Evaluating the Scientific Impact of XSEDE

Fugang Wang
Gregor von Laszewski*
Timothy Whitson
Geoffrey C. Fox
Indiana University
Smith Research Center, Ste 150
Bloomington, Indiana, U.S.A.

Thomas R. Furlani
Robert L. DeLeon
Steven M. Gallo
Center for Computational Research
University at Buffalo, SUNY
701 Ellicott Street
Buffalo, New York, 14203

ABSTRACT

In this paper we use the bibliometrics approach to evaluate the scientific impact of XSEDE. By utilizing publication data from various sources, e.g., ISI Web of Science and Microsoft Academic Graph, we calculate the impact metrics of XSEDE publications and show how they compare with the rest from the same field of study, or the peers from the same journal issue. We explain in detail how we retrieved, cleaned up, and curated millions of related publication entries. We then introduce the metrics we used for evaluation and comparison, and the methods how we calculate them. Detailed analysis results of Field Weighted Citation Impact (FWCI) and the peers comparison will be presented and discussed. We also explain briefly how the same approaches could be used to evaluate publications from a similar organization or institute, to show the generalization and ubiquitous nature of the presented evaluation approach.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Theory, Measurement

Keywords

Scientific impact, bibliometrics, h-index, Technology Audit Service, XDMoD, XSEDE

1. INTRODUCTION

*Corresponding Author.

PEARC18 ... 2018,USA

http://dx.doi.org/TBD ...$15.00.

To identify the impact on *scientific advancements enabled by enhanced cyberinfrastructure*, it is important to conduct a comprehensive analysis of achievements that can be attributed to the use of advanced infrastructure, such as provided by the Extreme Science and Discovery Environment (XSEDE) [5, 17].

We use the bibliometrics approach to evaluate the scientific impact of XSEDE. By acquiring related publication and citation data from multiple sources we calculate various metrics that show the impact of the publications and how they are compared to the peers that were published in the same journals, or in the same field of study. By crunching millions of publication data entries we normalized the citation count by field of study for comparison to eliminate the impact from the difference of field of study. We also introduced a novel approach to compare the target publications group with their peers published in the same publication venue, to further show how the target publications group performs compared to their peers within the same publication venue.

2. RELATED WORKS

Bibliometrics based analysis has been the most commonly used, a de facto standard way to evaluate the research impact of an individual, a research group, or even an organization. Publication count and citation count based metrics provided an effective way to show the quantity and quality, and the impact they generated, of scientific research activities. For instance it was used to evaluate the quality of research in UK [16, 15]. Most popular college/university rankings use citation based bibliometrics as an important factor to evaluate their research qualities, e.g., the overall publication count and citation count in a certain year or a year range; number or papers published in certain top journals; number of highly cited papers judging by if the citation count of a paper is on the top one percentage, etc.

As a virtual organization similar to XSEDE, Compute Canada also used bibliometrics based analysis to evaluate their impact of research [1].

Some previous work also studied the impact of TeraGrid [6], the early version of XSEDE, by analyzing the publications for one specific reserach allocation quarter, which involved very limited number of researchers and publications. Our work is unique in that it provides a *comprehensive* analysis superior in data volume, and with noval analyses approach such as Field Weighted Citation Impact and journal

publication-based peers comparison.

In addition to the more intuitive direct measuring of publication and citation count, some other derivative metrics such as h-index [12] and g-index [11] combines both publication and citation count to generate one metric. I10-index [2] on another hand measures only the count of those publications received at least ten references by other publications. In our evaluation we calculated such metrics for various XSEDE research entities to show the impact and comparison of the entities on the same level, e.g., individual, project, research field of study, organization, etc. The results are presented on the XDMoD scintific impact portal [4].

Usage based metrics[7, 8] were also proposed, in which it uses the usage, such as views and downloads, instead of more formal citations, of publications to evaluate the impact. However the actual use of this approach may be limited due to the fact that the usage data may not be available from a publisher, or different publishers may have different cretiria to measure the usage data thus created an inconsistent comparison base for papers published in different publishers. In our study we did not use such metrics in the evaluation due to this reason.

3. METHODOLOGY

We will first introduce the methodology we have used in the bibliometrics analysis to evaluate the scientific impact of XSEDE publications. This includes the clarification of the dataset and data sources used in the analysis; the approaches to define and calculate the various metrics; and the information of the software and service framework developed to facilitate this evaluation study.

3.1 Dataset and data sources

Several data sets and sources are involved in this study. These include:

- XSEDE publications (also include those from TeraGrid). These data are from two sources. One is from the user-submitted data from XSEDE user portal; another is from the past TeraGrid/XSEDE project reports to NSF. For the latter we extracted the publications appendix from the report text and then parsed the publication records before putting them into structured data into a database. The raw entries were over 20 thousand.

- Microsoft Academic Graph (MAG) data during the same time period (2005 - 2016) as the XSEDE publications. This dataset were retrieved with the API provided by Microsoft. The data were then cleaned up and put into a MongoDB database. This dataset has about 58 million entries.

