# CiteTracked: A Longitudinal Dataset of Peer Reviews and Citations

Barbara Plank[1] and Reinard van Dalen[2]

[1] IT University of Copenhagen, Denmark
[2] Rijskuniversiteit Groningen, The Netherlands
bplank@itu.dk

**Abstract.** Scientific dissemination is of central importance for the scientific process. This paper presents CiteTracked, a dataset of peer reviews and citation statistics covering scientific papers from the machine learning community and spanning six years. We describe and analyze the data collection of over 3,000 published papers, their peer review texts and citation counts, and depict possible usage directions. The dataset aims at fertilizing novel interdisciplinary work between fields such as scientometrics, information retrieval, computational linguistics and natural language processing to study the scientific publishing process.

**Keywords:** Peer reviews · citations · NLP for science

## 1 Introduction

Researchers around the globe continuously contribute invaluable information to the world's knowledge by publishing their findings in conferences and journals. These findings are subject to the scrutiny of the scientific review process. Peer reviewing is an essential component of the scientific process. Leading conferences and journals use peer reviewing to decide which manuscripts to include in their proceedings and journals. The reviewing process is of vital importance, yet the process itself is often subject to debate. For example, a recent experiment to examine the consistency of the review process observed that reject/accept decisions were following a disagreement rate of 26% on a random sample of a tenth of the papers that went through the review process twice (e.g., [6, 3, 4]).

Typically, reviews are accessible only to the authors of a manuscript, and to the few selected individuals who organize the scientific venue. Studies on the qualitative and quantitative properties of peer reviews had been limited. Few selected top-scientific venues recently started to make the peer reviews publicly accessibly. An example is NeurIPS (the conference on Neural Information Processing Systems, previously named NIPS) and the OpenReview initiative. Such initiatives contribute to opening up the largely covert process of peer reviewing. This starts a recent surge of interest in the study of peer reviews, e.g., [1, 13].

Once a paper is published, citations can be used to estimate the importance of a paper, as it encodes the implicit judgement of the importance of a paper by the community. Citation statistics hence can provide a valuable signal to

study the scientific impact. Efforts in understanding such signal typically resort to modeling citation networks and hence typically refer to the past [7]. Some work exists on predicting future impact, e.g., by correlating textual properties of papers (such as content from its title or abstract) to citation statistics.

In this paper, we present a novel corpus that provides a possible link between these research strands and enables prediction of scientific impact and the study of peer reviews. Our corpus called CiteTracked contains over 3,000 papers and over 12,000 reviews from the NeurIPS conference spanning the last six years.

## 2 A Dataset of Peer Reviews and Citation Statistics

### 2.1 Peer review collection and meta-data

The dataset contains 12,260 peer reviews and meta-data for a total of 3,427 papers published in the NeurIPS proceedings (http://papers.nips.cc) from 2013 to 2018. The meta-data includes information such as author names, abstract, title, a link to the original paper, and for a subset of editions also the paper presentation type (event type: oral, spotlight or poster) and author feedback.

**Table 1.** Overview of the dataset. *Indicates that a few papers were removed.

| Conference edition | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | Total |
|---|---|---|---|---|---|---|---|
| Number of papers | 360 | 411 | 403 | 568 | 679 | 1,006* | **3,427** |
| Average number of reviews per paper | 3.1 | 3.1 | 3.8 | 5.7 | 2.9 | 3.1 | 3.6 |
| Total reviews | 1,132 | 1,278 | 1,536 | 3,240 | 1,977 | 3,097 | **12,260** |
| Total author feedback | 359 | 408 | 403 | n/a | n/a | n/a | – |
| Even type available | ✓ | ✓ | ✓ | n/a | n/a | n/a | – |
| Average tokens: | | | | | | | |
| Reviews | 376 | 353 | 330 | 290 | 298 | 327 | – |
| Summaries | 36 | 35 | 41 | 90 | n/a | n/a | – |
| Author feedback | 629 | 660 | 644 | n/a | n/a | n/a | – |

An overview of the data set is provided in Table 1. First we notice that the data follows the general trend of increasing publication volume at computer science venues. There was an (almost) steady growth in papers, from 360 in 2013 to 679 in 2017, with a big step in 2018 (1,009 papers).[3]

For all years except 2015 and 2016, for each paper an average of 3 reviews per paper was solicited. In 2015, an average of 4 reviews per paper was implemented, while in 2016 this amounted to 6 reviews per paper. There is no further numerical scoring publicly available besides the review text, except for a single year (2016, in which reviewer confidence scores are available as well).[4]

[3] Note that we had to exclude three papers from the 2018 edition due to review pages which were inaccessible.

