Available online at www.sciencedirect.com

SCIENCE DIRECT®

Information Processing and Management 42 (2006) 1408–1421

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Measures of international collaboration in scientific literature: Part I

Abraham Bookstein [a,*], Henk Moed [b], Moshe Yitzahki [c]

[a] *University of Chicago, 1010 E. 59 St., Chicago, IL 60637, United States*
[b] *Centre for Science and Technology Studies, Leiden University, The Netherlands*
[c] *Bar-Ilan University, Israel*

Received 16 March 2006; accepted 16 March 2006

Abstract

Research evaluating models of scientific productivity require coherent metrics that quantify various key relations among papers as revealed by patterns of citation. This paper focuses on the various conceptual problems inherent in measuring the degree to which papers tend to cite other papers written by authors of the same nationality. We suggest that measures can be given a degree of assurance of coherence by being based on mathematical models describing the citation process. A number of such models are developed.

1. Introduction

A goal of Scientometric research is to develop and evaluate causal models of scientific productivity. Much of this effort involves a combination of carefully constructing datasets that are sensitive to the objectives of the research and incisive use of statistical techniques. But ultimately, for this effort to succeed, it is necessary that we develop measures that accurately capture the content we are exploring.

An example we will now analyze further is the measure defined by Bookstein and Yitzahki (1999, 1999a) that was designed to indicate the degree to which various language groups tend to rely on papers written in the same language, as reflected in the literature they cite. The underlying act being analyzed here is simple: a member of a particular language community made a decision to cite a paper from the same, or different, language community. We wanted to measure the degree of bias toward one's own language in a way that is reasonably free of influence from irrelevant, but possibly confounding, variables.

But focusing on language communities is only one of the possibilities to which measures such as that developed by Bookstein and Yitzahki (1999, 1999a) can be applied. An interesting similar application is to explore the degree to which authors' *nationalities* influence the papers they cite: in parallel with the tendency of authors to cite papers written in the their own language, we now wish to quantify the degree to which authors tend to

[*] Corresponding author.

0306-4573/$ - see front matter © 2006 Published by Elsevier Ltd.
doi:10.1016/j.ipm.2006.03.007

cite papers of the same nationality. More generally, we are interested in measuring the strength of linkages between pairs of nationalities, of which the linkage to one's own nationality, as referred to above, is a special case. These issues have gotten a great deal of attention. A discussion of the substantive issues involved, as well as a survey of the pertinent literature, appears as Chapter 24 (Do US scientists overcite papers from their own country, pp. 291–300) in Moed (2005).

This problem, in addition to its substantive interest, raises some issues of methodological interest. One complicating difference between these two superficially identical problems is that, whereas a scientific paper is almost certainly written in one language, it may have authors from multiple countries. Thus the task of assigning a paper to a nationality is not quite identical to that of assigning a paper to a language group: while one can unambiguously register that a given English-language paper cites a paper also written in English, if a source paper with authors from multiple nations cites another such paper, the problem of how to register this transaction must somehow be resolved. When constructing a measure of the strength with which one nation prefers to cite papers of a second nation, we must take care that the measure not be distorted by this complexity.

It is convenient to divide the measures we are proposing into two classes: In this paper we develop the first of these classes, which is based on models most directly extending the approach taken by Bookstein and Yitzahki (1999, 1999a): these concentrate on situations in which problems of multiple labelling, while recognized, can be overlooked; we shall refer to these as "simple-choice" measures. The second class treats the contrasting situation, in which we explicitly take into account collaboration choices that somehow involve a compromise between authors from different categories (for example, different nations); we shall refer to these as "cooperative-choice" measures, and examine them in a follow-up paper (Bookstein, Moed, & Yitzahki, 2006).

The strategy we adopt below for defining measures will be to construct highly simplified models describing how the choice of papers to cite is made, and to define our measures in terms of the parameters defined by the model. We introduce explicit models to guide us in isolating the quantities of interest for our measure and to separate them from other, confounding quantities. We are currently testing the various measures proposed here – experimental results will be reported in subsequent papers.

