend their

> own interests etc. This response is to suspend judgement on the matter of science's ability to arrive at truth and turn from an analysis of science in the abstract to science as a social phenomenon with actors, processes and norms. For some, the social influences on scientific knowledge are only to blame for those scientific endeavours that went wrong [34]. Some see the analysis of the social aspects as largely undermining the epistemic status of scientific knowledge, (e.g. [36], [10]). Very radical sociology of science does not separate the social processes from the epistemic questions at all. The 'Strong Programme', originating with David Bloor and Barry Barnes (the 'Edinburgh School') and Harry Collins (the 'Bath School'), considers all knowledge as resulting purely from social processes (cf. [6], [4] and [12]). There is no special epistemic status of scientific knowledge. Moderate positions in the sociology of science look with interest at the social processes and acknowledge the importance of those processes for scientific knowledge but leave epistemic questions of science and truth separate, thus not denying the possibility of science to find truth. Science's ability to find the truth, however, can no longer be evaluated by external criteria about science but only via the production of science. The quality of science is no longer assessed via the truth of scientific statements but in how far science follows quality criteria within the social process itself.

In this report we focus on quality in science from the social process perspective. We consider scientists' conceptions of science, what they value and how they make choices regarding quality. We embed this discussion into two models of quality in science, the ideal model deriving from a normative conception of science and the metric model deriving from administrative management of science. In our data on scientist's conceptions of quality and their negotiations of the two models in their daily lives, it will become clear that these two models, rather than supporting each other, create dissonance for scientists. We finish by discussing recommendation systems for quality in science to help scientists deal with an ever increasing amount of scientific production.

Chapter 2

Ideal Science

Whatever the right answer to the philosophical questions of science, science has continued to exist and scientific findings are still taken seriously, even if often with more skepticism than used to be the case. What has become clear, however, is that science is a social endeavour conducted by scientists. Much of this social study of science has resulted in some form of relativism, cf. [37], [8]. If science is simply another social endeavour, its output cannot possibly have special epistemic status. But is this a necessary conclusion? Could science not be a social endeavour with special epistemic status? A social endeavour to find the truth?

This brings us back to quality. Good food was defined via its function, i.e. to be healthy or to be tasty. We can conduct research to find out which foods are healthy (have a long term health effect) and we can almost instantaneously say whether food is tasty. Science's function is to find truth but we seem not to have any external assessment to judge whether a scientific finding *is* the truth since the failure of the formalisation exercise of logical positivism. Formal criteria guaranteeing truth and certainty cannot be found. Scientists are humans and science is a human endeavour. It is influenced by social processes. Also it seems untenable for social processes only to account for failed research programmes, e.g. [33]. Proper, good science is also governed by social influence. So, do we necessarily have to subscribe to scientific relativism or anarchy as [16] or [36] have us believe?

Although there are no (formal) criteria to assess whether a scientific finding is true we might be able to use notions of "good" and "bad" (or simply "quality") as a proxy for truth. Good quality science might increase the chance of the findings being true with bad quality decreasing this chance.

Let us look at some positions in the sociology of science that, although not advocating that science finds the truth, neither exclude the possibility of truth as an outcome of science.

2.1 Gilbert's Model Based View of Science

Gilbert [21] argues that the idea of paradigms in science put forward by Kuhn [32] is too strong and that instead the idea of 'sets of models' might be employed to characterise scientific research. He argues that science advances by models and that models are justified by their similarity and their fit to other, accepted models. Each scientist modifies existing models and thus produces new models, appropriate for his or her specific needs. Although akin to Kuhn's paradigms of normal science the idea of models is different in that models are a loosely related family distributed over the members of a research network, rather than the one fixed paradigm of a research community.

This view accounts much better for the actuality of science in which disagreement does not necessarily lead to revolutions, because a research community can cope with keeping competing models (for a while at least). The model, rather than the unified paradigm view, accounts for the possibility of disagreement within a field, the import of models from other sciences (e.g. exchange theory as an import of economic theory into sociology) and the actually observed developments of fields which undergo changes in foci without a revolution or major shift.

Despite the social aspect of the analysis, i.e. model construction informed by the needs of the researcher, advancement of science towards truth is at least not excluded from the analysis. The practice of science relies on the assumption that there is a truth out there. Procedures are put in place to support science as the search for truth, constructing the scientist as the "messenger relaying the truth from Nature" [21, p. 285].

The quality of research is evaluated according to its potential access to truth. To preserve the link of a piece of research with truth, research must be shown to be anonymous, it is 'anyone's research'. Rhetorically this is expressed by research findings being written in the passive and motives, in particular personal motives of the researchers, being left out. The anonymity of the researcher relates to the universalism norm (see Section 2.2). Truth is determined by the constitution of Nature and anyone who knows the proper procedures can learn the truth. The 'proper procedures' are a set of methods agreed by scientists in a specific field and taught to the next generation of scientists.

Furthermore, the citation of other knowledge claims constitutes an important part of the scientific endeavour. Citation, according to Gilbert, has two functions. First of all, past, accepted knowledge, methods and data are used to justify present knowledge claims, methods and data. Secondly, it shows that the author values the cited research as a contribution to knowledge. The scientific community can reject a new knowledge claim on both counts, either rejecting the cited knowledge completely or rejecting that it justifies the current claims. This is similar to the rejection of any argument where either the premises or the inference from the premises to the conclusion need to be challenged.

Finally, the knowledge claims need to be evaluated by the research community. This is related to the organised scepticism norm (see Section 2.2). Often peer review is the first instance of knowledge evaluation but often peer review might well only guarantee a minimal standard with citation as the real test of a knowledge claim.[1]

Although Gilbert [21] suspends judgement of whether science is a way to get to the truth, he sees the practice of science as behaving as if this was the goal. Scientific findings are judged according to an agreed set of methods, independent of who conducted the research, and scrutinised and extended by the community. By constructing models scientists construct representations of an external world with the overlap guaranteeing at least intersubjectivity if not objectivity.

There are several current processes supporting this model, for example institutionalised scholarship and teaching. The next generation of scientists is taught the methods, theories and models of the current generation and can build on this past knowledge. Also citation as an essential part of a scientific paper supports this model as it enables the reader to position knowledge claims within a context of other knowledge claims. Together with citation, peer review ensures that models are well embedded in knowledge but also contribute something new to the scientific endeavour.

2.2 Merton's Norms of Science

Another important contribution to moderate sociology of science is Robert K. Merton's influential work on the interactions of science and culture [41]. Merton was interested in the sociology of scientists rather than epistemic questions concerning science. He proposed a set of norms of science which distinguish it from other social endeavours:

1. COMMUNALISM: Scientists openly share findings with colleagues. Scientific discoveries are owned by everyone — or as [21] puts it, it is 'anyone's science'.

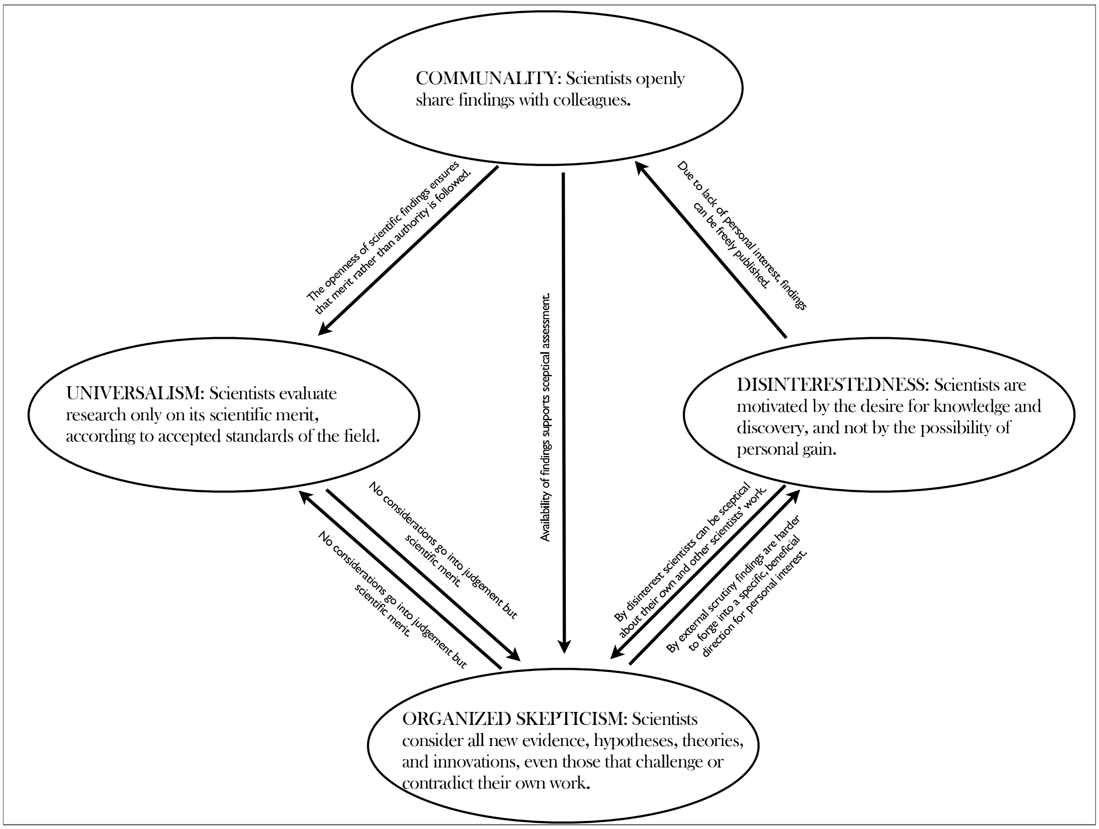
[1] For more research on peer-review processes see [51], [52] and for citation see [56].

2. UNIVERSALISM: Scientists evaluate research only on its merit, i.e., according to accepted standards of the field. No considerations onf class, race, sex or status go into the evaluation of a piece of research.

3. DISINTERESTEDNESS: Scientists are motivated by the desire for knowledge and discovery, and not by the possibility of personal gain.

4. ORGANIZED SKEPTICISM: Scientists consider all new evidence, hypotheses, theories, and innovations, even those that challenge or contradict their own work.[2]

These norms present a set of criteria that demarcate the social endeavour of science from other social endeavours. A piece of research violating any one of these norms should not be deemed science (for example the case of the MMR research withdrawn from the *Lancet* due to a self-interest of the researcher for a particular outcome of the study; cf. [22], [42]).


COMMUNALITY: Scientists openly share findings with colleagues.
UNIVERSALISM: Scientists evaluate research only on its scientific merit, according to accepted standards of the field.
DISINTERESTEDNESS: Scientists are motivated by the desire for knowledge and discovery, and not by the possibility of personal gain.
ORGANIZED SKEPTICISM: Scientists consider all new evidence, hypotheses, theories, and innovations, even those that challenge or contradict their own work.
The openness of scientific findings ensures that merit rather than authority is followed.
Availability of findings supports sceptical assessment.
Due to lack of personal interest, findings can be freely published.
No considerations go into judgement but scientific merit.
No considerations go into judgement but scientific merit.
By disinterest scientists can be sceptical about their own and other scientists' work.
By external scrutiny findings are harder to forge into a specific, beneficial direction for personal interest.

Monday, 29 October 12

Figure 2.1: The support network between Mertonian norms of science.

The norms mutually support each other and are also supported by processes within the science community and institutionalised science. For example, for most purposes

[2] This list is taken from [1].

scientific work is peer reviewed, i.e. assessed by other scientists whether it is worthy to be taken into the body of knowledge. This process supports organised scepticism as well as communality of scientific research. The peer review process is usually 'doubly blind', both the reviewers and the authors remain anonymous. The anonymity of the authors supports the norm of universality as the research will be accepted purely on its scientific merit. The anonymity of the reviewers supports the norm of disinterestedness, e.g. it undermines the possibility that reviewers do someone a favour.

Recent research has also shown the importance of disinterestedness for the peer review process as such. Squazzoni et al [51] ask the question whether peer review should be remunerated in some way (e.g. monetary reward). They perform experiments mimicking the processes of peer review and find that any external reward for peer review would actually be detrimental to the quality. This finding is mostly explained by the 'crowding out effect' of external rewards, cf. [18], making peer review no longer a norm driven but a gain driven activity. Peer review is used for many scientific processes such as publications and funding. Research funding is particularly sensitive to the norms of disinterestedness. Whereas public funding is usually seen as 'safe', more and more research is funded by private companies leading to potential conflict of interest, e.g. suppression of results contrary to the company's interest.

