Against the Grain

September 2010

MESUR: A Survey of Usage-based Scholarly Impact Metrics

Johan Bollen

Indiana University School of Informatics and Computing, jbollen@indiana.edu

Recommended Citation

SSRN's CiteReader technology, developed with **ITX Corp.**, scans a full-text PDF file and captures the references found in it. Those references are then verified through a combination of technology and human review. The verified references are parsed into smaller metadata fields and then matched against other articles in the **SSRN eLibrary**. It not only provides interesting data on who is citing whom and how often, but it also provides a research timeline allowing readers to easily go backward and forward in a subject matter. The References and Citations pages are freely available for the reader to follow the flow of the literature within and across multiple disciplines.

Interestingly, approximately 13% of **SSRN's** 3.9 million Citations are linked to working papers within the **SSRN eLibrary**.

Eigenfactor™

The **Eigenfactor™ Algorithm** provides a methodology for determining the most important or influential authors and papers in a network. The algorithm computes a modified form of the eigenvector centrality of each node in the network under the basis that important nodes are connected to other important nodes. This is the basic concept behind **Google's PageRank algorithm**.


most ranking algorithms → eigenFACTOR

journal x   journal x
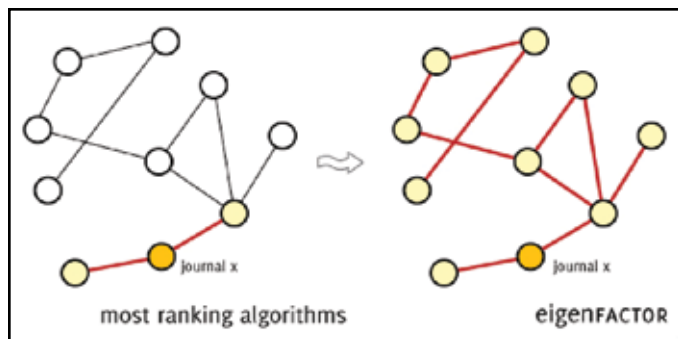
Gregory J. Gordon is President and CEO of **Social Science Research Network (SSRN)**, a leading multi-disciplinary online repository of working and accepted paper research in the social sciences and, recently, humanities. In addition, **SSRN** provides a variety of electronic distribution and conference management services. 🐦

Eigenfactor™ Scores have previously been used to rank scholarly journals, and the scores are freely available at *http://www.eigenfactor.org*. Within **SSRN**, we use article-level citation data to extend the **Eigenfactor™ Algorithm** to the author level and will apply it to the paper level in the near future. **CiteReader** calculates the number of times each paper in the **SSRN eLibrary** database has been cited by other papers in the **eLibrary**. This data is then used to construct an author citation network, where each author is a node.

At a more technical level, the **Eigenfactor™ Scores** can be seen as the outcome of two conceptually different, but mathematically equivalent, stochastic processes. The first process is a simple model of research in which a hypothetical reader follows chains of citations as she moves from node to node ad infinitum. An author's **Eigenfactor™ Score** is the percentage of the time that she spends with this author's work in her random walk through the literature.

The second process is an iterated voting procedure. Each author divides one vote equally among those authors she cites. In subsequent rounds, each author divides her current vote total, as received in the previous round, equally among those authors whom she cites. This process is iterated indefinitely until we reach a steady state where the number of votes doesn't change. An author's **Eigenfactor™ Score** is the percentage of the total votes.
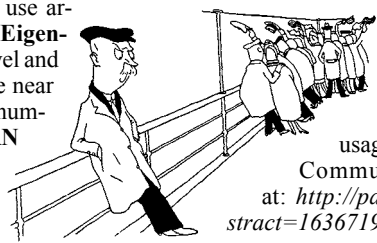
A more detailed discussion of **Eigenfactor™** usage within the **SSRN** Community is available at: *http://papers.ssrn.com/abstract=1636719*.

There are numerous methods for determining which articles you should read, and they have varying levels of success. Article-level metrics, especially in SSH, provide the best opportunity for finding the latest, most impactful research. For example, you can use downloads when you need currency, citations for more established areas, and **Eigenfactor™** for broader impact on a community. No one measure is perfect, and having a variety to choose from will allow you to use the best one in each situation. Approaching any measure with a reasonable degree of skepticism and minimal amount of cynicism is also a good thing.

When I think about the benefits of article-level metrics and the focus in many circles attributed to IF I remember a quote from **Max Planck**:

A new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die, and a new generation grows up that is familiar with it.