- For the publication venues with at least 10 XSEDE papers appeared in them, we retrieved all the publications from them during the same time period (2005 - 2016) to facilitate the peers comparison study. This dataset has about 2 million entries.

3.2 Field Weighted Citation Impact Analysis

Field Weighted Citation Impact (FWCI) metric is proposed as part of the snowball metrics [10]. It calculates the average citation count of a target group of publications based on their field of science, and then compare that with the average citation count of the whole field of science in the same time period. The result is a ratio

$$fwci = avg(CC_group)/avg(CC_field)$$

So a fwci value greater than 1 means the interesed publication group had higher citation comparing to the expected value of the field of science, while a value less than 1 means the average citation the group received was less than the expected value by the field.

In this study we followed the following process to calculate the FWCI values for the XSEDE publications.

1. Query every raw XSEDE publication by title against the MAG data set, and verify the matching ones by checking other properties such as published year. After this process we have found the verified matching records in the MAG for all valid XSEDE publications. During this process we enabled elasticsearch [?] to improve both the accuracy and performance of the query due to the size of the dataset.

2. For each of the 58 million MAG data, we use the assigned field of study values to trace up to the top levels of the fields along with other related data from MAG. This process narrowed down the 30k different assigned fields of study to 19 overall top level fields of study as defined in the MAG dataset. One thing to notice was that each publications were assigned to multiple science fields in the original publication records, and the final top level science field category of a publication may not be unique as well. However as a lot of researches and publications are themselves multidisciplineary we think such results are valid and acceptable. In the following analysis we counted a publication in all the top level science fields we found following the tracing up process.

3. Once we have for each and every publications in the MAG dataset, we can calculate the average citation count by each top level field, for all the MAG publications and XSEDE publications respectively. And following that we could calculate the ratio to get the FWCI values.

3.3 Metric for Journal Publication-based Peer Comparison

We followed this process to obtain the needed data for the analysis.

1. We started this analysis by querying all XSEDE publications against the third party data source - ISI Web of Knowledge [3]. The XSEDE data, as explained before, contains the publication entries extracted from past TeraGrid/XSEDE reports to NSF, and the publication data from XSEDE user portal. Both are user-submitted data or compiled from user-submitted data, thus this query and verification process ensures the quality and accuracy of the dataset. The final results were about 9 thousands verified publications at the time of the study.

2. From this verified publications list, we find the subset of all the publication venues with at least 10 XSEDE publications published on them. And for each of the

publications published on these venues, we retrieve from ISI Web of Knowledge the extended metadata to get the exact volume and issue number of the publication venue where the publication was published. The reasons why we chose a threshold value 10 to identify a publication venue subset are:

(a) This ensures the statistical significance of the analysis results.

(b) This eases the data retrieval work substantially. While we have 1400 distinct publication venues identified from all the verified XSEDE publications, the subset when we use 10 as the minimum number of publications appeared on the venue was cut to 120 publication venues.

(c) The actual involved XSEDE publications in the peeers comparison was about 5 thousands, or about 56% of all the verified ones. This represents a good portion of all the data.

3. For all the 120 publication venues, we retrieved all the publications data published on them during the same time period as the TeraGrid/XSEDE publications (2005 2016).

Based on these data we can form the suitable comparison peers groups, which are the each single journal issue (or journal volume when no issue data available for some publications) if an XSEDE publication appeared on it. For each comparison peers group, we rank the citation count of each publication (including the XSEDE ones and the peers). The calculated percentile ranking values serve as the basis of the peer comparison study. The comparison is between publications in such journals that we identified as XSEDE papers and those that were not.

To apply the percentile ranking to the field of science of XSEDE publications among the journal issues where each publication was published, we aggregate them based on FOS, based on the project field of science data obtained from XSEDE central database (XDcDB). We then calculated the average and median percentile rank for each field of science.

To identify the FOS for each publication, we followed this process:

1. Find the FOS information out of past XSEDE quarterly reports as this information may have explicitly been associated with them.

2. Find the FOS information from the project data in the XDcDB. Similar to the case mentioned in MAG data pre-processing, it is possible that one project is associated with more than one FOS. In such cases we counted the publications of the project for all involved fields.

(a) Some publications from XSEDE quarterly reports were identified only by the project proposal number. We mapped them to the project charge number and account id used internally within the XSEDE central database.

(b) For user uploaded publications data via the XSEDE user portal, a project charge number was associated with each publication.

(c) Identify from the charge number the FOS as defined in the XDcDB.

