Measuring Direct and Indirect Authorial Influence in Historical Corpora

Moshe Koppel
Dept. of Computer Science
Bar-Ilan University
Ramat-Gan, Israel
moishk@gmail.com

Nadav Schweitzer
Dept. of Computer Science
Bar-Ilan University
Ramat-Gan, Israel
nadavsh1@gmail.com

Abstract

We show how automatically-extracted citations in historical corpora can be used to measure direct and indirect influence of authors on each other. These measures can in turn be used to determine an author's overall prominence in the corpus and to identify distinct schools of thought. We apply our methods to two major historical corpora. Using scholarly consensus as a gold standard, we demonstrate empirically the superiority of indirect influence over direct influence as a basis for various measures of authorial impact.

KEYWORDS citation index, clustering, information retrieval, ranking authors, reference identifier

1 Introduction

Many textual corpora include documents of historical importance that span centuries and even millennia. Often authors of documents in such a corpus refer to previous authors represented in the corpus. Familiar modern examples of such corpora include scholarly papers in some given discipline. Older examples include corpora of legal decisions in various cultures and languages. It is often of great historical importance to find measurable answers to the following questions:

1. How much (direct or indirect) influence did author X have on subsequent author Y?
2. What is the overall importance of author X in the corpus as reflected in influence on subsequent authors in the corpus?
3. Can we identify distinct schools of thought among authors in a corpus based on the flow of influence within the corpus?

In this paper, we will provide answers to all these questions and apply them to actual historical corpora.

This work differs in several ways from standard work in citation indexing (Garfield 1955, 1979) and link analysis (Brin & Page 1998, Kleinberg 1999). First, standard citation indexing work assumes that citations are marked in the text. In this paper, we make no such assumption; rather, we consider a typical historical corpus in which cross-references are not marked and hence need to be extracted from the texts. Second, our object is to measure the influence of specific *authors*, rather than specific *documents*. A document citation matrix is binary and sparse – that is, one document either cites another or not, usually not. But we are concerned here with the *number* of citations from the entire oeuvre of one author to another. As we will see, this permits a richer theory of influence.

While previous work in citation analysis and link analysis has focused on the overall prominence of a document, we measure here the influence of one author on a given other author. We distinguish between two types of influence that one author has on a subsequent author: direct influence (roughly the proportion of the later author's total citations that point to that particular earlier author) and indirect influence (taking into account citations of intermediate authors who cite the earlier author).[1]

We parlay this measure into a measure of the overall prominence of an author. In this context, the distinction between direct and indirect influence corresponds to that between standard impact factors based on citation counts (Garfield 2006) and eigenvector centrality methods, such as PageRank (Brin & Page 1998).

One advantage of applying such measures to historical corpora is that these corpora are the subject of a great deal of scholarship, so that experts can provide us with a consensus gold standard against which to evaluate our methods. We will show empirically on two historical corpora that eigenvector centrality methods yield better results than simple citation counts, thus lending more credence to the disputed claim

[1] We note that, like all other work in the field, we use the measurable notion of citation as a proxy for the looser notion of influence. Of course, it is an imperfect proxy: influence is not always expressed through citation and, furthermore, some citations might be negative or mere matters of courtesy.

that such sophisticated methods offer significant added value over simple methods (Davis 2008, West et al 2008).

In the following section, we briefly sketch related work. In Section 3, we describe our corpora and consider the challenges involved in identifying citations in free text. In Section 4, we show how to measure both the direct and indirect influence of one author on another. In Sections 5 and 6, we will use these measures to, respectively, quantify author prominence and identify distinct schools of thought. In each case, using scholarly consensus as a gold standard, we find that indirect influence is a better basis for measurement than is direct influence.

2 Previous Work

The bibliometric literature (Garfield 1955, 1979) offers a variety of measures of the overall influence of a given document in a corpus based on the number of citations of that document in the corpus. Formally, these measures are based on analysis of adjacency matrices that indicate whether there is a citation from document i to document j. Typically such matrices are binary, sparse and cycle-free (since documents are ordered chronologically). The analysis of these matrices has proved to be useful for navigation and search, as well as for identifying trends.

While most early work focused on measures of the prominence of individual documents in a document corpus, there has also been work on the prominence of items in other types of corpora. For example, there are numerous proposed measures of the influence of journals related to Garfield's (2006) impact factor and a variety of proposed measures of author prominence based on Hirsch's (2005) h-index. (See Bornmann et al. (2008) for a discussion and comparison of many of these variations.) Like the early work on document prominence, most of this work has used fairly similar measures of impact, based on normalized citation counts.