Or as a scholar reminded me the other day, new ideas progress forward funeral by funeral … 🐦

MESUR: A Survey of Usage-based Scholarly Impact Metrics

by **Johan Bollen** (Associate Professor, Indiana University School of Informatics and Computing) <jbollen@indiana.edu>

Introduction

Metrics of scientific impact are frequently defined as a function of the number of citations received by a particular scholarly publication.

The commonly used **Thomson-Reuter's** journal Impact Factor (IF) epitomizes this approach. The IF is calculated by dividing the number of citations received by the articles in a journal by the number of articles that appeared in same journal. The IF thus represents the average number of citations to articles published in a journal which is used as an indicator of the influence or impact of journals.

The IF is, however, not the only conceivable citation-based impact metric. Other citation-based metrics have been introduced in the past five years to indicate various facets of impact such as author-impact, cf. **h-index** (**Hirsch**, 2005), journal influence, cf. **PageRank** (**Bollen**, 2006) and **Eigenfactor** (**Bergstrom**, 2007), and various other citation-derived indicators, e.g., **Leydesdorff** (2007). Many of these indicators are now commonly used to assess the impact of individual scholars and their publications.

In spite of its general acceptance, scholarly assessment from citation-data is, however, subject to a number of limitations that originate from the inherent properties of citation data. First, it can take anywhere from six months to several years to publish an article and for it to become "citable." Citation data is therefore subject to extensive publication delays and may for that reason be a delayed indicator of current scholarly activity. Second, citation data by its very nature is focused mostly on authors of journal publications. As a result, citation data does not fully represent the activities of communities that either do not publish and/or publish in different formats and venues, e.g., social sciences and humanities.

Citations have their origin in the world of print but many if not most scholarly publications are now published and accessed online. As users access the scholarly literature via online services, their activities are generally tracked and recorded in server log data. These records, referred to as usage data, provide detailed information on how scholarly resources affect the scholarly community through their usage.

Usage data may confer several significant advantages over citation data as a foundation for scholarly assessment. First, usage data can be recorded immediately after online publication and during all stages of scholarly activity. It thus provides a rapid, yet comprehensive indication of scholarly activity. Second, usage data can be recorded for a wide variety of participants in the scholarly communication process, not merely those who publish journal articles, and can in principle be recorded for any online resource including books, data files, software, images, and sound files. Third, usage data is recorded at a very large scale that may exceed the magnitude of all existing citations by several orders of magnitude. Its sheer scale can compensate for higher noise levels and lead to a more reliable assessment of scholarly activity and impact.

For the above reasons, usage data has generated considerable interest in the past ten years, cf. the success of the **COUNTER** project. The potential of usage data is clearly significant, but to arrive at systems of usage-based scholarly assessment a number of challenges must be addressed:

1) The lack of recording standards: there exist few standards for the recording of article-level usage data. The latter contains details on the time of the event, the user, and the resource that the event pertained to, and therefore a great deal of variability in recording formats can occur and prevent its correct interpretation.

2) The lack of representative usage data: usage data is generally recorded by specific institutions for specific sets of resources and communities. The result is usage data that pertains to one particular community and set of scholarly resources, but from which few "global" conclusions, e.g., a general impact ranking of articles or journals, can be derived.

3) The lack of suitable metrics: a myriad of citation-based impact metrics has been proposed for articles, journals, and authors. A similar number could possibly be defined for usage data. However, it is not clear which of these metrics provide the most valid and reliable indicators of specific facets of scholarly impact.

The MESUR Project

The **MESUR** project seeks to address the above mentioned challenges by a research program that is focused on exploring the vi-

Johan Bollen is an Associate Professor at the **Indiana University School of Informatics and Computing** and is also the Principal Investigator of the **MESUR project**. **Dr. Bollen's** current research interests include usage data mining, complex networks, computational sociometrics, informetrics, and digital libraries. 🐢

ability of usage- and network-based metrics of impact from large-scale, aggregated, and representative usage data.

The **MESUR** project started in 2006 at the Digital Library Research and Prototyping Team at the **Los Alamos National Laboratory's Research Library** with a grant from the **Andrew W. Mellon Foundation**. Under the direction of the Principal Investigator **Johan Bollen** and co-Principal Investigator **Herbert Van de Sompel**, it started an ambitious research program that proceeded along the following three lines:

1) Creation of a large-scale usage data set that pertains to a wide variety of user communities and scholarly resources by aggregating otherwise separately recorded usage data sets from the world's most significant publishers, aggregators, and institutional consortia (link resolvers).

2) A research program to determine the overall structural and network properties of usage data, in particular with the objective of establishing a foundation for impact metrics that do not merely rely on usage- or citation-counts but also take into account the contextual, structural features of scholarly activity.