2. Structure of problem

To set the stage, we first discuss the nature of the data with which we are working. We will be referring below to two conceptually distinct types of sets:

1. The first will be a universe of source documents (or, for that matter, any set of objects making choices), each of which selects, or cites, a subset of items from a second universe, the universe of potentially citable items. Typically, one or more labels will be assigned to each source item, and we construct sets of source items by means of these labels (for example, a source document might be labelled by the nationality, or nationalities, of its author(s), and a set defined as all papers in which at least one of the labels is "United States").
2. The second set to which our measures refer is the total universe of items from which our citations are chosen; we refer to this universe, and various subsets taken from it, as *target* documents. Typically, the universe of target documents will be labelled in the same manner as those in the source universe, and sets of target documents are similarly defined. The universes of source and target documents are conceptually different, and a set of source items may differ from the identically labelled set of target documents.

Our goal is, given a set of source documents, to assess the influence, as revealed by its citations, of a specific set of target documents. By specifying the problem abstractly, as quantifying the strength with which two sets are linked, we allow the researcher a maximum degree of flexibility. Conceptually, our measures are not concerned with how the sets are constructed; in practice, neither the source document sets nor the target sets will be chosen arbitrarily, but rather will reflect the content being researched. Consider, for example, the construction of a national-preference measure. Depending on what the researcher is intending to illuminate, he may define a target (or source) set as consisting of documents *all* of whose authors are from a specific country; or else, the set may consist of documents with at least one author from said country. The method is equally valid (if, perhaps, not equally wise) for either decision.

Categories other than nationality can also be used to label items. Universities, disciplines, individual authors, etc., may be used for other applications – in that sense, the reader may interpret the term "nation" as a metaphor for the classes being studied. Similarly, though we may refer, for simplicity, to individual "nations" (e.g., Germany, France, US, etc.), we can as well label papers according to aggregates of nations (Europe, all nations but that of the source paper, etc.). Also, the sets of source documents being studied need not be defined using the same categories as for the target sets: we might be interested in the nationalities of the target documents, but the specific responsible institutions of the source documents.

To develop a measure, we begin with a single document taken from a source set. To guide us in quantifying the influence of various sets of target items, we create a very simple model describing how that source document chooses the papers that it cites from the target universe. We then define a value for the influence that the target set exerts on the source document in terms of the model parameters. The measure we report will then be some average of these values over all documents in the source set. We will not discuss in detail the nature of the averaging process, but we explore some statistical concerns in Appendices B and C.

By depending on an explicit model, this approach ensures a consistency and coherence among our measures, and helps us tease out those aspects of the process that are important from those that are irrelevant. The defining characteristic of a simple-choice model is that a value is defined in terms of a source set and a target set, without further consideration of any underlying structure.

3. Preliminary model

Our measures will be defined in terms of families of parameters, and we shall describe the notation as needed. (To assist the reader, we summarize the notation for these parameters in Appendix A.) But we are trying to maintain a consistent structure for our notation, and it may assist the reader if we first give an overview of this structure.

The notation will generally have to identify source and target components. We will try to use the same basic symbol for members of the same family, relying on different organizations of subscripts to distinguish specific members of the family. We use the colon (':') to separate the index value defining the source document from the index values denoting the target sets; this reflects the very different roles they play in the analysis. The source papers used to construct our database of citations will appear to the left of the colon; the indices to the right of the colon refer to the classes to which our citations are assigned. Occasionally two target sets are required; in such instances, we will separate the values by a pipe symbol ('|') – as special cases are considered, we simplify the notation by reducing the number of subscript values made explicit.

