Author-Level Eigenfactor Metrics: Evaluating the influence of authors, institutions and countries within the SSRN community

Jevin D. West[*] Michael C. Jensen [†‡] Ralph J. Dandrea [§]
Gregg Gordon [¶] Carl T. Bergstrom [‖**]

July 20, 2012

Abstract

Abstract. In this paper, we show how the Eigenfactor® score, originally designed for ranking scholarly journals, can be adapted to rank the scholarly output of authors, institutions, and countries based on author-level citation data. Using the methods described herein, we provide Eigenfactor rankings for 84,808 disambiguated authors of 240,804 papers in the Social Science Research Network (SSRN)—a pre and post-print archive devoted to the rapid dissemination of scholarly research in the social sciences and humanities. As an additive metric, the Eigenfactor scores are readily computed for collectives such as departments or institutions as well. We show that a collective's Eigenfactor score can be computed either by summing the Eigenfactor scores of its members, or by working directly with a collective-level cross-citation matrix. To illustrate, we provide Eigenfactor rankings for institutions and countries in the SSRN repository. With a network-wide comparison of Eigenfactor scores and download tallies, we demonstrate that Eigenfactor scores provide information that is both different from and complementary to that provided by download counts. We see author-level ranking as one filter for navigating the scholarly literature, and note that such rankings generate incentives for more open scholarship, as authors are rewarded for making their work available to the community as early as possible and prior to formal publication.

[*]Department of Biology, University of Washington, Seattle, WA, jevinw@u.washington.edu
[†]Harvard Business School, Boston, MA, mjensen@hbs.edu
[‡]SSRN, Rochester, NY
[§]ITX Corp, Rochester, NY, rdandrea@itx.net
[¶]SSRN, Rochester, NY, gregg_gordon@ssrn.com
[‖]Department of Biology, University of Washington, Seattle, WA, cbergst@u.washington.edu
[**]Santa Fe Institute, Santa Fe, NM

Keywords. Eigenfactor Metrics, Author-Level Eigenfactor Score, SSRN, Author Rankings, Institutional Rankings, Citation Networks

1 Introduction

Since 1927, when two chemistry professors proposed using citation counts to make subscription decisions for university libraries [15], citation tallies have been used to estimate the academic influence and prestige of articles [25], authors [16], journals [13], departments [17], universities [20], and even nations [22]. But citations are not independent and isolated events. Rather, they form a network of interrelations among scholarly articles [7]. The structure of this network reflects millions of individual decisions by academic researchers about which papers are most important and relevant to their own work. In our efforts to extract the wealth of information from this network of citations, we can do better than simply tallying the raw number of citations: we can explicitly use information about the network structure in order to reveal the importance of each node (paper, author, journal or institution) within the citation network as a whole.

In this paper, we develop an Author-Level Eigenfactor[1] Score as a network-based measure of an author's influence within the Social Science Research Network (SSRN[2]). The SSRN corpus was selected as our data source because we were able to successfully disambiguate all authors in this community [3] and because we successfully dealt with the version challenge[4]. At the time the data for this paper was extracted from SSRN, this scholarly community consisted of 265,253 paper groups from 126,456 authors who either cited or received citations from other SSRN authors[5]. We then use Author-Level Eigenfactor Scores

[1] The Eigenfactor® Project is sponsored by the Bergstrom Lab in the Department of Biology at the University of Washington in Seattle, WA, USA. Rankings, algorithms, visual tools and maps of science are freely available at `http://www.eigenfactor.org/`.

[2] The Social Science Research Network is a pre- and post-print archive devoted to the rapid dissemination of scholarly research in the social sciences, business, law and humanities. More information can be found at `http://www.ssrn.com/`

[3] The process of author disambiguation is an ongoing effort at SSRN. Papers are received individually from authors and in bulk from other sources. Those papers that are received in bulk are reviewed by SSRN staff and each paper's authors are manually disambiguated as the papers are added to the corpus. Individual papers submitted by authors are submitted using the author's own account and therefore do not require disambiguation. SSRN's customer service team merges accounts when an author reports having two accounts or SSRN otherwise becomes aware of such duplicate accounts. Instances of false positives (where two accounts have the same name are incorrectly merged) are extremely rare.

[4] Pre-print and post-print archives such as SSRN often house multiple versions of a paper. For this study, we were able to successfully group all paper versions into 'groups' so that all authors receive the full credit of a citation.

[5] Citation data for this paper is based on SSRN CiteReader statistics as of March 14, 2011. Care was taken in this study to protect all authors' personal information. Only citation, article, institutional and download information were extracted from papers by SSRN authors. There are over 50,000 papers — primarily law papers — in the SSRN that have no formal bibliography. We could not include these in the analysis for this version of the paper, but as the references in the footnotes in these papers are extracted by CiteReader, they will be

Table 1: The 15 different disciplines represented in the Social Science Research Network. Paper counts are as of July 2012.

Research Network	**Number of Papers**
Accounting Research	18,184
Cognitive Science	6,202
Corporate Governance	15,207
Economics Research	260,198
Entrepreneurship Research and Policy	20,876
Financial Economics	97,981
Health Economics	5,664
Information Systems and eBusiness	12,335
Innovation Research and Policy	1,433
Legal Scholarship	134,691
Management Research	47,314
Political Science	44,941
Social Insurance Research	5,259
Sustainability Research and Policy	3,639
Humanities	24,385

to rank authors as well as institutions and countries associated with this set of scholars.

2 Methods

2.1 The Citation Network

The SSRN archive comprises 265,253 papers representing 126,456 unique authors across a number of disciplines in the social sciences and humanities, with particular focus in economics, law, and business (Table 1). For each paper, we extract the authors and their primary institutional affiliations, and the works cited in the article's references and footnotes. This includes 5,567,472 references. These references in each paper in the SSRN database can then be used to create large networks where the links represent references to other SSRN papers or citations from other SSRN papers; the nodes can represent either papers, authors or institutions. For this study, the nodes are authors and the links are citations between authors.

The authors in the SSRN network that we examined constitute a subset of the total number of authors. Authors with only abstracts were not included. For an author to be included in the network, he or she had to have at least one paper with a full text pdf file attached and either be cited by another SSRN author or cite another SSRN author or both. This subset of the SSRN

included in the rankings at SSRN.com. We *acknowledge* that omitting these papers creates a field bias.

network consisted of 4,468 institutions, 84,808 authors, 162,185 papers, and 1,465,082 references. Note the hierarchical structure of the data: authors are affiliated with one or more institutions; papers are affiliated with one or more authors; citations are directed among papers. To illustrate this basic – but also complicated – structure, Figure 1 shows a hypothetical example for a much smaller citation network with 10 authors, 8 papers and 3 institutions. The colored ellipses represent institutions, the numbers are individual authors and the rectangles labeled with letters are papers. The paper, author and institution relationships are combined in this figure, but they can be disaggregated to show only the papers (Figure 2), authors (Figure 3) or institutions[6]. In this study, we compute rankings based on the author-level network.

2.2 Eigenfactor Scores

The Eigenfactor® Algorithm provides a methodology for determining which nodes in a citation network are the most important or influential. The algorithm does this by computing a modified form of the eigenvector centrality of each node in the network [6]. The intuition behind eigenvector centrality is that important nodes are those which are linked to by other important nodes; while this may sound circular, importance scores can be calculated recursively according to this principle. While we apply this approach to citation networks, there are many other applications. For example, this basic concept is at the heart of Google's PageRank algorithm [23].

The Eigenfactor scores can be seen as the outcome of either of two conceptually different but mathematically equivalent stochastic processes[7]. The first process is a simple model of research in which a hypothetical reader follows chains of citations as she moves from author to author. Imagine that a researcher goes to the SSRN and selects an author at random. After (optionally) reading the article, the researcher selects at random one of the other authors who is cited by the present author. The researcher repeats this process ad infinitum. Eventually, her download patterns reach a steady state[8]. An author's Eigenfactor score is the percentage of the time that the researcher spends with this author's work in her random walk through the literature.

The second, equivalent, process is an iterated voting procedure. Each author begins with a single vote and passes it on, dividing the vote proportionally based on those authors whom she cites. In other words, if she cites two authors – author A one time and author B two times – she would distribute 1/3 of her vote to author A and 2/3 of her vote to author B. After one round of this procedure, some authors will receive more votes than others. In the second

[6] We include the paper and author-level figures to emphasize certain features of each network that are mentioned throughout this paper. Also, when developing network metrics, we find these sample networks to be invaluable in understanding the properties of network-based metrics.

[7] See "rate view" at http://www.mapequation.org/mapdemo/index.html for a demo of this process

[8] So long as the citation matrix is irreducible and aperiodic; we ensure these via the "teleportation" procedure discussed below.

