Comprehensive Evaluation of XSEDE's Scientific Impact using Semantic Scholar Data

Fugang Wang
fuwang@indiana.edu
Indiana University
Bloomington, IN, USA

Gregor von Laszewski
laszewski@gmail.com
Indiana University
Bloomington, IN, USA

Geoffrey C. Fox
gcf@iu.edu
Indiana University
Bloomington, IN, USA

ABSTRACT

The United States science and engineering community faces multiple challenges related to funding and funding policies for science and engineering. A framework is needed to evaluate the impact of scientific facilities and instruments. In this paper, we demonstrate such an activity through our comprehensive work evaluating the scientific impact of XSEDE using the Semantic Scholar Data. In contrast to other studies, our study includes the bibliographic references of all recorded papers related to XSEDE over the entire performance period till March of 2021. This makes this study unique and distinguishes it from our earlier work while using (a) over 180 million papers as a comparison to our peer analysis, (b) include all publications reported, and (c) conduct the study repeatedly over several years.

CCS CONCEPTS

• **Information systems → Information systems applications**; *Information retrieval*; Data management systems.

KEYWORDS

XSEDE, Scientific Impact, Peer Comparison, FWCI

ACM Reference Format:
Fugang Wang, Gregor von Laszewski, and Geoffrey C. Fox. 2021. Comprehensive Evaluation of XSEDE's Scientific Impact using Semantic Scholar Data. In *Practice and Experience in Advanced Research Computing (PEARC '21), July 18–22, 2021, Boston, MA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3437359.3465601

1 INTRODUCTION

We use a bibliometrics approach to evaluate the scientific impact of XSEDE. This requires identifying the XSEDE related publications, defined as those that were a direct or indirect scientific output from using XSEDE resources. We obtained them from the XSEDE central database, which includes data uploaded by the users to the XSEDE User Portal [2]. A portion of the publication data come from our extensive work from parsing the publication appendix section from the historical TeraGrid/XSEDE reports to NSF before the XSEDE user portal supporting the publication uploading feature.

We externally verify each of these publications and obtained citation data for those publications from third-party website or services, e.g., using Web of Science [3], Google Scholar [1], Microsoft Academic Graph [11], and Semantic Scholar dataset [10], the one source we focused in this study. The validation was important as many entries were not properly reported or only included preliminary citations. Through our curation process the data has been significantly improved and becomes usable for such a large study.

With the number of publications (after verification from external sources) and the citation data, we can then derive other well-known scientific impact metrics, such as H-index [9], G-index [5], i10-index [7], etc., for the different levels of XSEDE entities such as users, organizations, projects, fields of study. However more importantly, we can do various more sophisticated analysis such as a unique peers comparison [14][15] and a Field-Weighed Citation Impact (FWCI) [4] study. This paper focuses on the last two.

2 ARCHITECTURE

Our architecture is based on the premise that it can be applied to the analysis of any virtual organization. Furthermore, it should be possible to analyze bibliometrics for several organizations in parallel. Hence a general architecture is needed that provides scalability while being able to integrate comprehensive access to compute and data services supporting the analysis and is best suited or is available for it. This may include Kubernetes, Virtual Machines, Hadoop, Spark, and others. The architecture calls for functionality that can be hosted dynamically upon demand on cloud providers or even clusters such as available in XSEDE or departmental clusters. They are coordinated by utilizing cloud and compute abstractions as pioneered by Cloudmesh in collaboration with NIST [8, 12, 13]. A *Scientific Impact and Bibliometric Middleware Coordinator* facilitates the orchestration of compute and data infrastructure while allowing custom queries and metric analysis to be conducted on selected data sets based on bibliographic information that is fed to the system through its API. As we rely on reusable Interfaces via REST on all levels the architecture promotes reusability by users not only through the access within our own portal, but also through other interfaces such as Python APIs, Jupyter notebooks, or R-studio. Security to the system is controlled through common authentication and authorization frameworks and needs to be customized (for our study with XSEDE we integrated authentication with XSEDE security services). The comprehensive architectural vision we outline here is depicted in Figure 1.

