

Reputation or peer review? The role of outliers

Mario Paolucci¹, Francisco Grimaldo², and Jordi Sabater-Mir³

¹ Institute of Cognitive Sciences and Technologies, Italian National Research Council
Via Palestro 32, Roma, 00185, Italy, mario.paolucci@istc.cnr.it

² Departament d'Informàtica, Universitat de València
Av. de la Universitat s/n, Burjassot, 46100, Spain, francisco.grimaldo@uv.es

³ Artificial Intelligence Research Institute, Spanish National Research Council
Campus UAB, Bellaterra, 08193, Spain, jsabater@iia.csic.es

Keywords: Peer review, reputation, agent-based simulation

1 Introduction

Research, as a social process aimed to the discovery of natural and social truth, has been ever growing since the adoption of the scientific method. It has been a growth in numbers: scientists, papers, subjects; an increase in expenditure (in absolute, although maybe not in relative terms); an intensification in effects on the everyday life; and so on.

An ecosystem on its own, the research social system evolved in time and reacted to its growth by adapting its practices and by developing new ones. If a researcher could be blissfully unaware of the existence of it at the beginning of last century, peer review as a process became common knowledge after the second world war. There was a very simple reason for that: editors could not sustain the needs of journal publishing, in time and in field coverage.

In recent times, the combined effects of growth and technological changes put the research system, and peer review especially, to a crossroad. Should the peer review and journal publication system stay on its path, or should it take a new direction? On the course of ICT-mediated collaboration, could the research ecosystem be done with peer review and embrace the selection mechanism of online social networks, that is, a reputation system ([1])?

The interest of this comparison is manifold. First, one could argue that reputed referees should not be substituted by non-experts but it is a fact that the growth in the number of submissions threatens the sustainability of a system that already faces difficulties obtaining the cooperation of senior researchers and involves a high percentage of junior researchers.

Second, drafts need to be available in order to be read and assessed but on-line repositories are more and more used and solve any physical publication space limitation. Third, the citation and reputation systems can conceptually overlap as both recognize quality but the first one can be biased towards the aggregate reputation of the journal and develops more slowly due to the delay introduced by the publication time. Last but not least, reputation might be stronger as a drive to decide what to read since paper level scores could better match the

heterogeneous quality of manuscripts published in a journal and result in a fairer visibility of them.

2 Research questions

The core question we aim to answer is how the traditional peer review system compares to a distributed, crowd-sourced evaluation. Such a distributed system was not available when peer review first came out, but it could easily put into practice today. Should we promote such a change? Or instead, does peer review have substantial advantages against crowdsourcing? We think that an answer to this research question should be obtained from a computational approach, also taking into account the historical path of science. We have thus developed a computational model of science production and consumption to answer the question: how do different quality distribution influence the consumption of science under different filter mechanisms?

3 The model

To ground the discussion and provide computational answers to our research questions, we have built a simple model of paper production. On the top of paper production, we apply either a simulated peer review process, or a simulated reputation mechanism. Both of those mechanisms drive the reading process. Scientists, in one case, read papers as prioritized by journal's quality. In the other case, they read papers as prioritized by their reputation.

Following the approach used by other paper production models [?], we consider as main entities: the scientist (in the role of author and reader) and the paper (possibly co-authored and available to be read). Scientists and papers are generated with a random initial *quality*. In our NetLogo implementation, we use random generators of positive quality following either a *Zipf* or *exponential* distribution. Scientists are also characterised by their *productivity*, that is, the number of papers that they produce in a simulation step. The same value drives also the number of papers read by a scientist. Being capable of reading only a limited amount of papers, scientist must choose. The choice is driven by different mechanisms in the two submodels detailed below.

3.1 Sources of randomness

All models of science, including models of peer review as the current one, must deal with the problem of paper quality. In this paper, we choose the simpler approach of generating quality from an aleatory distribution. This entails a choice for that distribution; in most applicable definitions, quality is path-dependant, including at the very least both veracity (a paper without mistakes is a good paper) and novelty (a clean derivation of an existing result will not have the same impact as the first derivation of the same). It could be argued that impact

should also be considered - although impact can only be meaningfully measured a posteriori, adding even more intricate path dependence to the picture. With all those factor to consider, the question to be asked is what are the measurable indicators of this inaccessible quality. Most models of science use citations as a proxy to quality, estimating the cause by the effect. While this is also an oversimplification, we are going to adopt it for the purposes of the present work. However, while there is a consensus on the distribution of citations being right skewed, with many papers collecting little citations and few papers collecting many citations each, the fit on empirical data is never perfect. Different fields, different journals, different time scales can give rise to different fits. This inspired us to apply two different distributions to our simulation. These are the power law and the exponential.

4 Results

To compare the effect of a reputation-based system with a peer review based one, we run a set of joint simulations. Each run constitutes of two joint simulations: a set of author and papers, with random generated quality, is produced, and on that set the two approaches to filtering are executed.

