Prestige Bias in Double-Blind and Single-Blind Peer Review

Submitted by: Jainabou Barry Danfa

SI699: Big Data Analytics Mastery Course

University of Michigan

School of Information

Winter 2020

Table of Contents

Introduction and Motivation

The scientific method is used to acquire and justify knowledge in the ever-developing scientific field. Studies are peer-reviewed to ensure their validity and reproducibility. These reviews can either be single-blind, where the reviewers know the authors of the research, or double-blind, where the author information is hidden to the reviewers. There is no consensus on the "best" review method, but the main area causing debate is around the prestige of the author and its affect on the review of the paper. Questions that commonly come up are:

1. Is double-blind anonymity even possible? (Citations, writing style, ect)
2. Does the prestige of the authors affect the reviewer's decision on the paper?
3. Does the type of review affect the papers reviewers want to review?

The International Conference on Learning Representations (ICLR) is a globally renowned peer-review conference on artificial intelligence and deep learning. With an acceptance rate of 31.4% makes this conference very competitive. In 2018, they moved from a single-blind review process to a double-blind process.

The focus of this study is looking at the prestige of authors and its effect, if any, on the acceptance or rejection of the paper based on the review type in ICLR submissions.

Peer Review Process

Scholarly peer review works to evaluate new research by other experts reviewing and critiquing the new research. This process serves as a filter for quality work and scholarship. Authors submit their work to a conference and/or journal to be reviewed and hopefully published. The editor or conference chairs conduct an initial review of the papers and rejects those of poor quality. Papers are then assigned to reviewers either at random or through a bidding process. He assigned reviewers then approve, eject, or request modifications to the paper. A rebuttal period may be present, where authors can challenge a reviewer's decision. Then the editor or conference chairs make a final decision on the paper.

Single-blind peer review is a more traditional and common approach to peer reviewing, where the reviewers know the identity of the authors while the reviewer's identity is concealed. This type of review removes pressure from the reviewer by concealing their identity, allowing them to critique without fear. However, this type of review can cause bias to affect the reviewers view on the work since the authors identity is known. Double-blind review attempts to eliminate bias and discrimination by concealing the authors identity as well. This has proven more difficult to implement with the vast amounts of information online and papers being published prior to acceptance in journals.

One critique of allowing the reviewers identity to be concealed in both single- and double-blind reviewers is the risk of low-quality or unmotivated reviews. To mitigate critique, Open review identifies the authors and the reviewers, usually on a online open forum. This allows for transparency in the process and allows all parties to get recognition for their contribution in the review. However, this method could cause reviewers to give less “negative” reviews and cause conflicts of interest.

With the rise of online discourse, post-publication peer review has the research paper already published on a forum and allows for the public to join in on the discourse about the research. His allows for all the stages of the peer review to be available to the public. This method of review is particularly useful in computer and data science due to the limited supply of reviewers and the time needed to reproduce code to ensure the integrity of the results. Some critique this method of review, citing that it can create false positive or negative narratives about the paper since the work is already published and has public comments that the reviewers can see. [1]

ICLR Conference

The International Conference on Learning Representations was founded in 2013 by Yoshua Bengio and Yann LeCun. The conference focuses on advances in Deep Learning and Artificial Intelligence[3]. The founders sought to experiment with the model of publication, allowing for full transparency in the review process, except for the reviewer's identity. With the rapidly growing field of research, having papers released online and open discussion was intended to accelerate innovation in the deep learning and artificial intelligence areas. The general reviewing process is the following:

1. Authors submit papers to ICLR for review on a public forum (CMT or OpenReview)
2. Reviewers bid for papers to review. Affinity scores are used to help with paper bidding correlating reviewer's expertise to the papers area. Each paper is assigned 3 reviewers.
3. Reviewers post their full review to the forum
4. Authors can address reviewers' comments and modify papers during discussion/rebuttal period
5. Area Chairs and Reviewers make final decision on acceptance/rejection of paper

At any stage in the process, the public can comment on the forum and engage with reviewers and authors. If the format is singe-blind, the authors identity is known in the forum and to the reviewers when bidding for papers. In the double-blind format, the information is anonymized until the review period is over. Various changes took place in the reviewing process, that are highlighted in the table below.