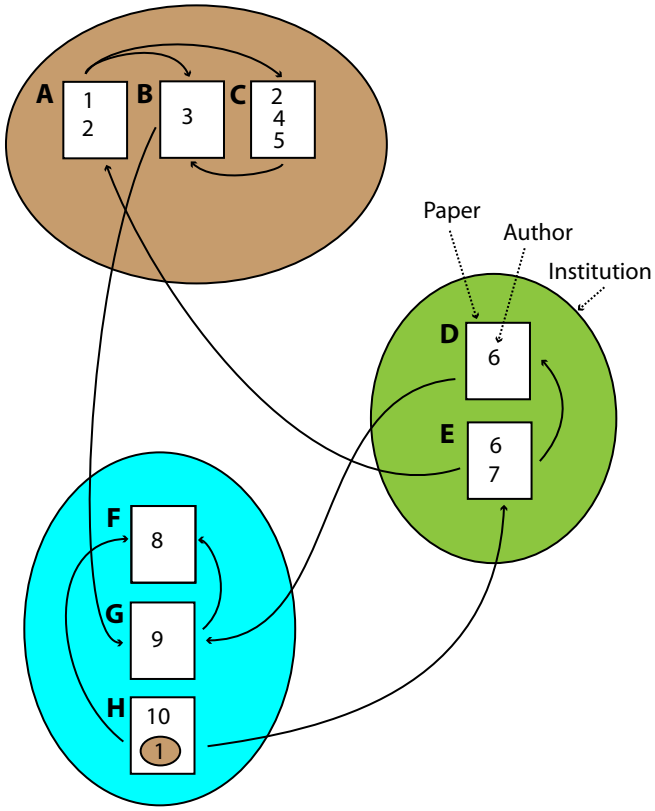

Figure 1: An example citation network among authors, papers and institutions. The large colored ellipses represent institutions. The white rectangles (labeled with letters) within each ellipse represent papers. The numbers within the rectangles represent individual authors. Many of the papers are multi-authored. For example, paper *C* has three authors (*2,4,5*). Authors are affiliated with the institution in which a given paper is located, unless indicated otherwise by coloration. For example, Author *1* is associated with the brown institution even though paper H appears in the blue ellipse. The arrows represent citations. There are 10 citations, 8 papers, 10 authors and 3 institutions in this citation network.

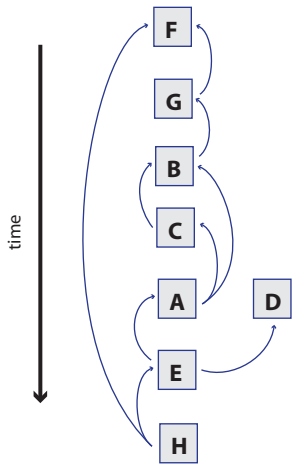

Figure 2: Paper citation network corresponding to Figure 1. Just as in that figure, the rectangles represent papers and the arrows represent citations among those papers. Paper F is the oldest paper in the example and paper H is the most recent paper written. Many of the papers cite multiple other papers but only cite backwards in time. Because of this time constraint, paper F cites no papers in this network and paper H receives no citations. Therefore, older papers in this type of network typically receive larger number of citations than newer papers. (*Note:* Albeit rare, there are scenarios when a paper will cite forwards in time. This occurs when a paper A cites an older version of paper B and paper B contains a newer version than paper A. Because we group all versions of a paper into a 'group', this could be considered a citation forwards in time.)

round, each author passes on her current vote total, as received in the previous round, again dividing this quantity equally among those authors whom she cites. This process is iterated indefinitely. Eventually, we reach a steady state in which each author receives an unchanging number of votes in each round[9]. An author's Eigenfactor score is the percentage of the total votes that she receives at this steady state.

Eigenfactor Scores have previously been used to rank scholarly journals [1, 28], and the scores are freely available at `http://www.Eigenfactor.org`. Here we extend the Eigenfactor Algorithm to the author level, and apply it to the SSRN database. The SSRN data tallies the number of times that each paper in the SSRN database has been cited by each other paper in the SSRN database since the inception of the database. From this data we can construct an *author citation network*—a directed network in which each author is a node and a

[9] Again we require irreducibility and aperiodicity.

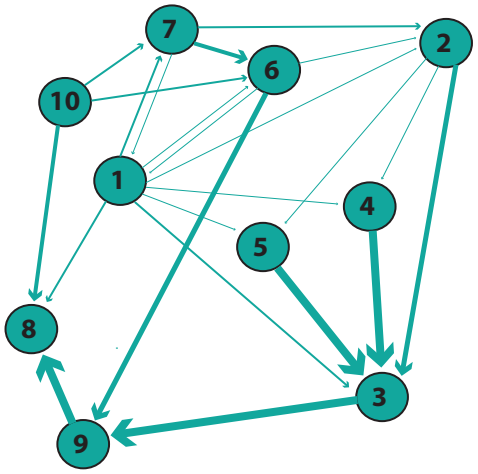

Figure 3: Author citation network corresponding to Figure 1. The circles represent authors and the arrows represent citations among the authors. The weight of each directed arrow indicates the relative fraction of citations from the source author to the recipient author. For example, the citation weight from author *9* to author *8* is twice the weight of that from author *10* to author *8*. This is because author *9* cites *only* author *8* whereas author *10* cites multiple authors.

weighted, directed edge connects author 1 to author 2 if any paper by author 1 cites any paper by author 2.

2.3 Creating the weighted cross-citation matrix

From the citation database developed by SSRN, we begin by extracting those citations from SSRN papers that reference other SSRN papers[10]. At the time of the analysis, this set of papers features 84,808 unique authors. From these authors and citations, we create a 84,808 by 84,808 square *cross-citation matrix* **R** that tallies the raw number of times that the SSRN papers of each author cite the SSRN papers of each other author, where

$$R_{ij} = \text{citations from author } j \text{ to author } i.$$

[10]At present, SSRN records only those citations listed in the references. Thus, we have missed references from legal scholars, who often include references in footnotes. SSRN is in the process of tallying these footnote references. These references will increase the number of references by approximately 75% and will disproportionately affect law authors.

When constructing $\mathbf{R}$, we omit all self-citations[11] by setting the values along the diagonal of this matrix to zero. We ignore self-citations in order to minimize the incentive for opportunistic self-citation. In the data used for this analysis, there were 21,470 authors who cited at least one of their own SSRN papers (25.4% of all authors)[12]. Before their removal, those citations consisted of 6.4% of all weighted citations.

Prior to calculating Eigenfactor scores, the citation matrix $\mathbf{R}$ must be normalized to divide credit among authors of multiple-authored papers and to scale by the number of outgoing references from each paper. We treat these steps in turn.

Dividing credit by the number of authors. The number of authors on a scholarly paper varies widely both within and between fields. As de Solla Price notes [8], if every author on a paper were to receive full credit for each citation that the paper received, this would cause some papers to be counted multiple times in the bibliometric tally, whereas others would be counted only once. Similarly, authors who tend to work as parts of large teams would be correspondingly overvalued. Such factors can have a major influence on both cardinal and ordinal rankings [9, 14].

We follow de Solla Price's proposed solution: the credit for a paper "must be divided among all the authors listed on the byline, and in the absence of evidence to the contrary it must be divided equally among them. Thus, each author of a three-author paper gets credit for one-third of a publication and one-third of the ensuing citations." [8].

Dividing credit by the number of outgoing references. Papers also vary widely in the number of outgoing references that they confer upon other articles. In order to correct for these differences, in our choice of weights we divide each reference by the number of outgoing references that each paper confers, such that each paper contributes a total citation weight of 1.0 that is shared among all of the papers that it cites.

Assigning credit across multiple versions of a paper. Pre-print archives such as SSRN tend to house multiple versions of the same paper. It is not unusual for each one of these versions to receive unique citations and the final published paper may receive only a modest fraction of the total citations received by all versions.

Thus instead of counting citations only to the final version of a paper, SSRN groups all variants of the same paper together into a "version group," and tallies the total number of citations to each version group. We do this for both the

[11] It should be noted that these are self-citations between authors and not papers. Therefore, a citation from a multi-author paper to another multi-author paper is not removed. Only the citation between the same author is removed.

[12] There were 143 authors that (1) only cited themselves and no other authors in the SSRN and (2) only received citations from themselves. This indicates that they did not co-author any of their self-cited papers with any other SSRN authors.

citing paper and the *cited* paper. This requires a lot more work in building the citation network, but we think this versioning step is critical for capturing the full credit to an author.