PEARC '21, July 18–22, 2021, Boston, MA, USA

ACM ISBN 978-1-4503-8292-2/21/07.
https://doi.org/10.1145/3437359.3465601

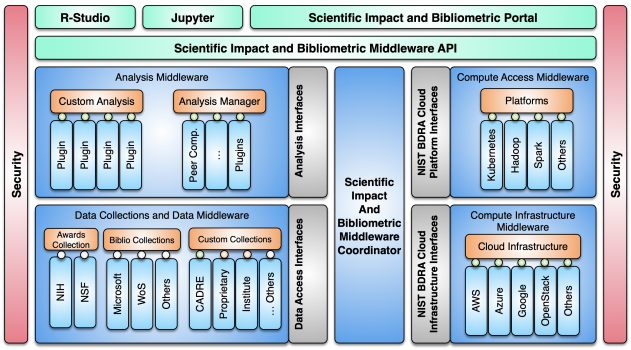

Figure 1: Architecture of the comprehensive scientific impact metrics analysis

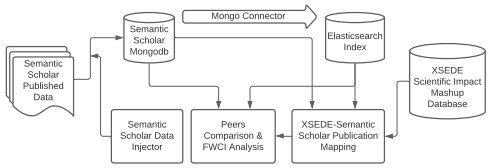

Figure 2: Analysis Workflow using Semantic Scholar Dataset

2.1 XSEDE Scientific Impact Analysis Workflow with Semantic Scholar Dataset

To adapt our general architecture to the needs of XSEDE we have implemented an analysis workflow as part of the Scientific Impact (SI) and Bibliographic Middleware Coordinator while relating it to XSEDE specific data and workflow processing needs as shown in Figure 2. To integrate and utilize the Semantic Scholar data, we added a customized workflow to integrate the data to our main mashup database - the *XSEDE Scientific Mashup Database* that we use to evaluate the XSEDE system's scientific impact. It contains data from the XSEDE user portal (XSEDE entities and publication data) and Web of Science (the citation data for the publications), NSF awards database, among others. We regularly update this data mashup and generate the updated scientific impact metrics, such as H-index, G-index, and other metrics for the XSEDE entities on different levels, such as user, project, organization, and field of study. However, for this comprehensive study, we integrated a new data source – the *Semantic Scholar Dataset* [10]. The workflow starts with the download of the *Semantic Scholar Published Data* Dataset, which contains about six thousand compressed text files (as of the end of February 2021).

A *Semantic Scholar Data Injector* carries out the necessary curation and writes the data into our *Semantic Scholar MongoDB* database. The Semantic Scholar publication data has *186 million* entries, thus imposing challenges for efficient queries during the analysis. To facilitate efficiency we added an *Elasticsearch Indexing* process that creates indices of the MongoDB data to improve the full-text query performance when relating to specific attributes (e.g., query title, journal name, field of study, etc.). This process is facilitated by a *Mongo connector* utility that synchronizes the data from the MongoDB database to elasticsearch. A *XSEDE-Semantic Scholar Publication Mapping* process queries all XSEDE publications against the elasticsearch to create a mapping of publication IDs between the two data sources. Finally, the main analysis module, the *Peers Comparison and FWCI Analysis*, conducts the data analysis for the peers comparison and the FWCI [4].

3 RESULTS AND DISCUSSION

In this section we demonstrate the analysis conducted and provide discussions about them. The results are organized into three parts, the *peers comparison study* (Section 3.1), the *Field-Weighted Citation Impact (FWCI)* analysis (Section 3.2), and the *highly cited papers* analysis (Section 3.3).

3.1 Peers Comparison Study

A publication's impact is usually judged by where it is published, and how many times it is cited by other publications. We extended the evaluation method as described in [16] and introduced the peers comparison study of papers [14][15], in which we compare a paper's relative impact compared to other peers published in the same journal issue. We calculate the citation count percentile ranking score of a target publication, which shows the relative standing of the paper's citation count among all the peers in the same venue. E.g., a percentile ranking score of 60 means the paper has got more citations than the 60% of papers published in the same venue (the same issue of the journal where it published).

In this study, we have compared *12461* XSEDE publications from *231* publication venues against their peers. Each publication venue has at least 10 XSEDE publications published on them to avoid the possible statistically insignificant results if too few publications appear in a publication. Figure 3 shows the average percentile ranking score for each publication venue, sorted in descending order. The horizontal red line is the baseline of score 50. Hence, Figure 3 shows that for the majority of the 231 publication venues the XSEDE publications had received more citations compared to an average peer.

Figure 4 shows the histogram of the averaged percentile ranking score for each publication venue. The distribution skews towards the right side suggesting that the XSEDE publications, in general, received more citations than the average peers.