ICLR Submission Process by Year

Conference Year	*Double-Blind?*	*Open Bidding?*	*Discussion/Revision Period?*	*Two-round review process?*
2016	No	Yes	7 days	No
2017	No	Yes*	6 weeks	Yes
2018	Yes	Yes*	6 weeks	No
2019	Yes	Yes*	4 weeks	No
2020	Yes	Yes*	2 weeks	No

Note: ** No changes were mentioned for bidding in the years subsequent to 2016, so it is inferred that the practice is still occurring.*

The reason for switching the review format from single-blind to double-blind was not publicly disclosed on the ICLR website.

Datasets & Collection

OpenReview is a peer review workflow system that enables publication venues to explore varying types of openness[3]. ICLR has been using OpenReview since 2017 and the review data is available using REST API for 2017-2020 data.

Microsoft Academic Graph is a open database on authors, institutions, journals, citations, ect. that is updated on a weekly basis[4]. This will be used to gather information about the authors of the papers to determine a measure of prestige. The measures that can be acquired from this that can measure the prestige of an author:

- Citation Count: *"The estimation uses a statistical model that leverages both the local statistics of individual publications and the global statistics of the entire academic graph."*
- Saliency Rank: *"Weighting on each citation based on the factors of the citing sources, including the reputation and the age of each citation"*

- Prestige Rank: *"publication size normalized saliency of an author, an institution or a publication venue, and present it alongside with saliency to give a more holistic perspective of the rankings"*

Each of these measures have benefits and drawbacks to get at the question "how known is this author and their work". All of these measures will be in our model to get at prestige of each individual author.

Methods

Past Approaches

In a study by Andrew Tomkins, Min Zhang, and William D. Heavlin, analysis was completed to determine if the method of peer review gives advantage to papers with famous authors from high-prestige institutions [5]. The measure of famous was an author with at least 3 accepted papers at the same conference they conducted the study at. They found that single-blind reviewers bid 22% less on papers, prefer papers from top universities and companies, and ae more likely to submit more positive reviews for papers with a famous author or a top university or company compared to double-blind reviewers.

Methodology

Data will be gathered from OpenReview about each paper submission from 2017-2020. The fields that will be used for each review is in the table below:

ICLR Submission Data Fields

Field	*Description*
Forum Id	Unique identifier for each submission
Submission Title	Title of paper submission
Authors	List of authors on paper submission
Keywords	Domains related to the paper submission
Official decision	Accept(poster/workshop), reject for the submission
Official review	Text of official review following the decision
Official decision confidence	Confidence level of area chair for decision
Official decision title	Title of decision review

Field	*Description*
Affiliations	The institutions that the authors are affiliated with by submission email domain name
Reviewer ratings	List of ratings (scale varies over years) for each of 3 reviews
Reviewer reviews	List of text for each of 3 reviews
Reviewer review titles	List of text for each of 3 reviews title

With the list of all authors, the citation count and publication count from Microsoft academic graph was obtained for each author with a 5-year window prior to the submission of the paper. This term will allow us to get a measure of the prestige at the time of submission. The author information was then aggregated for ach paper, yielding the average citation, average publication, max citation, and max publication for all authors in the paper. The sentiment score for each official review was obtained by using VADER (Valence Aware Dictionary and sEntiment Reasoner) [6] analysis tool.

To test the effect of the publication and citation scores, a logistic regression model was used with the outcome being the official review decision. The prestige measures [average citation, average publication, max citation, and max publication] were used interchangeably in the formula below:

$$decision = prestige\ measure + single_{blind_{dummy}} + prestige\ measure * single_{blind_{dummy}}$$

