Citations and Readership are Poor Indicators of Research Excellence: Introducing TrueImpactDataset, a New Dataset for Validating Research Evaluation Metrics

Conference or Workshop Item

How to cite:

Herrmannova, Drahomira; Patton, Robert; Knoth, Petr and Stahl, Christopher (2017). Citations and Readership are Poor Indicators of Research Excellence: Introducing TrueImpactDataset, a New Dataset for Validating Research Evaluation Metrics. In: 1st Workshop on Scholarly Web Mining, 10 Feb 2017, Cambridge, UK, ACM, pp. 41–48.

Version: Version of Record

Link(s) to article on publisher's website:
http://dx.doi.org/doi:10.1145/3057148.3057154
https://dl.acm.org/citation.cfm?id=3057154

Citations and Readership are Poor Indicators of Research Excellence

Introducing TrueImpactDataset, a New Dataset for Validating Research Evaluation Metrics

Drahomira Herrmannova[1]
d.herrmannova@open.ac.uk

Robert M. Patton[2]
pattonrm@ornl.gov

Petr Knoth[1]
petr.knoth@open.ac.uk

Christopher G. Stahl[2]
stahlcg@ornl.gov

[1]The Open University, Milton Keynes, UK
[2]Oak Ridge National Laboratory, Oak Ridge, TN, USA

ABSTRACT

In this paper we show that citation counts and Mendeley readership are poor indicators of research excellence. Our experimental design builds on the assumption that a good evaluation metric should be able to distinguish publications that have changed a research field from those that have not. The experiment has been conducted on a new dataset for bibliometric research which we call TrueImpactDataset. TrueImpactDataset is a collection of research publications of two types – research papers which are considered seminal work in their area and papers which provide a survey (a literature review) of a research area. The dataset also contains related metadata, which include DOIs, titles, authors and abstracts. We describe how the dataset was built and provide overview statistics of the dataset. We propose to use the dataset for validating research evaluation metrics. By using this data, we show that widely used research metrics only poorly distinguish excellent research.

CCS Concepts

•**Applied computing** → **Digital libraries and archives;**
•**Information systems** → *Information retrieval;*

Keywords

Information Retrieval, Scholarly Communication, Publication Datasets, Data Mining, Research Evaluation

Acknowledgements

This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan[1].

SWM '17, February 10, 2017, Cambridge, United Kingdom
© 2017 ACM. ISBN 978-1-4503-5240-6/17/02…$15.00
DOI: http://dx.doi.org/10.1145/3057148.3057154

1. INTRODUCTION

The question of how to evaluate research outcomes and impact is very difficult to answer and despite decades of research, the problem is still largely unsolved. Under the current system published research should receive ideally a fair review by experts in the area and be given scores according to a generally accepted set of standards and rules. This process however often does not work [27, 20]. This is largely due to the enormous and ever-growing number of papers being published every year, which was estimated to be over 1.5 million in 2008 [14] and over 1.6 million in 2011 [17]. Much work has been done in this area with the aim to develop research evaluation methods capable of simplifying or completely automating this task. Recent years have seen the emergence of many new directions, such as altmetrics [23, 6], webometrics [3, 13] and semantometrics [15, 10]. Despite the fact that this research area attracts so much interest and new methods are constantly being developed, there exists no ground truth or reference dataset for assessing the usefulness of the existing methods. As a consequence, the authority of these methods is often established axiomatically. For example, the two best known metrics, the Journal Impact Factor (JIF) [7] and the h-index [12] were both proposed and are widely used without empirical evidence demonstrating that they measure what they intend to measure.

When talking about research evaluation, research impact and research excellence, most people usually refer to the volume of change produced in a particular field (how much did a piece of work move the field forward), rather than referring to the educational (or other types of) impact generated.

[1] http://energy.gov/downloads/doe-public-access-plan

Seminal papers are a representative of the first type, while survey (literature review) papers of the second type. Indeed, the definition of the word *seminal* according to the Oxford Dictionary is “strongly influencing later developments” while the definition of the word *survey* is “a general view, examination, or description of someone or something”, which matches our understanding of the difference between these two types of papers. Hence seminal papers should perform better under research evaluation than survey papers, which by definition do not generate change in the field. Therefore, we argue that metrics used in research evaluation should be able to discriminate seminal papers from survey papers. However, our study shows that existing research evaluation metrics only poorly distinguish between these two types of research works, demonstrating the significance of the problem of applying them in practice. The work presented in this paper is conducted on a new dataset of seminal and survey papers which we call TrueImpactDataset. This dataset was built from data collected in an online survey. We asked the respondents to provide two references from their research area – a seminal and a survey publication. We share this dataset with the research community to help the development of new research evaluation metrics. The dataset consists of metadata (which include DOIs) of 314 research papers from different scientific disciplines – 148 survey papers and 166 seminal papers.