Computing the weighted citation matrix. Assume that SSRN authors have unique identifiers $\{1, 2, \ldots ...n_{\text{authors}}\}$. From the raw citation matrix $\mathbf{R}$, we construct a *weighted* cross-citation matrix $\mathbf{Z}$ such that Z_{ij} gives us the weighted number of times that author j has cited author i.

Per the discussion above, the weights are determined as follows. Take a paper X with m authors $x_1, x_2, \ldots, x_m$ that cites a paper Y with n authors $y_1, y_2, \ldots, y_n$. Let $c(X)$ be the number of references in the bibliography[13] of paper X. Then this citation from paper X to paper Y contributes weights

$$\omega = \frac{1}{c(X)} \frac{1}{m} \frac{1}{n} \tag{1}$$

for each author j of paper X to each author i in paper Y. The entry Z_{ij} is the sum of all weights as calculated above for all citations from author j to author i. And if the paper has multiple versions, the above refers to the version group, not any individual paper in the group.

2.4 Calculating Eigenfactor Scores for Authors

The Eigenfactor Algorithm models a random walk on the author citation network. This random walk is described by the column-stochastic form of the weighted citation matrix $\mathbf{Z}$. Thus to calculate Eigenfactor Scores, we first normalize $\mathbf{Z}$ by the column sums (i.e., by the total number of outgoing references from each author) to create a column-stochastic matrix $\mathbf{M}$, which can be written as

$$M_{ij} = \frac{Z_{ij}}{\sum_k Z_{kj}} \tag{2}$$

Following Google's PageRank approach [23, 19], we define a new stochastic matrix $\mathbf{P}$ as follows:

$$\mathbf{P} = \alpha \mathbf{M} + (1 - \alpha)\mathbf{A}, \tag{3}$$

where

$$\mathbf{A} = \mathbf{a}.\mathbf{e}^T, \tag{4}$$

where $\mathbf{a}$ is a column vector such that a_i = (number of articles by author i) / (number of total articles written by all authors in the database) and $\mathbf{e}^T$ is a row vector of 1's.

Under our stochastic process interpretation, the matrix $\mathbf{M}$ corresponds to a random walk on the citation network, and the matrix $\mathbf{P}$ corresponds to a Markov process, which with probability α follows a random walk on the author citation network and with probability $(1 - \alpha)$ "teleports" to a random author,

[13] This includes all references in the bibliography—those to SSRN papers and those to non-SSRN papers.

proportional to the number of articles published by each author. We teleport to an author with probability proportional to the number of articles (version groups) written by that author in order to avoid over-inflating the influence of authors with small numbers of articles (version groups) and under-inflating the influence of authors with large number of articles (version groups). We define the weight of each author as the leading eigenvector of $\mathbf{P}$. We compute the leading eigenvector of the matrix $\mathbf{P}$ (with teleportation) rather than using the leading eigenvector of $\mathbf{M}$ (without teleportation) for two reasons:

1. The stochastic matrix $\mathbf{M}$ may be non-irreducible or periodic. Adding the teleport probability $1-\alpha$ ensures that $\mathbf{P}$ is both irreducible and aperiodic, and therefore has a unique leading eigenvector by the Perron-Frobenius Theorem [21].

2. Even if the network is irreducible without teleporting, rankings can be unreliable and highly volatile when some components are extremely sparsely connected. Teleportation keeps the system from getting trapped in small nearly-dangling clusters by reducing the expected duration of a stay in these small cliques.

However, the teleportation procedure introduces a small but systematic bias in favor of rarely-cited authors, because these authors are visited occasionally by teleportation. The Eigenfactor Algorithm corrects for this directly. Our final author rankings will not be given by the author eigenvector $\mathbf{f}$ but rather by the product of $\mathbf{M}.\mathbf{f}$. Note that as the teleportation frequency α vanishes, $\mathbf{M}.\mathbf{f}$ converges to $\mathbf{f}$. We define the *Author-Level Eigenfactor Score* w_i of author i as the percentage of the total weighted citations that author i receives from our 84,808 source authors. We can write the vector of Author-level Eigenfactor scores as

$$\mathbf{w} = \frac{100\,\mathbf{M}\,\mathbf{f}}{\mathbf{e}^T\mathbf{M}\,\mathbf{f}} \tag{5}$$

2.5 Institutional Rankings

Like measures of mass or volume, the Eigenfactor score $\mathbf{w}$ is an additive metric. To find the Eigenfactor score for a group of authors, simply sum the Eigenfactor scores of the authors in the group. Thus, it is straightforward to use the author-level Eigenfactor scores to rank various departments, universities, or other institutions. By this approach, the Eigenfactor score assigned to an institution I_j is simply the sum $I_j = \sum_k w_k$, where w_k is the author-level Eigenfactor score of author k associated with institution. Moreover, when groups constitute a hard partition, i.e., each individual belongs to one and only one group, the group-level score computed in this way is the same value that one would get by computing Eigenfactor scores directly from a group-level citation matrix that weights each group member by its Eigenfactor score therein. We prove this in Appendix A.

Here, however, we are not dealing with a hard partition. Individual authors may be associated with multiple institutions. We could recover the equivalence result by assigning fractional credit to each institution with which a multi-institutional author is associated. But such an approach may bias results against institutions that are able to recruit high-prestige authors with multiple affiliations. Thus we have opted to grant credit for the full author-level Eigenfactor score to each and every institution with which an author is affiliated. As a result, the group-level Eigenfactor scores that we report here—computed by summing author-level scores—differ slightly from the scores that would result operating directly on the group-level citation matrix.

3 Results

3.1 Author Rankings

The Eigenfactor Algorithm was independently coded by the Eigenfactor team and the SSRN team in two different programming languages in order to insure the integrity of the results.

There were 84,808 authors ranked by Eigenfactor Scores[14]. The top twenty authors and their institution affiliations are shown in Table 2. The rankings for an additional 30,000 authors can be found at SSRN.com. When summed together, the top twenty authors accounted for 7.5% of the total Eigenfactor Score for all authors. The mean Eigenfactor Score for all 84,808 authors is 0.12 with a standard deviation of 0.97. The author-level Eigenfactor scores can be interpreted in the following way: if one were to randomly follow citations from author to author in the SSRN database for a very long time, 0.688% of the time would be spent at literature authored or coauthored by Andrei Shleifer (see Methods for alternate explanations). That is a substantial proportion, given the 84,808 authors in this citation network. It should be noted that many of the top 20 authors posted their papers on SSRN at the inception of SSRN. This gives these authors an advantage in accumulating downloads and probably citations, as well.

Columns four, five and six indicate the total citation weight given to other SSRN authors, the total citation weight received from other SSRN authors and the number of papers authored or co-authored by each author, respectively (CT_o = Outgoing Citation Weight, CT_i = In Citation Weight, Art = Articles in SSRN). The numbers in these three columns are not integer-valued because of the way citation and article credit are divided among multi-authored papers (see Methods, Equation 1 and Equation 6).

The cumulative distribution of Eigenfactor Scores for the top 10,000 authors is shown in Figure 4. The authors are ordered on the x-axis from highest ranked to lowest ranked (i.e., author 100 was the author that ranked 100th by Eigenfactor Score). The shape of the curve shows that most of the Eigenfactor comes from a small percentage of the authors. The dashed lines indicate the

[14]Note: The Eigenfactor scores are multiplied by 100. See Table 2 caption.

Table 2: The highest-scoring 20 of 84,808 authors in the SSRN ranked by their Author-Level Eigenfactor Scores. Note that these rankings reflect centrality and influence within the SSRN community, rather than in academia generally. A complete list for the top 30,000 authors is available at SSRN.com. EF = Eigenfactor $\times 100$, CT_o = Outgoing Citation Weight, CT_i = In Citation Weight, Art = Articles Written.

