Peer Comparison of XSEDE Publication Data

Gregor von Laszewski*
Fugang Wang
Geoffrey C. Fox
Indiana University
2719 10th Street
Bloomington, Indiana, U.S.A.

David L. Hart
Computational and Information Systems Laboratory National Center for Atmospheric Research
P.O. Box 3000
Boulder CO 80307-3000

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

We present a framework that compares the publication impact based on a comprehensive peer analysis of papers produced by scientists using XSEDE resources. The analysis is introducing a percentile ranking of citations of the XSEDE papers compared to peer publications in the same journal that do not use these resources. This analysis is unique in that it is a comprehensive study in which all reported published papers are compered to peer publications selected from within the same issue of the same journal. From this analysis, we can see that papers that utilize XSEDE resources are cited statistically significantly more often. Hence we find that reported publications indicate that XSEDE resources exert a strong positive impact on scientific research.

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, bibliometric, h-index, Technology Audit Service, XDMoD, XSEDE

1. INTRODUCTION

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) [3, 12]. Many recent science and engineering innovations and discoveries are increasingly dependent on access to high-end computing resources [**?**]. The demand for high-end resources is met by large-scale compute resources located in geographically dispersed locations that cannot typically be supported by any single research group. Accordingly, dedicated large-scale computing facilities play an important role in scientific research, in which resources are shared among groups of researchers, while the facilities themselves are managed by dedicated staff. The National Science Foundation (NSF) has supported such facilities for many years. XSEDE allocates resources to approved projects, which represent a substantial financial investment by NSF. Thus, justification for their use is warranted and questions regarding the scientific impact of these resources naturally arise. In previous work [**?**] we focused on the creation of a framework that collects bibliometric data and analyzes them with respect to a number of metrics. However, this work did not yet include a mechanism to compare publications with peers not using such resources.

In this paper, we significantly enhance our previous work by comparing publication impact based on a comprehensive peer analysis carried out on papers produced by scientists using XSEDE resources. 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. This analysis is unique in that it is a comprehensive study in which thousands of published papers are compared to peer publications selected from within the same issue of the same journal. From this analysis we can see that papers that utilize XSEDE resources are cited statistically significantly more often.

The paper is structured as follows. First, we review some portions of our previous work and relate it to the work reported in this paper (Section 2). We review our design and the architecture of our framework supporting this effort (Section 3). Next we introduce a journal-based peer metric that allows us to compare any resource provider's related publications with publications not using the resources (Section 4). To demonstrate general applicability of this method, we introduce in the next section a peer analysis of XSEDE (Section 5) publications. We present important statistics about this metric for XSEDE. Finally, we present our conclusions (Section 6).

2. RELATED WORK

Although a number of related studies have been conducted

*Corresponding Author.

submitted to ... 2015,USA

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

[11, 5, 7, 6, 4] our work is unique in that it provides a *comprehensive* analysis superior in data volume to other studies and is focused on the analysis of XSEDE data. More information about related work can be found in [**?**, **?**]. Furthermore, we are happy to engage in collaborative efforts to enhance this work or to integrate other related work by contacting us while targeting other resource providers.

3. BIBLIOMETRIC DATA ANALYSIS FRAMEWORK FOR RESOURCE PROVIDERS

The work described in this paper has been motivated by analyzing data related to XSEDE. The XSEDE data set includes publications from both XSEDE and its predecessor program, TeraGrid. For simplicity, we refer to them as *XSEDE* throughout this paper. As many other resource providers may have similar needs to analyze their data, this framework can be applied generally for other providers. In case this is desired, a custom integration can be performed and the analysis can be adapted by the team from Indiana University. It will mostly include new datasets for the bibliographic data. We are showing the generality of this approach in other work conducted for NCAR data [**?**].

3.1 Design

The design of our general framework must meet three main objectives. First, we need to be able to compare the impact of research conducted on these resources. In order to achieve this we need to provide Metrics, and data services that utilize these metrics and return an impact measurement (see Figure 1 on the left-hand side).