We use this dataset to demonstrate the significance of the problems of existing research metrics and show that using citations as a ground truth for novel research evaluation metrics is misleading. Furthermore, we discuss the parameters an ideal research evaluation data set should satisfy.

This paper is organized as follows. In Section 3 we explain how the dataset was created. Section 4 presents some statistics of the dataset and Section 5 the results of an experiment in which we examine the value of citations in predicting the class of a paper. In Section 6 we discuss our findings and the properties an ideal evaluation set should have. Related work is presented in Section 7.

2. METHODOLOGY

This paper aims to answer the following research question: “Are current research evaluation metrics sufficient for identifying highly influential papers?” A typical data analysis/statistics approach to answering this question would be to test the metrics on a ranked set of papers and to express the success rate of these metrics using an evaluation measure such as precision and recall. However, to our knowledge, there exists no such ground truth or a reference dataset that could be used for establishing the validity of research evaluation metrics. While there was an attempt at creating such a dataset (Section 7), this dataset wasn’t openly shared and so cannot assist in this task. Because building such dataset would require significant time and resources (Section 6) we were looking for an alternative way of validating the metrics.

As mentioned in the introduction, when talking about research evaluation, most people typically think of the amount of change produced in a research area (how much was the area pushed forward thanks to a given piece of work). This amount of change has been discussed and studied from different perspectives [32, 15, 30, 28, 19]. We were looking for a sample of research publications representing such work and we believe seminal research papers constitute such sample. To provide a clear comparison we were also interested in review publications (papers presenting a survey of a research area). While these papers are often highly cited [25, 2] they usually don’t present new original ideas. Our hypothesis states that metrics used in research evaluation should be able to distinguish these two types of papers.

To our knowledge, there currently isn’t any dataset which would categorize papers into these two categories. We were therefore left with creating such dataset ourselves. We have employed an online survey for this task. The format of the survey, the number of collected responses and other details are presented in Section 3.1. In the following section (4) we analyze the dataset to understand whether it is suitable for our purposes.

In order to answer our research question, we have designed a simple experiment. We chose citation counts and Mendeley readership as representatives of bibliometrics and altmetrics, as these two measures are both well known and are being used as measures of impact of the published research in many settings [24, 31]. We then classify the papers in the collected dataset into two classes (survey, seminal) using two models, a model using the papers’ citation counts and a model using their Mendeley readership (Section 5). We show that the model using citation counts outperforms the baseline by a small margin, while the model using readership doesn’t perform better than the baseline.

3. DATASET CREATION

This section describes the dataset and the process used to create it. The dataset is publicly available for download[2].

3.1 Initial data collection

The goal was to create a collection of research publications consisting of two types of papers, seminal works and literature surveys. The online form we have used to collect the references was composed of two sets of questions – questions about the respondent’s academic background (their discipline, seniority and publication record) and questions which asked for a reference to a seminal paper and one for a literature survey, both related to the respondent’s discipline. We have used the latest Research Excellence Framework (REF) units of assessment [1] as a list of disciplines when asking about the respondents’ academic background, because UK researchers are familiar with with this classification.

The survey was sent to academic staff and research students from all faculties of the Open University (to 1,415 people in total). We have received 184 responses (172 references to seminal papers and 157 references to survey papers), which represents a 13% response rate. The survey questions and email invitation are available online together with the dataset[2].

In case the respondents were unable to answer both questions (provide both references), it was allowed to answer only one. Ten respondents have only filled the questions related to their academic background but have not provided the references. We have removed these responses from the dataset which left us with 174 responses.

To make it easier to complete the survey we didn’t require the references to be in a specific format (e.g. a URL or DOI). The respondents were allowed to submit the references in any format they preferred (as a text, link, etc.). As a consequence, a few of the references were submitted in a

[2] http://trueimpactdataset.semantometrics.org

format which made it impossible for us to identify the papers (e.g. "Stockhammer (2004)"). We have removed these papers from the dataset. After removing empty and unidentifiable responses, we were left with 171 responses (166 seminal papers and 148 survey papers).

3.2 Additional metadata

Once the survey has been closed we have manually processed the data and collected the following information (by querying a search engine for the paper title and looking for a relevant page): a DOI, or a URL for papers for which we did not find a DOI, title, list of authors, year of publication, number of citations in Google Scholar and abstract. Where we had access to the full text we have also downloaded the PDF. We were able to download 275 PDFs and 296 abstracts. Due to copyright restrictions, the PDFs are not part of the shared dataset[3].

