The effects of homophily on the arbitrariness of peer review

Aliaksandr Birukou[1] and Elise S. Brezis[2]

[1] Springer-Verlag GmbH, Tiergartenstr. 17, 69121 Heidelberg, Germany
aliaksandr.birukou@springer.com
ORCID: 0000-0002-4925-9131
[2] Aharon Meir Center for Banking and Economic Policy, Department of Economics, Bar-Ilan University, Israel
elise.brezis@biu.ac.il
ORCID: 0000-0002-7954-8110

Abstract. This paper focuses on why innovative works are not highly ranked in the existing peer review process, and in consequence are often rejected.
We propose a model for studying this problem and illustrate it by an example. Our model is based on the assumptions that reviewers are different in their taste and display homophily, i.e., that projects with similarity to their own taste will be more appreciated, and be highly ranked.
Our model can explain the famous NIPS experiment showing that the ratings of peer review are not robust, and that changing reviewers can have dramatic impact on the review result.

Keywords: peer review, innovation, computational model.

1 Introduction

The process of peer review is known to have a bias against new ideas: new models and inventions are too often rejected. Despite a lot of research about peer review [1], and a lot of alternative models proposed (see, e.g., [2,3]), the hypothesis that it is harder for very innovative ideas to pass via the peer review process has not been thoroughly studied yet.

Some early research has claimed that grant-review committees are hesitant to risk funds on innovative or speculative proposals [4] and have included the story about Weneger's hypothesis of continental drift [5].

More recent examples include Ragone et al. [6] who studied the ability of peer review ratings to predict the future impact of paper. Their dataset included 9,000 reviews on ca. 2,800 papers submitted to computer science conferences. One of their conclusions was that there is a low correlation between peer review outcome and the future impact measured by citations. In addition to the results by Ragone et al., the NIPS experiment has shown that the ratings are not robust, e.g., changing reviewers can have dramatic impact on the review results [7].

In this paper we present a simulation model which describes the peer review process of grant proposals. The model is based on the concepts of homophily [8,9] implying that reviewers have personal bias, and ideas closer to one's mental model will get more attention and traction. In this paper, we assume that reviewers are different in their taste for innovation, and in consequence, they give grades according to how these projects are close to their own taste.

In this version of the paper, we present a numerical example of the model. Future work will include testing the model via simulations and comparing the results with real-world data, such as review ratings from computer science conferences or project/grant funding programs.

2 Peer review in computer science conferences

In this section we analyze which criteria are used for reviewers to make decisions during the peer review process at several computer science conferences. A-level conferences representing different fields (artificial intelligence, cryptology, computer vision) have been selected.

Table 1: evaluation criteria in computer science conferences

Group	Soundness				
Criteria	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**
Conference	Technical/Presentation quality	Clarity	Correctness	Meets CfP requirements	Experimental validation
NIPS[1]	X	X			
IJCAI[2]	X	X	X		
CRYPTO[3]	X		X	X	
ICCV[4]	X	X	X		X

Group	Contribution					Innovation	
Criteria	**(6)**	**(7)**	**(8)**	**(9)**	**(10)**	**(11)**	**(12)**
Conference	Potential impact	Significance of results	Opens new directions	Of interest to the experts	Importance / relevance	Novelty	Originality
NIPS	X					X	
IJCAI		X					X
CRYPTO			X	X		X	

[1] https://nips.cc/Conferences/2014/CallForPapers
[2] https://ijcai-17.org/MainTrackCFP.html
[3] https://www.iacr.org/docs/progchair.pdf
[4] http://im-lab.net/wp-content/uploads/PapersAndReviewProcess.pdf describes the review form

ICCV					X	X	

As one can see from Table 1, we can roughly group the 12 criteria used under three groups: *soundness*, dealing with the presentational and scientific validity aspects; *contribution*, responsible for the importance of the results; and *innovation*, showing how novel the results or ideas are.

3 The model

The way peer review works is that all reviewers are giving a grade to each paper/project, and the papers/projects with the higher average grades will be funded/accepted. In this section, we show that this leads to refuse projects with higher degree of novelty.

The main assumption of the model is that referees differ in their perception about inventions, due to homophily. Homophily is the notion that similarity breeds connections [8]. Applied to peer review, it means that reviewers are more likely to appreciate level of innovations similar to their own research tendency, and give grades according to how these projects are close to their own taste. So reviewers who are developing conventional ideas will tend to give low grades to innovative projects, while reviewers who have developed innovative ideas tend, by homophily, to give higher grades to innovative projects. In our opinion, the homophily element is one of the reasons why the variance between the grades given by two reviewers is so high (see the NIPS experiment [7], where a fraction of submissions went through the review process twice and the results differed significantly).

The second element leading to a high variance is that referees are not investing the same amount of time to analyze the projects (or equivalently are not with the same abilities). These two assumptions lead to the fact that good innovative projects are less accepted and often rejected.

3.1 Assumptions

1. We have k projects from which only h can be funded.
2. Referees are different in their subjective value of time, as well as their degree of homophily to the project.