Work on link analysis, motivated by the rapid expansion of the Internet, has led to a wealth of more sophisticated measures of document prominence (Page & Brin 1998, Kleinberg 1999). As in the early

work in bibliometrics, link analysis takes as its raw data sparse, binary adjacency matrices, but unlike that earlier work, the matrices are not necessarily cycle free (since web sites are dynamic and hence can cite each other). The crucial feature of the newer measures is the use of eigenvector centrality methods (Friedkin 1991) to recursively account for the influence of the influenced pages themselves. Many variations on these measures have been proposed for the purpose of improving precision and efficiency.

Such eigenvector centrality measures have also been applied to journals, for example in the Eigenfactors measure (Bergstrom 2007), based on the number of citations from one journal to another. The measure of journal impact is most relevant to the problem considered here in the sense that the underlying citation matrices are neither binary nor sparse. There has been some controversy regarding the added value for measuring journal impact of eigenvector centrality methods over simple citation counts (Davis 2008, West et al. 2010); the debate focused on the extent of the correlation between the measures rather than on the accuracy of the resulting rankings.

We are not aware of any work using eigenvector centrality methods to measure the impact of an author's complete oeuvre in a historical corpus.

3 Historical corpora

In this paper, we consider two historical textual corpora. Our first corpus consists of 40,689 legal decisions ("responsa") written by 80 experts in Jewish law over the past 1000 years. Each decision is written by a single author. The documents range in length from 5 words to over 80,000 words, with an average of 1,178 words. Citations from one author to another in this corpus are not marked. Nevertheless, using methods described below, we can estimate that the number of such citations exceeds 100,000.

The second corpus consists of United States Supreme Court decisions. This corpus consists of 54,808 documents (each representing an "opinion" regarding some case before the Court) written by 108 Supreme Court Justices, spanning a period of about 200 years. The documents range in length from 3 words to over 50,000 words, with an average of 2,052 words. While all nine sitting Justices rule on a

given case, there may be multiple opinions regarding the same case, each written by some subset of the sitting Justices. Citations from one Justice to another are marked in this corpus and the total of such citations is 154,572.

Identifying a reference from one author to another actually entails two separate challenges. First, we need to identify the presence of a reference and second, we need to identify the target of the reference. The difficulty of each of these tasks depends on the corpus. To illustrate, let's consider each of our two corpora.

In the case of the Responsa corpus, we first manually prepare a list of all names (and book titles) of authors we wish to consider. We search for all appearances of these names and titles, only some of which are actual citations (since many are generic acronyms and phrases). For example, the Hebrew abbreviation רע"א (RAE) might be a reference to one of the authors in our corpus (R. Akiva Eiger) or a pointer to the top of the left side of a folio page in a printed work or a pointer to section 271 of some book. Of 1102 appearances of that particular abbreviation in our corpus, 582 are actual references to the author Eiger and 520 are not. Following Kerner et al. (2011), we manually tag a subset of 10,000 apparent citations as either actual citations or as non-citations: initially, all those apparent citations that are in one of a number of standard forms (e.g., "as X writes") are marked as actual citations and all others are marked as non-citations; a human annotator then corrects the mistakes. Machine-learning methods are then used to more precisely identify regular expressions collocated with names and titles that indicate that they are actually citations. This method is simple (requiring fewer than 5 man-hours for the manual corrections) and quite generic; with some adaptation, it can be applied to any historical corpus. On a manually tagged test set, we find that our method identifies genuine citations with recall of 89.2% and precision of 91.4%.

In the case of Supreme Court opinions, references to earlier cases follow standard formats and are thus easily identified. It is non-trivial, however, to identify the author of a referenced opinion since there are often multiple opinions for a single case. Thus, a single case might produce a majority opinion, multiple concurring opinions and multiple dissents, each of which might be authored by one justice and

joined by one or more additional justices. To identify the specific opinion being cited, we manually design rule-based parsing methods to identify, for each case, the authors of the majority opinion, concurring opinions and dissenting opinions. Furthermore, for each reference to an earlier case, we determine if the reference is to a named author, or to a majority, concurring or dissenting opinion (when not specified, we assume the reference is to the majority opinion). On a manually tagged test set, we find that our method identifies genuine citations with recall of 90.1% and precision of 93.6%.

