Peer comparison of XSEDE and NCAR publication data

Gregor von Laszewski[1], Fugang Wang[1], Geoffrey C. Fox[1] David L. Hart[2]
Thomas R. Furlani[3], Robert L. DeLeon[3], Steven M. Gallo[3]

[1]*School of Informatics and Computing, Indiana University, Bloomington, Indiana, U.S.A.*
[2]*Computational and Information Systems Laboratory, NCAR, Boulder CO 80307-3000, U.S.A.*
[3]*Center for Computational Research, University at Buffalo, SUNY, Buffalo, New York, 14203, U.S.A.*
laszewski@gmail.com

Abstract—**We present a framework that compares the publication impact based on a comprehensive peer analysis of papers produced by scientists using XSEDE and NCAR resources. The analysis is introducing a percentile ranking based approach of citations of the XSEDE and NCAR papers compared to peer publications in the same journal that do not use these resources. This analysis is unique in that it evaluates the impact of the two facilities by comparing the reported publications from them to their peers from within the same journal issue. From this analysis, we can see that papers that utilize XSEDE and NCAR resources are cited statistically significantly more often. Hence we find that reported publications indicate that XSEDE and NCAR resources exert a strong positive impact on scientific research.**

Keywords-**Metrics; publications**

I. INTRODUCTION

We analyze the scientific impact on *scientific advancements enabled by enhanced cyberinfrastructure* for XSEDE and NCAR resources. We extend our previous work [1] by comparing publication impact based on a comprehensive peer analysis carried out on papers produced by scientists using XSEDE and National Center for Atmospheric Research (NCAR) resources [2], [3]. The analysis is based on a percentile ranking of citations of the papers derived from a comparison to peer publications in the same journal not using the resources. The result shows that papers that utilize XSEDE and NCAR resources tend to be cited more often than their peers are cited. We leverage for this work the data collection and processing architecture described in [1], [2]. To conduct the analysis, the general workflow includes obtaining the publication data for each XSEDE user, and then retrieving the citation data for each publication. Hence, the data is originally collected on a per-user and per-publication basis. As part of processing, the data are aggregated based on organization, XSEDE project/account, and FOS. By correlating the publication data with XSEDE usage and allocation data (for example the allocation amounts awarded by XSEDE), our intention is to determine if the analysis reveals patterns and trends in how XSEDE impacts the sciences and possibly to help better measure the return on investment (ROI) for NSF. Naturally, the same is applied to the NCAR data to demonstrate the generality of our methods.

II. METRIC FOR JOURNAL PUBLICATION-BASED PEER COMPARISON

We compared XSEDE and NCAR publications with their peers (more than one million publication) by looking at the citation count and relative ranking through the following process: (1) Identify publication venues of the reported XSEDE or NCAR publications; (2) Retrieve all publication data from those venues during the same time period; (3) Identify the peers comparison groups (All publications in same journal issues); (4) Calculate ranking scores of XSEDE/NCAR publications and their peers; and (5) Use the raw ranking scores to derive comparison metrics at different aggregation level (e.g. journal, Field of Study).

We defined a *performance score* based on weighted sum of the percentages of publications falling into each quarter by percentile ranking:

$$S = 1 * P_{Q_1} + 0.5 * P_{Q_2} + (-0.5) * P_{Q_3} + (-1) * P_{Q_4}$$

wher, P_{Q_i} is the percentage of publications falling into the top i quarter. P_{Q_i} gets the value in $[0, 1]$ and $\sum_i P_{Q_i} = 1$ for one aggregated subject (a journal or one FOS). As S has a range from $[-1, 1]$ and the possible values are symmetric, a positive value implies more publications appear in the upper half in ranking and negative means more in the lower half.

III. XSEDE PEER DATA ANALYSIS

We apply the procedure and then aggregate the raw percentile ranking scores at journal level, and calculate the average and median percentile ranking for each journal, as well as the resulting *performance score*. We include only those journals with at least ten[1] publications so the results have a higher statistical power. Table I compares the average and median rankings and citations received. The t-test shows that XSEDE publications are statistically higher ranking and have a higher mean citation rate than the non-XSEDE peer group.