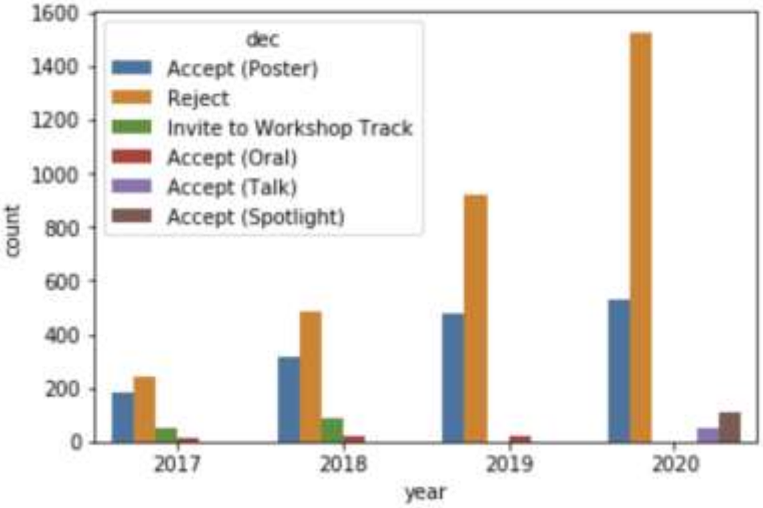
The decision metric was determined by aggregating the official decision, since they changed over the years. For the purposes of this paper, the accept was classified as Oral, Talk, and Spotlight. Reject was classified as Workshop, Reject, and Poster.

Analysis Summary

Exploratory Data Analysis

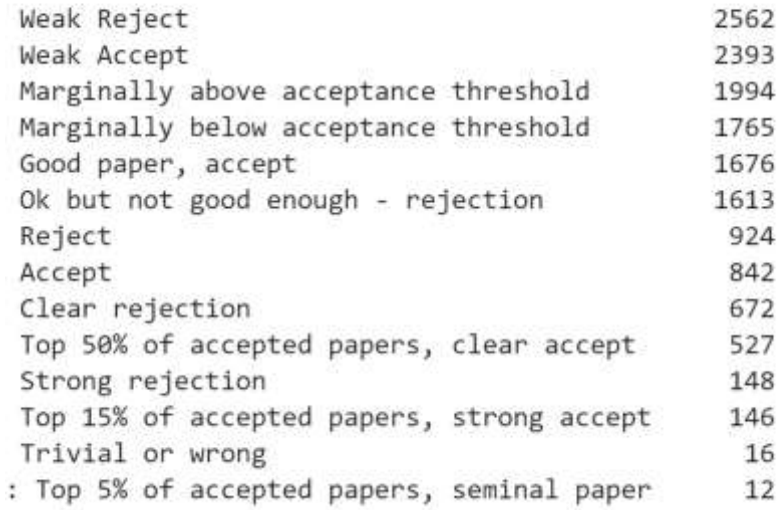
In our dataset, we have 5003 submissions over 4 years of data. The growth of this conference as risen substantially over time with 490 submissions in 2017 to over 2000 submissions in 2020. As we can see in the graph below, most papers submitted to this conference are rejected than accepted, with more rejections over time.

Official Submission Decision for each year



The most common keyword for the submissions were deep learning, reinforcement learning, and unsupervised learning. This is consistent with the domain of the conference. Of the individual reviews for each submission, we see that weak reject and accept are the most common scores merited by reviewers.

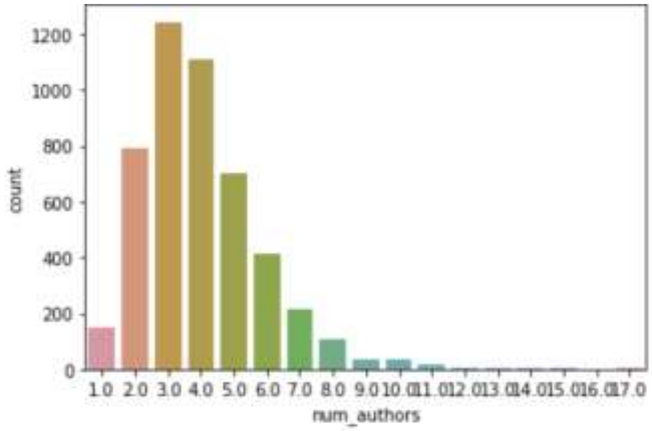
Individual review ratings for submissions across all years