4 Influence between authors

4.1 Direct influence

Once we have scanned our entire corpus and identified all references from one author to another, we can construct the direct influence matrix, M, representing the number of references from each author to each earlier author. We order the authors in the corpus chronologically and, for each $j<i$, we let m_{ij} be the raw number of references from author i to author j. Note that we assume throughout this paper that there is a natural chronological ordering of authors and that there are no cases of contemporary authors who each cite the other, so that m_{ij} is a strictly lower triangular matrix.[2] To prevent rows with no non-zero entries we add a small smoothing factor, c, to all values in the lower triangle. (Values not in the lower triangle are inherently 0, so we leave them unchanged.) A toy example of such a matrix is shown in Figure 1a. The fact that, unlike PageRank (Brin & Page 1998) and Eigenfactor (Bergstrom 2007), we work with lower triangular matrices, guarantees the existence of a unique eigenvector and allows for its efficient computation, as we shall see below.

Now we need to normalize the matrix to reflect the relative influence on a given author of each earlier author. The obvious method would be to normalize the rows to equal 1, that is, to divide each m_{ij} by $\sum_j m_{ij}$. However, since our matrices are lower triangular, this method suffers from both a technical

[2] Of course the veracity of this assumption varies from corpus to corpus. For example, in our Responsa corpus less than 0.4% of citations are from an "earlier" author to a "later" one. But in the Supreme Court corpus, 16% of all citations are above the diagonal. All these are simply ignored for our purposes.

deficiency and a substantive deficiency. The technical deficiency is an anomalous first row that sums to 0 and not 1 like other rows (not remedied by smoothing, which is applied only to the lower triangle). The substantive deficiency is that the simple normalization does not take into account that the number of potential referrers to earlier authors is greater than the number of potential referrers to later authors. Instead, for an n*n matrix, we normalize row i to add up to (i-1)/(n-1). Thus, the first row is not anomalous and the average value of the potentially non-zero values in each row is exactly 1/(n-1). The matrix shown in Figure 1a is thus normalized as shown in Figure 1b. This matrix captures the *direct* influence of author j on author i.

4.2 Indirect influence

Now we wish to measure the *indirect* influence of one author on another. Let V be the indirect influence matrix in which v_{ij} represent the indirect influence of author j on author i. Here is the key point:

The indirect influence of j on i should be proportional to the weighted average of the respective indirect influences of j on each author x, where the weight of x is determined according to the direct influence of x on i.

Formally:

For all $i \neq j$, $v_{ij} = d \sum_{1 \leq x \leq n} m_{ix} v_{xj}$ (for some positive constant d) (1)

For all j, $v_{jj} = 1 - [d * (j-1)/(n-1)]$ (2)

Note that the only non-zero terms in the sum are those for which $j \leq x < i$. Equation (1) guarantees that the i^{th} column of V is an eigenvector of the matrix $M^{(i)}$, where $M^{(i)}$ is identical to M except that $M^{(i)}_{ii} = 1/d$. Equation (2) is the only one that allows a solution for V such that the sum of the *non-diagonal* elements of row i is proportional to i-1 (as in M) and such that the sum of *all* the elements of row i is 1. Note that since M is a triangular matrix, for any value of d such that $0 < d \leq 1$, there is a unique solution V and we are able to recursively compute V efficiently, beginning with the first row.

Note that there is a single free parameter d in the above equations. This parameter determines the relative importance of the direct influence of j on i for computing the indirect influence of j on i. That is, when computing v_{ij} (the indirect influence of j on i), the coefficient of m_{ij} (the direct influence of j on i) in the product is v_{jj}. As d goes to 0, v_{jj} goes to 1 and thus the impact of m_{ij} on v_{ij} is very high. When d gets higher, v_{jj} gets smaller and so does the impact of m_{ij} on v_{ij}.

In Figure 1c, we show the indirect influence matrix V (with d=1) for the direct influence matrix shown in Figure 1b.