We present in figure (1) numerical results from a set of simulations that we run with 1000 scientists, 50 journals, 3 average authors per paper and 2 reviewers per paper. We run for 60 steps (years), a time window that we consider to be long enough (after 60 years, the field or subject modelled have probably changed so much that the model loses analogic validity).

For the exponential landscape, the results are quite stable and show a linear trend. Selecting papers to read through the reputation based evaluation system brings an advantage that accumulates linearly in time. On the opposite, runs under a Zipf landscape are wildly different. Peer review catches the bursts earlier, accumulating an instant dramatic advantage that is afterwards spent as reputation recovers. We have checked that this happens in correspondence to quality outliers - exceptionally good papers that the exponential landscape does not allow, that are taken over faster by the reputation system with some delay.

5 Discussion

We have proposed and described a model of paper publication to be studied under the effects of two different mechanism, peer review and reputation, under two different quality landscapes, one generated by an exponential and the other by a Zipf distribution. Preliminary results seem to point out that those systems differ in the rapidity with which they take up extreme events -in our case, papers that have an extremely high quality, that exploits the advantage of immediate recognition that comes with reputation systems. In absence of these extreme events, peer review performs better, and would recover the shock of an extreme

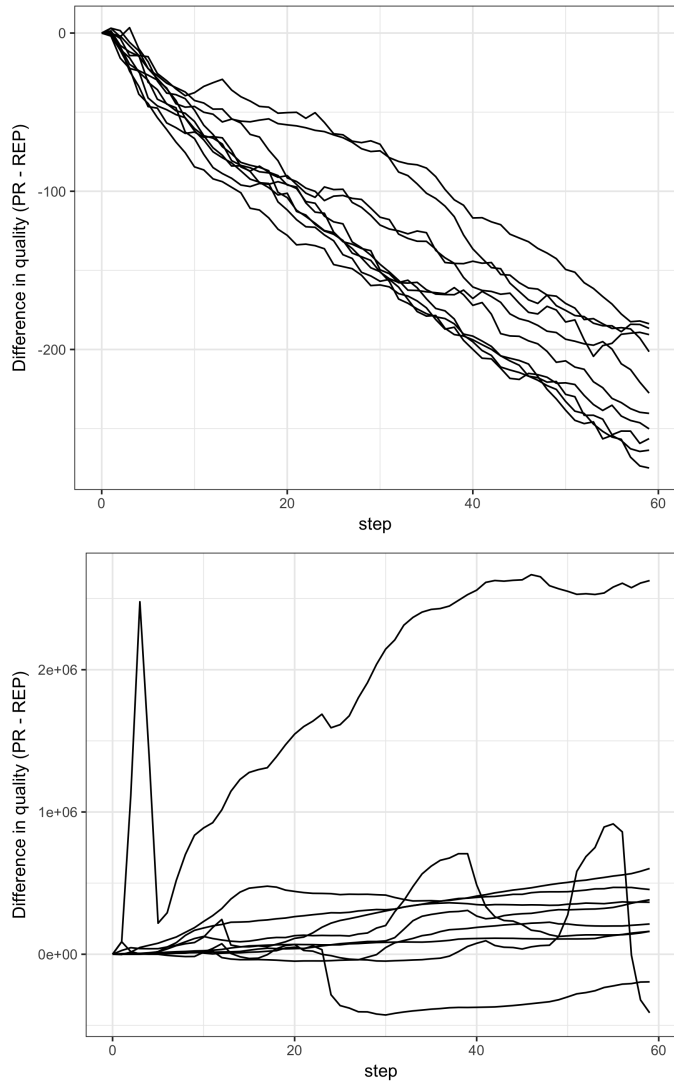


Fig. 1. Above: representative run of simulation in an exponential landscape ($\lambda = 0.09$). Below: ten runs with a Zipf landscape (exponent 1.88). Reputation performs better in an exponential landscape; outliers favour peer review in the Zipf landscape.

event in time. The interpretation of this result depends on the interpretation of quality that, we remember, is only observable a posteriori. If aiming to optimise the diffusion of quality, our simulation suggests that one should prefer a peer review system in established fields, where there it is unlikely that an extreme event will shake the field and where knowledge will be built incrementally. To the contrary, one should apply a reputation-based selection mechanism in fields that are open to disruptive innovation.

Acknowledgements

We acknowledge the support of the COST Action TD1306 "New frontiers of peer review" (www.peere.org), the FuturICT 2.0 project, the Cimplex project and the MILESS project (TIN2013-45039-P - financed by FEDER).

References

1. I. Pinyol and J. Sabater-Mir. Computational trust and reputation models for open multi-agent systems: a review. *Artificial Intelligence Review*, 40(1):1–25, 2013.