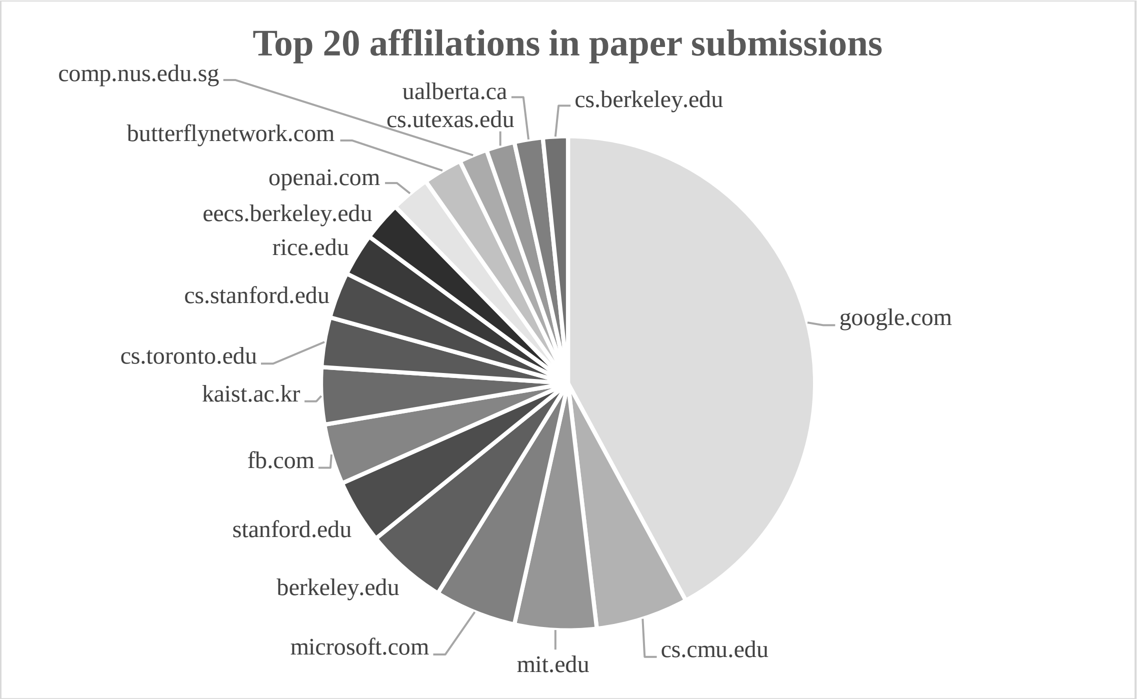
```
 Weak Reject                                      2562
 Weak Accept                                      2393
 Marginally above acceptance threshold            1994
 Marginally below acceptance threshold            1765
 Good paper, accept                               1676
 Ok but not good enough - rejection               1613
 Reject                                            924
 Accept                                            842
 Clear rejection                                   672
 Top 50% of accepted papers, clear accept          527
 Strong rejection                                  148
 Top 15% of accepted papers, strong accept         146
 Trivial or wrong                                   16
: Top 5% of accepted papers, seminal paper          12
```

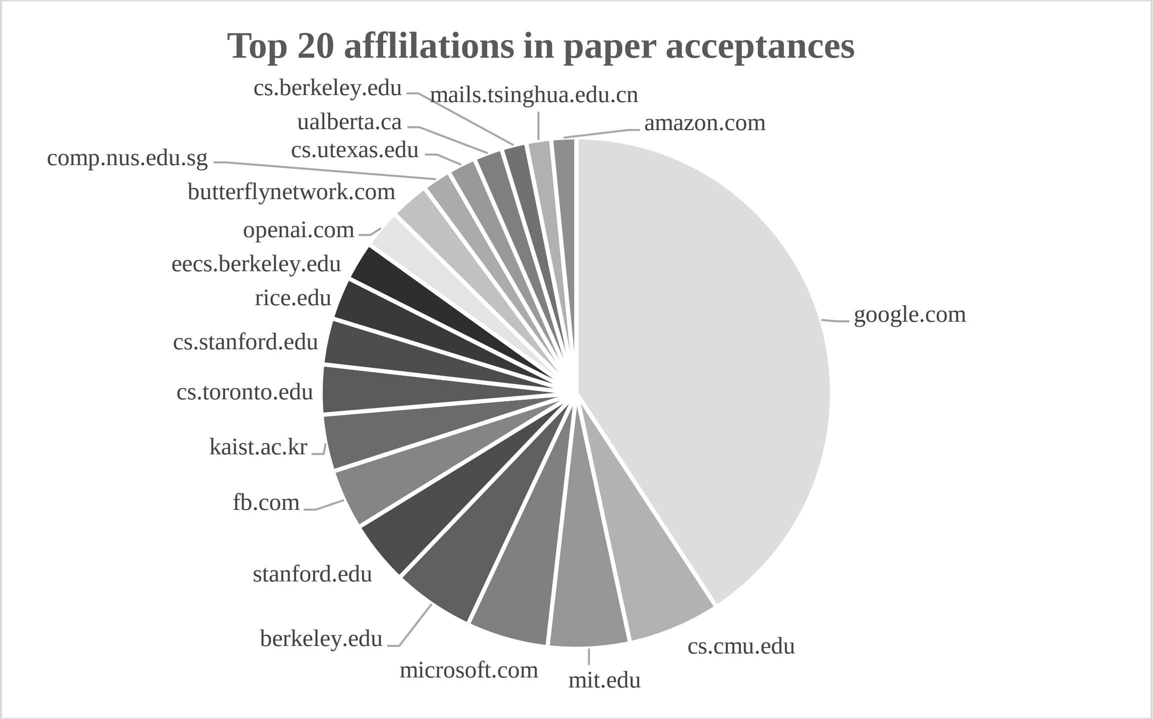
After conducting the Microsoft Academic Graph search, 4842 out of the 5003 papers were matched. The remaining 161 papers did not receive a match on Microsoft Academic API. With the papers that matched, the average number of authors per paper was around 3.5.