To obtain additional metadata we have used the DOIs, or title and year of publication for papers without a DOI, to query the Mendeley API[4]. We were mainly interested in the number of readers of each paper. The dataset contains a snapshot of the Mendeley metadata we were working with. We were able to find 137 out of the 166 seminal papers and 123 out of the 148 survey papers in Mendeley.

Furthermore, using the Web of Science (WoS) API[5] we managed to retrieve additional information for the seminal and survey papers indexed by WoS. We queried the WoS API using publication DOIs, if the document was in the system we obtained a full list of publications citing the paper in question and publications cited by the paper. This list included minimal metadata. In order to get full citation information we then had to query the API for each individual (citing and cited) result.

4. DATASET ANALYSIS

To ensure the collected dataset is suitable for our purposes we looked at several statistics describing the dataset including statistics of publication age, distribution across disciplines and citation and readership statistics.

4.1 Size

The size of the dataset is presented in Table 1. The row *DOIs* shows for how many papers in the dataset we were able to find a DOI and the row *DOIs in WoS* how many of these DOIs appear in the Web of Science database. The number of additional references which we collected using the WoS API is shown in the row *Citing & cited references.*

The rows *Authors total* and *Authors unique* show the total number of authors of all papers in the dataset and the number of unique author names. To count the unique names we have compared the surname and all first name initials, in case of a match we consider the names to refer to the same person (e.g. J. Adam Smith and John A. Smith will be considered the same person).

Responses	171
Seminal papers	166
Survey papers	148
Total papers	314
DOIs	256
DOIs in WoS	109
Authors total	1334
Authors unique	1235
Abstracts	296
Citing & cited references	19,401

Table 1: Dataset size.

4.2 Publication age

Figure 1 shows a histogram of years of publication with survey and seminal papers being distinguished by color. The seminal papers in the dataset are on average older than survey papers, by about 9 years. This shows survey papers might age faster than seminal papers, which is consistent with our expectations. An explanation for this could be that literature surveys theoretically become outdated as soon as the first new piece of work is published after the publication of the survey. Because the seminal papers are on average older this also means these papers had more time to attract citations. This is another reason to expect seminal papers to be distinguishable by citations and readership as features. Descriptive statistics of years of publication both sets are presented in Table 2.

	Seminal	**Survey**
Mean	1999	2008
Min	1947	1975
Max	2016	2016
25%	1995	2005
50% (median)	2002	2010
75%	2010	2013

Table 2: Descriptive statistics of publication age for both types of papers.

4.3 Disciplines

Figure 2 shows a histogram of papers per discipline. To assign papers to disciplines we have used the information we got about the respondents' academic background. The respondents have also provided a short description of the research area related to the two references (e.g. "molecular neuroscience", "combinatorics", etc.), however as these descriptions are more detailed and there is little overlap between them we haven't used these in our analysis.

The distribution of papers per discipline is to a certain degree consistent with other studies, which have reported Computer Science and Physics to be among the larger disciplines in terms of number of publications, however Medicine and Biology are typically reported to be the most productive [4, 5]. The distribution is therefore probably more representative of size of faculties of the Open University than of

[3] As there are Copyright Exceptions for text and data mining in some countries, such as in the UK, we are happy to provide the PDF documents for these purposes to researchers residing in these jurisdictions upon request.

[4] http://dev.mendeley.com

[5] http://ipscience-help.thomsonreuters.com/wosWebServicesLite/WebServicesLiteOverviewGroup/Introduction.html