3) Conducting a large-scale survey of usage- and citation-derived metrics to explore their properties as indicators of the various different facets of scholarly impact.

The **MESUR** project executed this research program with support from the **Andrew W. Mellon Foundation** from 2006 through 2008. In 2009 the PI moved to the School of Informatics and Computing at **Indiana University**, where the project continued, supported by a grant from the **National Science Foundation**. In 2010 the **Andrew W. Mellon Foundation** granted an award for maintaining the activities of the **MESUR** project and to support a planning process aimed at investigation models to evolve the project to an open, community-supported, sustainable framework.

MESUR's Usage Data Collection

MESUR began collecting its usage data in 2006 by negotiating data sharing agreements with a large variety of institutions that provide access to scholarly resources. To assure coverage of various types of usage the project took care to include as many different types of data providers as possible. **MESUR's** usage data providers therefore include some of the world's most important publishers, aggregators, and institutional consortia. In the period 2006 through 2008, **MESUR** achieved data sharing agreements with the following providers: **BioMed Central**, **Blackwell Publishing**, the **California Digital Library**, **California State University**, **EBSCO**, **Elsevier** (**Scopus** and

ScienceDirect), **Emerald**, **Ingenta**, **JSTOR**, **Mimas-zetoc**, **Thomson-Reuters** (**Web of Science**), and the **University of Texas**.

The usage data that was provided to the **MESUR** project was recorded in a variety of data formats, but was required to at least contain the following data fields and be recorded at the article-level:

1) Unique *event identifier*

2) A unique *session identifier* that indicates whether user requests occur within the same browser session.

3) A *date/time stamp* of the user request to the second.

4) A unique *document identifier* and/or sufficient *metadata* to uniquely identify documents.

5) A *request type* identifying the type of request issued by the user, e.g., "view abstract," "download PDF," etc.

All usage data was shared under agreements that contained strong guarantees of user, institutional, and provider privacy.

At the end of 2008 the **MESUR** project had collected more than one billion usage events (individual user requests) pertaining to nearly 50 million documents and about 100,000 serials (of which most are not scholarly journals). This usage data was recorded in the period 2002 to 2007.

All data arrived in the native format in which it was recorded by the usage data provider. After transferring usage data to the **MESUR** servers, it was extensively processed and normalized. Document identifiers were deduplicated to merge usage from different providers that pertained to the same document.

Mesur Results

The **MESUR** usage data contains extensive information on the nature and context of individual usage events. In particular, the session identifier identifies sequences of user requests that occurred within the same user session. This allows the reconstruction of user clickstreams — i.e., the sequence of how a user moves from one article (and journal) to the next in a session.

To investigate the structural properties of usage data one can combine these clickstreams to calculate the overall probability that users who visit one article will move on to another particular article. When calculated for all pairs of articles (and journals) a map of science results that shows the prevailing paths that users follow in their online activities as they move from one article (and journal) to the next. A sample of this map is shown in Fig. 1. Each circle represents a journal which is colored according to its domain classification given by the *Getty Institute's Arts and Architecture Thesaurus*. Pairs of journals are connected by

a thin line if there exists a high probability that users will move from one journal to the next in their online clickstreams. (See Fig. 1 below.)

More details on the **MESUR** map of science can be found in **Bollen** (2009a).

After calculating a usage-based network for nearly 100,000 serials (about 30,000 of which are actual scholarly journals) of which the network shown in Fig. 1 is but a small sample, the **MESUR** project retrieved and defined a myriad of metrics that exploit the structure of this network to assess various facets of the value of particular journals. For example, some journals in the mentioned map of science are not highly used but form crucial connectors between otherwise separated domains.