3.4 Software Architecture Supporting the Study

We have developed a software framework supporting the study, which includes data acquisition, cleanup, processing and presentation. The framework is based on a distributed set of software services. The service-oriented system is a layered architecture consisting of components for:

- A data layer that retrieves publication and citation data from external sources. This includes data from the ISI Web of Knowledge; Microsoft Academic Graph; Google Scholar, and even NSF award database.
- Business logic layer that deals with:
  - parsing and processing while correlating data from various databases and services, such as the XSEDE central database (XDcDB).
  - metrics generation and an analysis system for different aggregation levels – users, projects, organization, field of science.
- a presentation layer using a lightweight portal in addition to exposing some data via a RESTful API [19].

Due to the use of the Software as a Service (SaaS) approach, our framework is expandable as we are able to integrate new services and data resources as required. Hence our framework can be adapted to other resource providers as demonstrated in [18]. Obviously, Adaptation could mean that we simply have to change the bibliometric data, which could mean that we need to integrate new data sources and curation services.

3.4.1 Service Integration into XSEDE and XDMoD

Our current framework for XSEDE includes services that are motivated by our initial findings from XSEDE bibliometric data. A RESTful service is integrated into XSEDE User Portal as part of the publication discovery service.

The various impact metrics of different levels of XSEDE entities - person, project, organization, field of study - as well as part of the analyses are available on the XDMoD scientific impact portal [4].

4. RESULTS AND DISCUSSIONS

4.1 Field Weighted Citation Impact Metrics

First we show the calculated FWCI values in Figure 1. The plot listed the FWCI for the top level fields of science as defined from the MAG data. Each data point also has the number of XSEDE publication as well as the number of all publications in that field. The red veticle lines indicates the points for FWCI=1. The figure shows all fields but one (political science, with only 3 publications) had FWCI values greater than 1, with majority fields having much higher values.

In figure 2 we display the similar data but shows and sorts the data based on the number of XSEDE publications in the field. This shows better the FWCI for the fields that majority of XSEDE publications fell in.

Figure 3 compares the expected citation count and the actual average citation count for each field. This again shows how XSEDE publications received much higher citations.

In figure 4 we show the extra citations XSEDE publications receive for each field, comparing to the expected value.

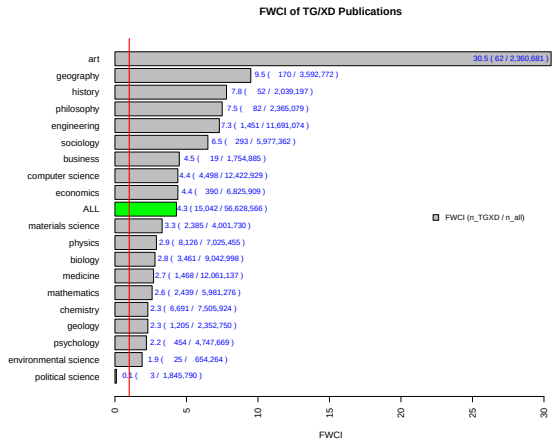

Figure 1: Field Weighted Citation Impact (FWCI) by Field sorted by FWCI.

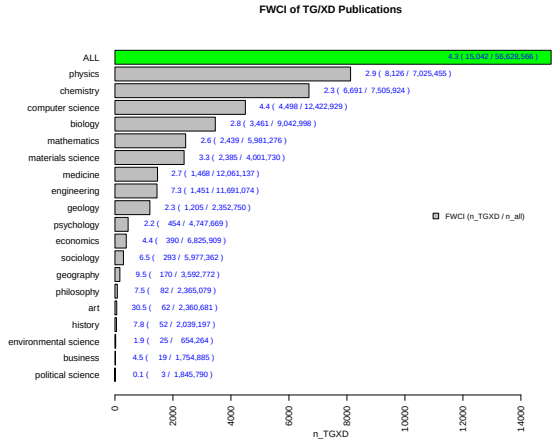

Figure 2: FWCI by Field sorted by Publication Count of the Field.

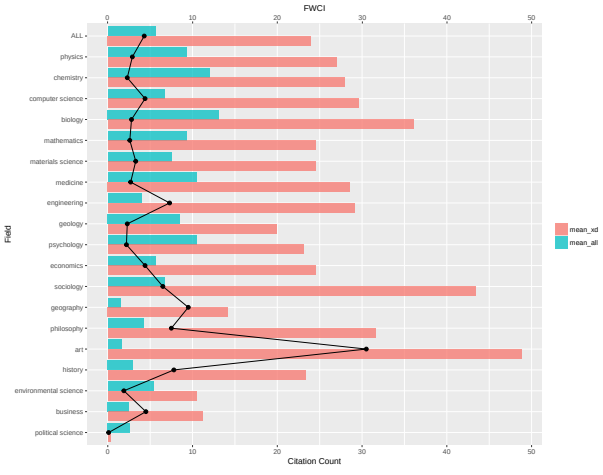

Figure 3: FWCI with Expected Citation Count and Actual Citation Count from XD Publications.