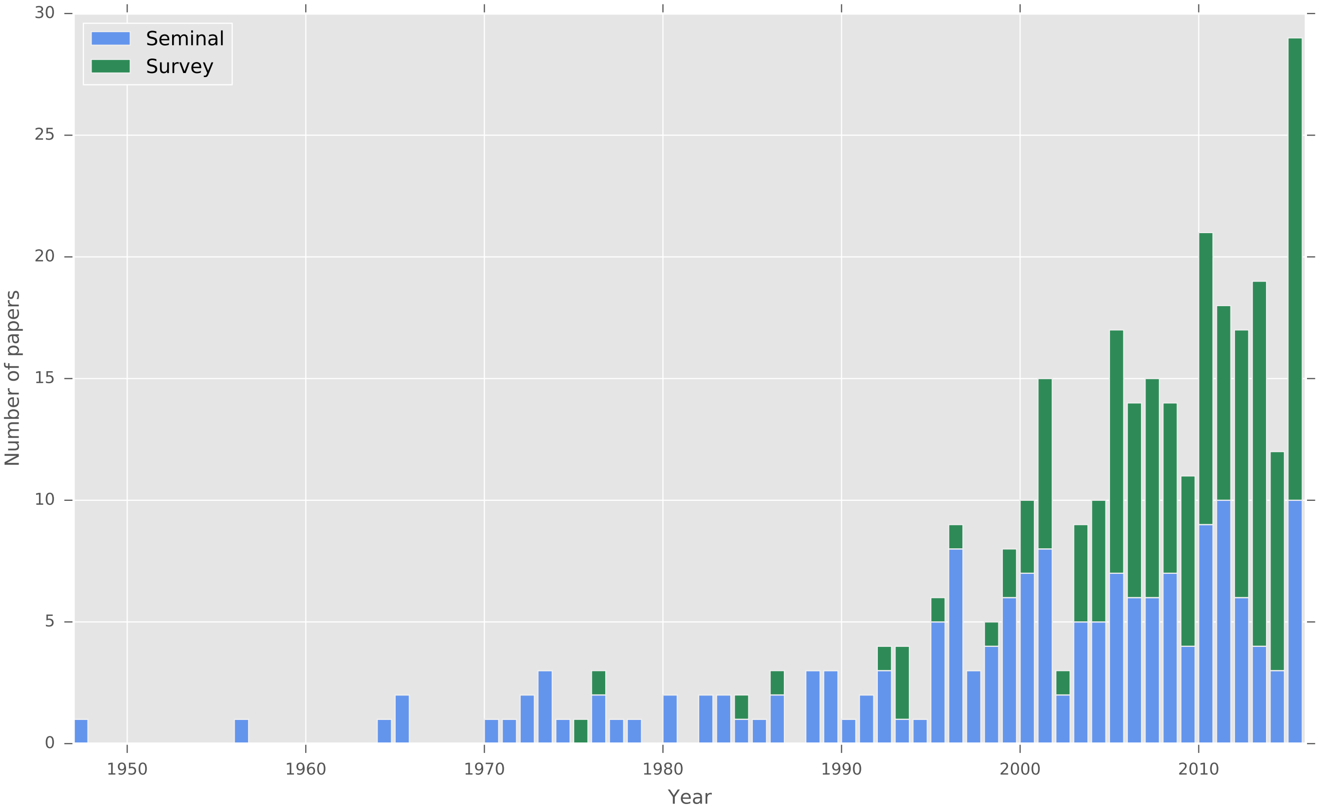

Figure 1: Histogram of publication years.

productivity of scientific disciplines in general, however, we believe this does not influence our study.

When answering the questions about academic background, 22 respondents have selected "Other" instead of one of the listed disciplines, these 22 responses provided us with 40 papers in total. We looked at the detailed description of these 40 papers, 9 of them are related to astronomy (the descriptions provided were "Binary stars", "Martian meteorites", "cosmochemistry", "Planetary sciences" and "planetology"), 4 could be classified as computer science ("virtual reality" and "Natural Language Understanding, Spoken Language Understanding"), the rest relate to different areas (e.g. "Microbial degradation of plastic", "MOOC", etc.).

4.4 Citations and readership

The dataset contains two basic measures related to publication utility – citation counts, which we manually collected from Google Scholar, and the number of readers in Mendeley, which we gathered through the Mendeley API. We also had access to the number of citations in Web of Science and while we couldn't make these data available together with the dataset, we provide an analysis of the WoS citations and a comparison with the other two metrics.

Table 3 shows basic statistics of Google Scholar citation counts and Mendeley readership of each paper in the dataset. We consider the readership of papers which we didn't find in Mendeley to be 0 (as papers are added to the Mendeley database by their readers). It is interesting to notice that while seminal papers are on average cited more than survey papers, this is not the case for readership, in fact survey papers attract more readers than seminal papers despite being on average younger (Section 4.2). We believe this is an important finding as readership counts are being more and more frequently used as a measure of impact complementary to citations [21, 18, 22]. We believe the fact that survey papers are more read than seminal papers, while being less cited, shows that readership is a measure of popularity rather than importance.

	Citations		**Readership**	
	Seminal	**Survey**	**Seminal**	**Survey**
Mean	2,458	519	229	358
Std	8,885	1,197	873	1,555
Min	0	0	0	0
Max	85,376	12,099	10,019	15,437
25%	78	24	4	8
50% (median)	249	109	41	42
75%	1,302	596	154	166

Table 3: Descriptive statistics of Google Scholar citation counts and of Mendeley readership.

