[ξ] KSZI [ξ] AKTÁK

MTA KIK TTO műhelytanulmányok

2013/7

Social networks as a potential source of bias in peer review

Sándor Soós, Zsófia Vida, Beatriz Barros,
Ricardo Conejo, Richard Walker

➲ http://www.mtakszi.hu/kszi_aktak/

Social networks as a potential source of bias in peer review

Sándor Soós[1], Zsófia Vida[1], Beatriz Barros[2], Ricardo Conejo[2], Richard Walker[3]

[1]Dept. Science Policy and Scientometrics, Library and Information Centre of the Hungarian Academy of Sciences
[2]Dept. Computer Science and Languages, University of Malaga
[3]Frontiers Research Foundation

Abstract

There is broad consensus that the optimal way of evaluating journal and conference papers, research proposals, on-going projects and university departments is through "peer review". However, it is also recognized that classical peer review can be expensive, conservative and prone to bias. Proposals for reform include the use of author-blind and results-blind review, the removal of traditional reviewer anonymity, and the introduction of open review and community review. To date, however, there have been few attempts to compare the effectiveness of different review systems. In this paper, we present a methodology to test the presence of "social network effects" deriving from authors' prestige (centrality), and from their respective positions in co-authoring networks. We describe a pilot test of our methodology on two databases of review results, the first from an open access publishing house with an open review process (Frontiers), the second from seven computer science conferences, that used classical review (WebConf). We found a number of differences between our datasets. In the Frontiers system author centrality has no influence over review scores. In WebConf, there is a small but significant correlation between the two. This is not necessarily a sign of bias: authors with high centrality will plausibly produce better papers than authors with low centrality. In neither dataset did we find any link between author-reviewer distance and review results. This suggests an absence of favoritism. In both systems, reviewers belonging to the same sub-community as an author gave higher review scores than reviewers belonging to different sub-communities. This suggests that in a fair review system, the reviewers assigned to a paper short come both from the authors' own communities and from outside.

Introduction

Most members of the academic community believe that peer review represents the best possible way of evaluating research proposals and the outputs of scientific research. However, it also widely recognized that current systems are prone to various forms of bias, as reviewed in an earlier publication.

In view of these findings, the European SISOB project is developing a methodology to systematically evaluate possible biases in different kinds of peer review system. As part of this work, we have developed a toolkit of techniques to detect social network effects on peer review outcomes. In this paper, therefore, we describe methods to detect

whether reviewer outcomes are affected by authors' prestige (their "centrality" in their respective communities), by their social relationships with reviewers (the distance between authors and reviewers in coauthoring and in author-reviewer networks) and by their membership of specific sub-communities.

We go on to present findings from a pilot test of our methods on two large databases of detailed peer review results - previous analyzed for "classical" forms of bias. The first provides an exhaustive description of authors, reviewers and review results for all papers (N=4550) submitted to Frontiers, an open access publishing house, in the period 2007-2012. The second provides equivalent data for 1204 contributions to seven conferences in computer science, in the period 2002 to 2011. In what follows, we will refer to this data as the WebConf dataset. Our methodology allowed us identify similarities and points of difference between the Frontiers and WebConf data

We find that there are no detectable correlations between author centrality and review scores in the Frontiers data; however the WebConf data shows small but significant correlations. Neither dataset shows any significant relationship between author-reviewer distance and review scores. This suggests an absence of favoritism. However, scores from reviewers belonging to the same sub-community as authors are significantly higher than scores from reviewers coming from different sub-communities. The effect detected is almost certainly too small to affect the outcome of the review process. It is possible, however, that biases in other peer review systems are stronger than those registered for Frontiers and WebConf. In such cases, the methodology presented here has the power to detect the bias.

Materials and methods

Hypotheses

The study considered three hypotheses concerning biases related to direct and indirect social relations between authors and reviewers (distance between authors and reviewers in co-authoring networks, distance between authors and reviewers in author-reviewer networks)

- Mean reviewer scores for papers by a given author are directly related to the lead author's position (centrality) in these networks
- Mean reviewer scores for papers by a given author are inversely related to the reviewer's distance from the lead author
- Reviewers belonging to the same sub-community as an author will give higher scores than reviewers belonging to different communities.

A preliminary study of our two datasets showed that author-reviewer networks were largely unconnected. In what follows, we limit our analysis to co-authoring networks.