A variety of network metrics can thus be calculated from the network of usage-based journal connections such that each embodies a different facet of a journal's impact. The **MESUR** project has calculated 39 of these metrics, some from the citation data in the Journal Citation Reports and some from **MESUR's** usage data. A comparison of the journal rankings produced by these metrics revealed a number of interesting properties of both existing and proposed metrics and the notion of scholarly impact itself. In fact, by calculating correlation coefficients between each pair of metrics we could visualize the similarities between metrics in a map of metrics that is shown in Fig. 2 above.
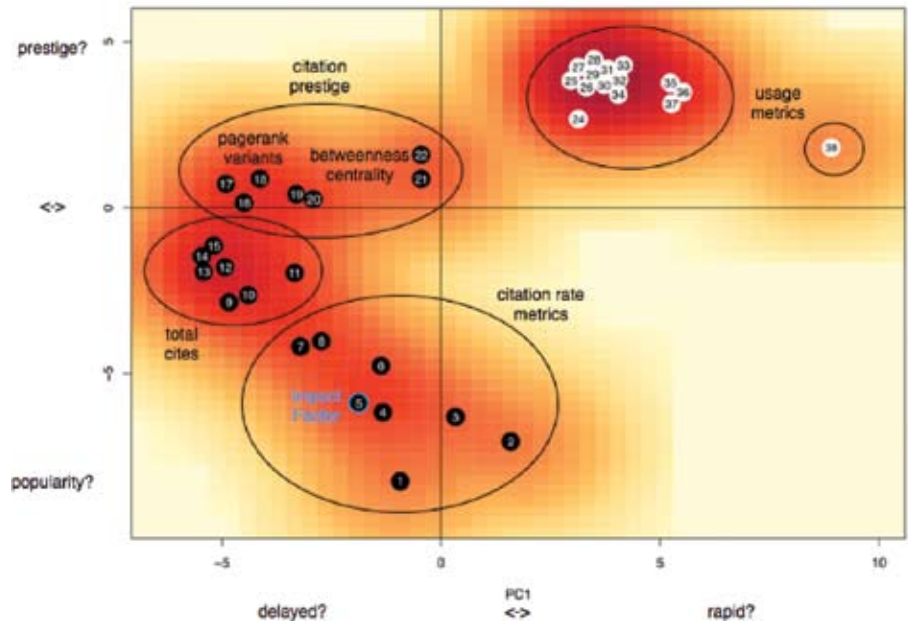

Figure 2. Annotated Map of Metrics as Produced by MESUR: Impact Metrics that Produce Similar Rankings are Positioned in Each Others' Vicinity. See Bollen (2009) for technical details and metric definitions.

More information on these results can be found in **Bollen** (2009b).

More information about the **MESUR** project including access to its maps of science and metrics can be found at the **MESUR** Website: *http://www.mesur.org/.*

Technical details on the **MESUR** project's mode of operation can be found in **Bollen** (2007a, 2007b, 2008).

Conclusion

The **MESUR** project is currently in its 4th year; over the past four years it has made significant contributions to the community's thinking on scholarly assessment. In addition, **MESUR** has pioneered the large-scale aggregation and normalization of usage data, defined minimal formatting and field requirements for article-level usage data, defined novel impact metrics, and created large-scale maps of science that can visualize current trends in science.

However, in spite of **MESUR's** progress and its compelling results more research and development are required to create a reliable and community-accepted system of usage-based scholarly assessments. The logistical requirements of large-scale usage aggregation in particular represent a significant burden. The lack of standards with regard to the recording, sharing, and normalization of usage data, as well the costs of negotiating tailored data agreements with a large number of usage data providers, needs to be addressed to secure the sustainability of the project in the future. This has become particularly pertinent as the **MESUR** project has accumulated a unique collection of data and results that represent a unique value to the scholarly community. The **Andrew W. Mellon Foundation** has for these reasons granted an award to the **MESUR** project in 2010 to conduct a planning process to investigate how the **MESUR** project could evolve to a more open, sustainable, and community-supported initiative.

References

J. E. Hirsch (2005). An index to quantify an individual's scientific research output.

Figure 1. Sample of MESUR's Map of Science Derived from Large-scale Usage Data.

PIRUS2 (Publisher and Institutional Repository Usage Statistics): Creating a Common Standard for Measuring Online Usage of Individual Articles

by **Paul Needham** (Cranfield University, UK)

and **Peter T. Shepherd** (Director, Project COUNTER, Edinburgh, UK) <pt_shepherd@hotmail.com> *www.projectCounter.org*

Introduction

One of the more significant developments since scholarly articles have been published online has been the growing role of institutional and subject repositories as hosts for these articles. The publishers of journals, though still the most important hosts, no longer have a monopoly of the distribution of these articles that they enjoyed in the print world. This trend has been given considerable further momentum by the Open Access movement, which encourages the free availability of the outputs of scholarly research, especially where that research has been publicly funded.

A reader searching online for a particular article may now find it in a number of different locations:

- the main journal publisher Website (e.g., **Elsevier's ScienceDirect**)
- a content aggregator site (e.g., **ProQuest**)
- a subject repository (e.g., **PubMed Central**)
- the author's local institutional repository (e.g., **Oxford University Research Archive - ORA**)

It is not the purpose of this article to argue the pros and cons of this highly distributed system for the publication of scholarly articles, still less to present the case for or against Open Access publishing. Rather, we accept that these trends are now well established and that any system for recording and reporting online usage of articles must take them into account. This makes the task of counting usage at a global level rather challenging. For a start, it will no longer suffice to record and report usage at the journal level: the journal as a package of articles is used by publishers, but not by repositories, which are organised on the basis of individual items. Then we have to consider the status of different versions of articles and which versions should be counted. Clearly, the accepted version of an article, or the published version of record has higher status than the author's initial draft, but does this mean that usage of the latter should not be counted at all, or does it mean that such usage should be weighted differently? These and other issues become highly pertinent in this increasingly heterogeneous publishing environment, and the aim of the **PIRUS** (**P**ublisher and **I**nstitutional **R**epository **U**sage **S**tatistics) project is to address them.