To fulfill these design objectives we will use a layers architecture in order to allow us to expand and adapt our framework to different resource providers. The design needs to be able to compare users and projects with peers the *use* the resources and with peers that *do not* use the resources. Thus we are addressing efficiency within the community and within XSEDE. The information derived from it may provide valuable input for the various user communities including users, allocation committees, the resource provider leadership, and the funding agencies, as well as external peers that can see the impact of the resource could have on their research. To show generality of this approach we have taken an analysis and applied it not only to XSEDE as a resource provider, but also to National Center for Atmospheric Research (NCAR) as a resource provider which we describe in other work [2].

3.2 XSEDE Design Considerations

For XSEDE we have identified specific design criteria that we are integrating into our architecture while leveraging and integrating with existing services. Specifically for XSEDE interested parties in this analysis are not only the users, but also the Resource Allocation Committee (RAC) and the XSEDE leadership. The bibliometric data services and the data mashup are largely hidden from these groups. The groups benefit form a number of preconfigured analysis that may be further customized or enhanced. The important issue is that the large amount of work to obtain the metric and to manage the bibliometric data is hidden from the users and exposed to them as Service.

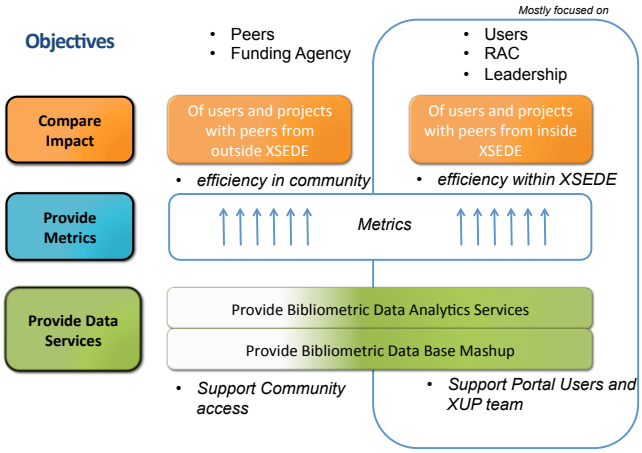

Figure 1: High level Objectives impacting the design of our framework

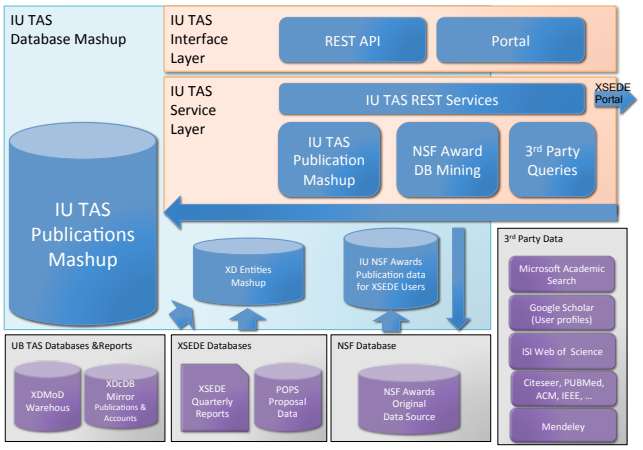

Figure 2: Architecture with the Interface, Service, Database Mashup, and Database Resources Layers

3.3 Architecture

To implement this service we designed a a layered architecture as depicted in 2. The main layers include (a) *Interface Layer* – easy to use interfaces for various communities, including an API, REST, and a Web GUI interface; (b) *Service Layer* – advanced services to bridge to a database mashup via sophisticated services and queries to the underlying database layer; (c) *Database Mashup Service Layer* – a sophisticated database mashup that contains the integration of data from a variety of data resources; and (d) the *Database Resources* that provide the underlying information for our service.

3.3.1 Service-Oriented Archotecture

The framework is based on a distributed set of software services. The service-oriented system consists of components for

- publication and citation data retrieval. This includes data from the NSF award database, Google Scholar, and ISI Web of Science;
- parsing and processing while correlating data from various databases and services, such as the XSEDE cen-

tral database (XDCDB), which stores all usage data for jobs run on XSEDE resources; and

- the XSEDE allocations database, which stores publication and grant funding information for PIs applying for XSEDE allocations.

The system also includes components for

- metrics generation and an analysis system for different aggregation levels – users, projects, organization, field of science (FOS) – as well as
- a presentation layer using a lightweight portal in addition to exposing some data via a RESTful API [**?**].