Following these guidelines, we can now describe our basic model. Suppose we imagine a paper in set s being generated. Further, suppose there are N candidates available for citation, of which the fraction α_{t} belong to the target class t, and that $P_{s:t}$ denotes the probability[1] that the source paper cite any specific paper in the target set. Then we can estimate the number of papers in t that it cites, call it $n_{s:t}$, by

$$n_{s:t} = (\alpha_t N)P_{s:t}, \tag{1}$$

since $\alpha_t N$ is the total number of papers available for citation that are in t, and $P_{s:t}$ is the fraction of these actually cited in the paper at hand. In terms of this model, a very tempting choice for a measure of the preference of a paper in source set s for a paper in target set t is $P_{s:t}$. Unfortunately, $P_{s:t}$ is a latent, non-measurable parameter. But its consequences can be measured. Thus, we estimate:

$$P_{s:t} = \frac{(n_{s:t}/N)}{\alpha_t}. \tag{2}$$

[1] Actually, a conceptual probability should be defined relating each source paper and each target paper. The probability indicated for the model is then some average, taken over the target and source sets, of the specific probabilities governing each choice. This is discussed in more detail in Bookstein and Yitzahki (1999, 1999a). Here we simply use this representative probability value as if constant for all cases encompassed by its index values.

We first note that in Eq. (1), $n_{s:t}$ is an expected value,[2] and the value it takes for a specific paper will fluctuate around this value. Thus, the computed value of $P_{s:t}$ will be an approximation. But this estimate for $P_{s:t}$ was derived on the basis of only a single paper in s; the reported value would be some average of these estimates taken over the set of all papers in s.

More seriously, we expect the estimated value of $P_{s:t}$ to have limited practical use. A problem is that $P_{s:t}$ will take a very small value, and in itself be difficult to interpret. This suggests it may be desirable to use as a measure of influence not $P_{s:t}$ itself, but a comparison of $P_{s:t}$ with a contrast set. This is the strategy we pursue in our fundamental model, and its variants. Below, we first define the model in very general terms; we then restrict the model to specific cases that may be of more practical interest.

4. Fundamental model

Taking into consideration the discussion of the preliminary model, we now measure the *relative* influence of target sets t and t' on a paper in the source set s. The most direct way to assess this is to define the measure, $\mu_{s:t|t'}$, simply as the ratio of the probabilities $P_{s:t}$ and $P_{s:t'}$, each as estimated over the set of source documents in s by Eq. (2):

$$\mu_{s:t|t'} = \frac{P_{s:t}}{P_{s:t'}} = \frac{n_{s:t/n_{s:t'}}}{\alpha_t/\alpha_{t'}}. \tag{3}$$

The measure $\mu_{s:t|t'}$ depends on the ratio of $n_{s:t}$ and $n_{s:t'}$. We expect each such n-value used to compute this measure will be the average over s of the respective n-values of the individual papers in s. Another possibility would be to compute for each paper in s the ratio of the n-values, and take the average of these over s. The distinction between using the ratio of averages and the average of ratios is discussed in Appendix B, and the possibility of using global statistics of cited items explored in Appendix C.

4.1. Alternative forms

Eq. (3) is expressed in terms of raw counts of citations. An equivalent form in terms of their corresponding fractions of citations follows trivially: Let n denote the total number of items cited in a paper. Then, since $\frac{n_{s:t}}{n_{s:t'}} = \frac{n_{s:t/n}}{n_{s:t/n}} = f_{s:t}/f_{s:t'}$, where the fs are the fractions of citations in a source paper that belong to their respective target classes. Thus we also have,

$$\mu_{s:t|t'} = \frac{f_{s:t/f_{s:t'}}}{\alpha_t/\alpha_{t'}}, \tag{3a}$$

a form that may be more convenient if the variability over s in the number of items cited is small, and can be ignored.[3]

At this point it may be appropriate to discuss a fine point regarding our data, alluded to in Section 1. We measured the preference of a source document for members of a primary target set by comparing this preference with that for a contrasting target set. It is important to realize, as we defined the primary and contrast target sets, that they may, in general, overlap. The simplest case will be when the overlap is null. For example, we may have restricted our target universe to papers, *all* of whose authors are from the same nation; any two such sets are necessarily disjoint. Or, broadening our target universe to include papers with multi-national authors, we may label each cited paper by the nationality of its senior author. For these cases, we may construct target sets based on the unique nationality label of each paper, and any pair of sets having different labels in this classification is disjoint.