(a) (b) (c)

$$\begin{pmatrix} 0 & & & \\ 4.01 & 0 & & \\ 0.01 & 2.01 & 0 & \\ 7.01 & 11.01 & 6.01 & 0 \end{pmatrix} \begin{pmatrix} 0 & & & \\ 0.33 & 0 & & \\ 0.03 & 0.63 & 0 & \\ 0.29 & 0.46 & 0.25 & 0 \end{pmatrix} \begin{pmatrix} 1 & & & \\ 0.23 & 0.77 & & \\ 0.13 & 0.34 & 0.53 & \\ 0.30 & 0.31 & 0.09 & 0.3 \end{pmatrix}$$

Figure 1. (a) Citation matrix (smoothed) (b) Normalized direct influence matrix (c) Indirect influence matrix

We computed the indirect influence matrix for both our real-world corpora. Although space does not permit display of the full matrix, in Appendix 1, we show both direct and indirect influence for a small but representative set of Justices. As can be seen, the extent of indirect influence of one Justice on another sometimes varies greatly from that of that Justice's direct influence on the other. For example, Anthony Kennedy cites decisions by Robert H. Jackson more frequently than he cites decisions by Harlan F. Stone. Nevertheless, the indirect influence of Stone on Kennedy is greater than that of Jackson. This is because some of the strongest direct influences on Kennedy, such as Harry Blackmun and Byron White, cite Stone significantly more frequently than they cite Jackson.

We shall see in the next two sections that indirect influence is a more reliable than direct influence as a basis for measuring author prominence and for identifying distinct schools of thought.

5 An author's overall prominence

5.1 Measuring overall prominence

Given the direct and indirect influence matrices, we wish to capture an author's overall prominence in the corpus. A straightforward way to capture an author's overall influence is to simply consider the total number of citations to that author or, more precisely, to capture the overall influence of author j by summing the values of column j in the direct influence matrix M. We will show empirically, however, that a better measure of the prominence of author j is the sum of values of column j in the *indirect* influence matrix V.

Measuring prominence as an aggregation of indirect influence is actually a form of eigenvector centrality, not fundamentally different than that used on binary document adjacency matrices by PageRank (Brin & Page 1998) and on non-binary journal citation matrices by Eigenfactors (Bergstrom 2007). Our indirect influence measure captures such prominence for the case of author citation matrices, which are not binary but which are cycle-free.

5.2 Adapting PageRank

Our method is similar to that of PageRank. Since PageRank is designed to work with binary matrices, we need to make some minor adjustments, similar in spirit to those made by Eigenfactors. PageRank begins with an n*n binary matrix indicating whether there is a link from i to j; rows are normalized to sum to 1, yielding the matrix M.

The vector R is the dominant solution to the equation

$R = d(M * R) + [(1 - d)/\mathrm{n}]*\mathbf{1}$

where d is a scalar "damping" factor and **1** is the unit vector.

The fundamental conceptual difference between our method and PageRank is that our adjacency matrix is lower diagonal and hence normalizing rows to sum to 1 is not natural. With our method, the natural normalization yields a unique eigenvector solution. For comparison purposes, we can adapt

PageRank to the problem of author prominence (rather than document prominence) in two different ways. The first possible adaptation is to simply substitute our (non-binary) author-to-author adjacency matrix M for the (binary) document-to-document matrix normally used by PageRank. The second possible adaptation is available if, in addition to knowing the number of references from each author to each other author, we also know specifically which documents by each author point to which documents by each other author. In this case, we can use PageRank to compute the prominence of each document and then set an author's prominence to be the sum of the respective prominence values of that author's documents. (Such summing is consistent with the probabilistic interpretation of PageRank.) For convenience, we call this latter method, PageRank*.

5.3 Empirical tests

One advantage of measuring author prominence in historical corpora is that scholars have studied these corpora for years and hence can provide consensus judgments regarding author prominence that can be used as a gold standard against which various methods can be compared. We thus test our automated methods for assigning prominence by checking the extent to which the resulting ranking of authors correlates with rankings assigned by domain experts.

We design our experiments as follows. For each corpus, obtain expert rankings of authors to be used as ground truth. Then we rank authors in the corpus using our method based on direct influence and indirect influence, respectively. In addition, we rank authors by applying PageRank, respectively, to the author reference matrix and, when available, to the document adjacency matrix. For each method, our measure of success is the proportion of author pairs <x,y> for which the method orders <x,y> correctly. (If according to the gold standard, x and y were tied, the pair is not considered in the measure.) We measure the success of each method on the first k authors in the corpus, for each value of k from 1 until n (the number of authors in the corpus). Note that it is to be expected that as k increases, results would eventually degrade somewhat, since the evaluation of the prominence of later authors is based on a diminishing set of subsequent authors.

Note that the measures we use employ some free parameters: the smoothing constant c for direct and indirect influence, the parameter d for indirect influence and for PageRank. We used development sets for each of our corpora (disjoint from the test sets) to optimize parameter values. For both corpora, we found that all sufficiently small values of c yield identical results and that the optimal value for d in indirect influence is 0.7 and the optimal value of d for PageRank is 0.9 (although, Page & Brin suggest 0.85).