Data

The study was conducted on two databases used in a previous study of potential bias in the peer review process. Briefly

Frontiers. The Frontiers database included details of all scientific papers submitted to the Frontiers Open Access Publishing House (N=8,565) between June 25, 2007 and March 19, 2012, the name of the journal to which the paper was submitted, the article type (review, original research etc.) the name and institutional affiliations of the authors and reviewers of specific papers, individual reviewers scores and the overall review result (accepted/rejected). 2,926 papers had not completed the review process at the time of our analysis and were thus excluded from the analysis. In other 1,089 cases reviewers had not assigned a numerical scores to the paper, and could not be considered. Our final analysis used 4,550 papers. Most of the papers in the database come from the life sciences. The majority of authors and reviewers come from Western Europe and Northern America. However, the database contains a substantial number of authors and reviewers from other parts of the world.

WebConf. The WebConf database included details of contributions (N=1204) submitted to seven computer science conferences (AH2002, AIED2003, CAEPIA2003, ICALP2002, JITEL2007, SINTICE07, UMPAP2011) held in the period 2 N002-2011. Each of these conferences was managed using the WebConf system, developed at University of Malaga and managed by one of the authors. The data included name of conference, type of contribution, name, gender and institutional affiliations of the authors and reviewers of specific contributions, individual reviewers scores and the final decision (accepted/rejected). All the papers in the database are in the area of computer science. Three conferences (CAEPIA2003, SINTICE2007 and JTEL2007) mainly involved authors and reviewers from Spain or from Spain and Portugal. The other conferences were international conferences, involving authors and reviewers from all other the world. This inhomogeneity may have introduced bias into our sample.

Construction of co-authorship networks

To identify co-authorship relationships, we constructed a list of all authors in the Frontiers and WebConf databases who could be unambiguously associated with an author with a Scopus ID (Frontiers: N=8,690; WebConf: N=2149). For each author, we generated a list of other authors with whom the author had previously published at least one paper referenced in the Scopus database. On this basis, we identified co-authorship relationships present in the two databases. This made it possible to construct two undirected co-authoring graphs (one unweighted – Frontiers, and one weighted – WebConf). Both graphs included a giant "connected" component (Frontiers: N = 15 842, WebConf: N=543;) and disconnected "islands". The subsequent study was restricted to the giant connected component.

Calculation of centrality indicators

For each author, we computed the following centrality measures showing the author's position in the coauthoring network:

- Degree centrality,
- Betweenness centrality,
- Closeness centrality,
- Eigen centrality
- Page Rank centrality.

Each of these indicators provides a different operationalization of the informal concept of "being central". The recursive computation of the indicators in the second group is computationally very costly, which made it impractical for large networks. After exploratory studies, we concluded that the notion of author centrality is best captured by the Page Rank centrality indicator. Figure 1 illustrates this approach, using the page rank indicator to show the size of nodes in a sub-community in the Frontiers network

Calculation of paper centrality indicators

Most scientific papers have multiple authors and it is plausible that the most salient authors for reviewers are those with "strong centrality". To capture this intuition, we calculated a "paper centrality" indicator for each paper, defined as the maximum value of a given centrality indicator associated with any of the authors of the paper. Formally:

$$C_{paper} = max(AC_i)$$

Where C_{paper} is paper centrality, and AC_i is the value of a particular centrality measure for the ith author of the paper.

Community detection and recalculation of centrality measures

To test whether review results are affected by the centrality of authors in individual sub-communities, we submitted the giant component of the two networks to two community algorithms: "edge betweenness community detection" and "fast greedy community detection". The edge betweenness algorithm is based on the idea that an edge with a high betweenness value represents a bridge between two sub-communities. Each iteration of the algorithm removes the edges with the highest betweenness values, forming isolated groups (communities) at an optimally selected level. The "fast greedy community detection" partitions the graph according to a modularity maximization procedure. Both methods extract subgraphs based on the connectedness of actors, resulting in sub-communities with dense within-, but sparse between-connectedness.

Analysis of the Frontiers data yielded two relatively large communities (each containing around 1000 authors), and numerous smaller communities (with sizes of the order of

100, 10 and 1). In the analysis of the WebConf data, the two algorithms identified the same number of similarly sized communities (30).