	Author	**EF**	$\mathbf{CT_o}$	$\mathbf{CT_i}$	**Art.**	**Institution**
1	Jensen, Michael C.	83.2	8.5	226.7	78.7	Harvard University
2	Shleifer, Andrei	68.8	14.1	210.0	78.0	Harvard University
3	Campbell, John Y.	53.1	11.2	150.5	61.0	Harvard University
4	Heckman, James J.	38.5	9.5	81.7	91.2	University of Chicago
5	Shavell, Steven	37.0	13.1	116.1	91.0	Harvard University
6	Glaeser, Edward L.	36.4	21.6	87.7	93.6	Harvard University
7	Hall, Robert E.	34.3	8.6	67.0	52.9	Stanford University
8	Barro, Robert J.	32.7	6.6	93.9	52.6	Harvard University
9	Acemoglu, Daron	32.5	17.1	104.2	94.2	MIT
10	Vishny, Robert W.	31.8	1.9	97.1	18.7	University of Chicago
11	Rajan, Raghuram G.	31.8	12.2	98.3	48.9	University of Chicago
12	Poterba, James M.	31.4	10.6	55.6	76.8	MIT
13	Zingales, Luigi	30.7	12.1	111.1	60.3	University of Chicago
14	Murphy, Kevin J.	30.4	4.0	58.8	19.0	University of Southern California
15	Stein, Jeremy C.	30.4	7.4	93.6	42.4	Harvard University
16	Feldstein, Martin S.	30.1	15.4	60.5	132.0	NBER
17	Shiller, Robert J.	29.6	7.8	62.3	61.6	Yale University
18	Harvey, Campbell R.	29.1	18.0	116.8	57.8	Duke University
19	Blanchard, Olivier J.	28.6	12.4	91.5	65.9	MIT
20	Lopez de Silanes, Florencio	28.0	3.4	97.2	22.4	EDHEC Business School

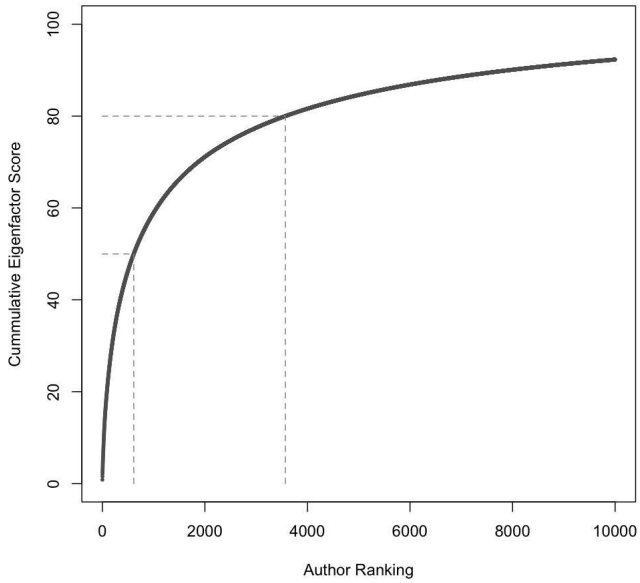

Figure 4: Cumulative Distribution of Eigenfactor score. The figure shows the fraction of the total Eigenfactor accounted for by the first 10,000 authors (92.82%) out of 84,808 total authors. The x-axis indicates author rank (i.e., author 500 is the 500th highest-ranked author ranked by Eigenfactor). The y-axis is the cumulative Eigenfactor score. The dashed vertical lines indicate how many authors account for 50% of the total Eigenfactor score and 80% of the Eigenfactor score. The 50% line crosses the x-axis at the author ranked 612 and the 80% line crosses the x-axis at the author ranked 3,569. The Eigenfactor scores at those 0.5 and 0.8 quantiles are 0.0296 and 0.0040, respectively.

authors at which 50% and 80% of the total Eigenfactor Score is attained. The top 612 authors account for 50% of the Eigenfactor Score, and the top 3,569 authors account for 80% of the Eigenfactor Score. The Eigenfactor Scores at those 0.5 and 0.8 quantiles are 0.0296 and 0.0040, respectively.

The Eigenfactor Score can be viewed as a form of weighted citation count where the weights reflect the prestige of the citing authors. Therefore, one would expect the Eigenfactor Scores to correlate with other weighted citation counts. Figure 5 shows a log-log plot of Eigenfactor Scores versus the total citation weight Ω for each author. We calculate total citation weight by simply tallying citations, and weighing each author's fractional share as we have done for the Eigenfactor scores, as given in Equation 1. Each author i receives citation weight ω from author j. Therefore, the total citation weight for author i is

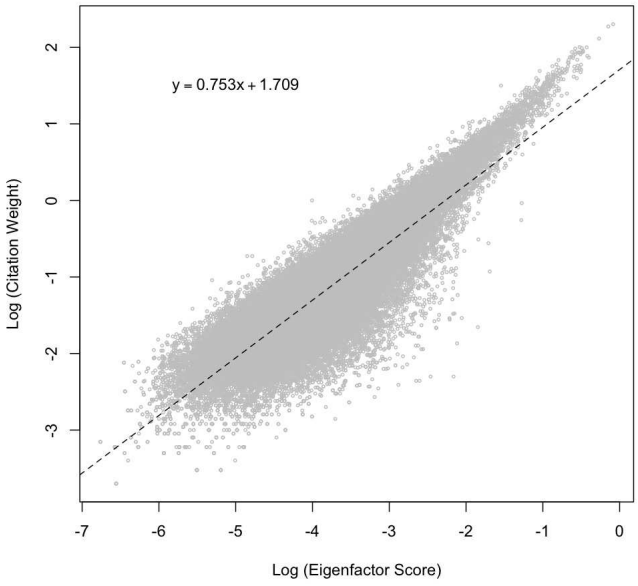

Figure 5: Relationship between Eigenfactor Score and total citation weight. The x-axis is the Eigenfactor Score. The y-axis is the total citation weight Ω_i for each author, i. The linear regression (dashed line) of log citation weight a on log Eigenfactor score b is given by the following equation: $a = 0.753b + 1.709$. ($\rho = 0.88$)

$$\Omega_i = \sum_j w_j \tag{6}$$

The dashed line in Figure 5 is a best fit linear regression line on the log data. Despite the relatively high correlation ($\rho = 0.88$), an Eigenfactor score near the middle of the distribution could be associated with a three-order of magnitude range of citation weights. In other words, two authors could have the same Eigenfactor Score but have a citation weight that was different by three orders of magnitude. The converse is even more extreme; authors with the same citation weight can have an Eigenfactor Score that varies more than four orders of magnitude. These differences result in very different ordinal rankings when authors are ranked by citations[15] or ranked by Eigenfactor score. See West et al. (2010a) for a more thorough discussion about these kinds of correlation coefficients.

[15]The differences are even greater when ranking by raw citations instead of citation weight.

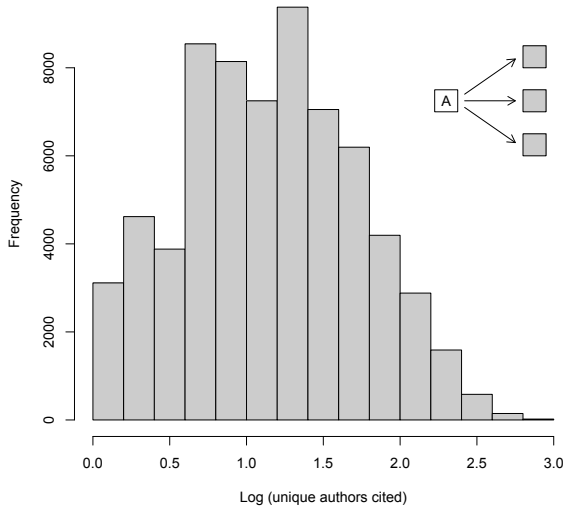

Figure 6: Unique Authors Cited. The histogram shows the number of different authors cited by each individual author in the SSRN. Most authors cite fewer than 100 different authors. The log scale on horizontal axis is base 10. The 10,268 authors that cited zero authors (but received citations) are not shown. Among the authors who cited other SSRN authors, the mean number of authors cited is 32.8 with a standard deviation of 53.1. The network schematic shows the direction of citations being tallied in this figure.

3.2 Network Sparseness

Author citation networks are typically very sparsely connected (i.e., the cross-citation network has many zero entries). However, there are well-connected authors that cite a relatively high proportion of all the other authors in the database. One contributor, Rene M. Stulz, cited 927 unique SSRN authors. Figure 6 illustrates this network sparseness. For the 73,471 authors that cited at least one author in the SSRN, we counted the number of unique authors cited. The distribution of this tally is shown in Figure 6. There are 67,338 authors (91.7% of authors that cite at least one paper) that cite fewer than 100 other different SSRN authors.