Table 4 shows a comparison of the number citations obtained from Google Scholar and from Web of Science. This table includes only those 110 papers (51 seminal and 59 survey papers) which appear in Web of Science. The higher

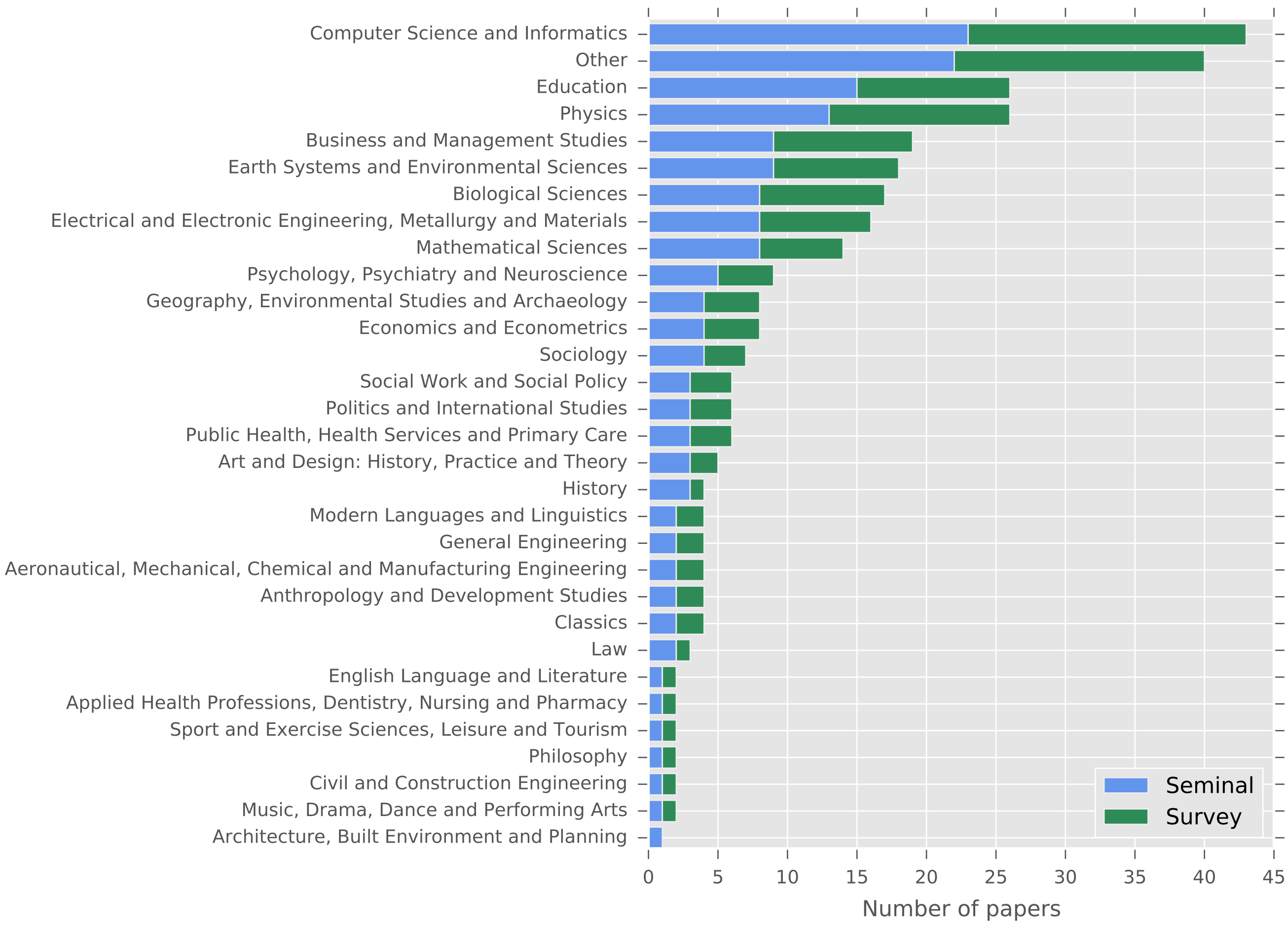

Figure 2: Histogram of publication disciplines.

citation numbers coming from Google Scholar are not surprising as Google Scholar's wider coverage of academic outputs is well known [9, 8]. This wider coverage is also demonstrated by the fact that we were able to find only 110 out of the 314 papers used in our study in Web of Science.

	Google Scholar		Web of Science	
	Seminal	**Survey**	**Seminal**	**Survey**
Mean	814	429	523	255
Std	1,599	566	926	373
Min	2	0	1	0
Max	8,246	2,446	4,753	1,709
25%	102	43	46	25
50% (median)	211	216	144	94
75%	929	612	677	354

Table 4: Descriptive statistics of citation counts acquired from Google Scholar and Web of Science.

This low coverage provided by Web of Science is quite striking, especially given the fact WoS misses more seminal papers. For example, a recent publication by Krizhevsky et. al. [16], a seminal deep learning paper which has caused a shift in the area of artificial intelligence/computer vision, is missing in WoS, but has attracted almost 8000 citations in GS since its publication in 2012. The reason for the paper not being indexed in WoS is probably the fact that the proceedings of the conference where the paper was presented aren't published through a major publisher but instead by the conference itself. We believe this is an interesting point at it shows important seminal work isn't always published by traditional routes in high impact journals. With the recent changes in scholarly communication towards Open Access and Open Science, it is reasonable to expect this to happen even more often in the future.