Due to 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 [**?**]. Obviously, Adaptation could men that we simply have to change the bibliometric data, which could mean that we need to integrate new data sources and curration services.

3.3.2 Service Integration into XSEDE

Our current framework for XSEDE includes services that are motivated by our initial findings from XSEDE bibliometric data. This includes specialized services focusing on user specific publication data as well as user, project, and field of science (FOS) views. Our mashup component aggregates the publication data mined from the previous component, in addition to those from XDCDB, and from other available external services. It also retrieves citation data for each publication from external services. Another essential service offered is to generate periodic metrics for users, projects, and FOS. THis is augmented with specific information that we relieve form data related to allocations and project proposals. These data is stored in the mashup database. As we are exposing our information through services it can also be integrated into the XDMoD system [9].

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.

4. METRIC FOR JOURNAL PUBLICATION-BASED PEER COMPARISON

While we focused in our previous work more on the internal analysis of data within XSEDE in regards to FOS, h-index [10], g-index [1] and i-index [8], in this work we extend the scope to a comparison with external peer publications and hence significantly expand our original analysis. To conduct this analysis, we introduce the definition of a metric that allows us to analyze and compare journal publications amongst peers, as well as identify a suitable peer group for the comparison.

Hence we considered for this work **all** publications that are uploaded through self-identification by users into the XSEDE portal, as well as XSEDE reports. We then identified from this set all journal publications and identified from that subset all journals that had at least ten self-identified publications from XSEDE. Although the number to meet this threshold is smaller, we found the restriction useful as it defines a peer group of scientists that publish repeatedly in these journals, thus making the comparison more meaningful. Once we identified such journals, we compared the citation count for each publication located in a journal issue that had at least one XSEDE publication. Our comparison is between publications in such journals that we identified as XSEDE papers and those that were not.

Next we introduce our metric that uses the percentile ranking of citations for the comparison of XSEDE publications with their non-XSEDE peers. The percentile ranking is based on the sum of all citations for a paper while at the same time ranking that number in four uniform percentile categories. Thus the papers in the first percentile have the most citations, the ones in the second have fewer, and so forth. We add the weighted sums of these counts. Thus the performance score is defined as:

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

where P_{Q_i} is the percentage of pubs falling into the top i-th quarter. The values for P_{Q_i} are in the interval $[0, 1]$ and $\sum_i P_{Q_i} = 1$ for one FOS. It is obvious to see that S has its value from $[-1, 1]$. A positive value implies more publications appear in the upper half in ranking and negative means more in the lower half.

5. XSEDE PEER DATA ANALYSIS

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, according to the self-reported categories defined in the XDCDB, and calculate the average and median percentile ranking for each field of science, as well as the resulting *performance score*. We include only those with at least ten[1] publications so the results have a higher statistical power.

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. Unfortunately, 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 FOSs.
   (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.

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

(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.

Through this data mashup we obtained enough data to conduct our analysis. We present our results in a number of graphs and tables. We first present some overall comparisons of the XSEDE and non-XSEDE citation comparisons. Figure 3 shows the kernel density of the distributions of XSEDE publications' percentile ranking and that of peers'. As expected, the non-XSEDE peer publications are symmetrically distributed by percentile ranking. The XSEDE publications are weighted to the higher percentile ranking. *This shows that XSEDE publications tend to be more highly cited.* Table 1 lists the average and median rankings and citations received of the two groups to evaluate and compare. We performed a T-test to test if the citation differences were statistically significant. 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.

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: $[10.80, 12.98]$
2. T-test for citation count:
   (a) T=7.057, df=2358.929, p-value=2.228e-12
   (b) 95% confidence interval: $[9.40, 16.63]$

Table 1: Basic statistics of XSEDE publications group and peers group

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

The results are depicted in Figures 6, 7, 8, 5, and 4.

NSF and XSEDE define a hierarchy of FOSs. In Figure 4, we show the top level FOS as defined by NSF. When we look to expand the FOS to the next level in the hierarchy, we find the results as depicted in Figure 8. The next level is shown in Figure 7. Each of these figures shows the list of FOSs in decreasing order by the performance score S.