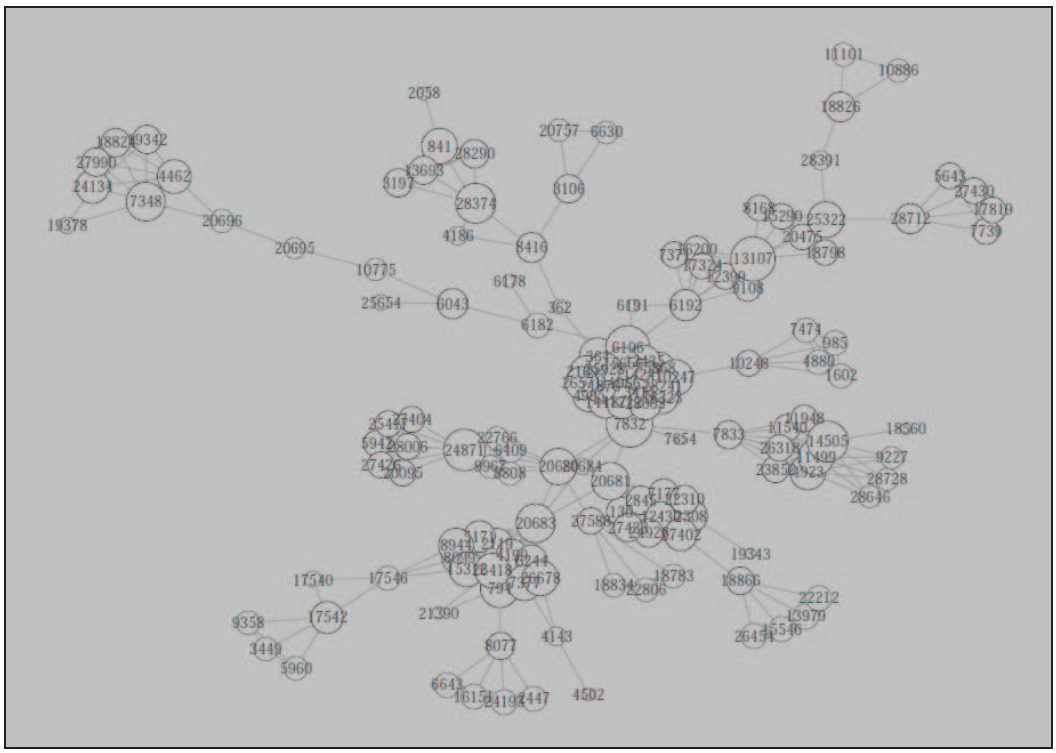

Figure 1: layout of a subcommunity in the Frontiers co-authoring network. size of node is proportional to page-rank centrality

Figure 2 shows the distribution of community sizes for the giant component of the Frontiers dataset. Figure 3 shows the equivalent data for the WebConf data, as determined by two different community detection algorithms.

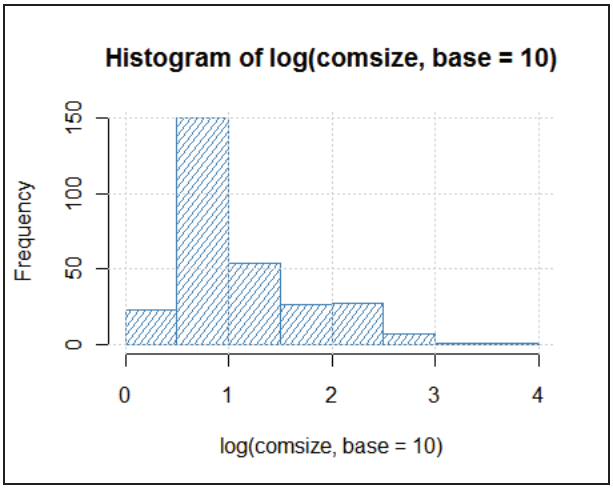

Figure 2: frequency distribution of communities in the giant component of the Frontiers database, by $\log_{10}$ of community size.

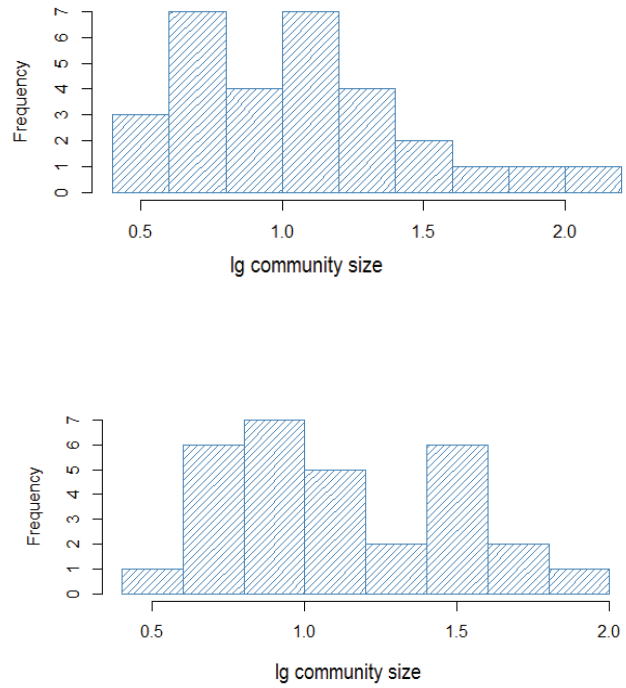

Figure 3: frequency distribution of communities in the giant component of the WebConf database, by $\log_{10}$ of community size. TOP: fast greedy algorithm; BOTTOM: fast edge betweeness algorithm