The availability of all the publications for each field makes it possible to calculate other interesting statistics, in addition to the above presented FWCI results. In Table 1 we display for each field the highly cited XSEDE papers (defined as top 1% and top 5% in citation count in that field) and the percentage of how many XSEDE publications fall into each category. The results show that for most fields a higher than expected percentage of XSEDE publications fall into the highly cited papers categories. E.g., when we consider all the publications and fields together, 5% XSEDE publications were in the top 1% highly cited group while 22.5% were in the top 5% highly cited group.

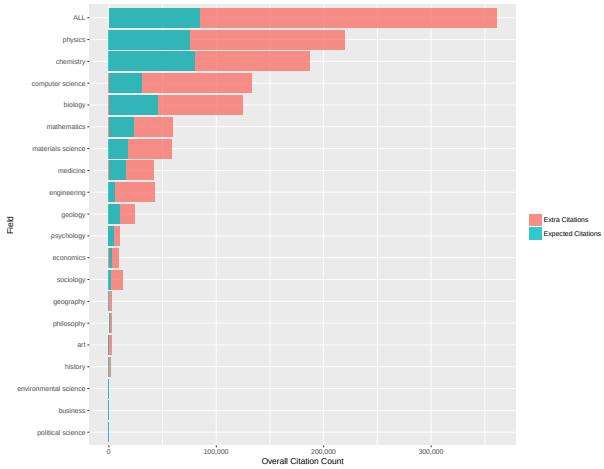

Figure 4: Extra Citation Count Achieved by XD Publications.

4.2 XSEDE Peer Data Analysis

Now we present the results from the peer compasiron study in a number of graphs and tables. Figure **??** shows the average percentile rank of XSEDE publications grouped by each publication venue. Figure **??** shows the similar but the median percentile rank values.

When we aggregate the results in fields of study instead of individual journal, we got the result as in Figure 7.

These show that for majority of the publications venues, or fields of science, XSEDE publications ranked at a higher percentile based on the citation count.

When we consider the overall comparison results, Figure 8 shows the distribution of the XSEDE publication's percentile rank in each 10% increment group. Again the result show the distribution skewed to the higher end, which means more XSEDE publications were ranked on the higher end.

Figure 9 shows the emperical cumulative distribution of the percentile ranks comparing to that of the peers group. Figure 10 shows the kernel density of the distributions of XSEDE publications' percentile ranking and that of peers'. As expected, the non-XSEDE peer publications are evenly distributed by percentile ranks with spike at 50% mostly coming from more recently published journal issues where most publications were not yet cited. The XSEDE publications are weighted to the higher percentile ranks side again. This again shows that XSEDE publications tend to be more highly cited comparing to their peers published in the same journal issue.

Table 2 lists the average and median rankings and citations received of the two groups to evaluate and compare. We used several non-parametric statistical tests to decide whether the population distributions are identical without

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

Field	# in top 1%	% in top 1%	# in top 5%	% in top 5%	# per 100,000	# XSEDE pubs
ALL	727	4.8	3380	22.5	26.6	15042
physics	292	3.6	1204	14.8	115.7	8126
chemistry	177	2.6	782	11.7	89.1	6691
computer science	223	5.0	1037	23.1	36.2	4498
biology	102	2.9	453	13.1	38.3	3461
mathematics	68	2.8	351	14.4	40.8	2439
materials science	108	4.5	446	18.7	59.6	2385
medicine	42	2.9	213	14.5	12.2	1468
engineering	111	7.6	414	28.5	12.4	1451
geology	33	2.7	183	15.2	51.2	1205
psychology	11	2.4	53	11.7	9.6	454
economics	26	6.7	101	25.9	5.7	390
sociology	18	6.1	63	21.5	4.9	293
geography	19	11.2	87	51.2	4.7	170
philosophy	11	13.4	31	37.8	3.5	82
art	15	24.2	39	62.9	2.6	62
history	6	11.5	19	36.5	2.6	52
environmental science	1	4.0	3	12.0	3.8	25
business	1	5.3	6	31.6	1.1	19
political science	0	0.0	0	0.0	0.2	3