In order to compare whether the two databases rank papers similarly we have correlated the citation counts, the results are presented in Table 5. Both correlations are weaker for seminal papers, however this could be caused by the age difference between the two types of papers as the databases might have a lower coverage of older publications. Overall, both Pearson and Spearman correlations are otherwise strong.

5. EXPERIMENT & RESULTS

In this section we present the results of the experiment the aim of which was to test whether citation or reader-

	Spearman	Pearson
Seminal	$0.8581, p \ll .001$	$0.6775, p \ll .001$
Survey	$0.9696, p \ll .001$	$0.9588, p \ll .001$
Overall	$0.9281, p \ll .001$	$0.7254, p \ll .001$

Table 5: Correlation between Google Scholar and Web of Science citation counts.

ship counts work as a discriminating factor for distinguishing seminal and survey papers. These two measures, and especially citation counts, are frequently used in research evaluation as proxies for scientific influence or quality. We believe a metric used for such purposes should be able to distinguish between these two categories of papers which represent two very different types of work and where only the seminal papers have contributed to a significant change in their respective discipline.

In order to test our hypothesis, we classify the papers into two classes (survey, seminal) based on their citation counts and based on their readership counts. As a baseline we use a model which classifies all papers as seminal, as that is the majority class. This baseline model achieves the accuracy of 52.87%.

To classify papers based on their citation and readership counts we use the following model: if the total number of citations (or the number of readers) for a given paper is higher than a selected threshold we classify the paper as seminal, otherwise as a survey. We find such threshold which minimizes the misclassification rate. We do this by calculating the misclassification rate for all thresholds in the interval $[0, max(citation\_count)]$ for the model using citation counts and $[0, max(reader\_count)]$ for the model using reader counts.

The threshold with the lowest misclassification rate is 50 for citation counts and 0 for readership counts, suggesting that the readership model doesn't outperform the baseline model. Table 6 shows the confusion matrix for the model using citation counts. The citation counts model classifies correctly 63.38% of the papers, which represents about 10% improvement over the baseline. We have calculated dependent t-test to confirm that the improvement over the baseline obtained by using citation counts is significant. The test confirmed the hypothesis with $t = -11.1269$ and $p \ll .001$.

		Predicted		
		Survey	Seminal	Total
Actual	Survey	41.22% (61)	58.78% (87)	148
	Seminal	16.87% (28)	83.13% (138)	166
	Total	89	225	314

Table 6: Confusion matrix for predicting the class of the paper using Google Scholar citation counts.

For a comparison we have repeated the classification experiment on the subset of publications found in WoS. First we classify these 110 papers using citation data obtained from WoS, then again using citations from GS. For the WoS citations the threshold with the lowest misclassification rate was found to be 605, for the GS citations it is 72. The baseline model (a model which classifies all papers as survey, because that is the majority class in this case) predicts correctly 53.64% of papers, the WoS citation model classifies correctly 60.91% of the papers and the GS citation model 59.09%. The improvement over the baseline provided by both citation sources was again found to be significant ($t = 4.9216$ and $p \ll .001$ for WoS citations and $t = 16.2999$ and $p \ll .001$ for GS citations).

Finally we classify the papers based on the number of GS citations received per year. Our reasoning for this is that we found the two classes differ in terms of age. Because one of the groups had more time to accumulate citations than the other, we normalize the GS citation counts of each paper by the number of years since the year of publication of that paper (1 for papers published in 2016, 2 for papers from 2015, etc.). In this case the threshold with the lowest misclassification rate is found to be 2. Table 7 shows the confusion matrix for this model. The accuracy of the model is 56.69% while the accuracy of the baseline model is 52.87% (classifying all as seminal). In this case the benefit of using citations is minimal, however the improvement is still statistically significant ($t = -6.5646$ and $p \ll .001$.

		Predicted		
		Survey	Seminal	Total
Actual	Survey	7.96% (25)	39.17% (123)	148
	Seminal	4.14% (13)	48.73% (153)	166
	Total	38	276	314

Table 7: Confusion matrix for predicting the class of the paper using citations per year.