Figure 7 illustrates the converse; it shows the number of unique SSRN authors citing each author in the database. In other words, for each author we look at the incoming citations (from other unique authors). Relative to the distribution shown in Figure 6, the mode of this distribution is shifted to the left, and the standard deviation is much higher (164.1). Most authors receive cita-

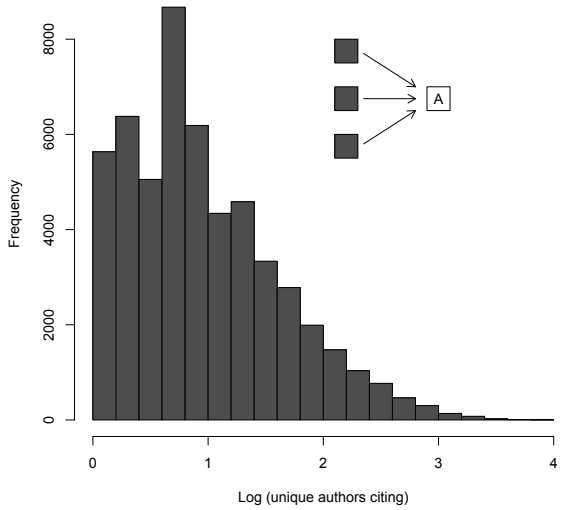

Figure 7: Incoming Citations from Unique Authors. The histogram above is the converse of Figure 6. It shows the frequency of different authors citing each author in SSRN (see the network schematic in the figure which shows the direction of citations tallied). Only the 53,253 authors receiving at least one citation are shown. Among these authors, the mean number of unique citing authors is 41.7 and the standard deviation is 164.1.

tions from relatively few other authors, but the distribution of citations received has a longer tail than the distribution of authors cited; some authors have been cited by a very large number of unique authors. For example, Andrei Shleifer has received citations from 9,672 different authors. Another way to think about it is that 7.6% of all authors in the SSRN community have cited Schleifer. This speaks to the centrality of Schleifer in this particular community.

Many authors either received no citations or gave out no references. There were 10,268 authors that received citations but gave out no references. These authors in a citation network are known as dangling nodes. Conversely, there were 24,602 authors that gave out references but received no citations. Authors that both gave out zero references and received zero citations were eliminated before the 84,808 x 84,808 adjacency matrix was created. Also, authors with zero articles and authors that have abstracts but no full text documents in the SSRN eLibrary were eliminated before the construction of the adjacency matrix .

Table 3: The top 20 research universities and other academic institutions ranked by the Eigenfactor scores of SSRN authors affiliated with the institution. There were 7,865 institutions ranked. A complete list will be available at SSRN.com. These rankings reflect centrality and influence within the SSRN community, rather than in academia generally. $\sum EF$ = Sum of SSRN Author-Level Eigenfactor Scores ($\times 100$) associated with that institution.

	Institution	$\sum$ **EF**
1	National Bureau of Economic Research	4,597.8
2	Harvard University	1,133.2
3	Centre for Economic Policy Research	1,128.2
4	University of Chicago	518.7
5	Institute for the Study of Labor	518.3
6	Massachusetts Institute of Technology	415.9
7	New York University	356.9
8	European Corporate Governance Institute	337.7
9	University of California, Berkeley	316.8
10	Columbia University	310.6
11	Stanford University	307.4
12	CESifo	293.7
13	International Monetary Fund	278.3
14	University of Pennsylvania	254.9
15	Princeton University	248.5
16	Federal Reserve Banks	243.7
17	Yale University	220.8
18	World Bank	185.9
19	Northwestern University	182.6
20	Government (USA)	161.1

3.3 Ranking Institutions

Thousands of institutions from around the world are represented in the SSRN database. Most of these institutions are universities or university departments, but there are other types of institutions such as research collaboratives (e.g., NBER, CEPR, ECGI, IZA, and CESifo). Using the author-level Eigenfactor Scores, these universities and departments can be ranked. This analysis was performed on 7,865 different institutions and constituent departments. Table 3 lists the top twenty institutions by Eigenfactor Score.

Institutions are one way to aggregate authors; countries are another. Countries are credited with a paper if the author of the paper is associated with an institution that lies within the country borders. There are 129 countries represented in the SSRN. The top twenty can be found in Table 4. The United States carries 72% of the total Eigenfactor for all countries. As with author rankings, it is important to understand this is a measure of centrality to the SSRN rather than a measure of the relative overall productivity of researchers in various countries. Some institutions have working paper series, which gives them an inherent

Table 4: The top 20 countries represented in the SSRN database, ranked by Author-Level Eigenfactor Scores. These rankings reflect centrality and influence within the SSRN community, rather than in academia generally. There were 129 countries represented in the database. $\sum EF$ = Sum of SSRN Author-Level Eigenfactor Scores ($\times 100$) associated with that country.

Rank	Country	$\sum$ EF
1	United States	13,124.0
2	United Kingdom	1,701.3
3	Germany	1,018.0
4	Belgium	393.2
5	Netherlands	208.1
6	Canada	192.4
7	Italy	186.8
8	France	179.9
9	Switzerland	144.6
10	Sweden	117.6
11	Spain	112.1
12	Israel	101.5
13	Australia	81.0
14	China	47.9
15	Denmark	30.3
16	Hong Kong	30.2
17	Norway	26.3
18	Singapore	24.9
19	Japan	24.3
20	Austria	20.0

advantage. Also, some countries benefit from the presence of collaborative research organizations such as NBER (National Bureau of Economic Research) in the United States or ECGI (European Corporate Governance Institute) in Belgium that may have hundreds of scholar associates who are full time faculty employed by various universities around the world.

3.4 Usage vs citations

Citation counts are not the only way to assess the quality or impact of scholarly work, and indeed they may systematically undervalue certain papers that are widely read by authors, students or practitioners but less often cited in the subsequent research literature [4]. In addition to tracking citations, SSRN has collected usage data as well, tracking every single download of every single paper in the archive since the archive's inception. We can use these data to rank authors by downloads. Table 5 lists the top 20 authors by this metric. Each time a paper is downloaded, the authors of that paper receive credit for that download. The credit is divided evenly among the authors, similar to how citation credit is distributed (see Methods), so that a paper with 3 authors and 300 downloads contributes a score of 100 to each author. Thus the "download

Table 5: The top 20 authors by downloads per author. The weight for each downloaded paper is distributed evenly among the authors (i.e., an author will receive half a download if they co-author a paper with one other author). The downloads shown in the table is the sum of this weight for each author.

	Author	Downloads	Institution
1	Jensen, Michael C.	325,974.0	Harvard University
2	Fernandez, Pablo	245,156.0	University of Navarra
3	Fama, Eugene F.	187,024.0	University of Chicago
4	Solove, Daniel J.	131,362.0	George Washington University
5	Velez-Pareja, Ignacio	107,687.0	University Tecnologica de Bolivar
6	French, Kenneth R.	90,433.0	Dartmouth College
7	Bebchuk, Lucian A.	88,217.1	Harvard University
8	Bruner, Robert F.	84,974.4	University of Virginia
9	Faber, Mebane T.	78,676.0	Cambria Investment Management
10	Bainbridge, S.M.	71,763.5	UCLA
11	Sunstein, Cass R.	70,571.3	Harvard University
12	Castronova, Edward	69,805.0	Indiana University Bloomington
13	Goetzmann, William N.	69,692.1	Yale University
14	Lott, John R.	65,280.5	University of Maryland
15	Meckling, William H.	64,982.3	University of Rochester
16	McGee, Robert W.	64,583.1	Florida Int. University
17	Lemley, Mark A.	62,386.6	Stanford University
18	Black, Bernard S.	60,447.9	Northwestern University
19	Lo, Andrew W.	59,930.0	MIT
20	Penman, Stephen H.	56,001.7	Columbia University

weight" for an author is simply the sum of the contributions from each paper on which that individual is an author[16].

Comparing Table 2 to Table 5, the top 20 lists change dramatically, indicating that downloads and citations provide different information. Researchers in bibliometrics have explored the relationships between citations and usage for several data sets; in general, citation measures and usage measures are positively correlated but provide complementary information about the influence of scholarly papers [18, 5, 3, 26]. Figure 8 is a log-log plot that shows author-level Eigenfactor Scores plotted against download weight.

We collected download information for the same 84,808 authors included in the citation network. When the credit is divided among the authors as explained above, the average download weight per author is 436.6 and the standard deviation is 2331.6. The maximum download weight attained up to this point is 325,974.

The Pearson's linear correlation coefficient between Eigenfactor and download weight is $\rho = 0.485$. Thus, we see that Eigenfactor scores provide considerable information above and beyond that available from download scores and

[16]SSRN has gone to great lengths to ensure that reported downloads are free of biases caused by bots, search engines, or gaming.

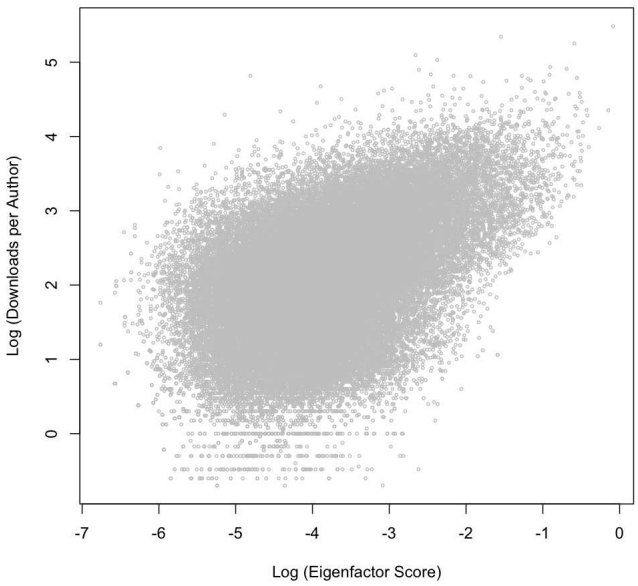

Figure 8: Downloads vs Eigenfactor Scores for SSRN authors. In this log-log plot, each data point represents an author and their corresponding Eigenfactor Score and number of downloads per Author (splitting credit among co-authors). There are 52,196 authors represented in this figure. Authors that have an Eigenfactor Score of zero or have zero downloads are not shown.