COUNTER as a Basis for Individual Article Usage Statistics

Currently the only widely implemented global standard for measuring online usage of scholarly information has been set by **COUNTER**, but until now the most granular level at which **COUNTER** requires reporting of usage is the individual journal. Demand for usage statistics at the individual article level has hitherto been low. This, combined with the unwieldiness of usage reports in an **Excel** environment, has meant that **COUNTER** has, until now, given a low priority to usage reports at the individual article level. Recent developments have, however, meant that it would now be appropriate to give a higher priority to developing a **COUNTER** standard for the recording, reporting, and consolidation of usage statistics at the individual article level. Most important among these developments are:

- Growth in the number of journal articles hosted by institutional and other repositories, for which no widely accepted standards for usage statistics have been developed.
- A Usage Statistics Review, sponsored by **JISC** under its Digital Repositories programme 2007-8, which, following a workshop in Berlin in July 2008, proposed an approach to providing item-level usage statistics for electronic documents held in digital repositories.
- Emergence of online usage as an alternative, accepted measure of article and journal value and usage-based metrics being considered as a tool to be used in the evaluation of research outputs.
- Authors and funding agencies are increasingly interested in a reliable, global overview of usage of individual articles.
- Implementation by **COUNTER** of XML-based usage reports makes more granular reporting of usage a practical proposition.
- Implementation by **COUNTER** of the **SUSHI** protocol facilitates the automated consolidation of large volumes of usage data from different sources.

Aims and Objectives of PIRUS2

The aim of **PIRUS2** is to specify **COUNTER**-consistent standards and protocols (as well as an infrastructure and an economic model) for the recording, reporting, and consolidation of online usage of individual articles hosted by repositories, publishers, and other entities.

In order to achieve this overall aim, the project will seek to meet the following main objectives:

- Develop a suite of free, open-source programmes to support the generation and sharing of **COUNTER**-compliant usage data and statistics that can be extended to cover any and all individual items in repositories.

MESUR: A Survey ...
from page 24

Proceedings of the National Academy of Sciences, 102(46), 16569-16572. DOI: 10.1073/pnas.0507655102.

Johan Bollen, **Marko A. Rodriguez**, and **Herbert Van de Sompel** (2006). Journal status. *Scientometrics*, 69(3), DOI: 10.1007/s11192-006-0176-z.

C. T. Bergstrom (2007). Eigenfactor: Measuring the value and prestige of scholarly journals. *C&RL News*, 68(5).

Loet Leydesdorff (2007). Betweenness centrality as an indicator of the interdisciplinarity of scientific journals. *Journal of the American Society for Information Science and Technology*, 58(9), 1303—1319.

Bollen J., **Van de Sompel H.**, **Hagberg A.**, **Bettencourt L.**, **Chute R.**, et al. 2009 Clickstream Data Yields High-Resolution Maps of Science. *PLoS ONE* 4(3): e4803. DOI:10.1371/journal.pone.0004803.

Bollen J., **Van de Sompel H.**, **Hagberg A.**, **Bettencourt L.**, **Chute R.**, 2009 A Principal Component Analysis of 39 Scientific Impact Measures. *PLoS ONE* 4(6): e6022. DOI:10.1371/journal.pone.0006022.

Johan Bollen, **Marko Rodriguez**, **Herbert Van de Sompel**, **Lyudmilla Balakireva**, and **Aric Hagberg**, The Largest Scholarly Semantic Network...Ever (poster). In *Proceedings of the 16th International World Wide Web* conference, May 2007. Note: Best Poster Award!

Marko. A. Rodrguez, **Johan Bollen** and **Herbert Van de Sompel**. A Practical Ontology for the Large-Scale Modeling of Scholarly Artifacts and their Usage. In *Proceedings of the Joint Conference on Digital Libraries*, Vancouver, June 2007.

Johan Bollen, **Herbert Van de Sompel** and **Marko A. Rodriguez**. Towards usage-based impact metrics: first results from the MESUR project, *JCDL 2008*, Pittsburgh, PA, June 2008. (arXiv:0804.3791v1, best paper finalist). 🐘