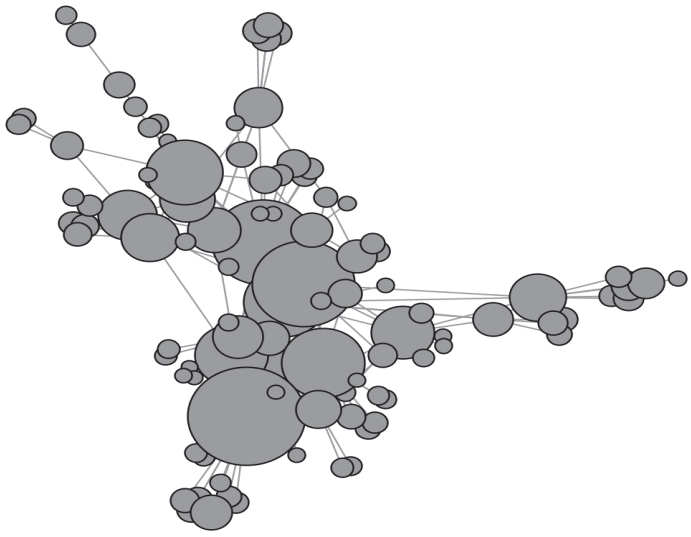

Figure 4: largest community detected by the "fast greedy" community detection algorithm when applied to the WebConf database –size of nodes is proportional to the Page Rank value

After application of the community detection procedure, centrality measures were recalculated using the methods described in the previous section.

Results

Author centrality affects review scores

- Design 1

For each of the two datasets, we began by calculating paper centrality indicators using the methods described above. We then computed the rank correlation between (a) different measures of paper centrality and (2) mean reviewer scores for the paper. The Frontiers dataset shows no sign of correlation. By contrast, the WebConf data show clear, but small correlation between centrality and score (see Table 1)

	bw	D	C	E	P	Score
bw	1	0.83	0.68	0.55	0.40	-0,01
d		1	0.64	0.53	0.41	0.01
c			1	0.87	0.20	-0,00
e				1	0.16	-0.01
p					1	-0.01

	bw	D	C	E	P	Score
bw	1	0.83	0.72	0.69	0.34	0.10
d		1	0.74	0.71	0.39	0.22
c			1	0.87	0.58	0.22
e				1	0.43	0.12
p					1	0,18

Table 1: cross-correlation between different measures of author centrality and mean review scores. TOP: Frontiers; bottom: webconf (bw: betweenness centrality; D:degree centrality; C: closeness centrality; e: eigen centrality; P: Page rank centrality)

- Design 2

In the light of the results for design 1, we conjectured that large-scale patterns might be suppressing possible network biases. We therefore conducted a second experiment, in which we focused on smaller communities within the whole graph.

First we identified sub-communities using the methods described earlier We then repeated the experiment described in Design 1 for each individual sub-community, The results were qualitatively similar to those obtained with design 1 (data not shown). As in the previous experiment, the Frontiers data showed no sign of correlation. The WebConf data showed signs of a small but significant correlation.

Author-reviewer distance affects review scores

- Design 1

To test our second hypothesis, we interpreted "distance" as the length of the shortest path connecting the first author of a paper and a specific reviewer. We then compared the distance between the author of a paper and the reviewer of the paper, against the score given by the reviewer to the paper. Analysis of the Frontiers dataset using this measure showed no correlation between the two variables either for Frontiers (rank correlation = -0,05). In the WebConf case, we applied a measure of "paper distance" analogous to the measure of paper centrality. Again we found no correlation (rank correlation=-0.06). We also tested a second method of analysis in which we tested the rank correlation between review scores and the author with the shortest path to the reviewer. No significant correlation was detected (rank correlation=-0.07)

- Design 2

In a second experiment we interpreted reviewer–author proximity as a binary variable, indicating whether the two actors belong to the same sub-community. Working with the same sub-communities extracted earlier, we partitioned author–reviewer pairs into two sets: (1) author-reviewer pairs in which the author and the reviewer belonged to the same subgraph, (value = 1) (2) author-reviewer pairs in which the author and the reviewer belonged to different subgraphs (value = 0)

For robustness, scores were represented as categorical variables. Values less than or equal to the mean score for the dataset were labeled "low"; scores above the mean were labeled "high". We then used a Chi-squared test to test for independence between co-membership and score category. Finally, as a last refinement, we used a simple one-way ANOVA and a two-sample t-test to compare groups (1) and (2) with respect to raw reviewer scores. Interestingly, scores were significantly higher when authors and reviewers were members of the same sub-community than when they belonged to different communities. The effect was larger for WebConf than for Frontiers, but relatively small in both cases.

