# Significance of scientific peer-review system: A case study of the Journal of High Energy Physics

Sandipan Sikdar
Indian Institute of Technology Kharagpur
Kharagpur,India 721302
sandipansikdar@cse.iitkgp.ernet.in

## ABSTRACT

The scientific 'peer-review system' has long been relied upon to determine the quality of a scientific contribution. It is primarily responsible for bringing quality research to the notice of the scientific community and also preventing flawed research from entering into the literature. With numerous cases coming up where the peer-review system has failed in correctly determining the quality of research, its need is now highly debated. The research community is hence reaching a consensus that although the peer-review system is important it is nonetheless flawed. A very pertinent question is then "Can peer-review system be improved?". We attempt to present an answer to this question by considering a massive dataset of around $29k$ papers with roughly $70k$ distinct review reports together consisting of $12m$ lines of review text from the Journal of High Energy Physics (JHEP) between 1997 and 2015. Our contributions are fourfold - (i) we review the performance of the editors and reviewers who are the most important pillars of the peer-review-system and leverage anomaly detection techniques to identify the under-performing ones (this work has been presented at **CIKM 2016**), (ii) we introduce a novel *reviewer-reviewer interaction network* (an edge exists between two reviewers if they were assigned by the same editor) and show that surprisingly the simple structural properties of this network such as degree, clustering coefficient, centrality (closeness, betweenness etc.) serve as strong predictors of the long-term citations (i.e., the overall scientific impact) of a submitted paper (this work was presented at **JCDL 2017**). (iii) we propose a framework based on genetic algorithms to recommend a referee group in case of a multiple referee system (this work is under progress), (iv) we propose a framework for recommending referees given a submission (we plan to accomplish this in future). We genuinely believe that our contributions would be immensely useful in improving the scientific peer-review system.

## CCS CONCEPTS

•**Applied computing → Digital libraries and archives;**

## KEYWORDS

citations; peer-review system; prediction; anomaly


*Conference'17, Washington, DC, USA*

DOI: 10.1145/nnnnnnn.nnnnnnn

**ACM Reference format:**
Sandipan Sikdar. 2016. Significance of scientific peer-review system: A case study of the Journal of High Energy Physics. In *Proceedings of ACM Conference, Washington, DC, USA, July 2017 (Conference'17),* 6 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

## 1 INTRODUCTION

Before the contributions of a paper are brought to the notice of the research community, it has to usually pass through a peer-review process, whereby, the correctness and the novelty of the paper is judged by a set of knowledgeable peers. The primary intent of which is to prevent flawed research from getting into mainstream literature [10].

**Debates on peer-review system:** The effectiveness of this system has been put to question in numerous cases ([9, 11, 13]) with flawed research being added to literature while significantly novel contributions being rejected. That the reviewers often fail to reach consensus ([6]) and that rejected papers are often cited more in the long run ([3]), have already been pointed out. Although there have been several proposals to make it more effective ([4, 7]), the research community is coming to a conclusion that although peer-review system is indispensable it is nonetheless flawed [2].

**Peer-review dataset:** We consider a set of around $29k$ papers along with roughly $70k$ unique review reports containing $12m$ lines of review text submitted to the Journal of High Energy Physics (JHEP) between 1997 and 2015. We would like to point out here that this dataset is unique as well as very rich and we do not know of any other work that presents such a large-scale analytics of an equivalent dataset. Informed with the details of the number of reviews per paper, the content of the review reports and the citation counts we perform, for the first time, a series of systematic measurements to determine whether the peer-review process is indeed able to correctly differentiate between high impact contributions and the rest.

**Entities in the peer-review system:** The effectiveness of the peer-review system is dependent directly on the knowledge and training of the editors and reviewers. The editor is responsible for identifying the correct set of referees who can give expert comments on the submission and also for taking the final decision whether a particular paper should be accepted or rejected. The assisting reviewers send their views on the paper in the form of a report. This report is an important part of the whole process as it not only forms the basis of the acceptance/rejection decision but is also sent to the authors for further improvement of the paper.