[4] We would like to note that beyond NeurIPS there are further venues such as ICLR which make review data publicly available, e.g., on the OpenReview platform. Col-

### 2.2 Collection of citations

We embarked on a manual effort to collect citation statistics over time. CiteTracked is intended to be an on-going dataset collection effort. The current release contains citation counts for all papers published up to 2018, i.e., for a total of all 3,427 scientific papers citation counts of 5 time spans are available.

The goal is to collect citation counts at different time intervals, with at least one such iteration per year for every paper (collected in a short time span) which was originally a manual effort and has now been semi-automatized. The recording of the citation scores started in April 2016 (citations1). The citation scores were further recorded in November 2016 (citations2), June 2017 (citations3), March 2018 (citations4), and June 2019 (citations5). Citation counts were collected via the bibliographic database provided freely by Google Scholar. In contrast to subscription-based services such as Web of Science of Scopus, Google Scholar provides higher coverage.[5] Citation statistics collected in this way provide an alternative to within-field citation networks [2].

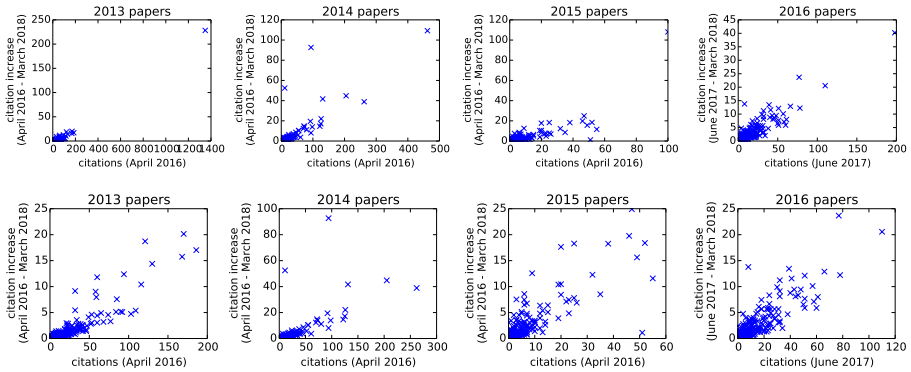


**Fig. 1.** Illustration of citation growth of papers per conference year for a subset of the data. Top: all papers; Bottom: most cited papers (outliers) removed from top plot.

Figure 1 shows the growth in absolute citations for each of the six years in our data collection. We observe the well-known non-linear growth rate [5]. Few papers receive an extremely high amount of citations. For example, the outlier on the right of the 2014 plot is the seminal paper by Sutskever et al., 2014 which introduces sequence-to-sequence models.

lecting the reviews from this source and respective citation counts for papers is currently beyond the scope of the current project. Recent work has started to collect such peer review data, cf. [1], including data from the Association of Computational Linguistics [2].

[5] https://libraryguides.helsinki.fi/metrics/citations

## 3 Analysis and Potential Use Cases

In this section, we provide a first quantitative and qualitative analysis of the corpus. We showcase the potential of using CiteTracked for data-driven analysis of linking peer reviews to scientific impact.

### 3.1 Analysis of most impactful papers and paper categories

In this section, we highlight the papers that received the most citations per year. We analyze whether the presentation type of a paper is linked to higher impact.