From Figure 7 we can identify that for most fields of science the XSEDE publications performed better than their peers. The average and median scores were higher than 50 and the score is positive. When looking at individual results, we see that astronomy and physics benefit most from using XSEDE. When looking at the fields that perform worst, we find fields such as Experimental and Theoretical Geochemistry, Geometric Analysis and Mechanical and Structural systems. Such fields are typically not dominated by simulation science and are less dependent on computational resources. Other fields such as Training include many other areas of training outside of supercomputing usage. We even find fields such as Computer and Computation Research to be less impacted. We certainly have to acknowledge in this

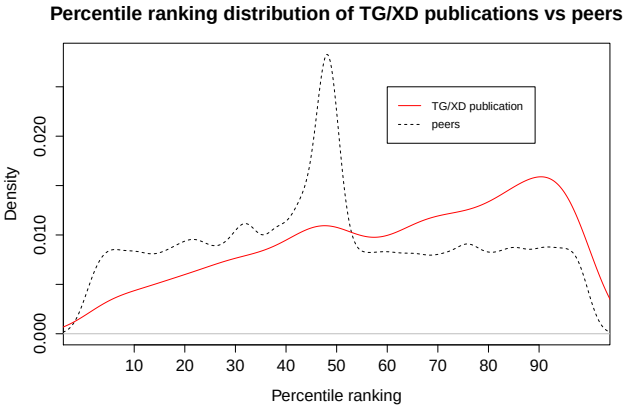

Figure 3: Kernel density of distributions for XSEDE publication percentile ranking versus peers.

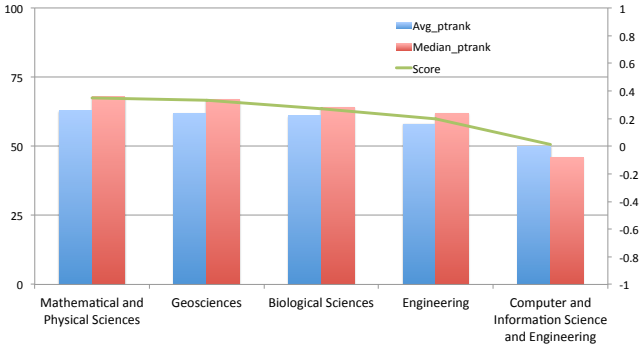

Figure 4: Peers comparison based on the topmost Field of Science category as defined by NSF.

case that many theoretical papers and papers not using supercomputers are published.

To show the percentile distribution in more detail for each journal, we present in Figure 6 a stacked barchart. Also here, as expected from our previous results, we see a positive impact in the percentile citation count for most of the journals. This is made obvious by looking at Figure 5. Here we also have included the average of the top 66 journals in which we found 10 or more XSEDE publications and find that the score metric is positive with a value of 0.284. Thus we can conclude that papers benefit from use of XSEDE resources, based on self-identification in reports and bibliographic upload to the XSEDE portal, resulting in their being more cited on average than their peers from the same journal.

6. CONCLUSION

We can see from the XSEDE results, research in the atmospheric sciences resulted in the highest score values using XSEDE resources.

However, the most important conclusion we make for our metric is that, for both XSEDE publications, the impact measured by a percentile score is positive and higher than their peers that have not used such resources.

We will further expand upon our metric and can define such values also for other groups defined in XSEDE or for other resource providers [**?**]. The important information we need is a reliable source that identifies publications that can be associated with the resource.