Table 6: The top 20 research universities and other academic institutions ranked by author downloads. There were 7,865 institutions ranked. A complete list will be available at SSRN.com. $\sum D$ = Sum of SSRN Author-Level download weights for authors associated with that institution.

	Institution	$\sum$ **D**
1	National Bureau of Economic Research	3,248,753.1
2	Harvard University	1,696,742.2
3	European Corporate Governance Institute	1,288,849.5
4	Centre for Economic Policy Research	992,061.1
5	University of Chicago	934,392.0
6	New York University	909,346.1
7	Yale University	780,232.1
8	University of Pennsylvania	695,644.3
9	Columbia University	693,639.0
10	World Bank	629,944.6
11	Stanford University	620,438.6
12	Institute for the Study of Labor	610,139.8
13	Massachusetts Institute of Technology	572,024.1
14	CESifo	565,851.0
15	International Monetary Fund	528,008.0
16	University of Virginia	489,823.1
17	Duke University	450,274.6
18	University of California, Berkeley	434,392.9
19	Government of the United States of America	413,465.0
20	George Washington University	413,285.8

vice versa. There are scholars that have a relatively high Eigenfactor Scores but few downloads; in many cases this occurs because the paper is available from other sources such as the NBER or CEPR servers and because NBER and CEPR charge non-members $5 for downloading NBER and CEPR papers on SSRN, while most of the rest of the papers on SSRN can be downloaded at no cost. There are also authors [such as Fairmain with his classic treatise "Fuck" [10]; see also [11] for the impact of that oft-downloaded article on institutional rankings] who have written papers that are downloaded a large number of times for various reasons but receive relatively few citations.

Within the SSRN database there are 22,475 authors that have an Eigenfactor Score of zero but have a nonzero number of downloads per author. This means that there are many papers in the SSRN that are downloaded and viewed but are not cited. There are no authors that have zero downloads but a nonzero Eigenfactor Score.

Usage data can also be used to rank institutions. Table 6 shows the top 20 institutions ordered by download weight. The $\sum$ **D** was calculated by summing the download weight for every author associated with each institution. As with author rankings, institutional citational ranks differ substantially from institutional download ranks.

3.5 The arrow of time

One particular challenge with iterative ranking algorithms at the paper level is the time-directionality of citation networks: any given paper cites only papers published earlier than it[17]. Therefore, a random walker following citations will progressively move backwards in time. One way to counter this effect is to bias the teleport process toward more recent publications [24]. In principle, the same problem could arise for author-level networks if they extend over sufficiently long time intervals; Alfred Marshall never cited Paul Samuelson. Random walks on the author network will tend to move backward in time and thus earlier authors may receive a disproportionate number of visits and thus disproportionately higher scores.

In practice, this does not turn out to be a major problem for the SSRN corpus, given its relatively narrow time window from 1995 to the present[18] and the fact that most authors with early papers in the database remain active in the community at present. Thus we do not need to employ any sort of time-biased teleport mechanism in the article-level Eigenfactor rankings that we compute for the SSRN. To check this we looked at the distribution of papers dates[19] immediately after teleport, one step after teleport, etc. If the random walk tended to drift back in time, we would see that, as we take more random walk steps, the distribution of paper dates would shift to earlier years. Figure 9 shows the distribution of authors' earliest papers (top panel), and most recent papers (bottom panel) after teleport (Step 0) and after a single step (Step 1) on the network[20]. After one step, the distribution of the oldest paper is shifted back in time, but this does not in and of itself indicate strong overall backward movement. In fact, the distribution of the most recent paper actually shifts forward in time after a step on the network. This means that the random walk process moves us toward authors with older papers in the database–but these same authors also have more recent papers as well. This is less counterintuitive than it seems. The random walk process moves toward authors who are more central and probably have more papers overall. Thus we should not be so surprised to see a broader range of dates for these authors.

[17] There are some exceptions with working papers where paper A cites paper B and paper B cites paper A, but this is rare.

[18] Authors can and do submit papers with dates earlier than 1995. As time goes on, more early papers will be uploaded to SSRN; however, if those earlier papers are from authors still active in the SSRN community, we don't expect our random walker to progressively move backwards in time.

[19] The 'paper date' is the first available date that we could find for each paper. The date would be the earliest of the following: (1) publication date, (2) date the paper was entered into the SSRN system, or (3) date shown on any citation that is matched to the paper. If a paper was entered in 2000 but has a publication date of 1975, then 1975 is the date used for the paper. If a paper was entered in 2000, has an unrecognizable paper date, but has a citation from 1975, then 1975 is used. The earliest date of these three scenarios is always used. This is especially useful when dealing with multiple versions of a paper.

[20] The distributions change very little for higher numbers of steps and thus are not shown.

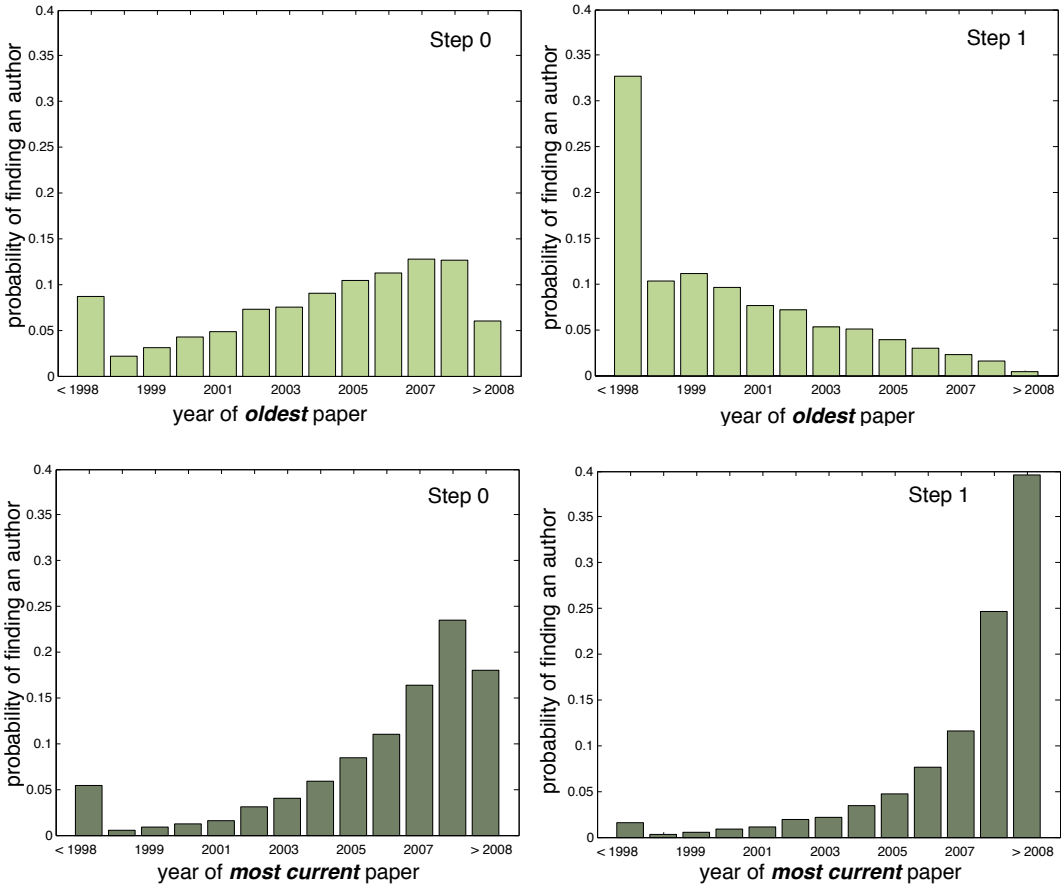

Figure 9: The probability distribution of finding an author with their oldest paper (top panels) and most current paper (bottom panels) in each of the last ten years, after teleportation (step 0) and one step after teleportation (step 1). After one step on the network, there is a higher probability of finding an author with a paper before 1998, but the probability is higher still of finding an author —possibly the same author — with a paper subsequent to 2008.