[2] More precisely, the number of items of a given category cited by a randomly chosen paper is a random variable, and the n in Eq. (1) is its expected value. Since the purpose of the model is to guide us in developing a reasonable measure rather than to begin a detailed probabilistic analysis of the citation process, we would not complicate our discussion by lingering on the distinction between random variables and estimates of their expectations.

[3] n varying substantially corresponds, in terms of our model, to $P_{s:t}$ varying significantly over the papers in s. In this case, using f-values defines a slightly different measure, in which values of μ are governed by the ratios of normalized P-values – that is, ratios of probabilities per item cited.

But more generally, it may be convenient, or demanded by substantive considerations, that the sets to which we assign items overlap. For example, if our universe is made up of all citable papers, a citable paper with authors representing two nations might only with difficulty be assigned to a single nation, and we may resolve the ambiguity by placing it in both sets. Thus, if our target is all papers with *at least* one Dutch author, then we can certainly construct a disjoint contrast set made up of papers with *no* Dutch author. On the other hand, the contrast set of all papers with, say, at least one German author need not be a disjoint contrast set, since it might include some papers with a Dutch author as well, and thus overlap the target set.

We want to emphasize that nowhere in the above derivation do we require that the targets sets being compared be disjoint. But if they do overlap, then we must be aware that the sum of all the α_t's, that is, $\sum_t \alpha_t$, need not equal one; and, similarly, the sum over t of $n_{s:t}$ may exceed the total number of items cited. For simplicity, we may prefer that our contrast set be disjoint from the primary target set, but this requirement is not strictly necessary.

Perhaps a numerical example will further illuminate the nature of the problem. Consider a target universe of three papers, R, S, and T. Suppose R is authored by two British scholars; S by one British and one American scholar; and T by two American scholars. Then we can create nationality sets in two ways:

1. We can define two classes for the purpose of evaluating a value of the μ-measure. One class includes all papers with *at least* one British author (call it class B), and the other class as all papers with *no* British author (denoted, in keeping with the above notation, as $\bar{B}$). With this classification, $B = \{R, S\}$, and $\overline{B} = \{T\}$; further, $\alpha_B = 2/3, \alpha_{\bar{B}} = 1/3$; and, of course $\alpha_B + \alpha_{\bar{B}} = 1$.
2. On the other hand, we might have, plausibly, defined a class as consisting of all papers with an author from a given country. For the example, one class, say B, would include all papers with a British author, and another, say A, would include all papers with an American author. Then, $B = \{R, S\}$, and $A = \{S, T\}$, with S a member of both classes. Now, $\alpha_B = 2/3$ and also $\alpha_A = 2/3$; thus $\alpha_B + \alpha_A = 4/3$.

These examples also make clear how the probability of a document's being cited influences the P values of the multiple sets which claim it as a member (e.g., document S in case 2 above). The group value $P_{s:t}$ is in effect an average taken over the probabilities of the specific papers in t, with $P_{s:t'}$ the corresponding average over the set t'. Thus, a paper in both sets contributes to the value of both P's, since it does influence the likelihood of each set's being cited. This issue is given more attention in Bookstein and Yitzahki (1999, 1999a).