6. DISCUSSION

We believe this study is novel in two ways. Firstly, we show on a real dataset of research publications that citation counts and reader counts don't work well in distinguishing important seminal research from literature surveys. While citations have been, especially because of their widespread usage, subject to much criticism for their unsuitability for research evaluation, this work is among the first to quantify the significance of this problem. In addition, our contributions include a novel dataset of 314 seminal and survey research publications, which is publicly available. We believe this dataset will be useful in developing new metrics. This work is however far from complete as a larger dataset would help to further the efforts in improving the way research is evaluated. We believe an ideal dataset for evaluating research metrics should meet the following requirements:

- Cross-disciplinary: A dataset containing publications from different scientific areas is important for two reasons. Firstly, publication patterns are different for each discipline, both in terms of productivity and types of outcomes (conference papers, journal papers, books, etc.). This is also important to enable detecting research which finds use outside of its domain.

- Time span: The dataset should also contain publications spanning a wider time frame. One of the reasons for this is that publication patterns are different not only across disciplines, but they keep changing also in time. Furthermore, some research publications only find use after a certain period of time, but nevertheless represent important research.

- Publication types: Different types of research publications (e.g. pure research, applied research, literature review, dataset description, etc.) provide different types of impact. This should be taken into account when developing new research evaluation metrics. For example, a publication presenting a system might not receive many citations, because it presents a final product rather than research others can build on. Such publication might however be very useful to its users and so have large societal impact.
- Peer review judgements: Finally, to provide a reference rank for comparing the research evaluation metrics to, the dataset should contain fair and unbiased judgements provided by experts in the area of the publication. These judgements should rate the publications based on an agreed set of rules and standards.

Creating such a dataset would require significant time and resources, both in terms of collecting a representative sample of publications and in terms of providing peer review judgements for these publications. While there was a recent effort to create such a dataset (Section 7), in this case the evaluation set contained only publications from one discipline (computer science) and the peer review judgements were not shared. Providing the peer review judgements could perhaps be a common effort and an existing open peer review system could be used for this task. This would still require selecting the reference publications, creating a set of rules according the which the papers in the set should be judged and ensuring fairness of the peer review.

7. RELATED WORK

The 2016 WSDM Cup Challenge [29] has been probably the biggest effort in this area up to date. The goal of the challenge was to provide a static rank for papers contained in the Microsoft Academic Graph (MAG) dataset [26]. The evaluation set has been built by computer science academics using pairwise judgement on a subset of the publications in MAG. We have provided an analysis of the evaluation method in our paper [11]. Unfortunately this evaluation set has not been published and so this effort does not extend beyond the challenge.

Our work is close to a recent effort by [28] in which the authors argue that not all citations are equal and that identifying which citations are important is necessary for better understanding of published research. They describe a supervised classification approach for classifying citations as important or incidental and explore a set of twelve features. This is well aligned with our research, as we show using all citations regardless of their meaning doesn't work well for distinguishing excellent research. As a future work we would like to test the model presented in [28] on our dataset to see whether important citations will help in distinguishing seminal research better.

8. CONCLUSIONS

In our study we show citation counts distinguish between seminal and survey papers only poorly. This issue is even more prominent when we consider the age of the publications – while the survey papers in our dataset are on average almost a decade younger than seminal papers, they are still very highly cited, in fact so much that they become indistinguishable from the seminal papers. This shows caution should be exercised when using citation counts as a proxy for scientific importance or quality.

In addition to quantifying the success rate when using citations for identifying excellent research, we also present and share[6] a novel dataset of 314 annotated seminal and survey research publications along with their metadata.

Finally, we show that the oldest and probably the best known citation database, Web of Science, which is frequently used in citation studies and to calculate metrics such as the Journal Impact Factor, omits many important seminal publications. We have used citation counts obtained from WoS to classify papers into the two categories and compared the results with classification using GS citations. This comparison has shown using citation data from WoS doesn't provide improved classification accuracy compared with citation data from GS.

9. REFERENCES

[1] Research Excellence Framework (REF) 2014 Units of Assessment. http://www.ref.ac.uk/panels/unitsofassessment/. Accessed: 2016-11-11.

[2] D. W. Aksnes. Characteristics of highly cited papers. *Research Evaluation*, 12(3):159–170, 2003.

[3] T. C. Almind and P. Ingwersen. Informetric analyses on the world wide web: Methodological approaches to'webometrics'. *Journal of documentation*, 53(4):404–426, 1997. DOI: 10.1108/EUM0000000007205.

[4] B. M. Althouse, J. D. West, C. T. Bergstrom, and T. Bergstrom. Differences in Impact Factor Across Fields and Over Time. *Journal of the American Society for Information Science and Technology*, 60(1):27–34, 2009.

[5] C. A. D'Angelo and G. Abramo. Publication Rates in 192 Research Fields of the Hard Sciences. In *Proceedings of the 15th ISSI Conference*, pages 915–925, 2015.

[6] F. Galligan and S. Dyas-Correia. Altmetrics: Rethinking the way we measure. *Serials review*, 39(1):56–61, 2013.