5.4 Results on the Supreme Court corpus

For the Supreme Court Justices corpus, our gold standard was taken from the ranking of Justices found in (Bader & Mersky 2004). Note that for this corpus, we actually know not only how often each author cites each other author, we also know which documents cite which documents. Thus, we can apply both adaptations of PageRank to this corpus.

In Figure 2, we show, for each of the four methods, the proportion of correctly ranked pairs over the k first authors in the corpus, for each value of k. For all values of k, indirect influence is a better basis for measuring prominence than direct influence. As can be seen, both versions of PageRank outperform total direct influence as a measure of prominence, but underperform total indirect influence.

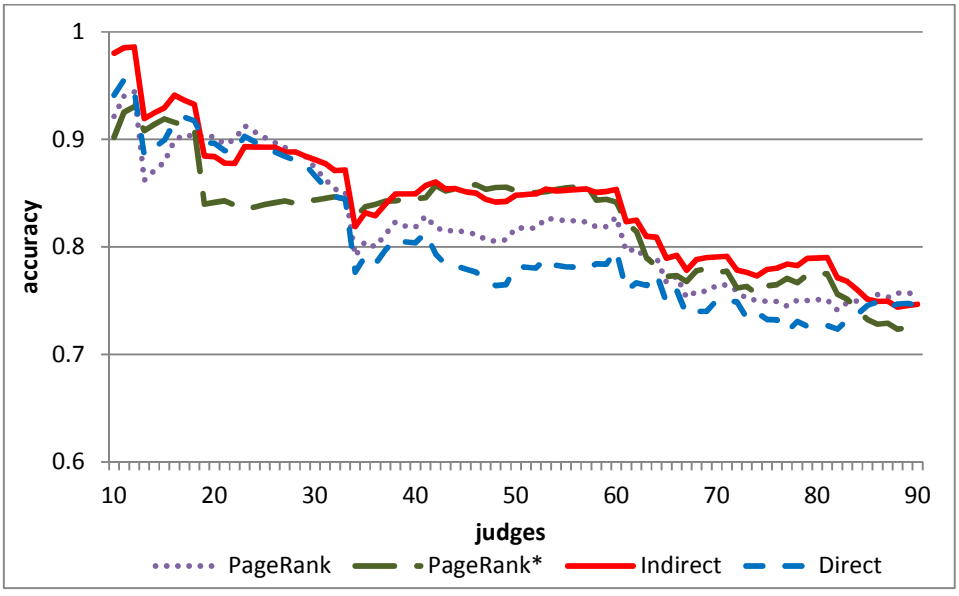

Figure 2. Proportion of correctly-ordered author pairs in the Justices corpus using total direct influence, total indirect influence, standard PageRank and PageRank*. The x-axis represents the number of authors being ranked.

5.5 Results on the Responsa Corpus

For the Responsa corpus, we asked three professional domain experts to manually rank all authors in the corpus in terms of overall prominence. (The rankings did not have to be strict: ties were allowed.) We use the aggregated results (average rank) as our gold standard. Note that for this corpus, we can estimate the number of times each author references each other author but we do not know which specific documents are referenced. Thus, we cannot apply PageRank to specific documents.

In Figure 3, we show, for each of the three applicable methods, the proportion of correctly ranked pairs over the k first authors in the corpus, for each value of k. As in the previous corpus, we find that for all values of k, indirect influence is a better basis for measuring prominence than direct influence. We also find that PageRank applied to authors yields essentially identical results to those of indirect influence. (Somewhat anomalously, we find results slightly improving for increasing k (in the region k>40), for all three methods.)