If our target sets are disjoint, a third form of Eq. (3) will occasionally be useful. If we are contrasting target sets t and t', we may want to use in our measure the relative proportions of citations to t and t', where we restrict our citations to only those two sets. That is, we may want to substitute in Eq. (3),

$$\frac{n_{s:t}}{n_{s:t'}} = \frac{n_{s:t}/(n_{s:t+n_{s:t'}})}{n_{s:t'}/(n_{s:t}+n_{s:t'})} = \frac{f_{s:t|t'}}{f_{s:t'|t}},$$

to get,

$$\mu_{s:t|t'} = \frac{f_{s:t|t'}/f_{s:t'|t}}{\alpha_t/\alpha_{t'}}, \tag{3b}$$

here, $f_{s:t|t}$ and $f_{s:t|t'}$ are defined in the obvious way by the first equation, with the pipe ('|') separating the primary target set from the contrast set.

But note that this equivalence, unlike the previous ones, does depend on the assumption that the sum $n_{s:t} + n_{s:t'}$ is indeed the total number of citations to one or the other class, an assumption that will be violated should the classes overlap (assuming the most straightforward counting rule).

This point may be clearer in an alternative derivation, based on first principles, which is in itself instructive. Here we again will assume non-overlapping classes. We again imagine a paper in s being generated. First we ask, if we restrict this paper's citations to only those items associated with t and t', what fraction of these belong to t? This is given (estimating with expected values) by

$$f_{s:t|t'} = \alpha_t P_{s:t} N/(\alpha_t P_{s:t} N + \alpha_{t'} P_{s:t'} N) = \frac{1}{1 + \frac{\alpha_{t'} P_{s:t'}}{\alpha_t P_{s:t}}}. \tag{4}$$

We can now use simple algebra to solve for the measure $\mu_{s:t|t'}$, which we defined as $P_{s:t}/P_{s:t'}$, to reproduce Eq. (3b). But note that, in the general case allowing overlapping sets, $\alpha_t P_{s:t} N + \alpha_{t'} P_{s:t'} N$ is *not* equal to the number of citations in the union of t and t' labelled sets, since some may appear twice; if such is the case, $f_{s:t|t'}$ would not be a genuine fraction. If the sets are constrained to be disjoint, this is not a problem.

We have noted several forms are available, depending on how the citations to a target set are quantified. We note in passing that the α parameters are also subject to rewriting, if the sets are disjoint:

$$\alpha_t/\alpha_{t'} = \frac{\frac{\alpha_t}{\alpha_t+\alpha_{t'}}}{\frac{\alpha_{t'}}{\alpha_t+\alpha_{t'}}}.$$

Thus we can use either the raw α values or the proportional values, given the restriction to t or t', when evaluating μ-values.

5. Special cases

The above, very general, measures of relative impact may well be required for many applications. Its appeal is that our groupings can be quite general, illuminating different aspects of the complex of international collaboration and dependencies. t and t' can easily involve either individual countries or sets of countries (e.g., to measure the relative impact of Europe (t) vs US (t') on Japan (s)). Or we can ask, whether a source European country (s), when not citing itself, more likely to cite another European country (t), or, say, the US (t')? We might want to understand this before we decide to treat European countries as a unified group in any subsequent analyses.

But for many applications, the above measure is likely to be too general, and require that we restrict the class values appearing in the μ parameter. These are important enough to deserve individual consideration.

5.1. Direct influence measure

An obvious special case is to assess how strongly a target set influences papers in a source set, relative to all *other* target papers. Such a measure may allow us to construct visual influence graphs showing linkages within a community of nations. This problem introduces nothing new, and reduces to the problem previously considered. Instead of developing measures that contrast s' specific preference for t over some specific target set t' (as done for $\mu_{s:t|t'}$), we contrast s' preference for papers in t over those not in t – that is, for papers in $\bar{t}$, using the bar to denote set complements. Since t and $\bar{t}$ are disjoint, all the forms considered above can be used. To be specific, we can define (simplifying the notation since the terminal t' is now implied):

$$\mu_{s:t} = \frac{P_{s:t}}{1-P_{s:t}} = \frac{f_{s:t}/(1-f_{s:t})}{\alpha_t/(1-\alpha_t)}. \tag{5}$$