[7] E. Garfield. Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, 122(3159):108–111, 1955. DOI: 10.1126

[8] A.-W. Harzing. Microsoft Academic (Search): a Phoenix arisen from the ashes?, 2016.

[9] A.-W. Harzing and S. Alakangas. Google Scholar, Scopus and the Web of Science: A longitudinal and cross-disciplinary comparison. *Scientometrics*, 106(2):787–804, 2016.

[10] D. Herrmannova and P. Knoth. Semantometrics in coauthorship networks: Fulltext-based approach for analysing patterns of research collaboration. *D-Lib Magazine*, 21(11/12), 2015. DOI: 10.1045/november2015-herrmannova.

[11] D. Herrmannova and P. Knoth. Simple yet effective methods for large-scale scholarly publication ranking.

[6] http://trueimpactdataset.semantometrics.org

In *WSDM Cup 2016 – Entity Ranking Challenge Workshop at International Conference on Web Search and Data Mining (WSDM)*, San Francisco, CA, USA, Feb 2016.

[12] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences of the United States of America*, pages 16569–16572, 2005. DOI: 10.1073/pnas.0507655102.

[13] P. Ingwersen. The calculation of web impact factors. *Journal of documentation*, 54(2):236–243, 1998. DOI: 10.1108/EUM0000000007167.

[14] A. E. Jinha. Article 50 million: an estimate of the number of scholarly articles in existence. *Learned Publishing*, 23(3):258–263, 2010.

[15] P. Knoth and D. Herrmannova. Towards semantometrics: A new semantic similarity based measure for assessing a research publication's contribution. *D-Lib Magazine*, 20(11):8, 2014.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[17] M. Laakso and B.-C. Björk. Anatomy of open access publishing: a study of longitudinal development and internal structure. *BMC medicine*, 10(1):1, 2012. DOI: 10.1186/1741-7015-10-124.

[18] N. Maflahi and M. Thelwall. When are readership counts as useful as citation counts? scopus versus mendeley for lis journals. *Journal of the Association for Information Science and Technology*, 67(1):191–199, 2016. DOI: 10.1002/asi.23369.

[19] R. M. Patton, C. G. Stahl, and J. C. Wells. Measuring Scientific Impact Beyond Citation Counts. *D-Lib Magazine*, 22(9/10):5, 2016.

[20] M. Peplow. Peer review – reviewed. *Nature News*, Dec 2014.

[21] H. Piwowar and J. Priem. The power of altmetrics on a cv. *Bulletin of the American Society for Information Science and Technology*, 39(4):10–13, 2013. DOI: 10.1002/bult.2013.1720390405.

[22] J. Priem. Altmetrics. In B. Cronin and C. R. Sugimoto, editors, *Beyond bibliometrics: harnessing multidimensional indicators of scholarly impact*, chapter 14, pages 263–288. MIT Press, Cambridge, MA, 2014.

[23] J. Priem, D. Taraborelli, P. Groth, and C. Neylon. Altmetrics: A manifesto. 2010. Accessed: 2016-11-07.

[24] REF 2014. Panel criteria and working methods. Technical Report January 2012, 2012.

[25] P. O. Seglen. Why the impact factor of journals should not be used for evaluating research. *BMJ: British Medical Journal*, 314(February):498–502, 1997.

[26] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-j. P. Hsu, and K. Wang. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th International Conference on World Wide Web*, pages 243–246. ACM, 2015. DOI: 10.1145/2740908.2742839.

[27] R. Smith. Peer review: a flawed process at the heart of science and journals. *Journal of the royal society of medicine*, 99(4):178–182, 2006.

[28] M. Valenzuela, V. Ha, and O. Etzioni. Identifying meaningful citations. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[29] A. D. Wade, K. Wang, Y. Sun, and A. Gulli. Wsdm cup 2016: Entity ranking challenge. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 593–594. ACM, 2016. DOI: 10.1145/2835776.2855119.

[30] R. Whalen, Y. Huang, A. Sawant, B. Uzzi, and N. Contractor. Natural Language Processing, Article Content & Bibliometrics: Predicting High Impact Science. *ASCW'15 Workshop at Web Science 2015*, pages 6–8, 2015.

[31] J. Wilsdon, L. Allen, E. Belfiore, P. Campbell, S. Curry, S. Hill, R. Jones, R. Kain, S. Kerridge, M. Thelwall, J. Tinkler, I. Viney, P. Wouters, J. Hill, and B. Johnson. *The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management.* 2015.

[32] R. Yan, C. Huang, J. Tang, Y. Zhang, and X. Li. To Better Stand on the Shoulder of Giants. In *Proceedings of the 12th Joint Conference on Digital Libraries*, pages 51–60, Washington, DC, 2012. ACM.