3.2 The model

The true value of a project is:

$$V_i = \alpha S_i + \gamma C_i + \beta I_i \tag{1}$$

where S represents the scientific soundness of the project, C the [scientific] contribution, and I is the innovative element of the project. Usually, funding commit-

tees suggest/propose many criteria used in projects funding and conference peer review [7], such as clarity, reproducibility, correctness, the absence of misconduct, novelty and value added. As shown in Section 2 for conference, we can group such criteria under the three groups described. More specifically for the project review, we define under S_i, criteria linked to soundness as clarity, reproducibility, correctness, the absence of misconduct; under C we include criteria linked to the impact and value added, while novelty-related criteria are under I_i. We order all the projects in an increasing value such that

$$V_1 < V_i ... < V_k.$$

The referees try to estimate these values. We denote U_{ij} the value given by the referee j to the project i. U_{ij} is a function of the time spent analyzing the project and its scientific soundness. It is also a function of the referee's opinion on how innovative the project is, and it is influenced by homophily.

We now present more specifically the way referees value project. First, we assume that the referees evaluate S_i without error, since usually there are no big debates about the 'soundness' of a project.

About the contribution and value added, there is usually a debate between referees. Indeed the contribution, C_i, is not easily evaluated. We define T_{ij} as the time that referee j takes to investigate the project i, and assume that if the time invested is higher than the contribution value, i.e., $C_i \leq T_{ij}$, then the referee can correctly estimate the true value of the project. However, if $C_i > T_{ij}$, then he/she does not appreciate the true value. In other words, we assume that the more time a referee spends analyzing the project, the closer he gets to the true value C_i; and the greater the difference between C_i and T_{ij}, the larger the error in valuation is.

While in general, the T_{ij} depends on both the reviewer and the project, we believe that it can be represented as

$$T_{ij} = T_j + \varepsilon_{ij},$$

where T_j represents the average time the referee j spends on review and ε_{ij} represents the project-dependent fraction of time. In the following, we set $\varepsilon_{ij} = 0$ for the same of simplicity.

About the innovative value of a project and the effect of homophily on the valuation, we assume that some of the referees are more innovative and have a tendency for more innovative ideas, while other referees are more orthodox in their essence and do not like unorthodox projects.

We call I_{ij} the *homophilic index* of scientist j w.r.t. project i, which is distributed normally on the range [0, Z]. We can compute homophily between the referee and the project as the similarity between the set of traits they have. Here we refer to cultural traits, which are characteristics of human societies that are potentially transmitted by non-genetic means and can be owned by an agent [12]. In general, similarly to T_{ij} we can split I_{ij} into two components:

$$I_{ij} = I_j + \gamma_{ij},$$

where γ_{ij} represents the homophily effect, while I_j represents the conformity, i.e., how receptive the referee is to innovative ideas. Even if conformity is just one of the cultural traits, we believe it is the most important one. When considering the

inventive element, I_{ij} , we assume that $\gamma_{ij} = 0$, i.e., homophily affects the valuation of referee in the following manner: (i) the more creative (or receptive to non-orthodox ideas) the referee is, the better he estimates the invention element; (ii) if the referee is more creative that the project proposed, he makes no error on the value; and (iii) the error is an increasing function of the difference between the true value and his creative possibilities. Therefore, we get that the valuation given by a referee is:

$$U_{ij} = \begin{cases} \alpha S_i + \gamma C_i + \beta I_i, & \text{for } C_i \leq T_j \text{ and } I_i \leq I_j, \\ \alpha S_i + \gamma T_j + \beta I_i, & \text{for } C_i > T_j \text{ and } I_i \leq I_j, \\ \alpha S_i + \gamma C_i + \beta I_j, & \text{for } C_i \leq T_j \text{ and } I_i > I_j, \\ \alpha S_i + \gamma T_j + \beta I_j, & \text{for } C_i > T_j \text{ and } I_i > I_j. \end{cases} \quad (2)$$

3.3 An example

In this position paper, we do not present a formal proof, and we present an example showing that innovative projects will not be chosen.

In the example we take k = 10 and h = 3. This is consistent with the acceptance rate for computer science conferences, where the median acceptance rate is 37% [10]. For the next version, we will take k = 100, and h = 30. We ignore α and β and γ at the moment, setting them to 1.

In Table 1, we present an example with 10 projects, and how they were ranked by the referees. There are two referees, which are different in their preferences. The first referee accepts to take time on each of the projects (His T_1 is 70, so that all projects with C lower than 70 will be judged accurately). But his homophilic index related to unorthodox views is low (His I_1 is 40, so that all projects with innovation index higher will not be judged accurately).

The second referee does not take much time (T_2=40), but his homophilic index is high (I_2 =120). The consequences of the fact that these two referees are quite dissimilar in the time to spend on referee and in their hemophilic index is that they will differ in their choice of projects.