1) T-test for ranking (Welch Two sample t-test)
   a) T=21.4134, df=2412.99, p-value<2.2e-16
   b) 95% confidence interval: $[10.80, 12.98]$

[1]For NCAR data, we used a value of five due to the smaller number of overall publications

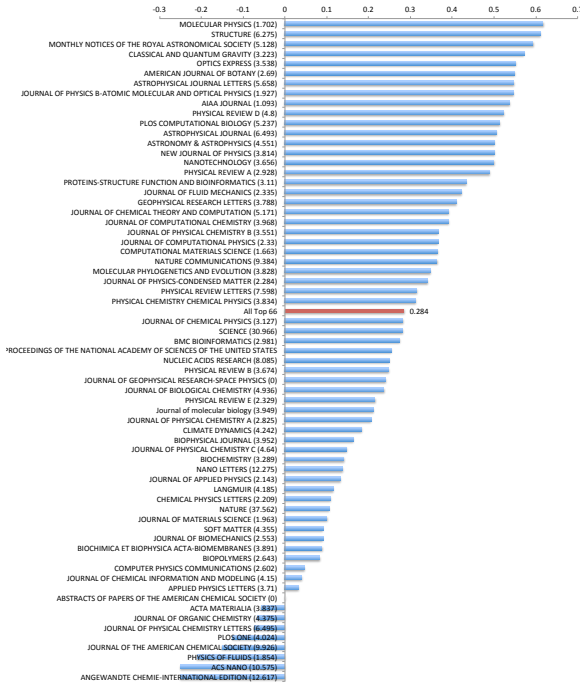|	Number of Publications	Rank		Citations	
		Average	Median	Average	Median
XD	2349	61	65	26	11
Peers	168422	49	48	13	5

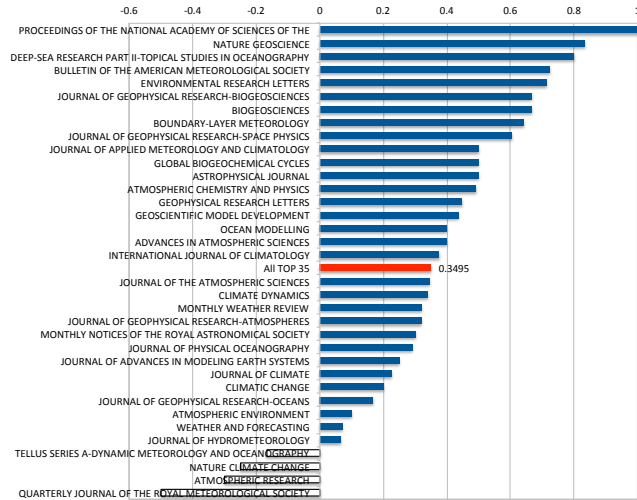Figure 1: The score of our peer comparison metric for XSEDE publications by journal.

Figure 2: Peer comparison based on Parent Field of Science from the original analysis of NCAR data.

2) T-test for citation count:

   a) T=7.057, df=2358.929, p-value=2.228e-12

   b) 95% confidence interval: [9.40, 16.63]

we included the performance scores of the top 66 journals in which 10 or more XSEDE publications published (see Figure 1) and find that the overal score metric for all the involved journals is positive with a value of 0.284. This implies that researches benefit from use of XSEDE resources as the resulting publications tend to be cited more than their peers from the same journal.

When aggregating the individual results on Field of Science (FOS) level we observe similar results that publications from most FOS performe better than their peers, while certain FOS such as Astronomy and physics benefit most from using XSEDE [2], [3]. Fields that perform not as good include Experimental and Theoretical Geochemistry, Geometric Analysis and Mechanical and Structural systems as they are typically not dependent on computational resources.

IV. NCAR PEER DATA ANALYSIS

We replicated the analysis with NCAR publication data of 900 records. Because the NCAR data set is smaller than the one from XSEDE, instead of looking at all journals that have at least ten publications, we reduced the value to five. This will give a large enough journal number to conduct our analysis in this case. The *performance score* as defined earlier is presented as in Figure 2 as a comparison with XSEDE. Again we see a positive value of about 0.35 indicating the better performance than the peers'.

V. CONCLUSION

The NCAR score is slightly higher than that of XSEDE as XSEDE has a wider range of FOS. Computational intense disciplines such as atmospheric sciences result in the highest score values using the resources. For both XSEDE and NCAR publications, the impact metric measured by a performance score (defined based on percentile ranking) is positive and higher than their peers that have not used such resources.

REFERENCES

[1] 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*, ser. XSEDE '14. New York, NY, USA: ACM, 2014, pp. 25:1–25:8. [Online]. Available: http://doi.acm.org/10.1145/2616498.2616507

[2] 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," Indiana University, Tech. Rep., 2015. [Online]. Available: https://github.com/cyberaide/paper-tas-2/blob/master/vonLaszewski-tas-cluster.pdf

[3] ——, "Peer comparison of XSEDE publication data," in *XSEDE2015*. Indiana University, poster, Technical Report. [Online]. Available: https://github.com/cyberaide/paper-tas-2/blob/master/vonLaszewski-tas-xsede.pdf