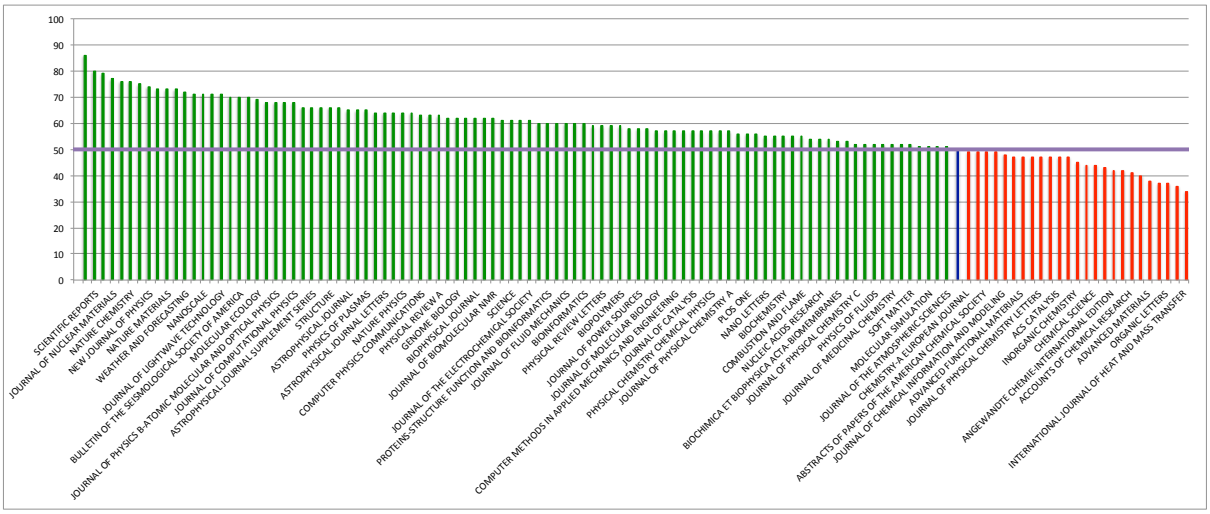

Figure 5: Average percentile ranking of XD publications by journal (by ISI)

assuming them to follow the normal distribution. We used Mann-Whitney-Wilcoxon test [14], Mood's median test [9], and Kruskal-Wallis test [13]. The results are as the following.

Wilcox test for citation count

- W = 1160300000, p-value < 2.2e-16. Alternative hypothesis: true location shift is not equal to 0

Wilcox test for percentile ranking

- W = 1090700000, p-value < 2.2e-16. Alternative hypothesis: true location shift is not equal to 0

Mood's median test for citation count

- p-value = 3.299883e-172

Mood's median test for percentile ranking

- p-value = 8.83052e-71

Kruskal-Wallis Test for citation count

- Kruskal-Wallis chi-squared = 1207.6, df = 1, p-value < 2.2e-16

Kruskal-Wallis Test for percentile ranking

- Kruskal-Wallis chi-squared = 632.35, df = 1, p-value < 2.2e-16

We performed a T-test to test if the citation differences were statistically significant.

Even though the distribution of the citation count of the XSEDE publication group we are study, and the peers group we are comparing to, are not necessarily normal distribution, due to the central limit theorem, when the sample size is

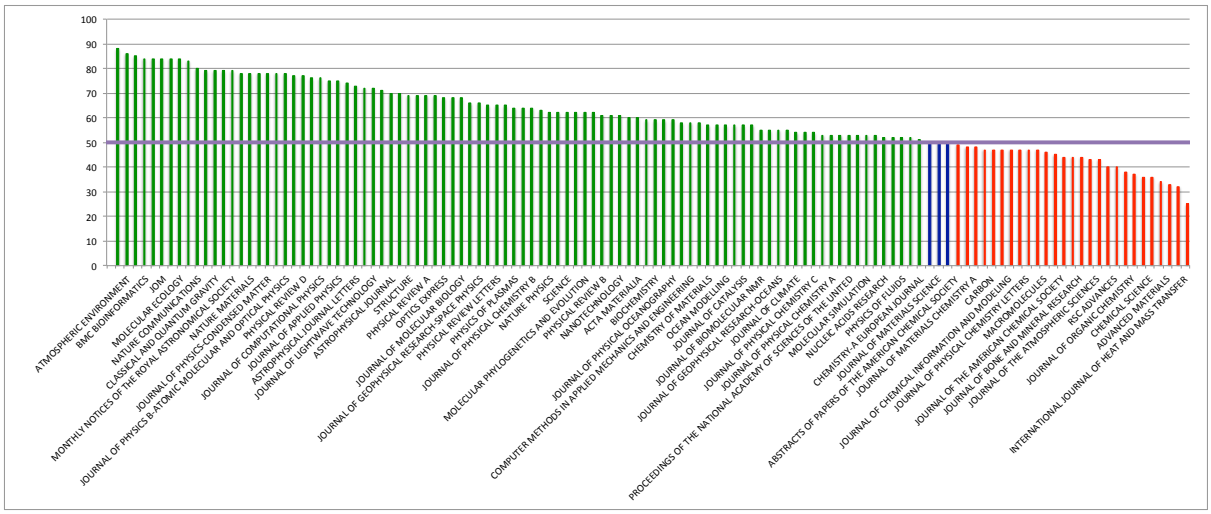

Figure 6: Median percentile ranking of XD publications by journal (by ISI)