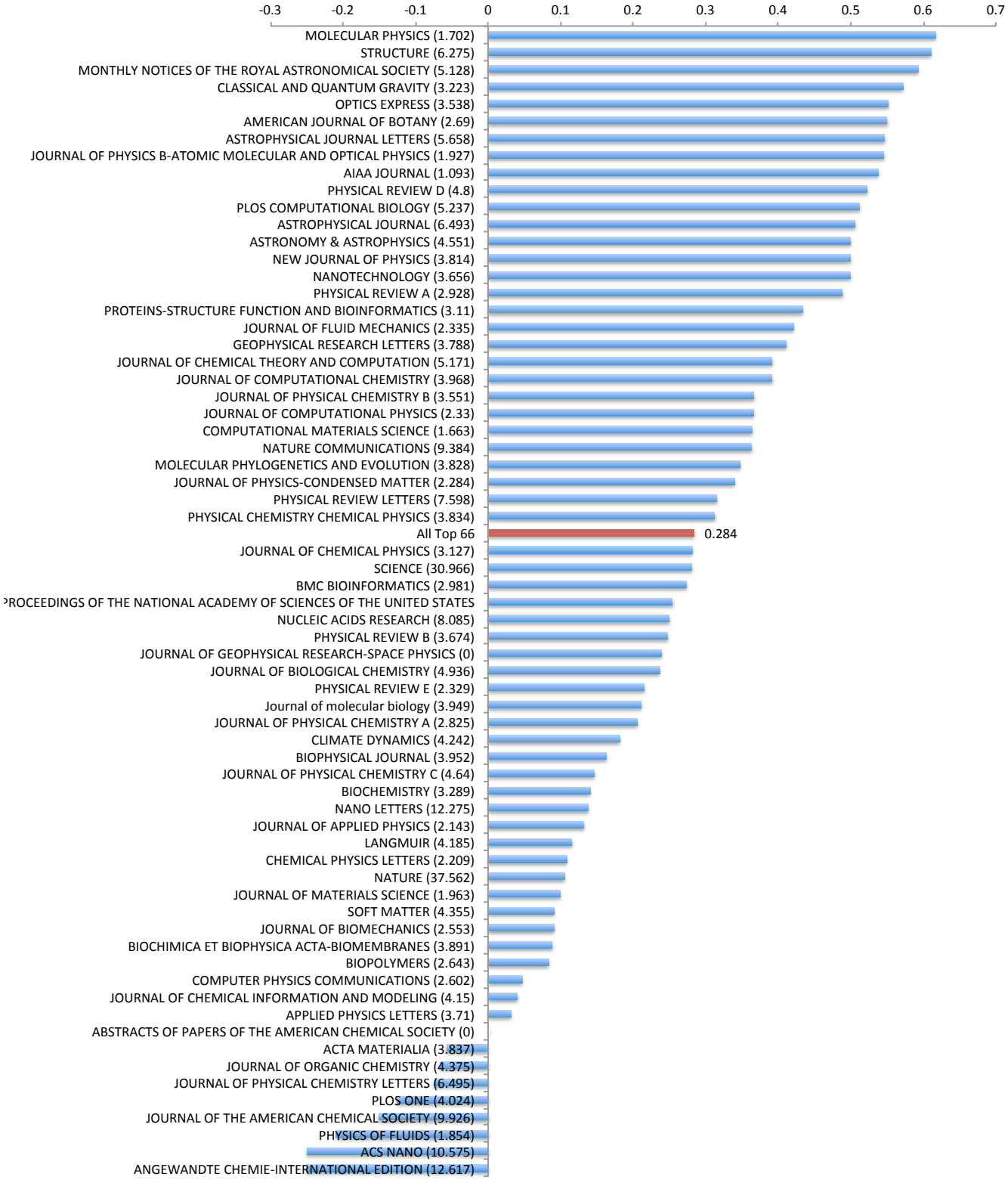

Figure 5: The score of our peer comparison metric for XSEDE publications by journal.