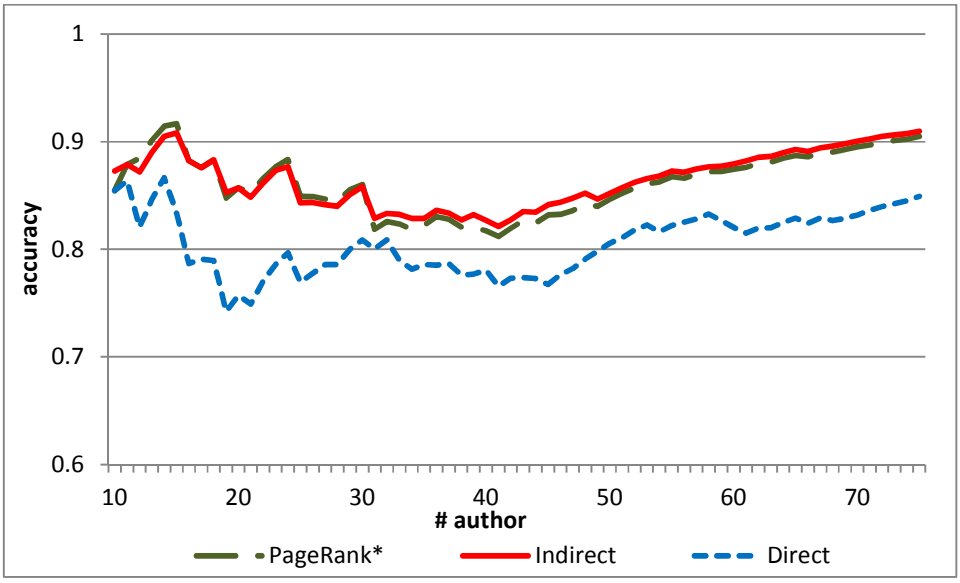

Figure 3. Proportion of correctly-ordered author pairs in the Responsa corpus using total direct influence, total indirect influence and PageRank*. The x-axis represents the number of authors being ranked.

These results illustrate the importance of measuring prominence using eigenvector centrality methods rather than simple citation counts, despite claims (Davis 2008) that the two methods yield highly correlated values.

6 Citation patterns and author clustering

6.1 Similarity Matrices

Another application of our influence matrices is the identification of distinct schools of thought based on citation patterns. As for our previous problem, there has been previous work (cf. Gibson et al. 1998, Flake et al. 2000) on this problem for the case of individual documents, the citation matrix of which is binary and sparse. We consider here the problem of identifying schools of thought among authors.

For sparse binary adjacency matrices, the natural way to identify clusters is apply a clustering algorithm directly to the binary adjacency matrix. The underlying rationale is that two authors probably ought to be in the same cluster if one (the later of the two) cites the other. As we shall soon see, this method does not work for non-sparse, non-binary influence matrices.

For non-sparse, non-binary influence matrices, we need an additional intermediate stage. Rather than identifying communities by applying a clustering algorithm directly to an adjacency matrix, we first measure influence across every pair of authors and then use that to measure the extent to which two authors were influenced (directly or indirectly) by the same previous authors. Clustering is performed on the resulting similarity matrix, as follows. Let M_j be the j^{th} row of the direct influence matrix M and, for any i<j, let $M_j^{(i)}$ be the j^{th} row of M truncated before the i^{th} element. That is, $M_j^{(i)}$ is a vector indicating the influences on author j of each of the authors in the corpus prior to author i. We define the similarity of two authors, i and j (i<j), as cosine($M_i^{(i)}, M_j^{(i)}$). That is, two authors are similar to the extent that they display similar citation patterns with regard to authors who are prior to both of them. Alternatively, we can use the identical method applied to the indirect influence matrix V rather than to the direct influence matrix M. Given the similarities of each pair of authors, we apply the n-cut clustering method (Dhillon et

al. 2007) to the similarity matrix to identify distinct schools of thought, the intuition being that authors in the same school cite the same earlier authors.

6.2 Experiments

For the case of the Responsa corpus, there is a scholarly consensus (e.g., Ta-Shma 2006) regarding the existence of two distinct authorial communities, namely, those authors residing in predominantly Christian countries (Ashkenazim) and those residing in predominantly Muslim countries (Sephardim). Taking this division as a gold standard, we used four methods to cluster the authors in the corpus into two communities:

1. A baseline method in which we apply n-cut clustering directly on the original citation matrix M (i.e., the "similarity" of two authors is simply the extent to which the later one is directly influenced by the earlier one)
2. Another baseline method identical to the above except that indirect influence matrix V is used.
3. Our method of using n-cut clustering on the similarity matrices, where two authors are similar to the extent that they are directly influenced by the same prior authors.
4. The same method but using indirect influence of prior authors.

In Figure 3, we show the purity of the clustering for each method. As can be seen, the baseline methods do not correspond to the consensus gold standard at all. The similarity-based methods are considerably stronger. Most significantly, we again note the superiority of indirect influences over direct influences. Basing similarity on direct influences yielded purity of 57.1%, while using indirect influences yielded purity of 91.0%. The identification of several specific exceptions appears to be of considerable academic interest to scholars in the field.