**Table 2.** Overview of two most impactful papers per conference year, including citation counts (retrieved June, 2019).

| **Year** | **Title** (**citations in parentheses**) |
|---|---|
| 2013 | 'Distributed Representations of Words and Phrases and their Compositionality' (13,295 citations), 'Translating embeddings for modeling multi relational data' (1,225) |
| 2014 | 'Generative Adversarial Nets' (9,580), 'Sequence to Sequence Learning with Neural Networks' (6,788) |
| 2015 | 'Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks' (10,079), 'Spatial transformer networks' (1,700) |
| 2016 | 'Improved Techniques for Training GANs' (1,929) , 'R-fcn: object detection via region-based fully convolutional networks' (1,374) |
| 2017 | 'Attention is All you Need' (2,136), 'Improved Training of Wasserstein GANs' (1,387) |
| 2018 | 'Are GANs created equal? A large-scale study' (145), 'Glow: generative flow with invertible 1x1 convolutions' (141) |

Table 2 shows the two most impactful papers per year. The most cited paper received over 13,000 citations. With regard to paper presentation mode (termed 'conference event type' in the original proceedings), we see that very impactful papers do not necessarily get one of the few oral presentation slots (oral or spotlight). Table 3 shows that in 2014 the general tendency clearly holds that orally presented papers got higher impact, which is less so for the other years. In 2013 both orals and posters received on average the same citation impact.

**Table 3.** Average citation rate for papers from 2013-2015 per event type.

| | 2013 | 2014 | 2015 |
|---|---|---|---|
| Poster | 68.1 | 45.0 | 41.8 |
| Spotlight | 39.4 | 91.0 | 58.6 |
| Oral | 68.1 | 286.6 | 73.8 |

### 3.2 Use Cases

Datasets like CiteTracked or PeerRead [1] can be used in a variety of ways.

*The language of peer reviews* The analysis of peer reviews can for example provide a more nuanced understanding of argumentation in the scientific process. For instance, [1] quantified how reviews recommending an oral presentation differ from those recommending a poster. A very recent study provides a dataset of ICLR reviews annotated for argumentation types [13] (Evaluation, Request, Fact, Reference, or Quote), which we use to train a bilstm-CRF. We analyzed the predicted argumentation types in reviews of top (and least cited) papers. As shown in Figure 2, top cited papers get more evaluative reviews (in 3 out of the 4 years). This analysis could be a starting point to analyze the stance of the review (and whether it is a potential 'advocate' for the paper).

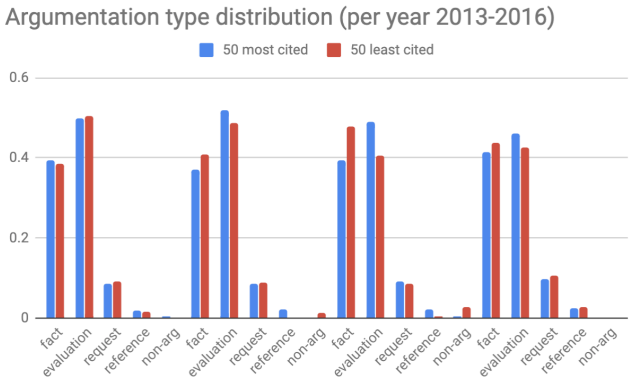


**Fig. 2.** Analysis of argumentation types predicted on reviews for the 50 most cited/least cited papers per conference year.

*Peer Reviews for Citation Impact* Another use case is to study *whether reviews are predictive of citation impact*. This is an aspect that, to the best of our knowledge, has not been studied yet. Therefore, in this paper we study whether we can successfully predict citation impact from peer reviews, and to what extent it is complementary to earlier work that relied on aspects of the paper itself.

In order to predict the impact of the scientific papers, we discretize time-normalized citation statistics into low, medium and high impact papers based on a boxplot and outlier analysis. We use the Upper Outlier Threshold (UOT). UOT is defined by adding up the Inter Quartile Range Rule (IQRR) to the third quartile. Papers with a growth rate above the UOT are therefore defined as high impact papers. Papers with a growth rate that is below the UOT, are defined as low/medium impact papers, which were further split up into low and medium impact papers based on an UOT analysis on the subsequent UOT analysis.

We create baselines which consider meta-data from the paper itself, namely the title and the abstract as done in earlier studies. We train a Support Vector Machine with commonly used features extracted from the text and titles of

the paper, i.e., word n-grams and character n-grams, including special features motivated by earlier work such as the average word length of the paper title, paper title length. In particular, this specific feature set includes the use of question marks in paper titles [8, 10], the use of colons in paper titles [9, 10, 8], the length of the paper title [11] and the number of authors of a paper [12]. The dataset (2013-2017) was split into: 60% training, 20% development and 20% test set. Model performance is reported in F1-score on the final test set. To put the results into perspective, we provide a random stratified baseline as well as models inspired by prior work, which only use title and abstract as indicators. We use review texts which may include review summaries (whenever available).