Merton's norms of science point towards a way in which the scientific endeavour, although social in nature, might still preserve its claim to be a social endeavour of a special epistemic nature. Note, as they are norms, they are the ideals of how science is conducted, not necessarily the actuality. Let us look at the norms in turn, how they might relate to quality and truth, and some problems and processes that undermine these norms.

COMMUNALISM: The connection between communalism and truth may best be expressed in Newton's "standing on the shoulders of giants" quotation, which acknowledges that scientific truth is not one person's truth but a communal effort which one person alone could not possible achieve all on his own. As we have seen in the above diagram, communalism is also essential as a structural support for the norm of organised scepticism, thus indirectly contributing to the truth via that norm. The main driver of communalism is the accessibility of research findings by other scientists.

Publication of research findings is one of the hallmarks of science. The main scientific publication outlets are books and scientific journals. Ideally, all research should be freely available to anyone interested. However, science must be funded somehow, the publication outlets need to pay editors and universities need to pay their research staff. In the case of universities there are different funding sources such as student fees, public funding, research project funding etc. For publishing houses, the only source of income is paid access to their publications. This payment restricts access, meaning communality of output is potentially undermined.

In the beginning of 2012 this was a hot debate. Research findings should be publicly available. At present, many research findings are published in proprietorial journals and access to articles is expensive at commonly around 30 US Dollars for access to one article for 24 hours. University libraries have subscriptions to journals through which academics inside institutions can have access but a) subscription might be patchy and b) people outside academia have no access or have to pay a lot of money.

Another increasing problem, resulting from the continuous reduction of public funding of research, is corporate funding of scientific research. Although possibly desirable that science keeps a link to industry and implementation, the privatisation of intellectual property undermines the communalism of the scientific endeavour.

UNIVSERSALISM: A conception of truth as "anyone's truth", with the scientist only being the messenger, is important in science. This is a direct example of the norm of universalism. As the scientist is only the messenger, it should not matter who brings the message of truth, only that it was arrived at in "proper ways".

In 1633 Galileo was forced to retract his findings, arrived at by the use of a telescope, corroborating the Copernican theory that the sun, rather than the earth, was the centre of the universe. Established religion did not condone such heresy. Sometimes the debate on intelligent design versus evolution is construed in similar terms. After the defeat of efforts to classify creationism as a science taught in public schools in the US, a new 'scientific version' of creationism entered the stage, *intelligent design*.[3] Although those seen to adhere to intelligent design are not usually part of the scientific community, the theory of intelligent design is now taught in parallel or as a criticism to evolutionary theory in state schools in some states of the US (e.g. Kentucky, Kansas and Ohio).

Many scientists see this as a misuse of the universalism paradigm. Intelligent design promoters argued that, in a reversal of the Galileo case, the link of intelligent

[3] See the rulings of the Supreme Court of the United States in 1987 Edwards v. Aguillard and in 2005 Kitzmiller et al. v Dover Area School District (see for example http://ncse.com/rncse/26/1-2/design-trial (accessed 06/01/2013)).

design to religion was unjustly used to discredit the theory and should be accepted as scientific knowledge under the principle of universalism.

Initially universalism was not targeted at the problem of the religious origin of research findings, however, but more traditional forms of discrimination. Violations to universalism on grounds of class, gender and race are hopefully a thing of the past. This kind of discrimination is unimaginable in today's science. Institutionally it is prevented (at least to some extent) by peer review of scientific work. However, there are two processes possibly undermining universalism that need to be kept in mind. The first one is the positive discrimination of scientific results coming from 'the scientific establishment'. The second one is the underrepresentation of certain groups in the scientific endeavour.

1. Positive Discrimination: In the introduction we saw the story of the research on MMR and Autism published in the *Lancet*. The paper was published in 1998 and it took until 2010 to withdraw the article. It might be argued that the high calibre of the *Lancet* provided kudos to Wakefield as belonging to the scientific establishment leading to less scrutiny of the findings.

2. Underrepresentation of Certain Groups: Although a scientific finding is not excluded from the literature because of the origin of the researcher, some sections of society are strongly underrepresented in science.

Feminist philosophy of science hones in on this point. Their claim is that science is as it is because it has been conceived and is executed mainly by men. This is a strong epistemic claim of the situatedness of the knower and situatedness of knowledge.

Even without subscribing to such a strong relativist standpoint a problem remains. Even if knowledge is not *per se* situated, what research focusses on might well be. For example for decades, criminology focussed on crimes that were almost exclusively executed by people of lower class origin (burglary, robbery, larceny etc.). For a long time, white collar crime was not on the radar of criminologists. One reason for this is that blue collar crime is more visible than the white collar variety. Another reason, however, could well be that the scientific establishment mainly consisted of middle class scholars preferring to shine the spotlight on the 'others'.

Opening up education has made access to scholarship much easier for underrepresented groups. However, problems of class (much more than gender, and probably more than ethnicity or race) still persist, denying certain groups access to the scientific establishment. This might have consequences in terms of research focus, away from interests of disadvantaged groups but also for the construction of those interests and the operationalisations used in research. For example, is an operationalisation of 'wellbeing' using 'gym membership' applicable to a member of a disadvantaged group? Are conceptions of wellbeing comparable at all? The possible consequences of class underrepresentation need to be considered.

DISINTERESTEDNESS: Many fraud cases in science result from a conflict of interest. For example Wakefield's research on MMR and autism was first condemned because of a conflict of interest when it was found out that he received payments from solicitors acting for parents who believed their children had been harmed by MMR vaccinations. His studies were later discredited due to lack of ethical approval and his misrepresentation of the children's illnesses. None of the research subsequently investigating a connection between MMR and autism and other illnesses could find any significant association.

In a study on fraud in science, reported in [15], it was found that the readiness to falsify research findings, in particular by young researchers (post-docs), was very high although the actual rate was not much higher than can be assumed for science in general.

> "In a sample of postdoctoral fellows at the University of California San Francisco, USA, only 3.4% said they had modified data in the past, but 17% said they were "willing to select or omit data to improve their results" [. . .]. Among research trainees in biomedical sciences at the University of California San Diego, 4.9% said they had modified research results in the past, but 81% were "willing to select, omit or fabricate data to win a grant or publish a paper" [. . .].' [15, p. 9]

However the study also found that,

> "Once methodological differences were controlled for, cross-study comparisons indicated that samples drawn exclusively from medical (including clinical and pharmacological) research reported misconduct more frequently than respondents in other fields or in mixed samples. To the author's knowledge, this is the first cross- disciplinary evidence of this kind,

> and it suggests that misconduct in clinical, pharmacological and medical research is more widespread than in other fields. This would support growing fears that the large financial interests that often drive medical research are severely biasing it [. . .]. However, as all survey-based data, this finding is open to the alternative interpretation that respondents in the medical profession are simply more aware of the problem and more willing to report it. This could indeed be the case, because medical research is a preferred target of research and training programs in scientific integrity, and because the severe social and legal consequences of misconduct in medical research might motivate respondents to report it. However, the effect of this parameter was not robust to one of the sensitivity analyses, so it would need to be confirmed by independent studies before being conclusively accepted.' [15, p.10]

The focus on (the number of) publications and the dependency of scientific careers on speedy output might well undermine the norm of disinterestedness. Although this is not a direct conflict of interest as in the case of Wakefield and the MMR trials, the scientists' personal interests in maintaining jobs or advancing careers might also impact on the quality of the scientific work produced and potentially undermine reporting of truth.

Another problem is the need for funding, in particular funding from political, private sector or corporate sources. We saw above that the privatisation of intellectual property potentially undermines the communalism of science. It also potentially undermines the disinterestedness. First of all, findings might be overemphasised if they are in the corporate, political or private institution's interest. They might also be suppressed if they undermine the intended message. For example, what happens if the findings of a commissioned piece of research are actually damaging to the funding institution? Should the institution be allowed to stop the publication of results after funding a piece of academic research? What if the findings are published nonetheless? It is unlikely that the academics will ever receive funding from the institution again and might even have difficulties with other, related, institutions if 'word get's round'. In order to not undermine the possibility of obtaining future funding, the academics might decide to not report the findings. Thus, private funding might well lead to conflicts regarding the reporting of truth.

ORGANISED SCEPTICISM: Organised scepticism can be seen as a consensus element of truth in science (e.g. [27]). Through rational debate and collective scrutiny, falsehoods will be uncovered and the movement of science towards truth guaranteed. This norm also supports the other norms. Only if scientific knowledge is communal can it be sceptically assessed and this assessment guarantees that violations to universality are discovered and personal interests of particular scientists undermined.

Processes like peer-review for conference or journal submissions and research grants are examples of organised scepticism. Papers are scrutinised by other scientists in the field. Importantly, the reviewers have no vested interest in publication (see [51] for a discussion of remuneration of peer reviewers).

There are some potentially negative outcomes of the process of peer-review. Ground breaking research might not be published if other scientists do not understand it or stick to a current paradigm. Also, findings or views disagreeing with a reviewer's own research might be unduly rated negatively (which is why items are usually reviewed by at least two reviewers). Another problem results from specific journal rejection policies. For example the intuitively sensible policy of only accepting work that has been endorsed by all reviewers might lead to excellent research not being published because of one reviewer's problem with it.

2.3 Summary

We have discussed two models of science in which science is analysed as a social endeavour without eradicating the potential of science having special epistemic status. The first one saw science as an endeavour in which scientists construct models. The scientist is the messenger of truth via these models and the families of models, whether they support or contradict each other, evolve through their communication and thus constitute a scientific discipline.

The second model is a normative account of science in which scientific knowledge is defined as knowledge obtained adhering to a set of norms. The norms are derived empirically from studying scientists. Although the norms governing ideal science support quality they can be seen as a minimum requirement of a piece of research. We have discussed their relationship to truth, some processes in place to support these norms and some problems associated with them. Let us next look at another framework or approach to quality in science, the idea of measuring quality of scientific output.

Chapter 3

Measured Science

In the ideal model of science we have a set of norms scientists need to observe. A breach of any of the norms will produce bad science (findings accepted on authority rather than merit) or something that cannot be deemed science at all (self-interested 'research' findings). But how can we make sure the norms have been adhered to? It is the purpose of the 'governance' of the scientific profession to ensure its output follows the norms. Governance has fallen from grace in the last decades and has largely been replaced by administration for many professions (see how [1] discusses governance (vs administration) as an addition to Mertonian norms), most notably education and health [23]. Also science has undergone an administrative turn since the 1980s posing the question of measuring quality in science.

3.1 Assessing and Measuring Academic work

The question of measuring quality in science has become more and more important since the 1980s. The UK is currently the most measured academic system but others, for example European countries, are following in its administrative footsteps. There are several processes underlying this increased importance, some due to a general culture, some specific to academia:

1. A general rise in auditing culture in the public sector.

2. The need to increase accountability and transparency for public funding.

3. An increased distrust of the professions to regulate themselves.

4. An increased distrust of science in general and academia in particular.

5. The feeling that academia had become self-serving and in an 'ivory tower'.

3.1.1 Research Assessment and Excellence

UK academia is one of the most managed academic systems in the world [25]. Given this strong reliance on measuring scientific outcomes, we use the UK system as a case study for discussing the metric model of scientific quality.

In the UK, scientific institutions, such as universities, used to be funded by public money, 'block grants'. The allocation of block grants was opaque. In 1985 the first Research Selectivity Exercise (RSE) was supposed to make the allocation more transparent by measuring the quality of research in institutions and funding accordingly. However, the RSE itself and the measures employed were so opaque themselves, it hardly helped transparency [9]. The RSE became the Research Assessment Exercise (RAE) but all that changed was that the measures and scales were refined but the assessments were essentially the same.

Academic quality assessments are based mainly on publications and publication measures, contextualised slightly by statistics on the general research environment. Although non-publication of research also violates the ideal science norms of communality and organised scepticism, Wells [57] argues that publishing is not research and might not even be a good proxy for research or research quality. Publications are just one particular way of presenting research. Lawrence [38] points out that traditionally papers were published to communicate scientific findings. For example, publications in *Nature* were traditionally short and written in an easily accessible language, understandable for the general reader. Longer, more detailed papers about the findings would later be submitted to more specialised journals. Now, however, papers in *Nature* are no longer understandable as their purpose is not to communicate research findings but simply to be accepted for publication in *Nature*.