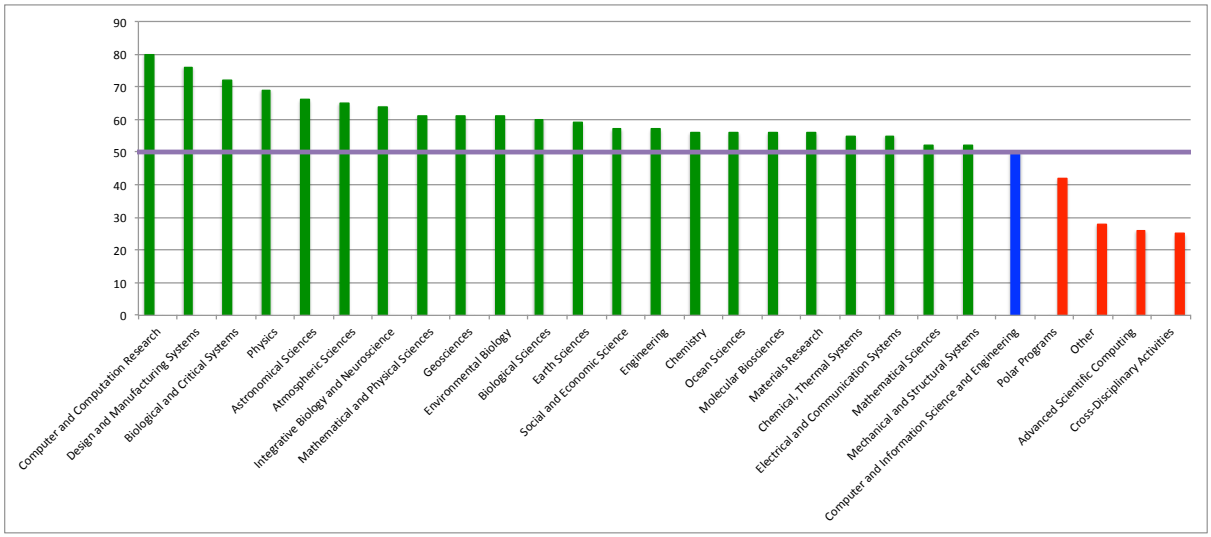

Figure 7: Average percentile ranking of XD publications by Field of Study (by ISI)

large enough, it is rational to use T-test to not only test if there is statistical difference of the two group, as having been shown by the several previous tests, but also quantifies the difference of the mean. The t-test results for both citation count and percentile ranking are as below.

- T=9.8328, df=5105.5, p-value< 2.2e-16, 95% confidence interval: [10.90, 16.32]

T-test for ranking (Welch Two sample t-test)

- T=25.412, df=5105.5, p-value<2.2e-16, 95% confidence interval: [9.07, 10.59]

The results show that the XSEDE group has a statistically higher citation ranking and a statistically higher mean citation rate than the non-XSEDE peer group.

4.2.1 *Journal peer comparison based on MAG data*

While we first incorporatd the MAG data in order to evaluate the field weighed impact of XSEDE publications, we could follow the same approach as done using ISI data to conduct the same peer comparison study. Figure **??** and Figure 12 shows the average and median percentile ranking for XSEDE publications by each journal using MAG data. The overall results are pretty similar to what we got from the study with ISI data.

Table 2: Basic statistics of XSEDE publications group and peers group

	Number of Publications	Rank		Citations	
		Average	Median	Average	Median
XD	5078	59	63	28	12
Peers	356464	49	49	15	5

5. CONCLUSION

We evaluated the scientific impact of XSEDE by examining its publications that resulted from using XSEDE resources. By curating the XSEDE publication data with cleaning up, verifying and correlating with various data sources,

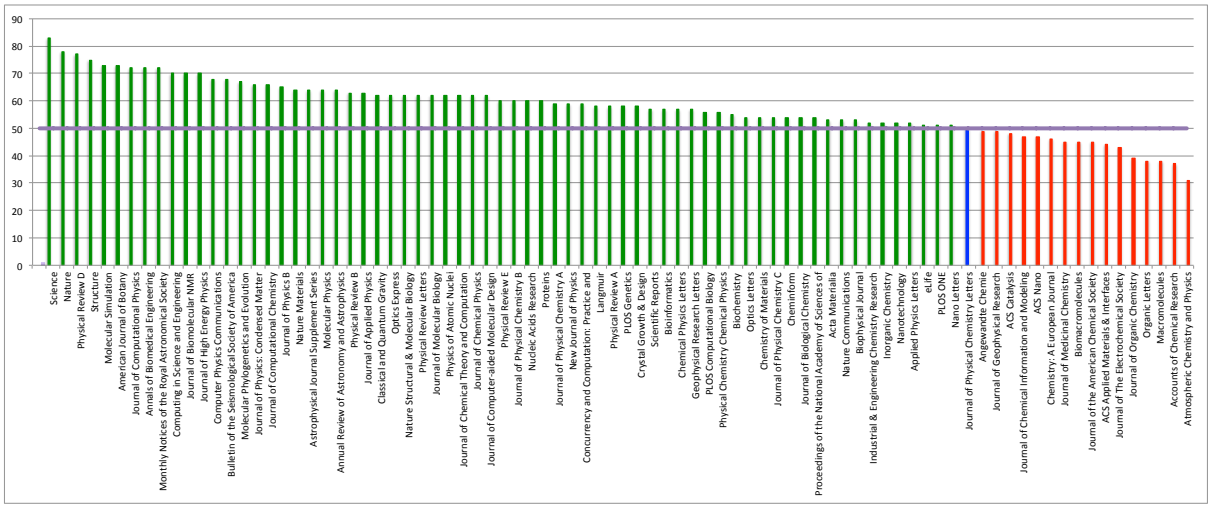

Figure 11: Average percentile ranking of XD publications by journal (by MS)