Number of Authors per paper for submissions across all years



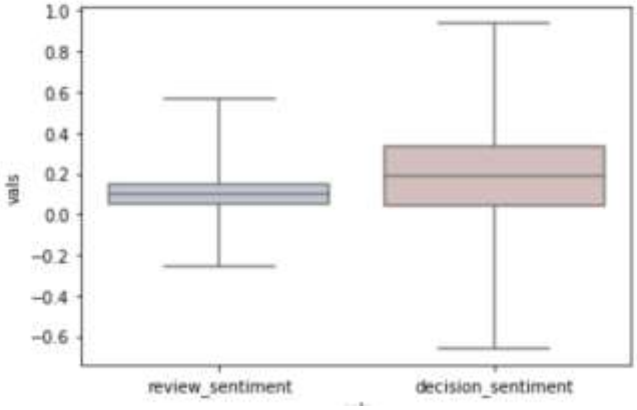
Looking at the affiliations of the authors for all submitted papers, after dropping gmail.com, we see that google and Microsoft have the highest authors submitting papers at this conference. When we look at accepted papers, the top affiliations do not change much, with the exception of amazon, which ranked in the accepted papers but not necessarily in the number of submissions.



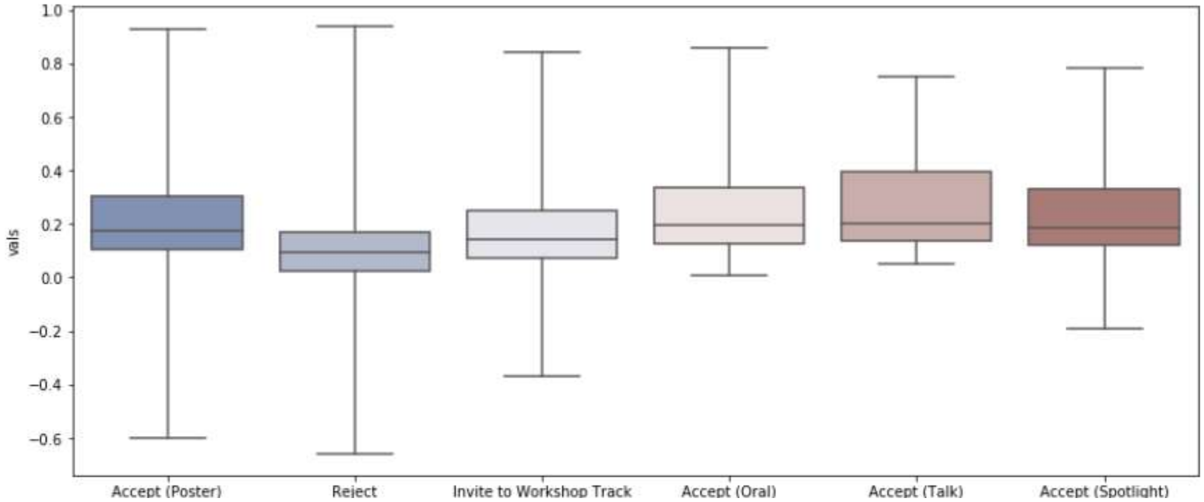


When taking the sentiment analysis of the official decision text, we that the reviewer's sentiment was much more neutral with less range compared to the official decision sentiment. Double-blind reviews tend to be slightly more negative compared to single-blind.