> "Once you start counting papers, scoring journals and measuring impact then the purposes of publication changes" [38]

Burrows [9] discusses the succession of assessment exercises and argues that the early ones, up to 1996, were merely trying to provide some transparency for funding allocation but that sometime between the 1996 and 2001 exercise, this had changed from being an assessment to feeding back into the research culture. As Peter Scott from the Institute of Education puts it in the *Guardian*, the first assessment lead to weeding out 'those self-regarding academics whose great monograph always seemed to be just out of touch at the end of the rainbow' but since then led to game playing rather than research improvement (Guardian, 5 March 2012).

But let us accept publication as a legitimate proxy for research. Even then the specific implementation of the RAE/REF have some perverse consequences. Although the RAE/REF guidelines do not mention any specific kind of publication as counting more than another, it is tacit knowledge that the gold standard is four journal publications.[1] In some disciplines there is an even further narrowing of publication standards to a small list of journals (those with a high Impact Factor (see discussion of this metric below)) and coming from within a discipline. This potentially has a host of negative consequences. The most obvious is the focus on journal publications leading to possible neglect of other outlets such as books (which often have more impact in terms of citations), conference presentations (which often contain more innovative work), government reports and policy advice (which can be seen as having wider impact). There are also more subtle consequences such as research tailored to journals which can be seen putting the cart before the horse, rather than focussing on the goal of communicating research finding. Also, if publications are chosen from a narrow list of journals, measured quality might decrease without quality necessarily decreasing. Given the fixed number of journals and an increasing number of publication submissions, in particular as more and more countries implement similar measures on the same list of 'top journals', it is likely that each UK department gets a smaller slice of the cake, thus falling in measured quality without any loss in actual quality.

Possibly the worst side effect comes from the submission policy of collaborative papers. Whereas a co-authored paper by authors from different institutions can be counted by both authors, if the co-authors are in the same department it can, generally, only be counted for one of them. This will certainly reduce within department collaboration, which undermines the collegiate spirit of scholarship.

Publication and Citation Measures

The most important and pervasive measures in academic life nowadays are publication and citation indexes. Citation can be seen as a proxy for the impact of a paper and, in turn, a researcher or a journal. There are two major metrics, the h-index relating to researchers and the Impact Factor relating to journals.

The h-index is a measure for both the productivity and impact of a researcher or

[1] Note that this focus is particularly strong in the natural sciences and engineering but by no means exclusive to those disciplines.

research group. It is calculated by ranking the papers of a researcher according to the number of citations and h is then the number of papers with the number of citations $\geq h$ [29]. Hirsch states that it is the best single measure of a person's scientific achievement and should be used to compare researchers on the same level of seniority. For comparing researchers at different stages the index is adapted to reflect the career stage by dividing the h-index by the number m, the number of years since the researcher's first publication. The h-index ha some positive features, such as taking into account the potentially long tail distribution of citations. It is however, flawed as a simple measure to compare scientists. Too much information is lost in reducing a scientist's achievement to the number h. In [53] intuitive examples are given to show that the h-index is far from useful for the comparison of researchers' achievements (p. 11). One problem is that all publications are discarded that have citations below h, meaning a scientist with 10 papers, each with 10 or more citations, has the same index value as a scientist with 100 papers, 10 of which have 10 or more citations. The other problem is that the number of citations below h is also discarded, meaning a scientist with 10 papers, one of which has 10 citations and the other 9 having 100, also has the same index value as the 10 paper, 10 citation scientist.

The h-index is heavily dependent on the database used. For example Thomson-Reuters' Web of Knowledge and Elsevier's Scopus give lower scores than Google Scholar as the former only cover selected journals (cf. [3]), Google Scholar in turn has the problem that on the web, author's names are not unique identifiers leading to false attributions of publications and citations.

These shortcomings are particularly problematic if the h-index is used as a single number to assess a scientist and many important decisions are based on it, like recruitment, funding, and career advancement (see [9], [53]).

There are also more general questions to be answered as to whether citation is an appropriate measure for the quality of a publication. To think so makes several assumptions about the nature of citation as a positive appreciation of the content of a paper. But there are different motives for citing, cf. [13]. A paper might get a host of citations for a mistake and we surely do not want mistakes to come top of our rankings. A paper might also simply be cited in disagreement or warning, a negative citation. Again, just counting citations is not sensitive to this use. Different kinds of papers get different numbers of citations. Very general and review articles get more citations than articles reporting new or very specific research findings without the former being of any better quality than the latter, just aimed at a broader set of readers. Thus, citations are neither an indication that the citer has thoroughly examined (or even read) the cited paper, nor that the outcome of this examination was positive (c.f. 'negative' or 'warning' citations). There are also some strategies for citation hunting, one of the games now played in the academic world, with researchers ensuring that these are maximised. For example an article entitled "A short history of SHELX", containing the following sentence:

> "This paper could serve as a general literature citation when one or more of the open-source SHELX programs (and the Bruker AXS version SHELXTL) are employed in the course of a crystal-structure determination." [50, p. 112]

The article received 2391 citations in 2009. Although this will make no difference to the h-index, which is deliberately robust to such outliers, it wreaked havoc with another metric of science, the Impact Factor (Impact Factor). Whilst the h-index is supposed to measure the quality of a researcher or research group, the Impact Factor is the equivalent measure for the quality of journals. It is calculated as the mean number of citations to articles in a journal over the previous two years. In the case of the SHELX paper mentioned above, the Impact Factor of *Acta Cristallographica A* rose from 2.051 in 2008 to 49.926 in 2009.

The journal Impact Factor is calculated as follows, using citation metrics.

$$IP = \frac{c}{n} \tag{3.1}$$

where c is the number of citations received by all articles published in the previous two years and n is the number of all citable items in the journal (e.g. papers, reviews). The Impact Factor has been criticised widely. It is prone to manipulation, for example by a journal reclassifying items as non-citable (forums, letters etc.).

We have already mentioned that scientists have different motives for citation, making citation not necessarily an endorsement [13]. A paper might get a host of citations for a mistake. It would be a travesty if the best strategy for an author is to publish mistakes in order to accumulate citations. Also, review articles generally get more citations than articles reporting new research findings without the former being of any better quality than the latter; they just attract a broader set of readers.

However nowadays the Impact Factor is used to judge the quality of single publications or researchers and as that, the Impact Factor is not a good measure at all and

[19], who conceived of the measure, warned against such a use of it. The Impact Factor is at best a crude measure of the quality of a journal but it is useless to judge the quality of papers in journals.

Although citation counts of papers can be seen as a proxy for their impact (keeping in mind the problems of citation hunting, negative citations and runaway effects of citing), using the Impact Factor as a proxy for the quality of a paper is riddled with problems, for example the inverse power law distribution of citations per paper makes the arithmetic mean of the citations of all papers as a measure of its quality, highly problematic. Arithmetic means and power law distributions are generally a bad combination as the mean does not capture outliers in a sample and a power law distribution is constituted by outliers, i.e. a very few cases having very high values.

The most colourful criticism of the Impact Factor is posed in a blog that concludes as follows:

> “If you use Impact Factors you are statistically illiterate.
>
> 1. If you include journal Impact Factors in the list of publications in your CV, you are statistically illiterate.
> 2. If you are judging grant or promotion applications and find yourself scanning the applicant’s publications, checking off the Impact Factors, you are statistically illiterate.
> 3. If you publish a journal that trumpets its Impact Factor in adverts or emails, you are statistically illiterate. (If you trumpet that Impact Factor to three decimal places, there is little hope for you.) If you see someone else using Impact Factors and make no attempt at correction, you connive at statistical illiteracy.”[2]

Adler et al [53] turn the question of citation counts from averages to probabilities. When comparing articles from two different journals, what is the probability that a randomly selected paper from the higher Impact Factor journal has more citations than the paper from the lower Impact Factor journal? They compare the *Proceedings of the American Mathematical Society* and the *Transactions of the American Mathematical Society*. Using the 2005 database of citations, the *Proceedings* had an Impact Factor of 0.434, the *Transactions* an Impact Factor of 0.846. However, drawing a random article from the *Transactions*, 62% of the time it will not have more citations than a randomly drawn article from the *Proceedings*, (see [53, p.11]).

[2] http://occamstypewriter.org/scurry/2012/08/13/sick-of-impact-factors/

There are also some more mundane problems with bibliometric and citation data worth mentioning briefly. Citation will be biased towards English language publications. As more people will be able to read a paper in the English language, citation might be a measure of academic impact but not necessarily of quality as high quality items might not be discovered due to language barriers. Also publication and citation have a time lag that might be different for different kinds of research and different in different disciplines.

The allure of the Impact Factor is easy to see: Given the amount of publications cluttering the scientific ether it seems impossible to judge each single one individually. Peer review certainly is a marker of quality but seems to ensure a minimum standard rather than high quality.

Measurement and Quality

Whereas peer review needs experts to judge, the main idea of the metric model of quality in science is to construct metrics that can be applied to scientific output by non-experts. We have seen citation measures and the Impact Factor as examples. The metric model is driven by competition; researchers need to increase their h-index in comparison to other researchers, getting a higher position in university league tables, such as the Times Higher Education World University Rankings (intrinsically comparative) or publishing in journals with higher Impact Factors than other journals. Another important measure in science is the impact, usually construed as the non-academic impact. Often scientific advances impact on the non-academic world much later than or in completely different ways than expected at their conception (e.g. the Laser, Internet, and Teflon).

One problem of metrics in general is that they need to use proxies for the measurement. Are publications really a good proxy for research? Publications are a necessary part of research in ideal science, for example, as allowing collaboration and collective scepticism, but they are not the research itself. What happens if we reduce research to publications, and worse, to publications in a specified set of top journals? It might crowd out new research (already peer review has this danger, see [51]). It might lead to less specialised science papers as authors are hunting for the larger readership that general and overview articles attract, see [53]. Depending on the discipline, it may mean that works in progress (conference submissions etc.) are not counted although they are arguably more valuable to the overall advancement of collaborative endeavour of science than the final journal product.

Chapter 4

Ideal Science, Measurement and Quality

In the previous two sections, we introduced two models of quality in science. The ideal model refers to how science should and should not be (deontic level) and the metric model constitutes an attempt to measure scientific output for the purposes of resource-allocation (pragmatic level). The metric model could be seen as a sub-model of ideal science in the sense that it lays out the general principles (output - 'originality, significance and rigour'; impact - 'reach and significance'; environment, structures supporting research, e.g. collaborations, staff development, research students, income, infrastructure and facilities), processes (research outputs, expert review), standards (e.g. up to four research outputs) and mechanisms (environment, including those assessed and those entitled to assess, e.g. RAE itself) for assessing science. If this were the case, the metric model would need to support the ideal model. Yet, rather than a supportive relationship, there is a direct tension between the models, as for example between the disinterestedness norm and the metric performance demands. It has been shown that measurement and proxies often lead to target hunting rather than quality production thus changing the essence of what they set out to measure. For example Goldstein [24] analyses evidence of target hunting in the US and UK education systems.

> "In both England and Texas, we see evidence that when learning outcomes are made the focus of targets, those who are affected will change their behaviour so as to maximize their ?results?, even where this is dysfunctional in educational terms." [24, p. 11]

Practices like “teaching to the test” will improve test scores but not learning outcomes. Similar practices of gaming have been shown for the English National Health Service in [5].

Besides the tension between the ideal and metric models, Anderson et al [1] identified some dissonance between scientists’ endorsement of normative standards in science (the ideal norms) and their own behaviour. In a survey of 3247 researchers some normative dissonance was found between scientist’s endorsement of normative standards in science and their own behaviour and substantial normative dissonance was found between perceptions of scientists’ own behaviour and that of other scientists.

The authors take the list of Mertonian norms of science discussed in Chapter 2 and add two of their own, elicited as relevant from previous focus group research [1]. They juxtapose the set of six norms with the corresponding counternorms; see Table 4. The two added norms are particularly interesting for our study. The first norm is Governance together with the counternorm of Administration. This can roughly be described as the processes by which the smooth running of science is ensured. Governance covers processes such as peer-review, open access, self-regulation. Administration, in contrast, is a top down process. Administrators, not necessarily *au fait* with science, govern the processes via targets and regulation. The second norm is Quality and the counternorm is Quantity. These two norm–counternorm pairs are directly concerned with quality in science, governance and administration as different processes to guarantee quality and quality and quantity as juxtaposed outputs of the scientific enterprise. The two are also related as often governance will lead to quality (for example peer review does not consider how many papers should be accepted, but only whether a paper is good enough to be accepted), whereas administration will lead to quantity, given that administration more often than not relies on the measurement of outputs. Substantive dissonance is found in particular regarding the need for high quantity output, leading to other norms, such as communality and disinterestedness, being undermined. Norms in science, just like other norms, are deontic rather than descriptive of actual behaviour. However, increasing non-compliance with a norm leads to a diminishing of its strength, leading in turn to more non-compliance [17].