Table 2: An example

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Rank	S_i	C_i	I_i	V_i	U_{i1} T_1 =70 I_1=40	U_{i2} T_2 =40 I_2=120	Average
1	40	0	0	40	40	40	40
2	50	30	0	80	80	80	80
3	50	40	0	90	90	90	90
4	50	50	20	120	120	110	115

5	55	40	80	175	135	**175**	**155**
6	30	80	66	176	140	136	138
7	70	65	42	177	**175**	152	**163**
8	45	75	60	**180**	**155**	145	150
9	40	60	80	**180**	140	**160**	150
10	70	80	120	**270**	**180**	**230**	**205**

3.4 Results of the referees ranking and projects chosen

Table 2 permits us to see the ranking of projects chosen by each referee, as well as their average. First, we see that Reviewer 1 will choose the 3 projects: 7, 8, 10. Then, we get that reviewer 2 will choose the 3 projects: 5, 9, 10.

What is striking is that both reviewers have in common only the project #10 (1/3!). Final choice of the projects using the average grade will include projects 5, 7, 10, while the best 3 projects are: 8, 9, 10.

So, this simple model permits to see that as emphasized by the NIPS experiment, the difference between referees is important. They all agree only on 30% of the projects.

The second result is that the referees should have chosen projects 8,9,10. In fact they have chosen, 5,7,10. A mistake of 66%!

4 Conclusions and future work

It is easy to see from our example that choosing the highest average grades, without taking into account the variance, might be a problem. In other words, such ranking-driven peer review leads to conformity, i.e., selection of less controversial projects. The results also resemble those of the NIPS experiment, as we see that reviewers disagree on the projects. This of course may influence the type of proposals scholars will propose, since scholars need to find financing for their research as discussed by Martin, [11]: "A common informal view is that it is easier to obtain funds for conventional projects. Those who are eager to get funding are not likely to propose radical or unorthodox projects. Since you don't know who the referees are going to be, it is best to assume that they are middle-of the road. Therefore, a middle of the road application is safer".

Interestingly enough, if the acceptance rate is indeed 3 projects out of 10, we end up selecting the projects 5 and 7, which are less innovative and clever than projects 8 and 9. Such low acceptance rate would still be higher than the ac-

ceptance rate in some H2020 calls[5] or 1-digit acceptance rate fashionable in some computer science conferences 10 years ago.

Our results from this example are similar to the facts presented in the literature. In future work we will simulate the model for bigger samples, possibly using real data from computer science conferences. Policy implications for improving the review processes will be presented in future work.

Acknowledgements

This research was funded by COST Action TD1306, New Frontiers of Peer Review (PEERE).

References

1. Squazzoni, F., Brezis, E., Marušić, A.: Scientometrics of peer review. **113**(1) (2017) 501–502
2. Kovanis, M., Trinquart, L., Ravaud, P., Porcher, R.: Evaluating alternative systems of peer review: a large-scale agent-based modelling approach to scientific publication. **113**(1) (2017) 651–671
3. Birukou, A., Wakeling, J., Bartolini, C., Casati, F., Marchese, M., Mirylenka, K.,Osman, N., Ragone, A., Sierra, C., Wassef, A.: Alternatives to peer review: Novel approaches for research evaluation. Frontiers in Computational Neuroscience **5** (2011) 56
4. Garfield, E.: Refereeing and peer review. part 3. how the peer review of ResearchGrant proposals works and what scientists say about it. Essays of an Information Scientists **10** (January 1987) 21+
5. Hallam, A.: Alfred wegener and the hypothesis of continental drift. Scientific American **232**(2) (1975) 88–97
6. Ragone, A., Mirylenka, K., Casati, F., Marchese, M.: On peer review in computer-science: analysis of its effectiveness and suggestions for improvement. Scientometrics **97**(2) (November 2013) 317–356
7. Francois, O.: Arbitrariness of peer review: A bayesian analysis of the NIPS experiment (July 2015)
8. McPherson, M., Lovin, L.S., Cook, J.M.: Birds of a feather: Homophily in socialnetworks. Annual Review of Sociology **27**(1) (2001) 415–444
9. Hirshman, B.R., Birukou, A., Martin, M.A., Bigrigg, M.W., Carley, K.M.: Theimpact of educational interventions on real and stylized cities. Technical Report CMU-ISR-08-114, Carnegie Mellon University (2008)
10. Malički, M., Mihajlov, M., Birukou, A., Bryl, V.: Peer review in computer science conferences published by springer. In: Eighth International Congress on Peer Review and Scientific Publication (PRC8), Chicago, IL, USA, September 10-12, 2017
11. Martin, B. 1997. "Peer Review as Scholarly Conformity" in Martin B. Suppression Stories, pp. 69-83.

[5] 1.8% according to https://www.linkedin.com/pulse/h2020-fet-open-18-chance-getting-funded-roy-pennings/

12. Birukou A., Blanzieri E., Giorgini P., Giunchiglia F. (2013) A Formal Definition of Culture. In: Sycara K., Gelfand M., Abbe A. (eds) Models for Intercultural Collaboration and Negotiation. Advances in Group Decision and Negotiation, vol 6. Springer, Dordrecht