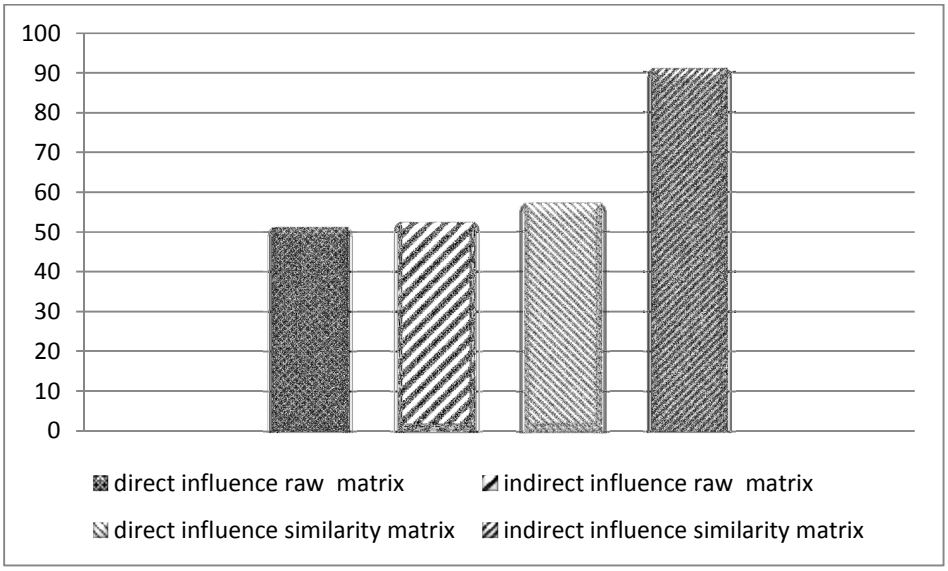

Figure 4. Purity of clusters for each of four methods of clustering authors in the Responsa corpus, using the Ashekenazi-Sephardi split as a gold standard.

With regard to the Supreme Court corpus, there is no clear consensus on how Justices should be divided. While the division into Democratic or Republican appointees seems a promising direction, we find that none of the proposed methods yields a clustering that corresponds closely to that division.

7 Conclusions

We have shown on two historical corpora that it is possible to automatically identify citations in free text and to use these citations to measure both direct and indirect influence of one author on another. These measures can in turn be used to measure the overall prominence of an author and to identify distinct schools of thought. Experiments show that our automated measures correlate well with expert-provided gold standards and that for these purposes our measure of indirect influence is a more effective foundation than is a simplistic measure of direct influence.

For this approach to be applied to historical corpora generally, a number of issues need to be resolved. First, corpora differ in the extent to which citations can be easily extracted. Although the methods we used for our test-bed corpora can be generalized, a number of aspects are corpus-specific and adaptations

will be necessary for use with different types of corpora. Second, our method assumes that a given pair of authors do not each cite the other. This is generally true for historical corpora and clearly false for contemporary corpora; for various corpora of an intermediate nature – such as the Supreme Court corpus considered here – our method involves ignoring citations from "earlier" authors to "later" ones. The cost of doing so remains unclear. Finally, testing on a wider variety of historical corpora is required to confirm the generality of our findings regarding the efficacy of our definition of indirect influence as a foundation for measuring author prominence and for identifying distinct schools of thought.

References

Bader, W. D. & Mersky, R. M. (2004). *The First One Hundred Eight Justices*, Hein Publishers: Buffalo, NY.

Bornmann, L., Mutz, R., & Daniel, H.-D. (2008). Are there better indices for evaluation purposes than the h index? A comparison of nine different variants of the h index using data from biomedicine. *Journal of the American Society for Information Science and Technology*, 59(5), pp. 830-837.

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

Brin, S. & Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, 30(1-7), pp. 107-117

Davis, P. M. (2008). "Eigenfactor: Does the principle of repeated improvement result in better estimates than raw citation counts?". *Journal of the American Society for Information Science and Technology* **59** (13): 2186–2188

Dhillon, I., Guan, Y. & Kulis B. (2007). Weighted Graph Cuts without Eigenvectors: A Multilevel Approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (PAMI), vol. 29(11):1944-1957.

Flake, G. W., Lawrence, S. & Giles, C. L. (2000). Efficient identification of Web communities. *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 150–160). Boston, MA: ACM Press.