If science is *defined* via a set of norms, widespread non-adherence to the norms leads to the outcome no longer resembling ‘science’, in the same way as changing the rules of how pieces move on a chess board results in one no longer playing chess but a different game. While this is not problematic *prima facie* because, first, the interpretation of the ‘game’ is dynamic and second, the new game might make better sense to

Communality: Scientists openly share findings with colleagues.	Secrecy: Scientists protect their newest findings to ensure priority in publishing, patenting, or applications.
Universalism: Scientists evaluate research only on its merit, i.e., according to accepted standards of the field.	Particularism: Scientists assess new knowledge and its applications based on the reputation and past productivity of the individual or research group.
Disinterestedness: Scientists are motivated by the desire for knowledge and discovery, and not by the possibility of personal gain.	Self-Interestedness: Scientists compete with others in the same field for funding and recognition of their achievements.
Organized Skepticism: Scientists consider all new evidence, hypotheses, theories, and innovations, even those that challenge or contradict their own work.	Organized Dogmatism: Scientists invest their careers in promoting their own most important findings, theories, or innovations
Governance: Scientists are responsible for the direction and control of science through governance, self-regulation and peer review.	Administration: Scientists rely on administrators to direct the scientific endeavour through management decisions.
Quality: Scientists judge each others' contributions to science primarily on the basis of quality.	Quantity: Scientists assess each others' work primarily on the basis of numbers of publications and grants.

Table 4.1: Norms and Counter-Norms of science. (Table reproduced from [1, p.6]).

what the players' goals are, it may become problematic if it leads to chaos (for example owing to not being based on 'shared' understandings or not being a game at all anymore). In this sense, norm adherence is crucial for the continuation of the 'game' but also depends on norms matching shared understandings about how the game should be played. It is also sensitive to the perception of others adhering to norms. If people think that most other people litter they will feel less inclined not to. It is thus an interesting finding that there is some dissonance between the attitude and behaviour of researchers with regards to ideal norms but also that researchers see themselves as mostly compliant to norms in science and other researchers as mainly non-compliant [1].

Anderson et al [1] identified a normative dissonance between scientists' subscription to ideal norms and their own (reported) behaviour. In our study, we have identified a tension between the metric model and scientists' own understandings of quality in science (Section 4.1). While the tensions noted above might be seen as problematic for the good practice of science on many levels (i.e. indicating confusion about what the good practice of science is, leaving it open to interpretation, as well as lack of communalism or shared understanding of what the good practice of science is), their existence does not necessarily mean the practice of bad science, but indicates that there is a need for creating the 'space' for the expression of the model of science that makes sense to all user communities involved.

4.1 Scientists on Ideals, Measurement and Practice

The first stage of this investigation into quality in science was to run a word elicitation study to explore how people describe quality and to compare this to previous findings relating to quality perception in other contexts [20]. Three sets of data collections were conducted regarding scientists' perceptions and understandings of quality: a medium scale survey, 20 individual interviews, and four focus groups.

4.1.1 Survey

In addition to the data collection technique used by Ghylin [20], we also draw upon approaches from psychology (e.g., [48]; [30], who used 'prototype' studies to explore concepts such as 'forgiveness').

Words relating to quality were collected using an online survey, run between August and October 2010. The questionnaire asked participants to 'write down the words and phrases that you associate with quality in science' (participants were also asked to perform the same task considering quality in a general context, but we do not discuss these finding here). Participants were encouraged to produce both positive and negative examples, and the online questionnaire enabled participants to produce an unlimited number of responses. On the following page of the online survey, participants were then presented with the quality words/phrases that they had previously supplied, and were asked to rate each on a 7-point scale ranging from "3 Very Positive: to "-3 Very Negative". The participants were also asked to complete a short demographic questionnaire.

Participants were recruited via UK universities through a personal network of contacts of the authors, with participants encouraged to distribute information to friends/contacts inside and outside universities. Students, staff, and members of the public were all encouraged to participate, in order to get a broad range of views on quality, especially in science (i.e. by including participants who are "inside" and "outside" science; cf. Matzlich [36]). Of the 225 participants who took part in this study, 170 provided age information (86 were aged 21-30 years, 51 were less than 21 years, 18 were 31-40 years, 12 were 41-50 years, 2 were 51-60 years, and 1 was 61 years and over); Of the 169 providing gender information, 90 were female, and 79 male; Of the 168 participants providing information about their highest level of education, for 74 this was a school-leaving qualification, for 43 this was an undergraduate degree, for 27 a taught post-graduate degree, and for 27 a research postgraduate degree.

Data processing

The 226 participants gave a total of 955 responses (words/phrases), a mean of 4.23 responses each. Of these responses, 750 were unique, with 90 shared by two or more participants (the most frequently response, "reliable", was provided by 23 individuals). 361 of responses were single (or hyphenated) words, whereas 389 were phrases, and in total the average number of words per response was 2.53.

As noted above, each of the responses was evaluated by the same participant on the following page of the survey on a scale of "3 Very Positive: to "-3 Very Negative". The most popular rating was 3 (Very Positive), which was given to 657 responses, followed by 0 (Neutral, 193 responses). The most popular negative rating was -3 (Very Negative), applied to 46 responses, followed by -2 (30 responses), and -1 (29 responses).

Coding of data

The next step of the processing of the data was to group items according to their semantic similarity to help to give insights on the data analysed in the following stages of this study. The Wmatrix corpus comparison tool [47] was then used to identify set phrases and multi-word expressions, with texts parsed using n-gram software [2] to identify less common quality-related 2-, 3-, and 4-grams. N-grams which did not relate to content (e.g., 'of the') and duplicate items were removed by hand. Items with a frequency of at least 2 were retained, giving a total of 312 science-related quality items. Grouping of the items into categories was performed by two researchers with expertise in text and content analysis, with both having an understanding of the literature relating to science and quality in order to take into account the context of the items under consideration (see [54]). The researchers coded the data separately, with differences in categories resolved through mutual agreement [11], [40]. This process resulted in 23 science quality topic categories.

Top positive items	rating	average freq.
reliable	3	23
accurate	3	21
valid	3	7
useful	3	6
precise	3	6
unbiased	3	5
peer reviewed	3	4
insightful	3	4
safe	3	4
relevant	3	4
logical	3	4
beneficial	3	3
verifiable	3	3
excellence	3	3
practical	3	3
well written	3	3
reliability	3	3
replicable	3	3
consistent	3	3
quality assurance	3	3

Top negative items	rating	average freq.
confused	-2.5	2
faulty	-2.5	2
plagiarism	-2.5	2
bad	-2	7
sloppy	-2	2
unverified	-2	2
poor	-1.7	3
inconclusive	-1	2

Table 4.2: Showing the most strongly rated positive and negative items which relate to quality in science.

Science Quality Topic Categories	Description	Number of items
appearance	appearance, elegant, presentation	6
clarity	well communicated, understandable,	14
	clear arguments	
context	context, subject, topic	3
correctness	accurate, precise, exact	19
depth	insight, thorough, detailed	17
ethics	unbiased, ethical, safe	11
evaluation/assessment	good, excellence, high	16
function/result	usable, practical, beneficial	14
information-oriented	information, informative	2
intelligence	intelligent, smart, clever	3
new	ground breaking, significant, surprising	22
new/relational	different	1
process-oriented	well designed, analysis, conditions	42
proof-oriented	believable, convincing	2
quantity	mass, quantity	2
relational	cited, contribution, recognised	13
resilience	durable, stands the test of time, solid	5
results/proof	testing, verifiable, accurate data	50
standards	conforms, guidelines, requirements	10
structure-oriented	arguments, designed, framework	16
trust-oriented	reputation, rigour, confidence	18
trust-oriented/category	professional, academic, expertise	6
value	value, worth, expensive	11

Table 4.3: The Scientific quality word groupings

Discussion of Survey

As can be seen from the results for both the most positive and negative items of the survey, and also from the human coded categories, what constitutes quality in science is largely framed using Mertonian norms. For example, 'reliable' and 'accurate' are by far the most popular positive descriptions of quality science; of the top 20 words presented in Table 4.1, only 'peer reviewed', 'practical', and 'quality assurance' can be seen to hint at any kind of external assessment of scientific quality. In the coded categories, we similarly find categories overwhelmingly reflecting a Mertonian perspective (correctness, depth, evaluation/assessment), but to a lesser extent also find items hinting again at external assessment (function/result, new, standards, and value).

4.1.2 Interviews

Twenty 30-minute semi-structured interviews with scientists from Physical and Engineering Sciences in the UK were conducted. Participants included professors, senior lecturers, lecturers, research fellows and research students. The interviews focused on quality as a collective process and on the mechanisms and structures through which quality in science is produced and constituted [7]; [35]. They were recorded and transcribed, each producing approximately 3,500 words of text. Transcriptions were subsequently discourse analysed focusing on regularities and irregularities in themes and arguments as well as the rhetorical strategies used and footing employed by participants, [58].

Summary

The thirty interviews produced the following findings:

1. **Quality is an ideal versus quality as practice** Quality as something that 'should' be done in particular ways listing criteria (both purpose-related e.g. novel, groundbreaking etc. and structure-related e.g. clear, without errors etc.) The comparison to quality as it should be and quality as it is, is usually implicit, i.e. this is how it should be, but whether this is followed or not is a matter of question. When items that are not the way they should be are considered high quality then blame is attributed to established measures for 'passing' them as quality.

2. **Quality as automatic, logical, natural and common sense** Quality is treated as an automatic and natural notion, which any 'logically-minded' person would

recognise and understand and which is sustained by an established system (norm). It is talked about as self-explanatory and given, and agency for the construction of quality on the part of scientists is not considered at all. In this line of argumentation, common sense is defined in terms of metrics (citation metrics, Impact Factor etc.). This is a common way of orienting to quality for many research students and some research fellows.

3. **Quality as feeling** This is related to the one above and below. Quality is something one can feel (using metaphors and senses). There is an interesting conjunction between 'feeling' and 'logic' sustained in a balanced relationship when it comes to recognising quality. Terms that refer to feeling usually come later in the flow of discussion (with the exception of one interviewee who used sense-related terms from the beginning of the interview).

4. **Quality in the distinction of personal and other (community, established) metrics** This distinction between 'personal' and 'established' is mainly made by senior lecturers and professors. Research fellows emphasise their personal understanding of quality but to a lesser extent refer to external metrics. Research students make this distinction but it is more related to their work rather than to the wider community and established measures (one exception).

5. **Quality metrics: 'not the best system but the best we have'** This is a common way of orienting to quality by research fellows, lecturers, senior lecturers and professors (plus one research student).

6. **Quality in market terms 'something you can sell'** This is approached as the norm by two research students and a research fellow and is approached as the conundrum of external quality assessment by more senior people (a research fellow, two senior lecturers, etc.; exception: one student).

7. **Quality defined in terms of its metrics** This refers to defining quality as what is used to measure it, e.g. citation counts "if a paper has many, that means 'quality'". This is primarily taken up by junior scientists.

8. **Quality in natural sciences as a moving goal post / relative** The approach is that there are no 'static' answers to problems in the natural sciences so what may be considered as high quality now may be disproved (or may not be equally regarded by everyone). This line of argument was mainly developed by research fellows.

9. **Quality is 'learned' through Didactic means and Experience** The issues of mentoring, supervision, role models, teaching, learning from peers and peer feedback were all part of the discussion. Research students were less reflexive than more senior people about these processes.

Research students and fellows were keener to adopt new technologies and methods in the way they conduct their day-to-day 'science'; professors were more 'traditional' and more time constrained.

Discussion

As we will see by focusing on the interview data, the practice model of science is supported by some processes in place to measure science (see extracts 1-3 below), supported by new processes (see extracts 4a & b, 5 below) and not supported (anti-supported) by some existing processes (see extracts 3-5 below). The extracts discussed below are presented as exemplar cases of the above. Participants ranged from senior lecturers to professors. Some of the professors also held administrative roles (including one Head of Department, two Deans and one Associate Dean and Director of Research Centre).