**Citation impact of accepted papers:** Assuming that citation count of a paper is representative of its overall quality, we observe that on average those papers which were accepted at JHEP after passing through the peer-review process, are cited more often compared to those which got rejected at JHEP and eventually got accepted at a different venue. While this is true for the majority, there are a few exception cases where either a rejected paper is found to receive high citations or an accepted paper is found to receive (almost) no citation.

**Characterizing and identifying anomalous reviewers and editors:** Ideally impactful papers should be accepted for publication while flawed works should be rejected. We quantify the impact of a paper by the citations it garnered. Thus, a paper getting accepted but managing to garner very less or no citation should be attributed to the anomaly of the system; similarly, a paper getting rejected by the peer-review-system but garnering large number of citations in the long run is also an anomaly. We further investigate the reasons behind the anomalous behaviors ([5]) of the reviewers and editors as they are the most important entities of the peer-review system. In fact we observe that **editors** who (i) are assigned papers more frequently, (ii) select reviewers from a very small set, (iii) assign themselves as reviewers more often (rather than assigning other reviewers) are often under-performers and hence anomalous. Similarly, for **reviewers** we observe the following behaviors to be anomalous - (i) frequent assignments, (ii) very small or very large delay in sending reports, (iii) reviewing papers in very specific topics, (iv) assignments from a very small set of editors or in some cases a single editor, (v) very high or very low proportion of acceptance, (vi) large delay in informing the editor about inability to review and (vii) often declining to review. Papers accepted by reviewers with such behaviors are often low cited while those rejected by them are often highly cited. Finally we use $k$-means clustering [8] to classify normal and anomalous editors and reviewers. We find 26.8% of the editors and 14.5% of the reviewers to be anomalous.

**Predicting future impact of papers:** We introduce reviewer-reviewer interaction network built as the one-mode projection of the editor-reviewer bipartite network and show that the network related structural features such as the degree, the clustering coefficient and the centrality values (closeness, betweenness etc.) of the reviewer nodes in the reviewer-reviewer network strongly correlate with the long-term citations received by the papers these reviewers refereed. We also build a set of supporting features based on the various characteristics of the papers submitted as well as the authors and the referees of the submitted papers. Based on the network features built above, we propose a supervised model which quite accurately predicts ($R^2$ = **0.79**, $RMSE$ = **0.496**) the long-term citation of a paper. In addition, if we also include the supporting features into the model we obtain further gains ($R^2$ = **0.81**, $RMSE$ = **0.46**).

**Recommending reviewer groups for multi-reviewer system:** We observe that in multi-refereed papers the referees often fail to reach consensus. As an immediate next step following the previous observations, we investigate the review reports of the multi-refereed papers. In fact in terms of report length, sentiment and content the referees differ in almost 30% of the cases on average across the two datasets. However, when we dig a little deeper, we find that the real impactful papers are multi-reviewed and the least-cited papers are single-reviewed. The multi-reviewer system fails due to several reasons (a). overburdening the reviewers (b). reviewers tendency to be too critical or too liberal. The discordance also occurs when such reviewers are grouped together, which perhaps leads to acceptance of paper with- out due diligence. We hypothesize that multi-referee systems fail due to lack of proper selection and the assignment of the referees. We hence proceed to propose a systematic scheme for recommending reviewer groups to the editor. More specifically, given a paper, its topic and a re- viewer pool with past information our algorithm is able to recommend a set of referee groups to assist the editor is as- signing referees.

**Developing a reviewer recommendation system:** Leveraging all the above observations, we finally plan to deploy a complete reviewer recommendation system that can assist the editors in assigning referees to a particular submission. Note that this system would differ from the one mentioned earlier in the sense that it will consider both the single and multi-reviewer cases at the same time.

## 2 RESEARCH OVERVIEW AND WORK DONE

As mentioned earlier our research contributions are fourfold - (i) characterizing and identifying anomalous editors and reviewers (presented at CIKM 2016) (ii) predicting future impact of papers (to be presented at JCDL 2017) (iii) recommending reviewer groups in multi-reviewer system (under progress) (iv) developing a full-fledged peer-review system (to be taken up in future) In this section we will elaborate on the first two while for the rest we defer our discussion to the next section since they are planned for future.