Table 4 shows the results. There are several take-aways. First of all, paper information is predictive of scientific impact. A classifier that only uses the paper title is able to achieve an average performance of .81 F1-score. This outperforms the random stratified baseline of .71 and confirms earlier findings. A closer look reveals that the model struggles to predict the mid class. It falls mostly back to the majority class (the low impact papers). Adding the paper abstract improves overall performance (from .81 to .83). Secondly and most importantly, the results show the potential of review texts. Review texts are predictive of scientific impact. The performance of a model based on review texts is higher than using only abstract or title, thereby confirming our hypothesis that reviews constitute valuable information for scientific impact prediction. A model which uses all information (title, abstract and reviews, indicated as 'all' in Table 4) result in an overall similar performance to reviews alone, but it improves prediction F1-score for the difficult mid class. This investigation shows the potential of learning from peer review texts.

**Table 4.** Results of predicting impact level of papers (F1 score).

| | low | mid | high | avg |
|---|---|---|---|---|
| title | .91 | .0 | .26 | .81 |
| abstract | .92 | .0 | .39 | .83 |
| reviews | .93 | .09 | .49 | .85 |
| all | .92 | .24 | .48 | .85 |

## 4 Conclusions

This paper introduces CiteTracked, a corpus of peer reviews from the NeurIPS conference enriched with citation statistics collected over several years. The current corpus contains 3,427 papers and over 12,000 reviews. We outline corpus collection, provide an initial analysis and discuss potential use cases that link work on bibliographic indicators, peer reviews and scientific publication impact.

## Acknowledgements

We would like to thank Stijn Eikelboom for help with improving the citation collection process and NVIDIA for supporting our research.

## References


1. Kang, D., Ammar, W., Dalvi, B., van Zuylen, M., Kohlmeier, S., Hovy, E., & Schwartz, R. (2018). A Dataset of Peer Reviews (PeerRead): Collection, Insights and NLP Applications. In Proceedings of NAACL.
2. Radev, D. R., Muthukrishnan, P., Qazvinian, V., & Abu-Jbara, A. (2013). The ACL anthology network corpus. Language Resources and Evaluation, 47(4), 919-944.
3. Greaves, S., Scott, J., Clarke, M., Miller, L., Hannay, T., Thomas, A., & Campbell, P. (2006). Nature's trial of open peer review. Nature, 444(7122), 971-972.
4. Shah, N. B., Tabibian, B., Muandet, K., Guyon, I., & Von Luxburg, U. (2018). Design and analysis of the NIPS 2016 review process. The Journal of Machine Learning Research, 19(1), 1913-1946.
5. Ruocco, G., Daraio, C., Folli, V. & Leonetti, M. (2017) Bibliometric indicators: the origin of their lognormal distribution and why they are not a reliable proxy for an individual scholar's talent. Palgrave Communications. 3:17064
6. Langford, John, & Mark Guzdial (2015). The arbitrariness of reviews, and advice for school administrators. Communications of the ACM 58.4: 12-13.
7. Weihs, L., & Etzioni, O. (2017). Learning to predict citation-based impact measures. In Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries (pp. 49-58). IEEE Press.
8. Hudson, J. (2016). An analysis of the titles of papers submitted to the uk ref in 2014: authors, disciplines, and stylistic details. Scientometrics 109(2), 871–889
9. Jacques, T. S. & N. J. Sebire (2010). The impact of article titles on citation hits: an analysis of general and specialist medical journals. JRSM short reports 1(1), 1–5.
10. Jamali, H. R. & M. Nikzad (2011). Article title type and its relation with the number of downloads and citations. Scientometrics 88(2), 653–661.
11. Subotic, S. & B. Mukherjee (2014). Short and amusing: The relationship between title characteristics, downloads, and citations in psychology articles. Journal of Information Science 40(1), 115–124
12. Vieira, E. S. & J. A. Gomes (2010). Citations to scientific articles: Its distribution and dependence on the article features. Journal of Informetrics 4(1), 1–13
13. Hua, Xinyu, Mitko Nikolov, Nikhil Badugu & Lu Wang (2019). Argument Mining for Understanding Peer Reviews. In Proceedings of NAACL.