There seems to be some alignment with the 'ideal' sense of quality (vs quantity) as an evaluative standard. For example, participants argue that publishing extensively is not an indication of good quality (but rather the opposite, see Extract 1 below) and are judgmental of colleagues or research cultures that practice this (see Extract 2). It is noteworthy that Extract 1 comes from an interview with a participant who also has an administrative role in the faculty.

Extract 1 (Interview 13)

R: [. . .] in in physical sciences to publish two or three papers a year is totally normal
I: mm
R: e:h but to publish sort of three or four good papers in a five year period that's what marks you out as a as a good researcher I think. ((and)) It's difficult to judge which ones are the good ones unless you are right in that field but I'm that's what I think should be what people should care more about is ◦the quality rather than the volume◦.

The participant orients to science as divided 'in physical science' and to publishing two or three papers a year as 'normal' (line 1). Yet, he distinguishes between papers and 'good' papers, which he treats as a criterion of one's own merit in research. The criterion or standard is: three or four in a five year period, which is in accordance with the standard set in the metric model (see Section 2.2). While the standard corresponds to the metric model in terms of the number of outputs, the quality 'good' requires field expert judgment (line 5). In the deontic that follows quality as 'that' is vaguely constructed as 'three or four good papers in a five year period' and is 'that' which is juxtaposed with quantity ('volume').

Extract 2 (Interview 11)

1 R: I referee quite a lot of papers that from groups is go: >not trying to be a racist or
2 xenophobic or anything but< there's a lot of groups in China who: publish a lot of
3 papers on very similar things like: there's a lot of repetition a:nd I think they have a:
4 very strong drivers there to publish as much as they can a:nd I receive a lot of papers
5 that I look at and say "well this has been done fifty times before" I could find identical
6 fifty papers
7 I: mm
8 R: showing exactly the same thing
9 I: mm
10 R: a:nd keep on they keep on sending them [...] there's no use no kind of driver to:
11 making something making something better out of this it's just that "we get that and
12 we then go and do something else". The kind of churn of publications rather than
13 actually being on a project and try to go towards
14 I: mm mhm
15 R: steps of advancement. I'm sure I'm as equally as guilty for a lot of this as many
16 other people w((h))ere you are driven by the number of publications you put out
17 I: mm mhm
18 R: I think higher quality is often when you don't just keep the churn of (.) material
19 coming focusing on making something better o:r improving what you've done

The extract seems to support Anderson et al's [1] finding of normative dissonance within and between scientists. According to this, not only subscription to norms is somewhat higher than norm adherence (lines 15-6), but other scientists' behaviour is considered as much more counter-normative than one's own. However, in lines 15-6, the speaker might be seen as disclaiming or warding off potential accusations of prejudice as was the case in lines 1-2, thus managing accountability rather than admitting a mismatch between attitude and behaviour with regards to quality.

In the next extracts, the ways in which the quality (vs quantity) norm has been applied indicate an uneasy relationship between scientists and metrics (Extracts 3, 4a&b, 5 below). For example, in Extract 3 a tension between established measures of quality assessment and measurement, and 'true' quality is indicated, challenging the foundations of 'universalism'. It is noteworthy that Extract 3 comes from an interview with a participant who mainly has an administrative role in the faculty.

Extract 3 (Interview 7)

66 R: I mean the good example recently that won the Nobel prize in physics (.) these two
67 guys in Manchester e:h discovered (.) graphene. A:nd fine >you know< the impli-
68 cations for that are very profound. This is wonderful material it could revolutionise
69 electronics a:nd best conductor known to mankind it's just e:h extraordinarily strong
70 material etcetera etcetera ok? That's fine. They did it by a bit of sellotape and pencilled
71 it (.) ok (.) is that? anyone can do that. ◦an eleven year old kid can do that◦. but that's
72 not really what it's about. …) So is that high quality? …) e:h the work is published
73 in the high quality journals (.) the: spin offs are enormous (.) absolutely huge. It's one
74 of the big areas of research (.) in a few years

75 I: mm

76 R: ok? So there's countless examples of this (.) in in in science in general and ◦in
77 physics in particular◦. A:h so yeah that's what I associate with quality. Otherwise it's
78 impossible to judge what is what good quality is. I mean you do read a paper and you
79 think "yeah this is superbly written"

80 I: mm

81 R: "it's very clear, it's thorough e:h it's these people have done a good job". You can
82 tell. They've sort of covered all the bases e:h and you say this is a high quality paper.
83 It's well written (.) well-presented. E:h (..) but I think in general I'd consider quality
84 to be a little bit more than that. ◦it's what it is (.) the relevance of the work◦

Extract 3 displays a construction of quality in science through a subtle contrast on which (and only on which) the speaker grounds (the) resort to official recognition of the scientific work as a quality assessment metric.

The exchange unfolds in response to Question 2 asking the interviewee to provide an example of high quality in their field.

The exchange is torn between quality constructed as scientific work having profound implications (impact) on the one hand, and well-conducted, well-written and well-presented scientific work, on the other, and, on this distinction the speaker grounds his association of quality based on established, 'recognisable' measures of quality (Nobel Prize, published in high quality journals, impact) disclaiming this grounding on that 'otherwise it is impossible to judge what good quality is'. Namely, while in lines 66-69 the speaker constructs quality as scientific work having profound implications, in lines 70-72 the speaker shifts footing and focuses to the scientific work entailed 'behind' producing science with profound implications. This formulation eventually escalates to a rhetorical question: 'so is that high quality?' (line 72). In response to the question the speaker draws on a list of 'accepted standards': published in high quality journals (output), enormous spin offs (impact), one of the big areas of research (impact), all of which subscribe to the metric model. The formulation oscillates between the norm of 'universalism' evaluating research on its merit through accepted standards, the accepted standards corresponding to the metric model and to particularism (questioning) the processes behind the output and orients to the former as 'it is impossible to judge quality otherwise'. To this end the speaker offers an alternative example of quality in science, which is constructed as the opposite emphasizing the research process (rather than the 'measurable' impact). Yet, this version of quality is not missing the 'measurable' aspects *per se* but rather, a third criterion: 'the relevance of the work' (line 84).

Therefore, the speaker oscillates between the norms of universalism (applied through processes and principles laid out in the metric model) and the norm of particularism (emphasizing the research process 'behind' the output), and offers a third option as a solution in practice (see also Extract 5). In this way, the speaker also manages accountability by presenting a balanced argument and by disclaiming the questioning of 'accepted scientific standards'.

In Extract 4a the participant orients to quality by drawing an explicit distinction between 'external' and 'internal' measures. The account seems to align with norms of ideal science with regards to 'quality' (vs quantity) but challenges the 'accepted scientific standards' which are 'in place' to represent the norm, thus the metric submodel.

Extract 4a (Interview 11)

1 I: and how do you develop 'cos you said you said there's the personal a:nd how do
2 you develop this sort of your own understanding? How do you learn about what
3 quality is

4 R: ((laughs))

5 I: and develop your own understanding? What's this (..) process about?

6 R: a:m (..) >I don't know I think< for me personally one of the biggest drivers was
7 actually one of my old bosses who: when I was a post doc at the university of Bristol.
8 Where the in that group the focus was on doing something well rather than pushing
9 out publications ◦my boss there was not very very interested in◦ publishing in par-

ticularly high Impact Factor journals o:r particularly pushing out as many papers as possible. He really just liked doing things correctly and you'll all rather than saying "well that looks fine (.) out it goes" he always said "could we do this extra or could we do that extra" so it was more: focused towards bei:ng sort of accurate. Am (....) conversely sort of here now funding ((laughs)) being much tighter these days and a lot more focus is put on the: kind of meeting the metrics that are being set by the funding bodies and the journals so the aim is maybe the driver there maybe when funding is tighter you more sort of judge yourself by the external metrics when i:t's (.) when you are left on your own you maybe more going towards your own internal sort of quality drivers ((inaudible)) what you want to do

Internal quality drivers seem to correspond to ideal norms, disinterestedness and organized skepticism in particular, while the external metrics, as set by the funding bodies and journals (part of the metric model), imply self-interestedness and dogmatism.

Extract 4b (Interview 11)

The extract then continues with the speaker explicitly arguing that ideal science is best represented by internal metrics that each scientist has than by 'accepted scientific standards' (the metric model) which are 'in place' to represent science.

I: mm. But from what you are saying they are not necessarily incompatible a:s

R: no::

I: at odds wi:th

R: no: (.) not necessarily you sort of what you want to do does overlap as to what you submit

I: mm

R: but i:f if say the only way you could see if you wanted to do a study of something and you need to do it pretty carefully and do a thousand (..) run thousand samples and do it this way and it will be quite formulaic maybe that (..) and then the grant proposal wants interesting science that's done that way and high impact fo:r society and developing this (..) if you thought you could still get the grant by: using your own internal metrics the:n you probably submit it. If you thought that you wouldn't get the grant unless you did something flashy a:nd quick and exciting you probably (.) you probably think about a:m being a bit more relaxed on your own internal quality metrics to try and get the grant. So maybe the sort of funding a:nd that really is the driver. °if you've got as much money as you want I think° a lot of scientists if they got a big grant they just do were happy to just carry on doing what they want

I: mm

R: maybe quality is best the external measures of quality become less important (.)°and the internal dominate more° yeah

The speaker not only argues that ideal science is better represented by internal metrics that each scientist has than by 'accepted scientific standards' (the metric model) which are 'in place' to represent it. He also implies that external metrics impact in particular encourage practices opposed to the scientific rigour scientists would normally apply (lines 29-32). A third way is negotiated as the model in practice: a flexible use of internal and external metrics. It is noteworthy that Extract 5 comes from an interview with a participant who also has an administrative role in the faculty.

Extract 5 (Interview 6)

I: and what do you what do you think about these ((R just mentioned)) the REF and the Impact Factor as metrics and assessment frameworks of quality?

R: e:m I think in some ways what I think about it >you know< it's the way we are going to be judged. And railing against the fact that it may not be >you know< I could have lots of debates about whether it's the right way of doing it but it's not going to make any difference to the way it's actually done

I: mm mhm

R: so I think it's a matter of making sure that for the university we fulfil those requirements but also for my own e:h I would always want to make sure I did what I saw as high quality research. So >for example< we have a list of journals that we should publish in. A:h I won't be using those journals simply because my research is in cancer >I'm in the department of electronic of electronic engineering< and consequently the top journals of electronic engineering include no cancer journals. Now if I were to publish my research in an electronic engineering journal they'd probably reject it. Because it's not in their remit. So that's where I would say I'd make the judgement because as far as I'm concerned the research has got to be in the right journal

Finally, Extract 5 displays an even stronger case of quality being negotiated at the juncture of personal criteria and external measures.

In the preceding lines the speaker made reference to the 'Impact Factor' and the 'REF' in response to a question on how standards of quality in science are set. The interviewer picks this up. The response which then follows is formulated along the lines of opinion versus fact, the scientist's own view of high quality versus established standards (the metric model). Initially, the speaker develops the argument that debating these standards will not change the way 'we' are going to be judged. This constructs these standards as debatable, the way 'we' are going to be judged as factual and positions the speaker as a powerless and a passive agent in this process, merely following these standards.

In lines 113-114, in the process of articulating how these standards are followed, the speaker draws a distinction between those standards and her standards, as the 'for the university but also for my own' analogy is not made. So while 'for the university' 'we' make sure we fulfil those standards, (but) 'for her own' the speaker shifts footing to talk in the first person introducing (the distinction) her own view of 'what I saw as high quality science'. So while the position in the debate over the quality standards such as the 'Impact Factor' and the 'REF' is passive yet implying that the former are debatable, the practice of high quality research is presented as a matter of rational ('not in their remit', 'right journal') and calculated ('make the judgment') personal judgment. This also presents a distinction between the (external) ways of judgment and the (personal) right judgment. (The speaker is a (better) judge of quality while the 'Impact Factor' and the 'REF' as mere requirements to be fulfilled). Most importantly, in so doing, the speaker attends to the norm of 'disinterestedness' (vs selfinterestedness) as it is relevance ('the right journal') and not self-interested motivation and pursuit of wealth that guides the challenge of 'accepted scientific standards'. This registers a tension between ideal and metric models and suggests a third model, the practice model of science.