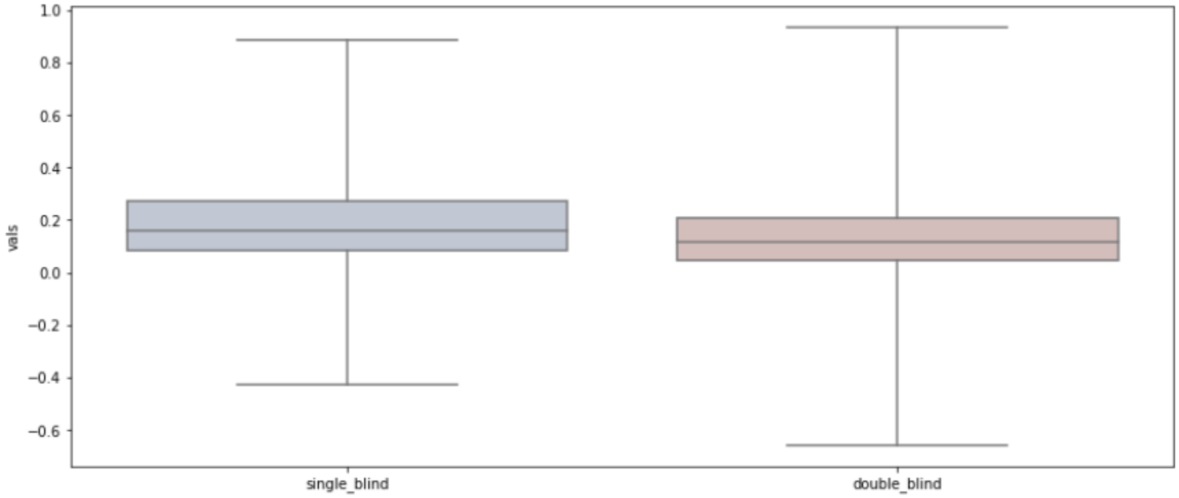
Sentiment Scores by official review and aggregated reviewers' reviews



Sentiment Scores by Official decision



Sentiment Scores by review type

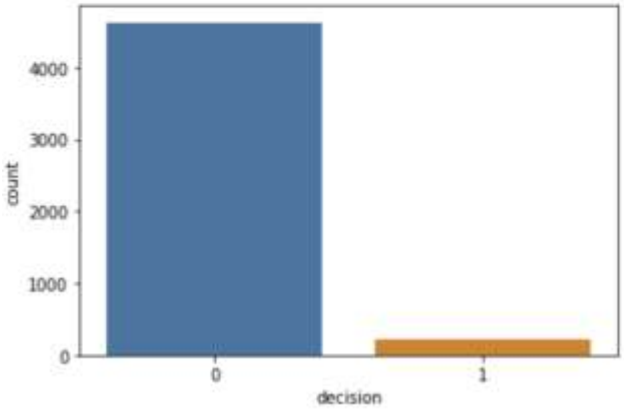


The slight decrease in sentiment of double-blind reviews hints at the interaction between the reviewers and the authors.