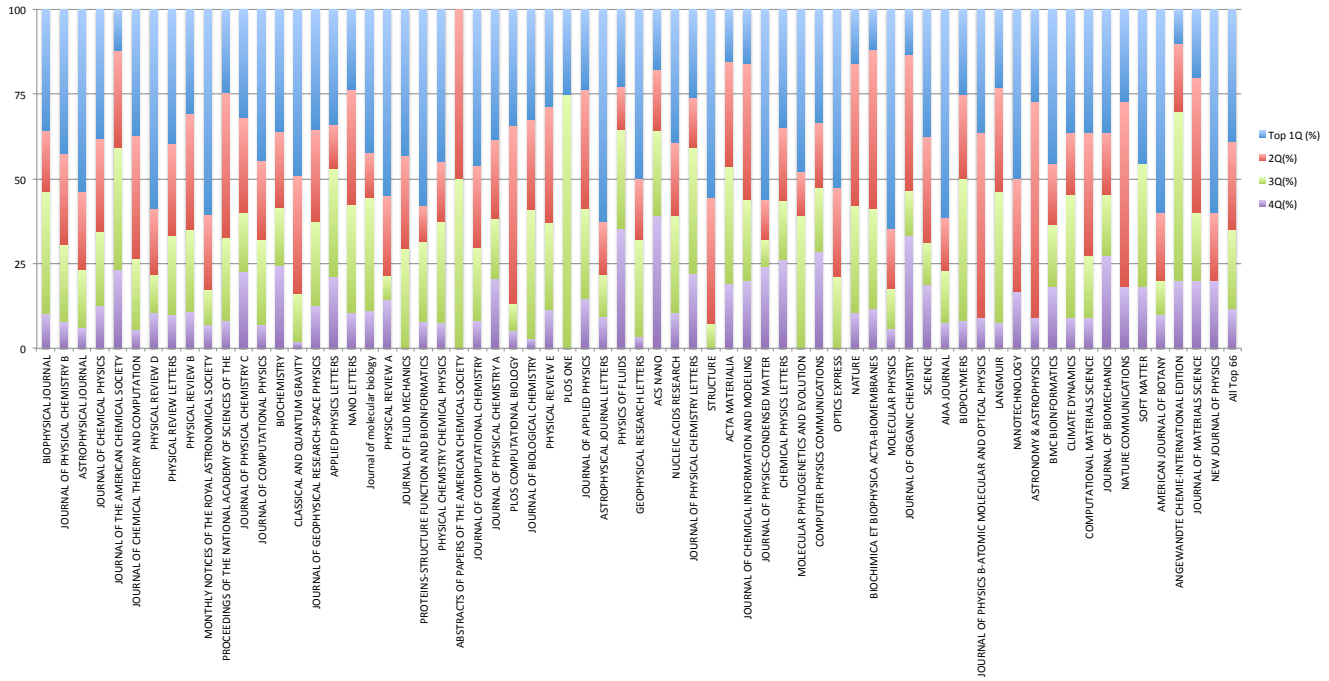

Figure 6: Percentile ranking by journal in a stacked barchart of XSEDE publications.

7. ACKNOWLEDGMENTS

This work is part of the Technology Auditing Service (TAS) 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.

8. REFERENCES

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

[2] NCAR. Web Page. URL: `http://ncar.ucar.edu`.

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

[4] 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.

[5] 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`.

[6] J. Bollen, H. Van de Sompel, A. Hagberg, and R. Chute. A principal component analysis of 39 scientific impact measures. *PloS one*, 4(6):e6022, 2009.

[7] 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`.

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

[9] T. R. Furlani, B. L. Schneider, M. D. Jones, J. Towns, D. L. Hart, S. M. Gallo, R. L. DeLeon, C.-D. Lu, A. Ghadersohi, R. J. Gentner, A. K. Patra, G. von Laszewski, F. Wang, J. T. Palmer, and N. Simakov. Using xdmod to facilitate xsede operations, planning and analysis. In *Proceedings of the Conference on Extreme Science and Engineering Discovery Environment: Gateway to Discovery*, XSEDE '13, pages 46:1–46:8, New York, NY, USA, 2013. ACM. URL: `http://doi.acm.org/10.1145/2484762.2484763`, `doi:10.1145/2484762.2484763`.

[10] 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.

[11] 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.

[12] 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`.