Overall, it could be argued that the interview data has highlighted that there is a tension between norms of ideal science, especially between quality (vs quantity) on the one hand and universalism (vs particularism), organised scepticism (vs organised dogmatism) and disinterestedness (vs self-interestedness) on the other. This tension is grounded upon the application of normative ideals of science in established, institutional practices, and could therefore, be classified under the dichotomy of governance vs administration [1]; in other words, the misapplication of the norm of 'governance' of scientific practice, which renders it tedious and counter-productive 'administration' instead, but which, nevertheless, scientists are resisting through alternative action or rationalisation, as seen in the extracts above.

These findings are supported by a follow up study where academics and researchers in the social and natural sciences discussed activities related to collaboration and assessing publications in focus group sessions (see section 4.1.3). The main patterns identified included first, a tension between the ideal and metric models, in particular as regards the norm of communality (vs secrecy) and the predicament of research output in the form of publications as the standard on which tenure and promotion in academia is based, encouraging self-interestedness. Namely, the more junior participants were in their academic career, the more cautious they were about sharing work. Mid career academics developed a line of disinterestedness 'in principle' and self-interest 'in practice'; in other words, that collaborations advance science while also ensuring personal gain (in terms of advancement in one's career or material returns such as employment and resources). Secondly, quality was juxtaposed to quantity with reference to citations. Namely, there was a tension with regards to the importance of the number of citations as an indicator of selecting work to read as well as a criterion of one's reputation, while participants drew a clear distinction between citations and relevance in deciding the worth of journals (see also Extract 5). Finally, participants commonly oriented to science as a divided enterprise. Namely, in cases of disagreement participants claimed to speak for their discipline or research field, rather than for science as a whole.

4.1.3 Focus Groups

A focus group study on scientists' attitudes and judgments about science-related information was conducted between June and July 2012. In particular the study focused on how academics made decisions in relation to their own work, and how academic tools (e.g., literature search, community website etc.) can be better matched to their requirements.

4.1.4 Participants and Recruitment

Participants were recruited at the University of Surrey via an email invitation to the staff and research student lists. After receiving initial indications of availability and interest in participation, participants were allocated into groups of 4 members each. In each focus group the aim was to include a range of senior and more junior academics and a variety of disciplines. Overall, 4 focus groups were held with 16 participants. As regards career level, the distribution was as follows: 2 Professors, 2 Lecturers, 2 Senior Fellows, 1 Research Fellow, 8 PhDs and 1 Research Assistant. As regards discipline, the distribution was as follows: engineering (various - 6), psychology (3), sociology (1), politics (1), music (1), tourism (1), computing (1), information systems (1), and health (1). An effort was made to ensure that both the career level and disciplinary background were represented in the composition of each of the focus groups in order to enhance multisubjectivity. It was expected that participants would adhere to the

Activity	Description
Introduction	Introduce the study to participants, supply the basic information (i.e., participant information sheet), and respond to any questions.
Consent	Explain to participants about providing their consent, respond to any questions, and ask them to sign the form.
Scenarios	The scenarios which form the basis of the study were presented one after the other, as a separate discussion, introduced by the moderator with the help of printed questions or prompt cards.

normative imperative of multisubjectivity according to which everyone is entitled to their own opinion. Particular attention was paid in the ways in which participants claimed that they were speaking from their point of view, which fences off the speaker and achieves sticking to one's own opinion, while being explicitly invited by the focus group schedule to reach agreement (intersubjectivity). This co-existence of multisubjectivity and intersubjectivity might be help to explain some of the results presented below. Finally, to avoid the influence of senior academics upon more junior academics, focus groups did not include colleagues within the same discipline.

4.1.5 Method and Research Process

The focus groups consisted of single one-hour discussions of issues relating to quality, trust and reputation. Participants were invited to discuss six scenarios which were provided as visual prompts in print form (cards with text and print-screen output). In asking participants to discuss the scenarios, it was expected that these could be run as a group sorting exercise (i.e. discussing the options provided and then sorting them). In order to enable the development of discussions on the given scenarios in as mundane a manner as possible, the moderator's role was restricted to introducing each of the questions/tasks and ensuring that participants completed each of the tasks in a timely manner (e.g., by encouraging groups to come to a decision on a question). The actual responses of the group in the final ordering of options were taken into account and are presented below, but it was mainly the exchanges in discussion, which were recorded (using a digital recorder), and their analysis that provided useful insights. The discussions were anonymised before analysis.

The focus group sessions consisted of the following activities:

4.1.6 Scenarios and Summary of Results

Scenario 1 invited participants to consider the importance of various activities in which members of an online academic community engage (e.g. upload, download, review, rate and comment on papers, follow other people's work and write in blogs), and then rank them on the basis of how important they are to the community.

This commonly led participants to making a distinction between personal and community benefit activities and to an initial attempt to distinguish between important and unimportant activities. The distinction between personal and community benefit activities was not clear-cut as participants observed how dissemination, such as uploading papers, can be beneficial at both a personal and a community level. The activity identified as strictly addressing the community level was 'uploading news or events'. While the distinction of important and unimportant activities was also challenging on this basis, overall, writing in blogs was considered of low importance for a community while uploading papers was considered as an important activity, and as a prerequisite for other activities to exist, such as the activity of commenting. Finally, field and discipline-related differences were disclaimed to assess importance of particular activities by one participant in two focus groups (following someone else's work and writing blog posts respectively, which were marginal otherwise).

Scenario 2 asked participants to imagine a situation where they were trying to become familiar with a different field in another discipline and based on the results of a literature search, which provided article citation, download and date information, decide which article they would read first to familiarise themselves with this new field.

Participants drew a distinction between citations and downloads and argued that the latter are not indicative of the centrality of a publication in a given field. Participants commonly advocated tracing citations through/down (chronologically) as a tracking strategy for finding landmark papers. Another distinction commonly made was between the recency and longevity of a publication with participants in the natural and digital sciences opting mainly for the former and participants in the social science for the latter (and for books/textbooks) as criteria for their selecting readings into a new area or field. In cases of disagreement, participants oriented to a view of science as fragmented in fields, disciplines and research areas.

Scenario 3 asked participants to imagine a situation where they had a day to finish revising a paper for a journal responding to the reviewers' comments asking them to make reference to a particular research area. The scenario asked them to choose one article only basing the decision on the author and provided different variations of authors' reputation and social distance for participants to choose from.

Participants were divided in terms of career level with junior people mainly opting for articles by highly-cited experts, and senior people mainly opting for articles from contemporaries. Mid career people would tend to emphasise the relevance of the article and indirect trust (through others) rather than the author's career level as a criterion in deciding which article to make reference to. Early to mid career participants also mentioned opting for a reference coming from the reviewer or in the reviewer's area, or published in the journal in question.

Scenario 4 asked participants to imagine a situation where at a dinner conversation with a colleague and friend after a conference, they try to decide which scientist is the most respected in their field by ranking them on the basis of their work and how this is considered a contribution to science. Participants were provided with different variations of scientists' publication, citation and download records, reputation and social distance.

Participants opted for reputation as a first reaction: 'a scientist who is commonly regarded as an expert, who was involved in the field at the start and published quite a few papers then and is regularly cited, but has only published occasionally in the field since then' or 'a scientist who has published many papers on the topic over many years, and who has received many citations, and is well-known and respected by scientists you are friends with' – and then for publications and citations. They also argued that quantity is not an indication of quality and familiarity does not necessarily mean respect.

Scenario 5 invited participants to consider journal status by comparing top journals and less prestigious journals in terms of worth assessed through citations.

Participants treated the grounds of the scenario as topical but were commonly critical of the criteria of comparison, yet acknowledging a distinction between impact factor and paper citations and arguing that the latter is more important than the former. Overall, participants topicalised the purpose of the publication and the relevance of the journal claiming that a journal's worth is eventually relative to those. Field and discipline-related differences were used as disclaimers (i.e. orienting to science as divided into different disciplines and research areas, and attributing difference of opinion to this difference) in stressing the importance of the relevance of a journal over its impact factor, claiming that 'metrics are not comparable across sciences' and may only be indicative when it involves extremes (i.e. 'top' vs 'bad' journal).

Scenario 6 presented participants with a situation where they were contacted about sharing valuable information for the purposes of joint publication or bidding and asked them to choose from options of potential collaborators varying their career level and social distance.

Participants based their responses on their own status: the more junior they were in their career, the more cautious about sharing. They also stressed the importance of personal trust. When social distance was small owing to personal contact or indirect trust (through others), participants argued that they would share with contemporaries when co-authoring papers and with more senior people for applying for funding. Mid career people developed a line of disinterestedness 'in principle' and self-interest 'in practice'; in other words, that these kinds of collaborations advance science while also ensuring personal gain (in terms of advancement in one's career or material returns such as employment and resources).

Overall, mixing levels and disciplines enabled the following patterns:

1. 'Science'/field/discipline specificity/difference related argument: Participants would commonly treat science divided into fields. They mobilised their field or 'coming from a different field' as a disclaimer in voicing an opinion different to the group or to other participants'. The argument also developed in cases of disagreement, to justify alternative views. These practices occurred in discussing scenarios 2 and 5 in particular (in FG2 it was mobilised by one participant in scenarios 3 and 4, in FG3 by one participant in scenarios 1 and 3 and in FG4 it was also mobilised in scenario 1 by one participant).

2. Position and positionality: Position, and in particular 'footing' (i.e. where one stands in relation to what he/she says), were particularly salient in the context of scenario 6. Namely, participants would orient to 'trust' issues in terms of their career level. Overall, PhDs were more protective of their data, more suspicious of potential collaborators and less willing to share. Positionality (i.e. speaking as expected of one; reflexively by invoking one's category as synonymous to experience with regards to publications, assessing the importance of publications and collaborations) was normally the case with participants who were of more senior level (scenarios 3, 5 and 6), explicitly personalising the scenarios as "these things happen to me" and then speaking from that position.

3. Purpose/Relevance: This was a common response to scenario 5. It was suggested that a quantitative basis for comparing journals is not sufficient and that better criteria for comparison are the purpose of publication and the relevance of

the journal.

4.2 Summary

Using the conceptualisations of 'models of quality' in science, we have identified three models of science: metric, ideal, practice. We have looked at processes in place to assess quality in science and have identified a tension between the metric and ideal models. We then examined empirical data (survey, interviews and focus groups) and focused on the model users – scientists – according to which there is a tension between the ideal and metric models and between scientists' understandings of quality in science and the metric model. It becomes apparent that this tension has an impact on the daily research lives of scientists. Whilst the theoretical associations with quality in science lean heavily towards the ideal science model, the interviews show very clearly that scientists are torn between this ideal of quality and the need for amassing publications for personal career advancement. Another clear tension exists between the openness of science and ensuring intellectual property rights. This was particularly important for younger scientists (see focus group results).

This tension is intensified by some processes at work in institutional science, often based on competition and market dynamics, such as a scarcity of funding resources compared to applications and academic positions compared to the number of PhDs. If the numbers of applications for funding as well as for jobs becomes very high, it is likely that rough quantitative measures are used for an initial selection, for example by looking at the length of the publication record on an applicant's CV. For research applications it might mean that applications focus on short term impact as it is easier to assess quickly.

There are also some negative consequences of the metric model of science. One problem of metrics in general is that they need to use proxies for the measurement. In Section 3.1.1 we discussed the problems of using publications and citations as proxies. But is it just that the proxies have to be improved? As discussed in Chapter 3 measurement and proxies can easily lead to target hunting rather than quality production [5], [24]. Although one might eradicate the worst outcomes by targets and measurement, the best is also often crowded out as people try to achieve the measured medium performance.

Overall, these contradictions and tensions indicate that different discourses are at work. Quality in interviews was negotiated at the juncture of internal, personal criteria for assessing quality and external, established measures. The internal criteria seem to correspond to Mertonian norms – the ideal model. The external measures seem to correspond to the metric model. This juncture indicates a tension or dilemma of expertise (Scientists versus University administration, REF, Journal Editors and Reviewers). Quality seems to be considered within this context and the negotiation constitutes a third way – the practice model.

One way to explain this is by making an analogy between quality and knowledge in science, drawing on the argument that knowledge is for use rather than just for contemplation, and actors – scientists in this case – have their own interests about how instruments work. So, instead of reaching closure in terms of what scientific practice is, actors can be seen as constantly seeking to extend culture in order to accommodate those interests, while interests themselves become the standards by which the products of scientific culture extensions are assessed [42, pp.4-5]. This argument seems to follow up to the latter point, whilst if it was indeed a straightforward case of actors? interests emerging as scientific standards, then tension would cease to exist.