4 Discussion

Ranking papers, authors, journals, departments, or institutions does not necessarily make the world a better place. Indeed, where ranking systems provide narrow-minded administrators and faculty with an excuse to avoid hard work and deep thought, they may even be harmful to the functioning of academia. Then why rank at all? While ranking for its own sake may or may not offer net benefits to the community, we believe that advances in scholarly ranking will have at least two consequences that benefit science greatly. We treat these in turn.

First, ranking provides incentives for early and open sharing of scientific information. The SSRN repository, like other on-line archives, performs an important function for the scholarly community. By enabling the distribution of working papers and by making author-submitted manuscripts at all stages easily available and at zero cost, SSRN reduces the time that it takes for an idea, first conceived in one scholar's mind, to become a part of the conversations among many scholars around the world.

Scholars respond to incentives like anyone else, and because scholars are rewarded for prominence and prestige, ranking systems can provide strong incentives [29]. One of the best ways an author can advance his or her score under the ranking system described in this paper is by uploading all versions of all papers to SSRN, as early as possible. In this way, authors have better chances of being read and subsequently cited. In this respect, ranking generates a positive externality for science.

Second, ranking furthers search [2]. Search engines such as Google have fundamentally changed scholarship by improving our ability to find the information that we value, rapidly and efficiently. Ranking algorithms such as PageRank lie at the heart of these search engines — effective search requires that we account for not only the match of search terms to target document, but also for the importance of the target document within a larger collection. Just as Google's PageRank algorithm helps with the discovery process on the world wide web by filtering search results, the Eigenfactor metrics described here can help with the discovery process within this citation network. Properly integrated with other search tools and algorithms, the Eigenfactor Metrics can help users to find important papers that may have been overlooked by other ranking methods based on downloads or reputation. Such applications in discovery provide a major motivation for the present work. It is our hope and belief that advances in ranking will serve our quest for more efficient search, helping academics sift through ever-growing volumes of information to find the hidden gems and lost papers that are valuable for their research endeavors.

Finally, it is important to recognize what these statistics do and do not represent. Eigenfactor is not a direct measure of quality. Rather, Eigenfactor is (as discussed above) one of a family of network centrality measures. The Author-Level Eigenfactor Scores presented here measure the *centrality* of authors within the particular network (SSRN) that we study. For example, notice that of the top 20 authors in Table 2, seven authors are associated with Harvard University.

While all of these individuals are influential academics by any measure, the preponderance of Harvard faculty at the top of the list probably reflects the origins of SSRN at Harvard and thus the centrality of this group of researchers in the broader network they have formed around themselves. The same caution should also be applied to the institutional rankings derived here.

For our purposes, these rankings are simply one of many filters that can be applied to a large, seemingly unmanageable data set. In this study, we used the entire SSRN corpus. This includes papers from the social sciences, economics, law, and business. However, rather than running the Eigenfactor Algorithm on the full network, we can apply the algorithm to any subset of the citation network, such as those authors affiliated with one particular institution or country, to get rankings specific to the interests of that group. Librarians and other collection agencies could analyze their own specific subscriptions. Departments and colleges could look at intra- or inter-college citations among their faculty to look at how closely their faculty are working together and who is most central to the collaborative work being done. Journal societies and associations could use these algorithms to find the active members—who is citing and who is being cited by their members. Other online archives like SSRN could find who is being read in their collections and what groups are contributing to their particular field. There are many ways that reference networks can be analyzed using the Eigenfactor metrics and related approaches. We believe that pre-print and post-print archives such as SSRN are extremely useful for the scholarly community and for the quick dissemination of new ideas and papers. Ultimately, we would like to use filters such as the author-level Eigenfactor scores developed here to build better search algorithms that help researchers mine the vast, and ever-expanding scholarly literature.

Acknowledgments

This work was supported in part by NSF grant SBE-0915005 to CTB.

References

[1] Bergstrom, C.T. (2007). Eigenfactor: Measuring the value and prestige of scholarly journals. *College and Research Libraries News*, 68(5):314–316.

[2] Bergstrom, C.T. (2010). How to improve the use of metrics: Use ranking to help search *Nature*, 465:871-872.

[3] Bollen, J. and van de Sompel, H. (2006). Mapping the structure of science through usage. *Scientometrics*, 69(2):227–258.

[4] Bollen, J. and Van de Sompel, H. (2008). Usage impact factor: the effects of sample characteristics on usage-based impact metrics. *Journal of the American Society for Information Science and Technology*, 59(1):001–014.

[5] Bollen, J., Van de Sompel, H., Smith, J., and Luce, R. (2005). Toward alternative metrics of journal impact: A comparison of download and citation data. *Information Processing and Management*, 41(6):1419–1440.

[6] Bonacich, P. (1972). Factoring and weighting approaches to clique identification. *Journal of Mathematical Sociology*, 2:113–120.

[7] de Solla Price, D. (1965). Networks of Scientific Papers. *Science*, 149: 510-515

[8] de Solla Price, D. (1981). Multiple authorship. *Science*, 212:986.

[9] Egghe, L., Rousseau, R., and Van Hooydonk, G. (2000). Methods for Accrediting Publications to Authors or Countries: Consequences for Evaluation Studies. *Journal of the American Society for Information Science*, 51(2):145–157.

[10] Fairman, C. M. (2006). Fuck. *SSRN eLibrary*, http://ssrn.com/paper=896790.

[11] Fairman, C. M. (2007). Fuck and law faculty rankings. *SSRN eLibrary*, http://ssrn.com/paper=971103. Working Paper Series.

[12] Garfield, E. (1955). Citation indexes for science. *Science*, 122:108 – 111.

[13] Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science*, 178(4060):471–479.

[14] Gauffriau, M. and Larsen, P. (2005). Counting methods are decisive for rankings based on publication and citation studies. *Scientometrics*, 64(1):85–93.

[15] Gross, P. and Gross, E. (1927). College libraries and chemical education. *Science*, 66(1713):385–389.

[16] Hirsch, J. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46):16569–16572.

[17] Kalaitzidakis, P., Mamuneas, T., and Stengos, T. (2003). Rankings of Academic Journals and Institutions in Economics. *Journal of the European Economic Association*, 1(6):1346–1366.

[18] Kurtz, M., Eichhorn, G., Accomazzi, A., Grant, C., Demleitner, M., Murray, S., Martimbeau, N., and Elwell, B. (2005). The bibliometric properties of article readership information. *Journal of the American Society for Information Science and Technology*, 56(2):111–128.

[19] Langville, A. and Meyer, C. (2006). *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton University Press.

[20] Liu, N. and Cheng, Y. (2005). The Academic Ranking of World Universities. *Higher Education in Europe*, 30(2):127–136.

[21] MacCluer, C. (2000). The many proofs and applications of perron's theorem. *Society for Industrial and Applied Mathematics*, pages 487–498.

[22] May, R. (1997). The Scientific Wealth of Nations. *Science*, 275(5301):793.

[23] Page, L., Brin, S., Motwani, R., and Winograd, T. (1998). The pagerank citation ranking: Bringing order to the web. *Technical report, Stanford Digital Library Technologies Project*, http://ilpubs.stanford.edu:8090/422/.

[24] Walker, D., Xie, H., Yan, K., and Maslov, S. (2007a). Ranking scientific publications using a model of network traffic. *Journal of Statistical Mechanics: Theory and Experiment*, page P06010.

[25] Walker, D., Xie, H., Yan, K., and Maslov, S. (2007b). Ranking Scientific Publications Using a Simple Model of Network Traffic. *Journal of Statistical Mechanics: Theory and Experiment*, 2007:P06010.

[26] Watson, A. (2009). Comparing citations and downloads for individual articles. *Journal of Vision*, 9(4):i, 1–4.

[27] West, J.D., Bergstrom, T.C., and Bergstrom, C.T. (2010a). Big macs and Eigenfactor scores: Don't let correlation coefficients fool you. *Journal of the American Society for Information Science and Technology*, 61(9):1800 – 1807.

[28] West, J.D., Bergstrom, T.C., and Bergstrom, C.T. (2010b). The eigenfactor metrics: A network approach to assessing scholarly journals. *College and Research Libraries*, 71(3):236–244.

[29] West, J.D. (2010). How to improve the use of metrics: Learn from Game Theory *Nature*, 465:871-872.

A Aggregating Eigenfactor Scores

A.1 Statement of problem

Networks are often composed of nodes that themselves can be arranged into larger groupings. For example, individuals belong to households, college teams are organized into conferences, and scholars are associated with research institutions. In these situations, it may be desirable to calculate eigenvector centrality scores not only for the individual nodes, but also for the groups into which they the individual nodes are assigned. How should these group-level rankings be generated? Should a group's eigenvector centrality score be defined as the sum of the eigenvector centrality scores of its members? Or should it be defined as the eigenvector centrality score associated with that group in a new, coarse-grained group-level network, in which each group is itself a node?