### 2.1 Characterizing and identifying anomalous editors and reviewers:

In the peer-review process each submission is assigned to an editor who in turn assigns one or more reviewers with the task of judging the quality of the contributions of the submitted paper. The reviewer submits a report to the editor who in turn takes the final decision as to accept or reject the paper based on the report. Therefore, the editors and the reviewers are the two important entities of the peer-review system and they are mainly responsible for ensuring that flawed research does not get into the literature while at the same time correctly identify impactful contributions for publication. So in our setting we define the following two cases to be anomalous -
(i) Accepted papers having low citation (research wrongly judged as impactful).
(ii) Rejected papers having high citation (quality research wrongly judged as flawed).
In this section we look into the anomalous behavior of the two important entities of the peer-review process: (i) the editors and (ii) the reviewers.

We begin by analyzing the anomalous behavior of the **editors**. We define the behavior of an editor to be anomalous if the papers assigned to her are on average cited less when accepted or are cited more when rejected. In specific, we investigate different factors related to the editor that can lead to such anomaly.

*2.1.1 Mean Editor Assignment Time (MEAT).* For each editor we obtain the time span (in days) between any two consecutive assignments and calculate the average time span between the two assignments. Formally, we define for editor $i$, $MEAT_i$ as

$$MEAT_i = \frac{1}{n-1} \sum(\delta_{j+1} - \delta_j)$$

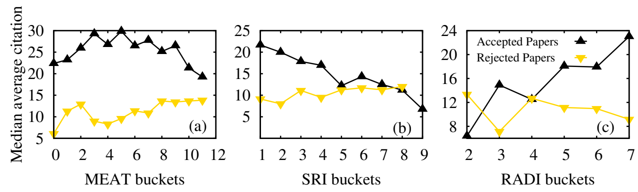


**Figure 1: (a) Median Average citation (MAC) versus *MEAT*. *MEAT* values are bucketed into 12 bins of equal size with range(1, 498.8).(b) MAC versus *SRI* and (c) MAC versus *RADI*. For both (b) and (c), the x-axis values are bucketed by values corresponding to (≥ 0 and < 0.1), (≥ 0.1 and < 0.2) and so on.**

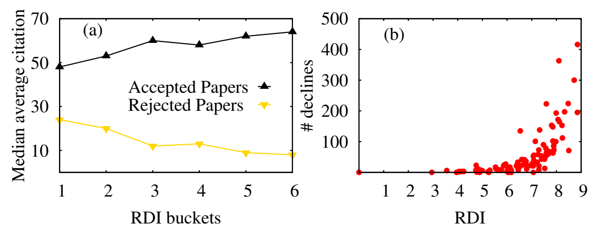


**Figure 2: (a) Median Average citation versus *SRI*. *SRI* values are bucketed by values corresponding to (≥ 0 and < 0.1), (≥ 0.1 and < 0.2) and so on. (b) *RDI* versus number of declines. Increasing trend indicates higher the *RDI*, higher is the number of declines.**

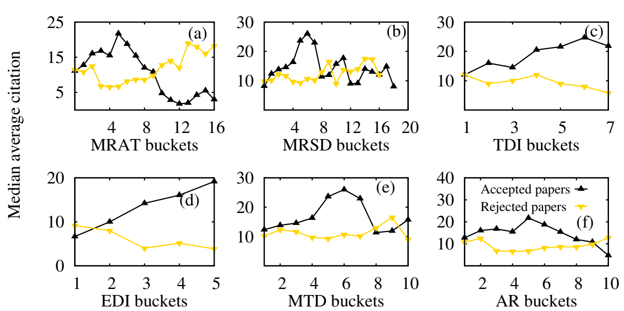


**Figure 3: (a) Median Average citation (MAC) versus *MRAT*. *MRAT* values are bucketed into 20 buckets of equal size with range(1,498.8),(b) MAC versus *MRSD* (c) MAC versus *TDI*, (d) MAC versus *EDI*, (e) MAC versus *MTD* and (f) MAC versus *AR*. For both (c),(d) and (e), the x-axis values are bucketed by values corresponding to (≥ 0 and < 0.1), (≥ 0.1 and < 0.2) and so on. For (b) and (f) values (x-axis) are divided into 10 buckets of equal size.**