Another way to explain this, thus, is through the binary of governance/administration, arguing that in the process of creating and establishing ways to assess quality, science metrics become a self-fulfilling prophecy, parts of the definition of quality. Yet, these parts do not develop organically as the tension indicates, and contingency measures are taken by scientists, e.g. open access approach to publications. The emergent model at work seems to be a compromise – the practice model – which ensures a rudimentary compliance with the metric model. This model is viable for producing quality according to scientists, and, thus, needs to be better supported by processes. Currently scientists are left to negotiate different conceptions of quality in science individually. Thus, while this model may emerge as the contingency strategy or ‘third way’ in academia, it cannot equally function as a solution to the tension observed for everyone involved, thus necessitating more widespread adoption of processes that support it.

Chapter 5

Smart Recommendation Systems

In this chapter we discuss possibilities of online systems supporting quality in science. We discuss four approaches. The first is an agent-based model of the influence of social structure and trust and develops a model of the influence of trust and reputation on quality [60]. The data we discussed in Chapter 4, in particular the survey and the interviews, fed into this model. One of the important findings in the data was that the norm of quality was seen as something learned vertically through mentoring and teaching as well as horizontally via peer interaction. This also corresponds to Gilbert's [21] model based view of science.

The second recommendation application is an implementation of a quality ranking algorithm relying on item quality (assessed socially) as well as the trust and reputation of users. The third investigates collaborations and whether and how the composition of the research group influences the impact of research output. The fourth tackles the problem of overcrowding that can be a detrimental result of recommendation systems.

5.1 Trust, Reputation and Quality of Information

In [60] a simulation model is developed looking at the influence of trust (and different kinds of trust) on the quality of information agents obtain about their environment. In science, agents have to interact with other agents (other scientists) and objects (scientific papers, blogs, reviews, books; in short, sources of information). In the simulation model the object of interaction is a piece of information about the environment, the local temperature. Information is varied locally so that one temperature report might be true in one place and false in another. In the simulation the environment is divided into a cold and a warm area. Agents are located randomly. Agents are connected to

	Steps	Error
Baseline (local)	10	0.1527
Social Influence (local)	10	0.0992
Social Influence (fully connected)	10	0.50
Trust (cumulative)	10	0.0667
Trust (relational)	10	0.1188
Influence	100	0.085
Trust (relational)	100	0.085 reducing to 0.06

Table 5.1: Simulation Results.

other agents via a social network that spans across the divide. Agents are heterogeneous with respect to their ability to gather evidence from the environment. Incorrect information from the environment is implemented as noise. Agents with a lot of noise pass on more incorrect information and should not be trusted by others. Trust is implemented in two ways. Trustworthiness can be the attribute of a person. If correct information is received from a source the trust in this source increases, e.g. 'I trust the BBC, they are held as a beacon of unbiased reporting'.

Trustworthiness can also be an attribute of a relationship. While I might trust the BBC because many people trust the BBC, the origin of my trust in a source can come from personal experience with a source. I might 'trust Adam, despite his reputation for lying' because in our past interactions he has always been truthful towards me.

The purpose of the simulation is to investigate the interrelationship between trust and the quality of information. It tests the two kinds of trust, cumulative and relational and several plausible sets of update rules for trust to see which ones are most effective. Table 5.1 shows the development of information error using different trust relationships. It shows that error is significantly reduced through social interactions and the development of trust. Individuals on the border between the hot and cold zone are trusted less than others due to their information being at odds with the information of agents further inside the areas. Thus, using relational trust and social network attributes is helpful for the assessment of the quality of information.

5.2 An Application of Quality, Trust and Reputation in Online Communities

Like the simulation described in the previous section that investigates the relationships between trust, reputation and quality, Liao et al. [39] develop an algorithm for quality recommendations that uses not only popularity data of items but also social network data of users of a system.

Ranking techniques and reputation systems are nowadays widely employed in e-commerce online services: buyers and sellers may give each other a score after a completed transactions, and in the long term this encourages good behavior. The algorithm is a novel and generalized ranking algorithm for bipartite systems to assign quality values to objects and reputation values to users. A bipartite system is a system consisting of both items (papers, songs, etc) and users. This algorithm, called QTR (Quality, Trust and Reputation), builds on the classical HITS algorithm [31] but it allows for multilevel connections between users and items and exploits the information coming from the users' social relationships. QTR is a generalized algorithm in the sense that it can be easily adapted to different situations (e.g. by giving more weight to a certain kind of action, or to a particular behavior of users).

The algorithm is tested on two datasets, the EconoPhysics forum online community and the Last.fm online radio and social network. They are both particularly suited for testing the generalised algorithm. The EconoPhysics data contains various user-item interactions (uploading a paper, downloading a paper, and viewing a paper's abstract) and thus benefits from the QTR's multilevel nature. The Last.fm data is not just a standard bipartite data set where users are connected with items they have listened to but it contains also information about social relations between the users, thus making it also a good candidate for testing the QTR algorithm.

The simulations on the real datasets show that depending on the choice of its parameters, QTR has the potential to produce results that are less prone to spamming and manipulation than results based on plain popularity of items or users. Furthermore, we show that social (or trust) relationships between the users can play a valuable role in improving the quality of the resulting rankings of items.

The results obtained from using QTR show that using the trust relationship between users improves the ranking of items compared with the HITS algorithm which only considers objects.

5.3 Predicting Impact

Science is intrinsically a collaborative exercise. Nowadays most science is conducted in teams, i.e. horizontal collaboration rather than the vertical cumulation that for example Newton's famous quote "standing on the shoulders of giants" refers to this.

Following the old saying that 'two heads are better than one', collaboration might well have intrinsic value for science. But who should collaborate with whom? The following study tries to shed some light on collaboration in the context of citation of resulting collaborative publications.

Roth et al [49] look at the influence of team composition on the citation of publications. We have seen the h-index above as a measure of a researcher's productivity, here we look at teams of researchers. The study relies on bibliometric data of journal citation from the Journal Citation Reports (JCR) in four disciplines from 1991-2010. The data comprises 108 journals from Artificial Intelligence (AI), 279 from Mathematics (maths), 64 from Nanoscience and Nanotechnology (nano) and 185 for Oncology (onco). The information extracted from the data is the year of publication (t), authors list (A), concepts list (C) and citation index (c) of each article. The concept list (C) is derived from the n-grams collected from article abstracts . There is a significant number of articles without keywords which are not classifiable. The average number of concepts is somewhere around 2 or 3 in all disciplines. The number of authors differs more profoundly between disciplines: 2.75 for AI, 1.86 for maths, 4.18 for nano and 6.34 for onco. The citation index c is a normalisation for the raw citation count for each article in order to take into account the fact that older articles have had a longer time to accumulate citations than newer articles, and the diverse citation levels across disciplines. It is calculated by dividing the final citation count by the mean citation count of all papers published during the same year t_i.

There are several measures correlating with a paper's impact. The number of citations increases with the size of the research group. In particular for maths, where the mean number of authors is relatively low, doubling the group size corresponds roughly to a doubled impact. In the other disciplines the boost lies between 20 and 40%. Similarly, the number of keywords is positively correlated with the number of citations. Looking at the age of group members, the success of the group is highest if the most senior author is of an intermediate age in terms of his or her activity in the field. Age of keywords used is similarly correlated with impact but also with group age. A strong correlation is also found between past visibility of authors and concepts. This is not too surprising due to reinforcement dynamics of preferential attachment in citation networks (again a Matthew Effect).

Finally, the social and cognitive arrangement of teams is found to have mild yet potentially counter-intuitive effects: an original configuration of team members, for one, does not seem to consistently correspond to a higher or lower impact, except perhaps in the case of very strong yet not complete originality. Semantic originality, or rather, lack thereof, seems to more consistently correlate with differential impact. On the concept side, the totally unoriginal teams (in terms of the concepts which are being gathered in a given article) generally achieve smaller impact: while repetition does not appear to pay, perfect originality does not seem to pay either.

The authors conclude:

> "In terms of data mining, the present results may help design a recommendation engine for papers at the time of publication: since the various explored dimensions are generally available in bibliographic notices, for a given field, it is possible to imagine that an algorithm combining these often independent indices could be able to make predictions on the possible impact of given papers when they get published." [49]

5.4 Crowd Avoidance and Diversity in Recommendation

Gualdi et al [26] develop a recommender system which deals with the common problem of "overcrowding". Recommender systems recommend items regardless of potential adverse effects of item overcrowding. While there are situations where an arbitrary number of users can be recommended the same item, in other situations this is not the case. For example, one cannot recommend the same restaurant to many people as it has limited space and service capabilities. In many other situations, the adverse effects of overcrowding are less visible but still exist in terms of unnecessary competition, over-exploitation of resources, and creating hypes. This shortcoming is addressed by introducing crowd-avoiding recommendation where each item can be shared by only a limited number of users. An alternative formulation is based on penalising items according to how many users they are already recommended to which consequently decreases the chances of these items being recommended to more users. These two approaches yield comparable results.

Note that crowd avoidance is a general concept which can be used also in situations where resources can be shared by an arbitrary number of parties and user satisfaction does not decay with the number of other users sharing the resource. Given a set of user score (or cost) values, one can always apply an occupancy constraint or penalty and see how this impacts the assignment of items to users. It is of particular importance to note that this new assignment is bound to be more diverse than the original one where no additional constraints were present. For example, if an individual item scores top for many users, it cannot be assigned to all of them if the occupancy constraint is sufficiently strong. Other items then have to replace it and the composition of the assigned items becomes more diverse. In this way, the crowd avoidance concept can help address a long standing challenge of information filtering: the lack of diversity. Through crowd avoidance, a recommender system can thus avoid the potentially detrimental effects of recommendations on the item ecology.

Using a standard way of testing recommendation algorithms it is shown that the introduction of the occupation constraints enhances recommendation diversity and, contrary to expectations, accuracy even in systems where overcrowding is not detrimental. A simple explanation for the unexpected accuracy improvement observed is based on correcting biases (whatever they are) of the recommendation method. A simple way to model artificial systems with crowd avoidance is suggested and approximate analytical results for item popularity distribution in these systems is obtained.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

In this report we surveyed the interconnections of science, quality, truth, norms and metrics.

In Chapter 1 we discussed general considerations about quality in science, truth and social aspects of enquiry. We looked a little bit into the history of scientific methodology and some philosophy of science in order to clarify why quality assessments of scientific outputs are so important but also so difficult.

In Chapters 2 and 3 we discussed different models of quality in science, a descriptive account from [21], a deontic account from [41] and a metric account currently used in academia in the UK. We pointed out what the three accounts contribute to science being an epistemically special endeavour and how they, rather than supporting each other, create a tension.

In Chapter 4 we presented a study on the perceptions of scientists regarding quality in science. Using the conceptualisations of 'models of quality' in science, we have identified three models of science: ideal, metric, practice. We have looked at processes in place to assess quality in science and have identified a tension between the ideal and metric models. We then examined empirical data (surveys, interviews and focus groups) and focused on the model used by scientists according to which there is a tension between the ideal and metric models and between scientists' understandings of quality in science and the metric model. It becomes apparent that this tension has an impact on the daily research lives of scientists. Whilst the theoretical associations with quality in science lean heavily towards the ideal science model, the interviews show very clearly that scientists are torn between this ideal of quality and the need for amassing publications for personal career advancement. Another clear tension exists between the openness of science and ensuring intellectual property rights. This was particularly important for younger scientist (see focus group results).

Overall, these contradictions and tensions within discourse indicate that different discourses are at work. Quality in interviews was negotiated at the juncture of internal, personal criteria for assessing quality and external, established measures. The internal criteria seem to correspond to Mertonian norms: the ideal model. The external measures seem to correspond to the metric model. This juncture indicates a tension or dilemma of expertise (Scientists versus University administration, REF, Journal Editors and Reviewers). Quality seems to be considered within this context and the negotiation constitutes a third way: the practice model.

One of the ways to explain this is through the dichotomy of governance / administration, arguing that in the process of creating and establishing ways to assess quality in science, metrics become a self-fulfilling prophecy, parts of the definition of quality. Yet, these parts do not develop organically, as the tension indicates, and contingency measures are taken by scientists, e.g. an open access approach to publications. The emergent model at work seems to be a compromise — the practice model which ensures a rudimentary compliance with the metric model. This model is viable for producing quality according to scientists, and, thus, needs to be better supported by processes. Currently scientists are left to negotiate different conceptions of quality in science individually. Thus, while this model may emerge as the contingency strategy or 'third way' in academia, it cannot equally function as a solution to the tension observed for everyone involved, thus necessitating more widespread adoption of processes that support it. In order to ensure quality despite measurement and measureability despite idealism, both scientists and institutions need to take the existence of tensions between the traditional models seriously and work towards an agreed practice to actively support the production of quality.