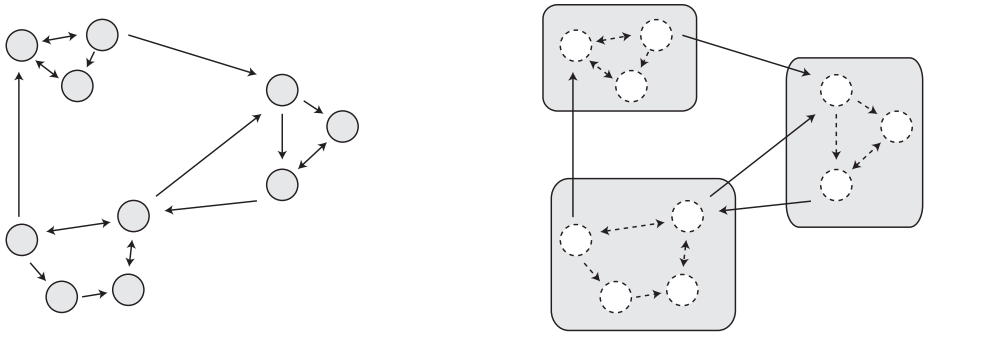

Figure 10: The Markov process ψ on individual nodes has a shadow process ψ' on groups.

The Eigenfactor algorithm provides a way to compute something like an eigenvector centrality score for a matrix that fails to be irreducible and aperiodic. Similar questions arise for Eigenfactor scores; how should the Eigenfactor score of a group be defined?

Here we demonstrate one strength of the eigenvector centrality and Eigenfactor approaches. As with simple citation tallies, one gets the same result whether one sums the scores of individual nodes to provide group-level scores, or whether one analyzes the group-level network assembled using a reasonable choice of aggregation procedure.

A.2 Eigenvector centrality scores

Setup: Let $\mathcal{I}$ be the sets of nodes $i = 1, 2, \ldots n$ in a weighted, directed graph. Let $\mathcal{G}$ be a hard partition of the nodes of $\mathcal{I}$ into $k = 1 \ldots g \leq n$ groups, such that each node i is member of exactly one group G_k. Let $\mathbf{P}$ be the Markov matrix associated with an irreducible and aperiodic Markov process ψ on the set $\mathcal{I}$ of individual nodes. (If ψ is the PageRank process, for example, $\mathbf{P}$ is irreducible and aperiodic by construction.) Let $\mathbf{P}'$ be the Markov matrix associated with this process when it is described at the level of groups instead of nodes.

Proposition 1 *Given the group-level Markov matrix* $\mathbf{P}'$*, the eigenvector centrality score of each group is equal to the sum of the eigenvector centrality scores of its member nodes as computed from the individual-level Markov matrix* $\mathbf{P}$*.*

Proof: By the Perron-Frobenius theorem, ψ has a unique stationary distribution λ. Construct a *shadow process* ψ' on the groups in the partition $\mathcal{G}$ such that the shadow process tracks the process ψ but records its present state as the group

index k rather than the node index i. Thus the shadow process ψ' is in state k if and only if the process ψ is in some state i such that $i \in G_k$. Moreover every node i belongs to one and only one group G_k because $\mathcal{G}$ is a hard partition.

If the process ψ begins at its stationary distribution λ, its shadow process ψ' can be described by an *aggregated matrix* $\mathbf{P}'$ as follows: let $P'_{k,m}$ be the conditional probability the Markov process ψ transitions from a node in group G_m to a node in group G_k in a single step, conditional on starting in G_m at a node selected with probability proportional to λ. This matrix is given by

$$P'_{k,m} = \sum_{i \in G_k} \sum_{j \in G_m} P_{i,j} \frac{\lambda_j}{\sum_{l \in G_m} \lambda_l}$$

In other words, to construct $\mathbf{P}'$ we compute the outlink weights for any group G_k by summing the outlink weights from each node i in G_k, where the contribution of node i in G_k is chosen proportional to λ_i.

Because ψ has a unique stationary distribution with probability λ_i on each node i, the shadow process ψ' has a unique stationary distribution on that places probability $\lambda'_k = \sum_{i \in G_k} \lambda_i$ on each group G_k. Thus by the Perron-Frobenius theorem the vector of eigenvector centrality scores for the aggregate matrix $\mathbf{P}'$ that we have constructed to describe ψ' is this vector λ'.

A.3 Eigenfactor scores

Eigenfactor scores are calculated by taking a link-weight matrix $\mathbf{Z}$, removing self-links, and normalizing to produce a column-stochastic matrix $\mathbf{M}$. This matrix is converted to an irreducible and aperiodic matrix $\mathbf{P}$ by a teleport procedure. The eigenvector centrality scores for the matrix $\mathbf{P}$ are given by the vector λ. To ameliorate the effects of the teleportation procedure, this vector λ is then multiplied once more by $\mathbf{M}$, and the scores are normalized to sum to 100. Here we will show that the Eigenfactor scores associated with a group-level aggregate matrix $\mathbf{M}'$ are equal to the sums of the Eigenfactor scores associated with the individual-level matrix $\mathbf{M}$.

Proposition 2 *Given the group-level Markov matrix* $\mathbf{M}'$*, the Eigenfactor score of each group is equal to the sum of the Eigenfactor scores of its member nodes as computed from the individual-level Markov matrix* $\mathbf{M}$*.*

Proof. Define the group-level Markov matrix $\mathbf{M}'$ as the aggregation to the group level of the individual level Markov matrix $\mathbf{M}$ with each node weighted according to its eigenvector centrality λ under the process $\mathbf{P}$:

$$M'_{k,m} = \sum_{i \in G_k} \sum_{j \in G_m} M_{i,j} \frac{\lambda_j}{\sum_{l \in G_m} \lambda_l}$$

Let w_i be the individual level Eigenfactor score associated with node i. We will show that the sum $\sum_{i \in G_k} w_i$ of the individual level Eigenfactor scores for

the nodes in a group G_k is equal to the group-level Eigenfactor score w'_k for that group, as calculated based on the group-level transition matrix.

Let the Eigenfactor normalizing constant $s = \sum_{i=1}^{n} \sum_{j=1}^{n} M_{ij}\lambda_j$. Now we note that the normalizing constant s' for the group-level Eigenfactor scores is equal to the normalizing constant s for the node-level Eigenfactor scores:

$$\begin{aligned}
s' &= \sum_{k=1}^{g}\sum_{m=1}^{g} M'_{km}\lambda'_m \\
&= \sum_{k=1}^{g}\sum_{m=1}^{g}\left(\sum_{i\in G_k}\sum_{j\in G_m} M_{i,j}\frac{\lambda_j}{\sum_{l\in G_m}\lambda_l}\right)\lambda'_m \\
&= \sum_{k=1}^{g}\sum_{i\in G_k}\sum_{m=1}^{g}\sum_{j\in G_m} M_{ij}\frac{\lambda_j}{\sum_{l\in G_m}\lambda_l}\sum_{l\in G_m}\lambda_l \\
&= \sum_{k=1}^{g}\sum_{i\in G_k}\sum_{m=1}^{g}\sum_{j\in G_m} M_{ij}\lambda_j \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij}\lambda_j = s
\end{aligned} \tag{7}$$

Thus we have

$$\begin{aligned}
\sum_{i\in G_k} w_k &= \sum_{i\in G_k}\sum_{j=1}^{n}\frac{100}{s}M_{i,j}\lambda_j \\
&= \frac{100}{s}\sum_{i\in G_k}\sum_{m=1}^{g}\sum_{j\in G_m} M_{ij}\lambda_j \\
&= \frac{100}{s}\sum_{m=1}^{g}\sum_{i\in G_k}\sum_{j\in G_m} M_{ij}\lambda_j \\
&= \frac{100}{s}\sum_{m=1}^{g}\sum_{i\in G_k}\sum_{j\in G_m} M_{i,j}\lambda_j\frac{\lambda'_m}{\sum_{l\in G_m}\lambda_l} \\
&= \frac{100}{s}\sum_{m=1}^{g} M'_{k,m}\lambda'_m \\
&= \frac{100}{s'}\sum_{m=1}^{g} M'_{k,m}\lambda'_m \\
&= w'_k
\end{aligned} \tag{8}$$

This final expression is simply the group-level Eigenfactor score based on the group-level Markov matrix $\mathbf{M}'$ and the group-level eigenvector centrality scores λ'. Equation 8 follows from Proposition 1. Thus we have demonstrated the proposition.