where $n$ is the total number of assignments to the editor $i$ and $\delta_j$ is the date of the $j^{\text{th}}$ assignment. In figure 1(a) we bin the editors based on the *MEAT* and calculate the median average citation of the papers assigned to the editors in each bin. We observe that for accepted papers very low or very high *MEAT* values lead to lower citations. An exact opposite behavior is observed for rejected papers. This indicates that editors who are assigned time and again (low *MEAT*) or rarely (high *MEAT*) often fail to judge the quality of the papers assigned to them.

### 2.1.2 *Self Review Index (SRI).*

**Self Review Index (SRI)** measures the fraction of papers for which the editor assigned herself as the reviewer. Formally, for an editor $i$, we define $SRI_i$ as

$$SRI_i = \frac{\varrho_i}{\rho_i}$$

where $\rho_i$ is the number of papers $i$ was assigned as editor while $\varrho_i$ is the number of papers $i$ assigned herself as reviewer. We observe that with increasing values of *SRI* the median average citation for accepted papers decreases while that for rejected papers increases (refer to figure 1(b)).

### 2.1.3 *Referee-Author pair Diversity Index (RADI).*

We observe that editors in numerous cases assign papers from a certain author to only a certain reviewer. To investigate whether this allows for less impactful research from this author getting accepted, we define a metric which we call **Referee-Author pair Diversity Index (RADI)**. Formally we define for editor $i$, the $RADI_i$ score as

$$RADI_i = -\sum_{j,k} p_{j,k} \log p_{j,k}$$

where $p_{j,k}$ denotes the proportion of times a paper from author $k$ was assigned to reviewer $j$ by the editor $i$. In figure 1(c) we bin the editors based on the *RADI* and calculate the median average citation of the papers assigned to the editors in each bin. We observe that more the diversity score higher is the citation of the accepted papers and correspondingly lower is the citation of the rejected papers.

### 2.1.4 *Referee Diversity Index (RDI).*

As a following step, we check whether an editor always chooses from a fixed set of reviewers or a diverse set of reviewers while making a paper assignment and, more importantly, does this influence the performance of the editor in terms of the impact of the reviewed paper. We define for each editor($i$) a metric called **Referee Diversity Index** (**$RDI_i$**) as -

$$RDI_i = -\sum_j p_j \log p_j$$

where $p_j$ denotes the proportion of times reviewer $j$ was assigned a paper by editor $i$. More diverse the set of reviewers higher is the score. In figure 1(b) we bin the editors based on the *RDI* and calculate the median average citation of the papers assigned to the editors in each bin. We observe that more the diversity score, higher is the citation of the accepted papers and correspondingly lower is the citation of the rejected papers.

The dataset allows us to find out the cases when the reviewer declined to review a paper on being assigned by an editor. We observe that editors with high *RDI* are also declined more often. In figure 2(b) we plot *RDI* value and the number of declines for each editor. An increasing trend indicates that more diversely the editor tries to select reviewers more she gets declined by the reviewers. This in many cases may force the editor to be less proactive and always select from a specific set of 'reliable' referees.

As in case of the editors, we also investigate different factors that could be indicative of anomalous behavior in **reviewers**.

*2.1.5 Mean Reviewer Assignment Time (MRAT).* This is essentially same as MEAT. For a reviewer $i$, we define $MRAT_i$ as

$$MRAT_i = \frac{1}{n-1}\sum(\delta_{j+1} - \delta_j)$$

where $n$ is the total number of assignments of reviewer $i$ and $\delta_j$ is the date of the $j^{\text{th}}$ assignment. In figure 3(a) we plot $MRAT$ (binned) and median average citation of the papers reviewed for each reviewer. We observe that papers reviewed by reviewers with low $MRAT$ (high frequency of assignment) tend to be cited less and increases as $MRAT$ increases. This is followed by again a steep decrease in citation. This indicates that the reviewers assigned very frequently are often less reliable while those assigned only occasionally are also not likely to correctly judge the quality of the paper.