Friedkin, N. E. (1991). Theoretical Foundations for Centrality Measure. American Journal of Sociology 96(6), pp. 1478-1504

Garfield, E. (1955). Citation indexes for science. A new dimension in documentation through association of ideas, *Science*, Vol: 122, No: 3159, pp. 108-111

Garfield, E. (1979). Citation Indexing: Its Theory and Application in Science, Technology, and Humanities. *Wiley*, New York.

Garfield, E. (2006). "The history and meaning of the journal impact factor". *JAMA* **295** (1): 90–3.

Gibson, D., Kleinberg, J. M. & Raghavan, P. (1998). Inferring Web communities from link topology. *Proceedings of the 9th ACM Conference on Hypertext and Hypermedia*, 20-24 June, Pittsburgh, PA, 225-234

Giles, C. L., Bollacker, K. & Lawrence, S. (1998). CiteSeer: An Automatic Citation Indexing System, Digital Libraries 98: *Third ACM Conf. on Digital Libraries*, ACM Press, New York, 1998, pp. 89-98.

Hirsch, J. E., 2005. "An index to quantify an individual's scientific research output". *PNAS* **102** (46): 16569–16572.

HaCohen-Kerner, Y., Schweitzer, N. & Mughaz, D. (2011). Automatically Identifying Citations in Hebrew-Aramaic Documents. *Cybernetics and Systems* 42(3): pp. 180-197

Kleinberg, J. (1999). "Authoritative sources in a hyperlinked environment", *Journal of the ACM* **46** (5): 604–632.

Ta-Shma, I. M. (2006). Creativity and Tradition: Studies in Medieval Rabbinic Scholarship, Literature And Thought. *Harvard: Cambridge, MA*.

West, J.D., Bergstrom, T.C. & Bergstrom C.T. (2010). Big Macs and Eigenfactor Scores: Don't Let Correlation Coefficients Fool You, *Journal of the American Society for Information Science & Technology*. 61(9): 1800-1807

Appendix 1:

11x8 matrices for the Supreme Court corpus showing direct and indirect influence, respectively, for a representative set of Justices (the full 108x108 matrix is too large to be shown).

	Owen Roberts	Harlan F. Stone	Robert H. Jackson	Sherman Minton	Thurgood Marshall	Byron White	Harry Blackmun	William Rehnquist
Lewis F. Powell	52	108	79	11	0	0	0	0
William J. Brennan	163	253	228	48	0	0	0	0
Thurgood Marshall	81	143	119	28	0	0	0	0
Byron White	108	184	157	34	284	0	0	0
Harry Blackmun	79	131	103	12	302	437	0	0
William Rehnquist	88	142	170	31	340	537	308	0
John Paul Stevens	82	155	159	19	414	606	424	538
Sandra Day O'Connor	44	67	85	10	268	358	277	359
David Hackett Souter	31	37	42	5	125	155	121	163
Antonin Scalia	29	52	88	10	203	284	220	284
Anthony Kennedy	25	37	47	3	160	201	151	204

Direct influence matrix (unsmoothed non-normalized raw values shown for clarity)

	Owen Roberts	Harlan F. Stone	Robert H. Jackson	Sherman Minton	Thurgood Marshall	Byron White	Harry Blackmun	William Rehnquist
Lewis F. Powell	0.011362	0.018121	0.011379	0.002208	0	0	0	0
William J. Brennan	0.012711	0.017802	0.012451	0.002849	0	0	0	0
Thurgood Marshall	0.011597	0.017125	0.011481	0.002819	0.377778	0	0	0
Byron White	0.01153	0.016852	0.011446	0.002679	0.009329	0.371296	0	0
Harry Blackmun	0.011333	0.016408	0.010764	0.001987	0.012964	0.017801	0.364815	0
William Rehnquist	0.01078	0.015419	0.012571	0.002625	0.012251	0.018261	0.010016	0.358333
John Paul Stevens	0.010207	0.015417	0.011516	0.002052	0.013456	0.018729	0.012369	0.014897
Sandra Day O'Connor	0.010083	0.01419	0.011566	0.002064	0.015827	0.020358	0.014708	0.018068
David Hackett Souter	0.011	0.014177	0.011338	0.001995	0.015602	0.018881	0.01373	0.017456
Antonin Scalia	0.009076	0.013352	0.012498	0.002071	0.014796	0.019756	0.014282	0.01749
Anthony Kennedy	0.009933	0.013889	0.011432	0.001741	0.016849	0.020643	0.014549	0.018523

Indirect influence matrix