When grouping the percentile ranking scores by the publication's identified *field of study*, we obtain the result as shown in Figure 5. Also here, we discarded the data points for a few fields of study due to the limited number (less than 10) of XSEDE publications. The results show that the average percentile ranking score is above 50.

In one of our previous studies [15], we have conducted the peers comparison analysis based on the data of more than 5000 publications from about 120 publication venues. Now with more data available over a longer period of time we could verify the trend due to its similar results. Hence we assess that the studies results validate each other and show the consistent performance of XSEDE publications.

3.2 Field-Weighted Citation Impact (FWCI) Analysis

Throughout the duration of XSEDE, we are producing periodically updated scientific impact metrics and publish them in a web portal [6]. When comparing the impact metrics between different entities, sometimes the comparison may not be fair and informative if they

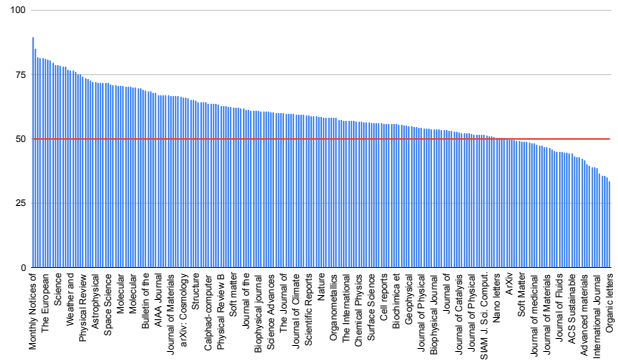

Figure 3: Percentile ranking scores averaged by publication venue

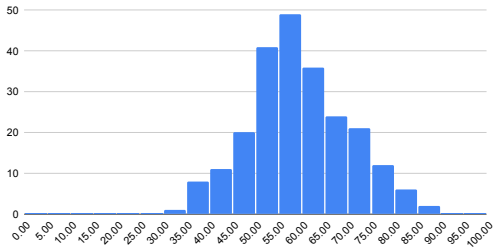

Figure 4: Histogram of average percentile ranking scores by publication venue

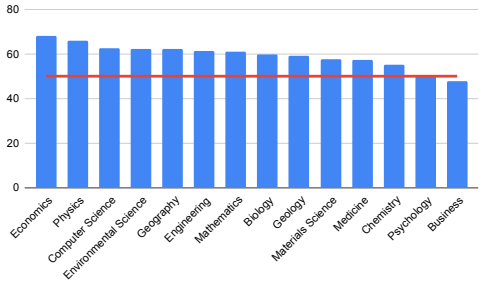

Figure 5: Percentile ranking scores averaged by field of study

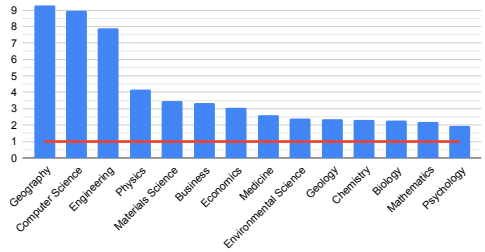

Figure 6: Field Weighted Citation Impact

belong to different fields of study. Publications from a certain field of study might receive in general higher citations compared to some similar publications but in another different field of study.

However, the FOS analysis from our peers comparison described in the previous section has eliminated this concern as we are comparing the *relative* standing of publications within the *same* field of study and summing them up relatively. Another approach, as proposed by the snowball metrics [4], is to use Field-Weighted citation to do the comparison and evaluation. In this approach, the citation data of a group of publications from the same field of study from a target group (here, all XSEDE publications) are compared to *all* the other publications from the same field of study.

$$FWCI = avg(CC_{group})/avg(CC_{field})$$

Although this may appear very straightforward, it is however in practice not that convenient to calculate, due to the fact that accessing all the publications and their citation data may not be an easy task and is limited to the available publication data sets. In our previous study [15] we used the Microsoft Academic Graph dataset [11], which had, at that time, about *58* million publication entries. For our new study we have significantly increased the available data set to *186* million publication entries while leveraging data from Semantic Scholar. With this new data set we repeated our FWCI analysis to obtain updated results as depicted in Figure 6. It lists the FWCI values for each field of study. The horizontal red line shows the baseline at 1. It shows that for all the FOS the *FWCI* values are greater than 1, which indicates that the XSEDE publications in each FOS received more citations than expected. For the *Geography, Computer Science, and Engineering* the publications received about 8 times more citations than expected from the field average. Clearly, research conducted by the researchers using XSEDE resources have a larger impact while using citation count as metric.