*2.1.6 Mean Report Sending Delay (MRSD).* We argue that the time taken by a reviewer to send back the review report could be an indicator of his performance. If a reviewer on average sends back the review very quickly it is highly likely that the review was done in a hurry. Similarly, if the report was sent after being reminded by the editor numerous times, it is also highly likely the review report could be anomalous. For a reviewer we calculate the time delay between the date of her assignment and the date she sent back the report for each of her assignments. To measure $MRSD$ we calculate the mean value of all the delays. Note that we do not consider the assignments which the reviewer declined. Formally, for a reviewer $i$, we define $MRSD_i$ as

$$MRSD_i = \frac{1}{n}\sum(\delta_i - \Delta_i)$$

where $n$ is the total number of assignments, $\Delta_i$ is the date of assignment and $\delta_i$ is the date when the report was received by the editor. On plotting against median average citation we observe a similar trend as was observed in case of $MRAT$ (refer to figure 3(b)). Papers reviewed by reviewers with low $MRSD$ value are often less cited, indicating that reviewers sending back their report very quickly often do it in a hurry and fail to correctly judge the quality of the paper while those taking very long to send report are prone to failure as well.

*2.1.7 Topic Diversity Index (TDI).* JHEP associates with each submission a set of keywords which roughly indicates the domain of the work. We use these associated keywords as a proxy for topic. For each reviewer, we segregate all the keywords of the papers reviewed by her which we call the keyword corpus for the reviewer. Formally for a reviewer $i$, we define $TDI_i$ as

$$TDI_i = -\sum_j p_j \log p_j$$

where $p_j$ is the proportion of keyword $j$ in the keyword corpus for reviewer $i$. We segregate the reviewers based on the diversity score and calculate the median average citation of the papers reviewed by them. We observe that the median average citation for reviewers with low $TDI$ are low mainly because the number of papers reviewed by them are also less. The value increases with increasing $TDI$ (refer to figure 3(c)). The reviewers with low $TDI$ are often the ones who have reviewed a very small number of papers while the reviewers with high $TDI$ are mostly assigned papers by a large number of editors.

*2.1.8 Editor Diversity Index (EDI).* Reviewers could be selected for review by a large set of editors or could only be selected by a single or a small set of editors. We check whether a reviewer selected by many editors is more reliable compared to one who is selected by a single or a very small set of editors. To this aim we assign each reviewer a score called Editor Diversity Index, $EDI_i$ which is defined as

$$EDI_i = -\sum_j p_j \log p_j$$

where $p_j$ represents the proportion of times reviewer $i$ was assigned by editor $j$. We segregate the reviewers based on $EDI$ and calculate the median average citation of the papers reviewed by them. We observe that as $EDI$ increases median average citation also increases (refer to figure 3(d)) indicating that reviewers assigned by multiple editors are often more reliable.

*2.1.9 Mean Time to Decline (MTD).* We further investigated the cases where the reviewer declined the assignment. In specific, we calculated the time delay (in days) between the date she was assigned and the date she conveyed her decision of declining to review. For each reviewer we define **Mean Time to Delay**, $MTD_i$ as

$$MTD_i = \frac{1}{d}\sum_j(\mu_j - \Delta_j)$$

where $d$ is the number of assignments that reviewer $i$ declined and $\mu_j$ and $\Delta_j$ are respectively the date of assignments and date of reply for paper $j$ by reviewer $i$. We segregate the reviewers based on their $MTD$ values and calculate the median average citation. We observe that the reviewers who delay often in reporting their decision to the editor of being unable to review usually tend to fail in judging a paper quality when they do review (refer to figure 3(f)).