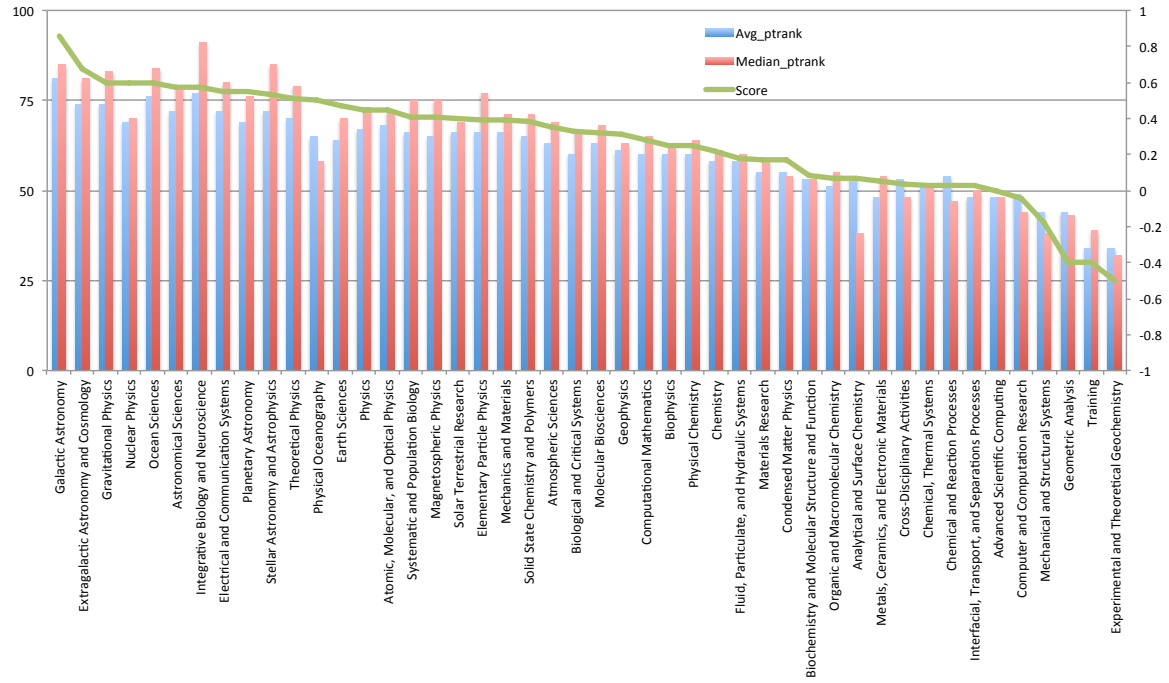

Figure 7: Peer comparison based on Field of Science of XSEDE publications.

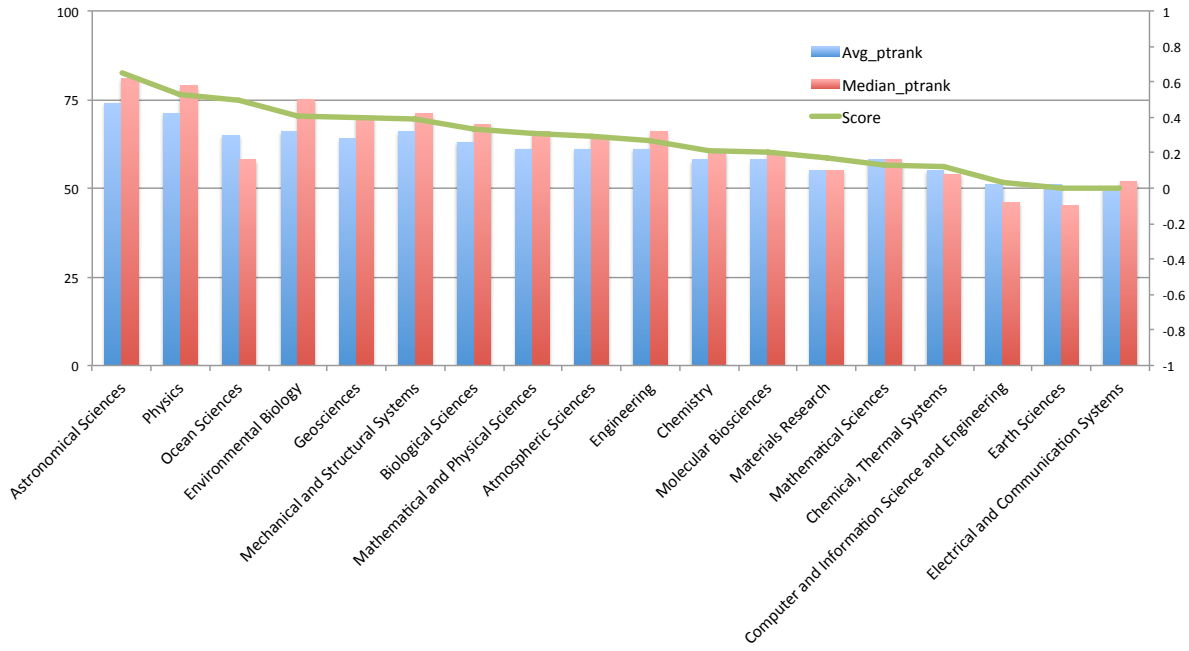

Figure 8: Peer comparison based on Parent Field of Science from the original analysis of XSEDE data.