Finally, in Chapter 5 we discussed models and recommendation systems for scientific quality. Given the fast proliferation of scientific (and non-scientific) information it will be helpful for scientists to have help in deciding on the quality of work. But what kind of quality do we want to support?

Recommendation systems could be tailored to the norms of science we discussed in Section 2. Recommendation systems might recommend other researchers with similar interests as well as events and publication outlets to support communality of science. Recommendation systems might provide an anonymous paper upload section for peer review commentary to support universalism and organised scepticism. Recommendation systems might provide guidance on funding sources and how to deal with conflict of interest in commercially funded research to support disinterestedness. Recommendation systems might provide a section for funders to look at existing research with research bids to be funded. It is difficult, however, to use norms for the recommendation of specific papers. How is one to know whether a scientist followed the norms in a particular piece of research? Because they are of a deontic nature they cannot be assessed as a function on the output. However, online communities might use reputation and trust relationships to enhance compliance to norms and make it easier for researchers to adhere to norms when conflicts arise between personal and normative demands.

Recommendation systems could also follow the metric model by recommending for example scientific output according to metric quality criteria. Citation of a paper, although still beset with problems (see Section 3.1.1), can be seen a criterion for impact and thus quality but it takes time for publications to accumulate citations, sometimes years. One of the ways to help scientists would be to find a proxy for a paper's quality *at the point of publication*.

Candidates we have seen for such recommendation were the Impact Factor of journals as a proxy for the quality of a paper at the point of publication. We have discussed the problems of the Impact Factor in some detail in Section 3.1.1. Downloads or views in an online system can be seen as impact and are faster accumulated than are citations. However, this was seen as a poor indicator of quality in the focus groups, when offered in Scenario 2 as a quality criterion, as it is easily skewed by, for instance, recommending one's paper to the compulsory undergraduate course one is teaching. Roth et al [49] cite their team composition metrics as a possibly better proxy for paper impact.

The QTR algorithm uses a social network and trust and reputation between users to enhance its recommendations and in the simulation discussed in Section 5.1 we saw the potential influence of such trust relationships on the quality of information. It could be seen as taking the best of both worlds, on the one hand using metric assessments of item quality but supplementing them with user reputation measures, thus taking into account also normative facets of quality and recommendation.

An important question we need to ask is whether there are possible repercussions from recommending something as of high quality. The biggest problem with recommendations following the norms of science is that we might support the narrowing of research fields by recommending outlets and collaborators with already high overlap. It might be that the best quality science is done at the edge when scientists with different interests meet each other.

Metric recommendation systems also have their dangers. Let us take citation as a proxy for quality. We have seen the Impact Factor of journals as a proxy for impact of a paper at the point of publication and discussed problems besetting this measure. The composition of the team of authors might fulfil the same purpose. However, people might a) try to compose their group according to the measure in order to be recommended and b) people might cite papers following those measures (e.g. with more authors), both regardless of actual paper quality.

Recommending papers with many citations is prone to speeding up the already existing Matthew effects in science (see for example [56] for an analysis of the Matthew effect in publication and [51] in peer-review) but this time the effect would not emerge but be enforced. This would lead to a change in the ecology as less and less items would get any recognition with all the focus on very few, top rated items. Recommendation systems need to be aware of the possibility of these kinds of runaway effects which are useful for selling in e-commerce systems (e.g. recommendations used in Amazon) but will potentially be very detrimental to science. The crowd avoidance mechanism discussed in Section 5.4 might help to avoid these problems to some extent by preserving diversity, at least partially undermining the Matthew effect.

Bibliography

[1] M.S. Anderson, B. C. Martinson, and R. De Vries. Normative dissonance in science: results from a national survey of us scientists. *Journal of Empirical Research on Human Research Ethics*, 2:3–14, 2007.

[2] S. Banerjee and T. Pedersen. The design, implementation, and use of the ngram statistics package. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, 2003.

[3] J. Bar-Ilan. Which h-index? – a comparison of wos, scopus and google scholar. *Scientometrics*, 74(2):257–271, 2008.

[4] B. S. Barnes and R. G. A. Dolby. The scientific ethos: A deviant viewpoint. *European Journal of Sociology*, 11:3–25, 1969.

[5] Gwyn Bevan and Christopher Hood. Have targets improved performance in the english nhs? *British Medical Journal*, 332(7538):419–422, February 2006.

[6] David Bloor. *Knowledge and Social Imagery*. Chicago University of Chicago Press, 1991.

[7] L. Boltanski and T. Pedersen. The design, implementation, and use of the ngram statistics package. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, 2006.

[8] Thomas Brante. Consequences of realism of sociological theory-building. *Jounral for the Theory of Social Behaviour*, 31(2):167–195, 2001.

[9] R. Burrows. Living with the h-index? metric assemblages in the contemporary academy. *The Sociological Review*, 2012.

[10] K. Knorr Cetina. Laboratory studies: The cultural approach to the study of science. In S. Jasanoff, G. Markle, J. Petersen, and T. Pinch, editors, *Handbook of Science and Technology Studies*. London: Sage, 1995.

[11] K. Charmaz. Discovering chronic illness: Using grounded theory. In B. Glaser, editor, *More grounded theory methodology: A reader*. CA: Sociology Press, 1994.

[12] H. M. Collins. The sociology of scientific knowledge: Studies of contemporary science. *Annual Review of Sociology*, 9(265-85), 1983.

[13] Susan Cozzens. What do citations count? the rhetoric-first model. *Scientometrics*, 15(5-6), 1989.

[14] Pierre Duhem. *Aim and Structure of Physical Theory*. Princeton University Press, 1954.

[15] Daniele Fanelli. How many scientists fabricate and falsify research? a systematic review and meta-analysis of survey data. *PLoS ONE*, 4(5):e5738, 2009.

[16] Paul Feyerabend. *Against Method*. London: Verso, 1975.

[17] Urs Fischbacher and Simon Gächter. Social preferences, beliefs, and the dynamics of free-riding in public good experiments. *American Economic Review*, 100:541–556, 2010.

[18] Bruno Frey. How intrinsic motivation is crowded out and in. *Rationality and Society*, 6(3):334–352, 1994.

[19] Eugene Garfield. Citation analysis as a tool in journal evaluation. *Science*, 178(4060):471–479, 1972.

[20] K. M. Ghylin, B. D. Green, C. G. Drury, J. Chen, J. L. Schultz, A. Uggirala, J. K. Abraham, and T. A. Lawson. Clarifying the dimensions of four concepts of quality. *Theoretical Issues in Ergonomics Science*, 9(1):73–94, 2008.

[21] Nigel Gilbert. From research findings to scientific knowledge. *Social Studies of Science*, 6(3/4):281–306, 1975.

[22] Nigel G. Gilbert and M. Mulkay. *Opening Pandora's Box: A Sociological Analysis of Scientist's Discourse*. Cambridge: Cambridge University Press, 1984.

[23] H. Goldstein and D. J. Spiegelhalter. League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal Royal Statistical Society A*, 159(3):385–443, 1996.

[24] Harvey Goldstein. Education for all: the globalisation of learning targets. *Comparative Education*, 40(1):7–14, 2007.

[25] Andrew Gray and Bill Jenkins. Government and administration: Too much checking, not enough doing? *Parliamentary Affairs*, 57(2):269–287, 2004.

[26] S. Gualdi, M. Medo, and Y. C. Zhang. Crowd avoidance and diversity in recommendation. *EPL draft*, 2012.

[27] Jürgen Habermas. *Truth and Justification*. The MIT Press, Cambridge, Massachusetts, 2003.

[28] Oswald Hanfling. *Logical Positivism*. Columbia University Press, 1981.

[29] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proc Natl Acad Sci USA*, 102(46):16569–16573, 2006.

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

[31] Jon Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[32] Thomas S. Kuhn. *The Structure of Scientific Revolutions*. Chicago University Press, 1962.

[33] Imre Lakatos. *The Methodology of Scientific Research Programmes: Philosophical Papers Volume 1.* Cambridge: Cambridge University Press, 1978.

[34] Imre Lakatos and Alan Musgrave, editors. *Criticism and the Growth of Knowledge*. Cambridge: Cambridge University Press, 1970.

[35] Bruno Latour. A textbook case revisited—knowledge as a mode of existence. In Edward J. Hackett, Olga Amsterdamska, Michael Lynch, and Judy Wajcman, editors, *The Handbook of Science and Technology Studies*. Cambridge, MA, US: MIT Press, 2007.

[36] Bruno Latour and Steve Woolgar. *Laboratory Life: The Construction of Scientific Facts*. Princeton University Press, 1979.

[37] Larry Laudan. *Science and Relativism: Some Key Controversies in the Philosophy of Science*. University of Chicago Press, 1990.

[38] Peter Lawrence. The heart of research is sick. Interviewer: Jeremy Garwood, February 2011.

[39] Hao Liao, Giulio Gimini, and Matús Medo. Measuring qualit, reputation and trust in online communities. *http://arxiv.org/abs/1208.4042*, 2012.

[40] John Lofland and Lyn Lofland. *Analyzing Social Settings: A Guide to Qualitative Observation and Analysis*. Belmont, CA: Wadsworth, 1995.

[41] Robert K. Merton. The normative structure of science. In Robert K. Merton, editor, *The Sociology of Science: Theoretical and Empirical Investigation*. Chicago: University of Chicago Press, 1973.

[42] Andrew Pickering. *Science as Practice and Culture*. Chicago: The University of Chicago Press, 1992.

[43] Karl Popper. *The Logic of Scientific Discovery*. Routledge, 1934/1959.

[44] Stathis Psillos. Living with the abstract: Realism and models. *Sythese*, 2009.

[45] Hilary Putnam. Problems with the observational/theoretical distinction. In Robert Klee, editor, *Scientific Inquiry*, pages 25–29. New York, USA: Oxford University Press, 1999.

[46] W. V. Quine. Two dogmas of empiricism. *Philosophical Review, Vol.60, No.1, pp. 20–43*, 60(1):20–43, 1951.

[47] P. Rayson. From key words to key semantic domains. *International Journal of Corpus Linguistics*, pages 519–549, 2008.

[48] E. Rosch. Principles of categorization. In E. Rosch and B. B. Lloyd, editors, *Cognition and Categorization*. Lawrence Erlbaum Associates, Publishers, (Hillsdale), 1978.

[49] Camille Roth, Carla A. Taramasco, Jean-Philippe Cointet, and Víctor A Bucheli. How citable is your team? impact as a function of socio-semantic dynamics. Technical report, QLectives working paper, Forthcoming.

[50] G. M. Sheldrick. A short hisotry of shelx. *Acta Crystallographica A*, 64(1):112–122, 2007.

[51] Flaminio Squazzoni, Giangiacomo Bravo, and Karoly Takas. Does incentive provision increase the quality of peer review? an experimental study. *Research Policy*, 2012.

[52] Flaminio Squazzoni and Claudio Gandelli. Saint matthew strikes again: An agent-based model of peer review and the scientific community structure. *Journal of Informetrics*, 6:265–275, 2012.

[53] Citation Statistics. Robert adler and john ewing and peter taylor. *A report from the International Mathematical Union (IMU) in cooperation with the International Council of Industrial and Applied Mathematics (ICIAM) and the Institute of Mathematical Statistics (IMS)*, 2008.

[54] A. Strauss and J. Corbin. *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Thousand Oaks, CA: Sage., 2nd edition, 1998.

[55] Bas van Fraassen. The pragmatic theory of explanation. In Bas von Fraassen, editor, *The Scientific Image*. Oxford University Press, 1980.

[56] Christopher Watts and Nigel Gilbert. Does cumulative advantage affect collective learning in science? an agent-based simulation. *Scientometrics*, 89(1):437–463, 2011.

[57] Peter Wells. The research excellence framework: And why it isn't. *The Center For Business Relationships, Accountability, Sustainability and Society*, 2012.

[58] M. Wetherrell. Positioning and interpretative repertoires: Conversation analysis and post-structuralism in dialogue. *Discourse and Society*, 9(3):387–412, 1998.

[59] John Worrall. Structural realism: The best of both worlds? *Dialectica*, 43(1-2):99–124, 1989.

[60] Michal Ziembowicz. Novel models of agency and social structure for trust and cooperation. Technical report, QLectives, January 2012.