*2.1.10 Acceptance Ratio (AR).* Acceptance Ratio (*AR*) of a reviewer is defined as the proportion of papers accepted by the reviewer. For a reviewer $i$, $AR_i$ is formally defined as

$$AR_i = \frac{a_i}{a_i + r_i}$$

where $a_i$ and $r_i$ respectively denote the number of papers accepted and rejected by reviewer $i$. We observe that reviewers with high *AR* often accept less impactful papers while reviewers with very low *AR* often fail to identify quality research (refer to figure 3(e)). Note that the reviewers are segregated based on their respective *AR* values while the median average citation is calculated. They are segregated into bins based on the *AR* values where typically the bins are ($\geq 0$ and $< 0.1$), ($\geq 0.1$ and $< 0.2$) and so on.

To detect anomalies we use the $k - means$ clustering setting with $k = 2$. In fact the anomalous editors and reviewers account for 26.8% and 14.5% of the total editors and reviewers respectively. **This work was presented at CIKM 2016 [12]**.

## 2.2 Prediction of future impact of research contributions

In this work we investigate several factors that could be indicative of the future impact of research papers. Note that the impact is quantified by the number of citations it accrues in future.

*2.2.1 Reviewer-reviewer interaction network.* We construct a reviewer-reviewer interaction network and show that its properties are linked to the future scientific impact of a paper (measured in terms of the

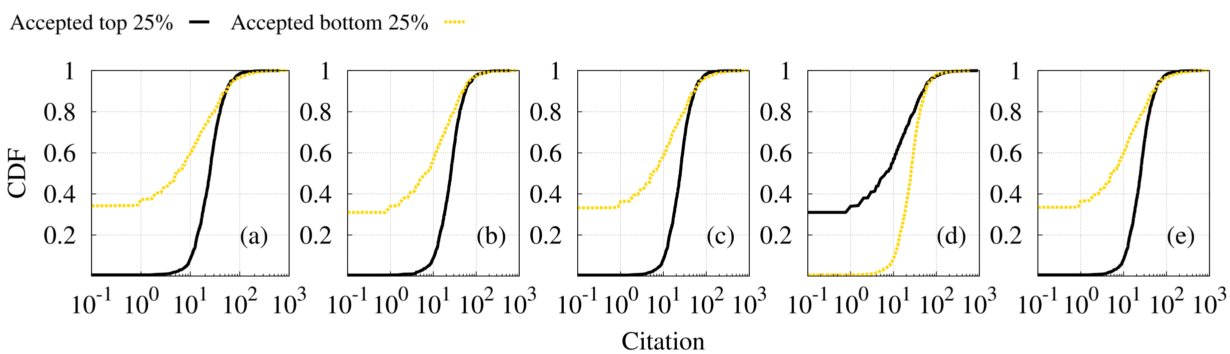


**Figure 4: Cumulative distribution function (CDF) of citations received by the papers (accepted) reviewed by referees in top 25% and bottom 25% reviewers ranked according to (a) degree, (b) betweenness centrality, (c) closeness centrality (d) clustering coefficient values and (e) PageRank in the reviewer-reviewer interaction network.**