Baseline

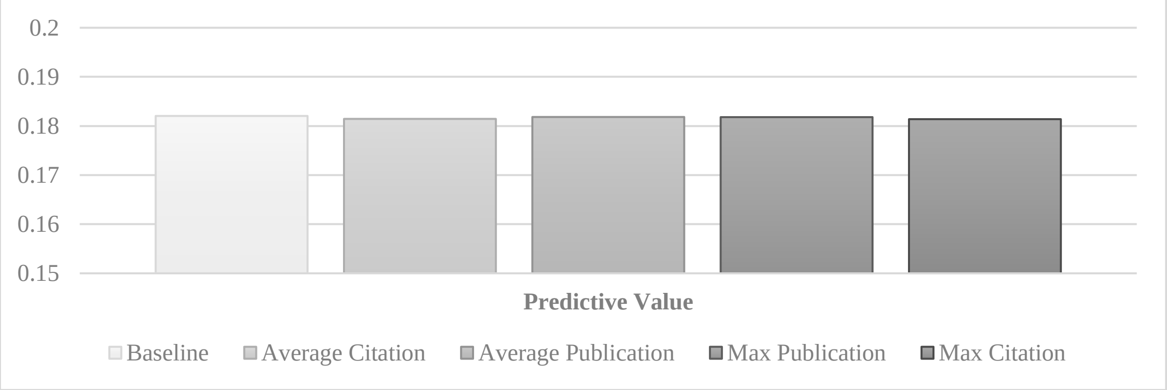
The baseline model used in this study was a basic model predicting class balance. The baseline model had a function value of 0.182%.

Class distribution in logistic regression model

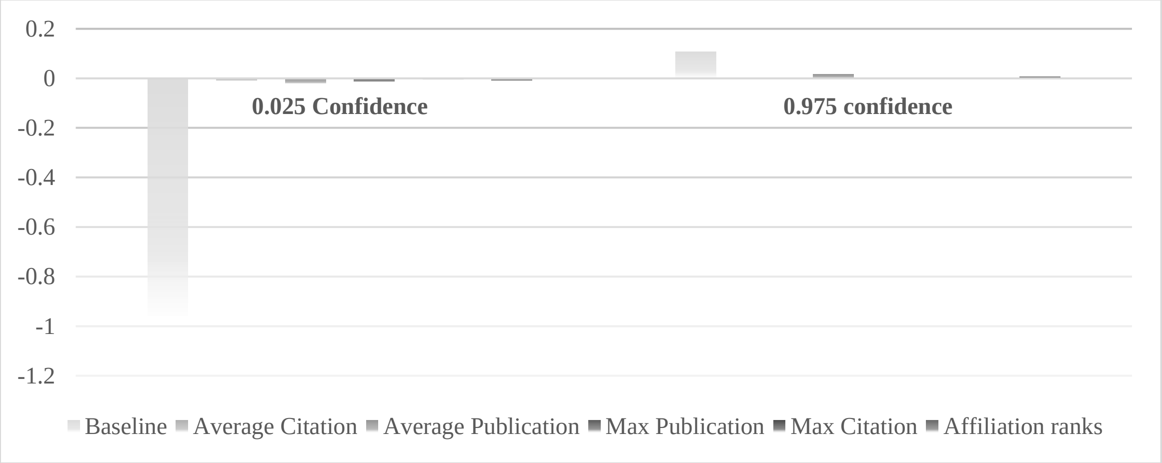


Analysis relative to baseline

When replacing the interaction term with each one of the model parameters, we get almost no difference in the model function value. Suggesting that addition of prestige measures, do not impact the decision in single-blind or double-blind cases.



Taking a deeper look into the interaction term confidence intervals, we see the same story with none of predictive value.



Conclusions

As we can see from the logistic regression results, the prestige of an author does not have an impact on the official decision based on the type of review. However, this conclusion is nuanced and requires further investigation and analysis. Future work on this topic can look into different measures of prestige and another conference to validate. Secondly, does this hint at the model of double-blind and its ability to be executed any differently than single-blind in the official review stage since most papers are posted on Arvix prior to review anyways. Thirdly, does the bidding process reveal bias that is not represented in this analysis? This study lends itself to more questions than definitive answers. However, with the rise of peer reviewing and new academic journals in rapidly developing fields, the answers to these questions are more important than ever.

Citations

[1] - [https://medium.com/syncedreview/cvpr-paper-controversy-ml-community-reviews-peer-review-79bf49eb0547]

[2]- [https://www.kdnuggets.com/2016/02/iclr-deep-learning-scientific-publishing-experiment.html]

[3]- . [https://blog.codeforscience.org/welcome-and-contratuations-openreview/].

[4]- [https://academic.microsoft.com/faq?target=ranking4]

[5]- [https://www.pnas.org/content/114/48/12708]

[6]- https://github.com/cjhutto/vaderSentiment