3.3 Highly Cited Papers

With access to the much larger publication dataset, we can also identify for each field of study the highly cited papers. We consider the top 5% and top 1% most cited papers for the results reported here. Hence, we can find out what portion of XSEDE publications fall into those groups. Also here, we excluded the few fields of study that had a very small number of publications identified as belonging to them (less than 10) due to the same reason as mentioned in the previous sections. Figure 7 shows the actual numbers, percentages for each field of study that fall into the top 5% and top 1% categories, as well as the number of XSEDE publications belonging to that field of study. The blue and yellow lines indicate the 5% and 1% baseline. All fields of study show a disproportionately higher percentage of XSEDE publications fall into the highly cited papers categories. As a example we find specifically that 38.4% of geography related papers were in the top 5% highly cited papers, and 13.7% were in the top 1% most cited papers. Table 1 summarizes the data for all the fields of study.

4 CONCLUSION

We have identified that our general architecture can be used to facilitate scientific impact analysis for virtual organizations. We have demonstrated its usefulness for the XSEDE community and significantly enhanced our analysis. This is facilitated by developing tools and methods and collecting publication and citation data to evaluate and track the scientific impact of XSEDE (or other similar virtual organizations). In addition to the periodically updated frequently-used scientific impact metrics that are available from our portal, we are capable of supporting unique peer comparison studies, as well as

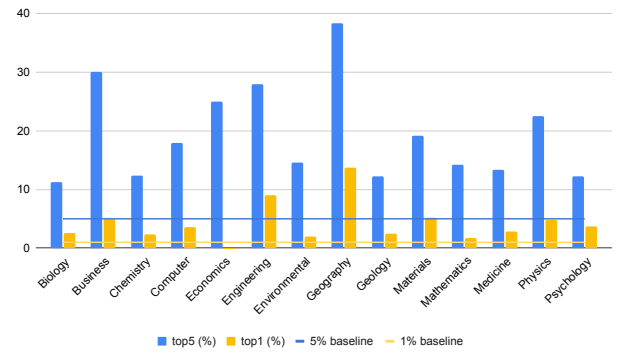

Figure 7: Percentage of XSEDE publications falling into the top 5% and top 1% for each field of study

Table 1: Highly Cited Papers Statistics (in top 5% and 1%)

Field	# in top 5%	% in top 5%	# in top 1%	% in top 1%	# XSEDE publications
Biology	213	11.3	50	2.7	1886
Business	6	30.0	1	5.0	20
Chemistry	609	12.4	114	2.3	4924
Computer Science	476	17.9	95	3.6	2658
Economics	7	25.0	0	0.0	28
Engineering	65	27.9	21	9.0	233
Environmental Science	76	14.6	10	1.9	521
Geography	28	38.4	10	13.7	73
Geology	70	12.3	14	2.5	571
Materials Science	615	19.2	165	5.2	3202
Mathematics	76	14.2	9	1.7	534
Medicine	1062	13.3	231	2.9	7978
Physics	1236	22.5	268	4.9	5504
Psychology	10	12.2	3	3.7	82

utilizing more and diverse dataset to be integrated in such analysis. This has been demonstrated by the integration of Semantic Scholar Data to conduct a comprehensive peers comparison study and Field-Weighted Citation Impact (FWCI) analysis. The results show that XSEDE publications had received more citations compared to their average peers in the same publication venue, or within the same filed of study. The FWCI analysis shows that XSEDE publications had received two times more citations comparing to the expected average in the same field. For some fields it was even 7-9 times higher. The highly cited papers statistics results also indicate that XSEDE publications have disproportionately higher percentage to fall into the highly cited papers categories. It is important to note that it is possible in our architecture to ingest new metrics to cover the diverse needs of the community to conduct different impact analytics. Lastly, in comparison to our earlier work, we increased the data from 58 million to 186 million by utilizing new dataset. With this considerably larger coverage, compared to our previous study, and with many more publications added from XSEDE, we were able to verify a trend of consistent impact of XSEDE using the peers comparison study, FWCI analysis, and highly cited papers analysis. These updated results and the previous one [15] validate each other and indicate that XSEDE's scientific impact is consistent across the years.