cumulative citation count). In specific, we find that the position of the assigned reviewer in the network (measured in terms of degree, centrality, clustering coefficient and PageRank) could be used to predict the long term citation of the paper. The reviewer-reviewer interaction network is created with each node representing a reviewer and an edge exists between two reviewers if they have been assigned by at least one common editor. Note that there are 4035 unique reviewers in the system each of which form a node in this network. In specific we look into the following properties -
(i)**Degree:** Degree of a node $v$ is the number of other nodes it is connected to in the network. A node with a higher degree in the reviewer-reviewer interaction network would indicate (i) assignment from multiple editors, (ii) assignment from a reputed editor (with large number of assignments) which in turn would indicate the reputation of the reviewer. To verify our hypothesis, we rank the reviewers based on their degree in the network and calculate the mean citation of the papers reviewed by the reviewers in the top and the bottom 25% of the rank list. We observe that the papers reviewed by the top 25% reviewers receive much higher citations than the those reviewed by the bottom 25% reviewers (refer to fig. 4(a)).
(ii)**Betweenness centrality:** Betweenness centrality of a node quantifies the position of a node based on the number of shortest paths the node is part of. For every pair of nodes in the network there exists a shortest path between them. Betweenness centrality of a node ($v$) is the fraction of all such paths that pass through $v$. In the reviewer-reviewer interaction network, a high centrality value would indicate assignment by multiple editors and that this node acts as a bridge between them. We again rank the reviewers based on the betweenness centrality values and calculate the average citation of the papers. We find that the papers accepted by the top 25% reviewers tend to be cited more compared to those accepted by the bottom 25% (refer to fig. 4(b)).
(iii)**Closeness centrality:** Formally closeness centrality of a node in a network is the inverse of the sum of length of its shortest path to all other nodes in the network. Hence higher centrality value indicates that the node is more closer to all other nodes in the network. In the reviewer-reviewer interaction network, a reputed reviewer will be assigned by multiple reviewers and hence will be closer to the other reviewers in the network. This is represented in fig. 4(c), where we show that the papers accepted by top 25% most central reviewers are cited more often compared to the bottom 25% reviewers.
(iv)**Clustering coefficient:** Clustering coefficient of a node is measured as the fraction of connections among the neighbors of the node. For the reviewer-reviewer interaction network, every reviewer assigned by a common editor is connected to every other reviewer in the network. A reviewer assigned by many editors would actually act as a bridge between two cliques and hence would have a lower clustering coefficient value compared to a reviewer who is part of a single clique (always assigned by a single editor). This is further demonstrated in fig. 4(d) where we observe that the papers accepted by reviewers having lower clustering coefficient tend to be cited more.
(v)**Page rank:** PageRank is a link analysis based algorithm that calculates for each node its relative importance within the network. Specifically, PageRank outputs a probability distribution which is used as the likelihood of a random walker to end up in a specific node. Further analysis indicates that the papers accepted by the top 25% reviewers (based on PageRank) are cited more often compared to those accepted by the bottom 25% reviewers (refer to fig. 4(e)).

2.2.2 *Supporting features:* We further look into set of supporting features which would further help in predicting the future impact. We next elaborate on each of them.
(i)**Paper based features:** In specific we look into two factors in this category (a) number of review rounds which is the number of rounds of review a paper went through before acceptance (b) team size i.e., the number of contributing authors.
(ii)**Review report based features:** We analyze if there are certain features that could be extracted from the reports sent by the reviewers which are indicative of the quality and hence long-term citation of the paper. The features we look into are - (a) length of the report in terms of number of words, (b) sentiment of the report

(positive/negative/neutral), (c) quality indicators extracted through LIWC text analysis tool.
(iii)**Author based features:** In specific we look into (a) reputation of the author measured in terms of the accept to submission ratio and (b) author productivity measured in terms of mean time between two successive submission.
(iv)**Reviewer based features:** We investigate whether there are certain reviewer based features which could indicate long term impact. Mainly we look into - (a) acceptance ratio which is measured as the acceptance to assignment ratio, (b) time since last assignment and (c) delay in submission of the report.
**Network features only:** Considering only the network features, we obtain the best result using support vector regression (RBF kernel) with parameters $C = 100$ and $\gamma = 0.01$. We perform a 10-fold cross-validation and obtain a high $R^2$ of **0.79** and a low *RMSE* of **0.496**.
**Network + supporting features:** Considering both the network and the supporting features we obtain a further overall improvement. In specific, using support vector regression (RBF kernel) we obtain a high $R^2$ of **0.81** and a low *RMSE* of **0.46**. The parameters were set as parameters $C = 100$ and $\gamma = 0.02$. We further calculate the $F$-Statistic values for all the features used in the regression task and observe that the network features, are in general, are more suited to the task of prediction.

Thus our system is correctly able to predict the citation rank of the paper. We believe our system could be useful in assisting the editors in deciding whether to accept or reject the papers especially in cases where the reports are contradictory. **This work was presented at JCDL 2017 [1]**.