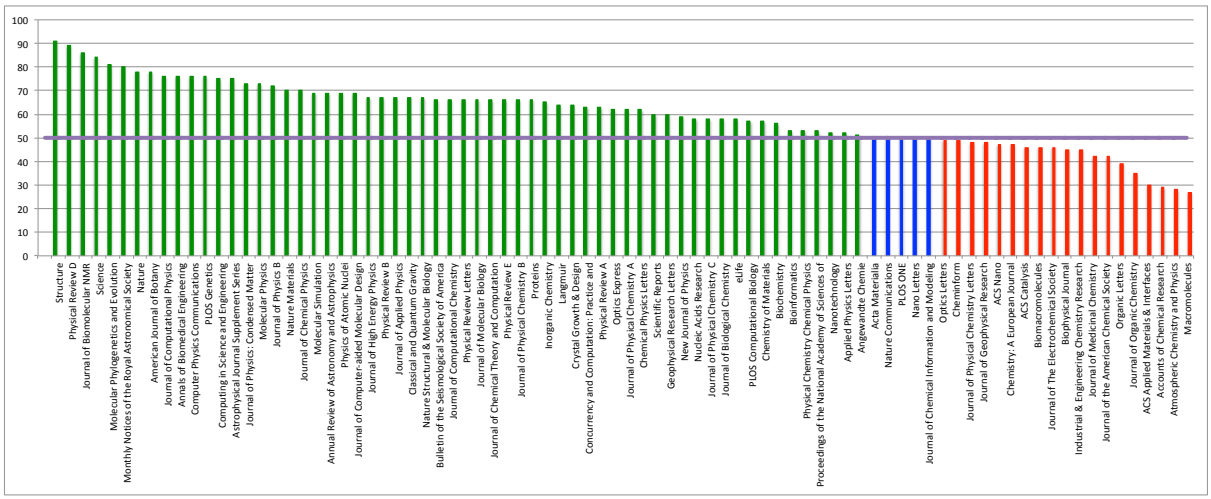

Figure 12: Median percentile ranking of XD publications by journal (by MS)

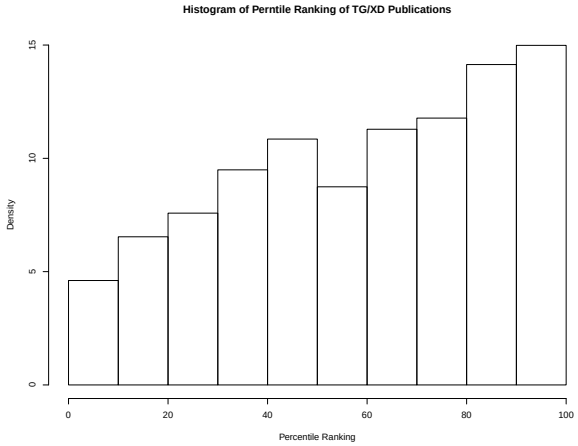

Figure 8: Histogram of Percentile Ranking

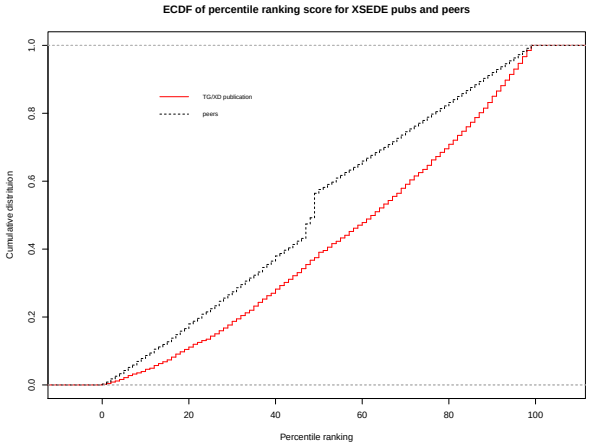

Figure 9: Empirical Cumulative Distribution of Percentile Ranks

we obtained substantially valuable data to compare and evaluate the impact of XSEDE publications. While using