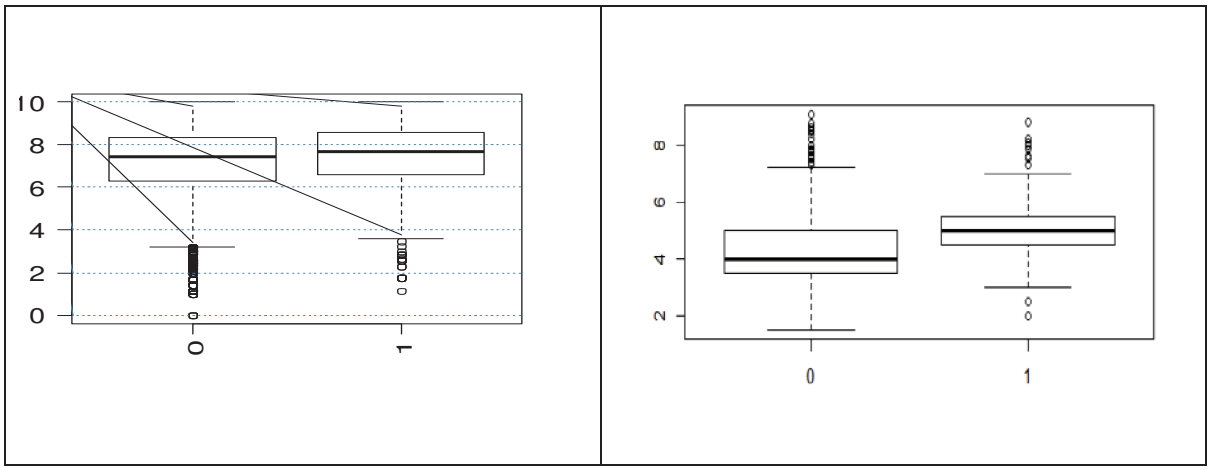

Figure 5: mean scores review scores when authors and reviewers are in different sub-communities (column 0) or in the same community (column 1). comparison between frontiers (left) and WebConf (right). the difference is larger for WebConf than for frontiers.

Discussion

Our study shows that the centrality of a paper's authors has no measurable effect on Frontiers review scores. By contrast, the data for the WebConf system shows a small but significant correlation. The correlation in the WebConf data can be explained in two ways. Either authors with high centrality produce better papers – a plausible supposition – or reviewers have a (small) bias towards papers produced by prominent authors - a result that would confirm previous suggestions that peer review suffers from "cronyism"[1].

Unfortunately, it is not possible to discriminate between these interpretations using observational data alone. As we have suggested elsewhere, the only way to make such a distinction would be through experiments, in which some observers have access to author names and some are blinded. A null result would suggest that observed differences are due to genuine differences in quality. *Vice versa*, rejection of the null hypothesis would be evidence of bias. Such work lies outside the scope of this paper.

In neither of our datasets were review scores correlated to author-reviewer distances in co-authoring networks. This suggests an absence of favoritism. However, in both systems, reviewers belonging to the same community as authors gave (slightly) higher scores than reviewers coming from different sub-communities - an effect that was stronger for WebConf than for Frontiers. This effect does seem to signal some kind of bias – perhaps because reviewers are most familiar with the language, style and concepts of their own sub-communities. Editors should take care that the reviewers assigned to an author include some who belong to the author's own community and some from elsewhere.

In both the Frontiers and the WebConf reviewing systems, the final decision to accept an author's contribution belongs to the editor and does not depend directly on the scores assigned by reviewers. Given that the potential biases detected in our study were very small, they probably had little effect on publication decisions. However, this is not necessarily true for all peer review systems. Our results suggest that the SISOB methodology has the power to detect potential issues if they are present.

We are currently working to implement the methods described in this paper into software tools, available through the SISOB workbench. The current version of the workbench is available at http://sisob.lcc.uma.es/workbench.

Acknowledgement

The work reported in this manuscript was supported by the European Commission under the FP7 Science in Society Grant No. 266588 (SISOB project).

---

[1] Travis, G.D.L. and H.M. Collins, *New Light on Old Boys: Cognitive and Institutional Particularism in the Peer Review System.* Science, Technology, & Human Values, 1991. **16**(3): p. 322-341.