## 3 FUTURE DIRECTION AND EXPECTED CONTRIBUTION

In future we are intended at developing reviewer recommendation system that could assist editors in assigning reviewers to a particular submission. Due to unavailability of any explicit ground truth developing such a system is indeed a difficult task. We have initially looked into a less constrained problem whereby given a submission we **(iii) recommend reviewer groups in multi-reviewer system** i.e., which set of reviewers to group such that overall performance of the system improves. The primary motivation behind designing such a system is the observation that multi-refereed papers the referees often fail to reach consensus. In fact we looked into several review reports of the multi-refereed papers and have been able to establish this lack of consensus among reviewers. However, when we dig a little deeper, we find that the real impactful papers are multi-reviewed and the least-cited papers are single-reviewed. The multi-reviewer system fails due to several reasons (a). overburdening the reviewers (b). reviewers tendency to be too critical or too liberal. The discordance also occurs when such reviewers are grouped together, which perhaps leads to acceptance of paper with- out due diligence. Contrary, we find that even when underperforming reviewers are grouped with performing reviewers, the overall quality of acceptance improves. From the above observations we hypothesize that multi-referee systems fail due to lack of proper selection and the assignment of the referees. We plan to propose a systematic scheme for recommending reviewer groups to the editor. More specifically, given a paper, its topic and a reviewer pool with past information our algorithm is able to recommend a set of referee groups to assist the editor is as- signing referees. In fact we plan to propose a genetic algorithm based scheme (shown to be very effective in identifying groups from a population such that the overall efficiency improves) aimed at recommending reviewer groups.

Note that the above system is only suitable for multi-reviewer system. But our initial results suggest that in many cases single reviewer is more effective than a multi-reviewer system. We hence plan to come up with a full-fledged **(iv) reviewer recommendation system** which can decide whether to assign a single or a multiple reviewer for a given submission and further recommend a referee(s) suited for the task. We also plan to deploy our system and further improve it based on feedbacks from the publishing house.

## REFERENCES

[1] 2017. *2017 ACM/IEEE Joint Conference on Digital Libraries, JCDL 2017, Toronto, ON, Canada, June 19-23, 2017.* IEEE Computer Society. http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=7988694
[2] Peter Bacchetti. 2002. Peer review of statistics in medical research: the other problem. *British Medical Journal* 324, 7348 (2002), 1271.
[3] R D Braatz. 2014. Papers Receive More Citations After Rejection [Publication Activities]. *Control systems, IEEE* 34, 4 (2014), 22–23.
[4] Hal Caswell. IMPROVING PEER REVIEW. (????).
[5] Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection: A survey. *ACM computing surveys (CSUR)* 41, 3 (2009), 15.
[6] S Cole, G A Simon, and others. 1981. Chance and consensus in peer review. *Science* 214, 4523 (1981), 881–886.
[7] Jonathan Graffy, Rosamund Bryar, Sally Kendall, and David Crook. 2006. Improving peer review. *Primary Health Care Research and Development* 7, 01 (2006), 1–2.
[8] John A Hartigan and Manchek A Wong. 1979. Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28, 1 (1979), 100–108.
[9] F J Ingelfinger. 1974. Peer review in biomedical publication. *The American J. of Med.* 56, 5 (1974), 686–692.
[10] J P Kassirer and E W Campion. 1994. Peer review: crude and understudied, but indispensable. *JAMA* 272, 2 (1994), 96–97.
[11] A S Relman and M Angell. 1989. How good is peer review? *The New Eng. J. of Med.* 321, 12 (1989), 827–829.
[12] Sandipan Sikdar, Matteo Marsili, Niloy Ganguly, and Animesh Mukherjee. 2016. Anomalies in the Peer-review System: A Case Study of the Journal of High Energy Physics. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016.* 2245–2250. DOI: http://dx.doi.org/10.1145/2983323.2983675
[13] R Smith. 2006. Peer review: a flawed process at the heart of science and journals. *Journal of the R. Soc. of Med.* 99, 4 (2006), 178–182.