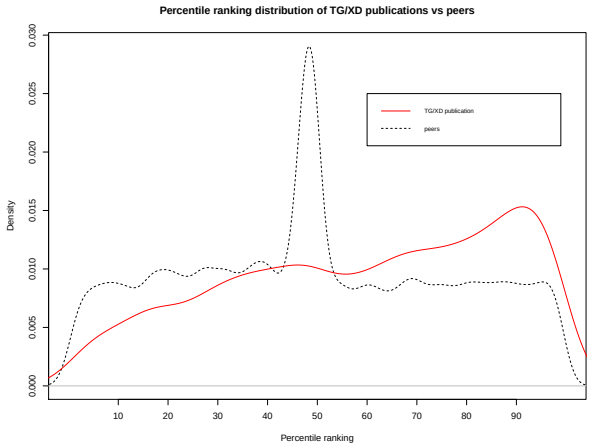

Figure 10: Kernel Density of the distributions of XSEDE publications' percentile ranking and that of peers'

two distinct analysis - Field Weighted Citation Impact analysis, and another novel journal publications-based peer comparison study, we found that XSEDE publications tend to be cited more than the non-XSEDE publications group. Various statistical tests show the results are statistically significant. The results from this study could give information to the XSEDE leadership team and the funding agency, to possibly impact the management and certain policy of the facility, e.g., to provide useful information to the resource allocation committee during selecting proposals to approve the usage requsts. While we showed the study extensively with XSEDE data, the approaches used could be ubiquitous and generalized to be used to evaluate similar facilities or groups. In fact we have done similar analyses for NCAR, BlueWaters, and Bridges using the developed methodology and software framework.

6. ACKNOWLEDGMENTS

This work is part of the XSEDE Metrics Service (XMS) project sponsored by NSF under grant number OCI-1025159. Lessons learned from FutureGrid have significantly influenced this work. Gathering publications was first pioneered by FutureGrid, influencing the development in the XSEDE portal. We would like to thank Matt Hanlon and Maytal Dahan for their efforts to integrate this framework into the XSEDE portal.

7. REFERENCES

[1] Compute Canada Report: Increasing Canadian Research Impact. Web Page. URL: https://www.computecanada.ca/compute-canada-report-increasing-canadian-research-impact/.

[2] i-10 index | google scholar citations open to all. URL: http://googlescholar.blogspot.com/2011/11/google-scholar-citations-open-to-all.html.

[3] ISI Web of Science. Web Page. URL: http://wokinfo.com/.

[4] XDMoD Scientific Impact Portal for XSEDE. Web Page. URL: https://sciimp.ccr.xdmod.org/xdportalpub/overview/.

[5] XSEDE. Web Page. URL: https://www.xsede.org/.

[6] J. Bollen, G. Fox, and P. R. Singhal. How and where the TeraGrid supercomputing infrastructure benefits science. *Journal of Informetrics*, 5(1):114–121, 2011.

[7] J. Bollen, M. A. Rodriguez, and H. Van de Sompel. MESUR: Usage-based Metrics of Scholarly Impact. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '07, pages 474–474, New York, NY, USA, 2007. ACM. URL: http://doi.acm.org/10.1145/1255175.1255273, doi:10.1145/1255175.1255273.

[8] J. Bollen, H. Van de Sompel, and M. A. Rodriguez. Towards Usage-based Impact Metrics: First Results from the Mesur Project. In *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '08, pages 231–240, New York, NY, USA, 2008. ACM. URL: http://doi.acm.org/10.1145/1378889.1378928, doi:10.1145/1378889.1378928.

[9] G. W. Brown, A. M. Mood, et al. On median tests for linear hypotheses. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. The Regents of the University of California, 1951.

[10] L. Colledge. Snowball metrics recipe book. 2014, 2014.

[11] L. Egghe. Theory and practise of the g-index. *Scientometrics*, 69(1):131–152, 2006.

[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*, 102(46):16569–16572, 2005.

[13] W. H. Kruskal and W. A. Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.

[14] H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.

[15] T. Penfield, M. J. Baker, R. Scoble, and M. C. Wykes. Assessment, evaluations, and definitions of research impact: A review. *Research Evaluation*, 23(1):21–32, 2014.

[16] P. Thomas and D. Watkins. Institutional research rankings via bibliometric analysis and direct peer review: A comparative case study with policy implications. *Scientometrics*, 41(3):335–355, 1998.

[17] J. Towns, T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw, V. Hazlewood, S. Lathrop, D. Lifka, G. Peterson, R. Roskies, J. Scott, and N. Wilkins-Diehr. XSEDE: Accelerating Scientific Discovery. *Computing in Science Engineering*, 16(5):62–74, Sept 2014. doi:10.1109/MCSE.2014.80.

[18] G. von Laszewski, F. Wang, G. C. Fox, D. L. Hart, T. R. Furlani, R. L. DeLeon, and S. M. Gallo. Peer comparison of xsede and ncar publication data. In *2015 IEEE International Conference on Cluster Computing*, pages 531–532, Sept 2015. doi:10.1109/CLUSTER.2015.98.

[19] F. Wang, G. von Laszewski, G. C. Fox, T. R. Furlani, R. L. DeLeon, and S. M. Gallo. 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, XSEDE '14, pages 25:1–25:8, New York, NY, USA, 2014. ACM. URL: http://doi.acm.org/10.1145/2616498.2616507, doi:10.1145/2616498.2616507.