Semantic Scholar has been publishing their dataset in monthly basis more recently. With the availability of this data we could utilize the proposed architecture and workflow to redo the analysis with updated dataset periodically. This provides a unique advantage that, instead of comparing the results from different data source which may not allow consistent comparisons, we can observe changes and trends over a common basis using the same data source, leading to continuous comparable analysis reports.

It is important to note that any future facility of the scale of XSEDE should have such an analysis as part of its reporting. Although we are able to and have published on our portal the scientific impact metrics for different level of entities, we have not included them in this paper. Additional work that can be conducted will also allowing us to draw relationships between funded projects as they are reported to within the XSEDE allocations process.

In case you are interested in obtaining an impact analysis for your organization, department, institute, facility or group, please contact `laszewski@gmail.com`.

ACKNOWLEDGMENTS

This work is part of the XSEDE Metrics Service (XMS) project sponsored by NSF under grant number OCI-1025159.

REFERENCES

[1] [n.d.]. Google Scholar. Web Page. http://scholar.google.com/
[2] [n.d.]. XSEDE User Portal. Web Page. https://portal.xsede.org
[3] Clarative Analytics. 2021. Web of Science. Web Page. http://wokinfo.com/
[4] Lisa Colledge. 2014. Snowball Metrics Recipe Book. 2014.
[5] Leo Egghe. 2006. Theory and practise of the g-index. *Scientometrics* 69, 1 (2006), 131–152.
[6] Gregor von Laszewski Fugang Wang. 2021. XSEDE Scientific Impact Metrics Portal. https://sciimp.ccr.xdmod.org/xdportalpub/
[7] Google. 2011. i-10 index | Google Scholar Citations Open To All. http://googlescholar.blogspot.com/2011/11/google-scholar-citations-open-to-all.html
[8] Wo. L Chang Gregor von Laszewski. 2019. *NIST Big Data Interoperability Framework: Volume 8, Reference Architecture Interfaces.* Technical Report. NIST. https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-9r1.pdf
[9] Jorge E Hirsch. 2005. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences of the United States of America* 102, 46 (2005), 16569–16572.
[10] Semantic Scholar. 2021. Semantic Scholar Dataset. http://s2-public-api-prod.us-west-2.elasticbeanstalk.com/corpus/download/
[11] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June (Paul) Hsu, and Kuansan Wang. 2015. An Overview of Microsoft Academic Service (MAS) and Applications. In *Proceedings of the 24th International Conference on World Wide Web* (Florence, Italy) *(WWW '15 Companion)*. ACM, New York, NY, USA, 243–246. https://doi.org/10.1145/2740908.2742839
[12] Gregor von Laszewski. 2021. *Clusmesh Manual.* https://cloudmesh.github.io/cloudmesh-manual/
[13] Gregor von Laszewski, Anthony Orlowski, Richard H. Otten, Reilly Markowitz, Sunny Gandhi, Adam Chai, Geoffrey C. Fox, and Wo L. Chang. 2021. Using GAS for Speedy Generation of Hybrid Multi-Cloud Auto Generated AI Services. In *IEEE COMPSAC'21.*
[14] G. von Laszewski, F. Wang, G. C. Fox, D. L. Hart, T. R. Furlani, R. L. DeLeon, and S. M. Gallo. 2015. Peer Comparison of XSEDE and NCAR Publication Data. In *2015 IEEE International Conference on Cluster Computing*. 531–532. https://doi.org/10.1109/CLUSTER.2015.98
[15] Fugang Wang, Gregor von Laszewski, Timothy Whitson, Geoffrey C Fox, Thomas R Furlani, Robert L DeLeon, and Steven M Gallo. 2018. Evaluating the Scientific Impact of XSEDE. In *Proceedings of the Practice and Experience on Advanced Research Computing*. ACM, 10. https://doi.org/10.1145/3219104.3219124
[16] Fugang Wang, Gregor von Laszewski, Geoffrey C. Fox, Thomas R. Furlani, Robert L. DeLeon, and Steven M. Gallo. 2014. Towards a Scientific Impact Measuring Framework for Large Computing Facilities - a Case Study on XSEDE. In *Proceedings of the 2014 Annual Conference on Extreme Science and Engineering Discovery Environment* (Atlanta, GA, USA) *(XSEDE '14)*. ACM, New York, NY, USA, Article 25, 8 pages. https://doi.org/10.1145/2616498.2616507