Of course, this result follows as well from first principles: beginning as usual with a set of source nations, s, and focusing on the target nation t, we assess the influence of t on s by

$$f_{s:t} = \alpha_t N P_{s:t}/(\alpha_t N P_{s:t} + \alpha_{\bar{t}} N P_{s:\bar{t}}), \tag{6}$$

easily yielding Eq. (5). The measure $\mu_{s:t}$ is asymmetrical in s and t, so it might be interesting to compare $\mu_{s:t}$ with $\mu_{t:s}$ to see if s' choice of t is requited.

5.2. Own nation bias

Bookstein and Yitzahki (1999, 1999a) examine the bias towards one's own language. A parallel bias would be the bias towards one's own nation. More generally, we might be given an arbitrary source set of interest, and hope to measure the strength of that source set's bias in favor of similarly categorized papers in the target set.

This measure results from a further specialization of $\mu_{s:t}$ to $\mu_{s:s}$. With simpler notation:

$$\mu_s = \mu_{s:s} = \frac{f_s/(1-f_s)}{\alpha_s/(1-\alpha_s)}, \tag{7}$$

where we use the simpler $f_s = f_{s:s}$. This result agrees with the measure defined by Bookstein and Yitzahki (1999, 1999a).

5.3. Within-group impact measures

Up to now, we have been concentrating on "binary" measures, which quantified the strength with which a specific source set was attracted to a specific target set. But an important application of bibliometrics involves determining the impact of a member of a set on the set as a whole – that is, the degree to which the totality of source documents tends to prefer a given target. Since the preceding arguments placed no restrictions on the source set, we may as well allow it to be the full set of source documents, and apply the approach taken above directly. For the most part, this involves little more than changing the notation to emphasize the special character of this problem.

So we now turn to the task, given a collection of nations and the totality of source documents associated with them, of assessing the impact of each member nation on the full community. The source set is now fixed; we shall refer to it as the superset, and give it the index value '0', placed as a superscript, to emphasize the special role it plays in this analysis. We are trying to assess the impact of a target set, t, of citable items, where the items in t are associated with a nation (or nations) in the superset. Our argument will not require that a citable item be assigned to only a single nation. A special case has the superset comprising the entire universe of items, but this restriction is not necessary; for example, we might ask the impact of each European nation on the research literature of Europe.

We distinguish two types of measures.

5.3.1. Global contrast set

We proceed using notation similar to that already used above:

- We continue to denote the fraction of citable items including an author from a specific member nation, indexed by t, by α_t. But the fraction of citable items including any author from the superset being analyzed, indexed by 0, is now denoted by $\alpha^{(0)}$. We use a superscript to emphasize the special role played by the superset in this analysis. If the superset is the entire universe of items, of course $\alpha^{(0)} = 1$.
- We let the parameters $P_t^{(0)}$ denote the intrinsic probability that an author in the superset 0 cite a paper written by an author in member nation t of the superset, with $P_0^{(0)}$ denoting the probability of citing any paper in the superset. If the superset is the whole universe, then trivially, $P_0^{(0)} = 1$. (Previously, we denoted $P_t^{(0)}$ by $P_{0:t}$.)
- Let $f_t^{(0)}(f_0^{(0)})$ denote the average fraction of citations in a paper by an author in the superset 0 that cites a paper belonging in set t (or superset 0). Again, if the collectivity is the full universe, $f_0^{(0)} = 1$. In terms of the earlier notation, $f_t^{(0)} = f_{0:t}$.
- We define $n_t^{(0)}$ and $n_0^{(0)}$ in accordance with the above pattern.

Then, proceeding as before, given a paper in the superset, we estimate:

$$n_t^{(0)} = \alpha_t N P_t^{(0)},$$

and,

$$n_0^{(0)} = \alpha^{(0)} N P_0^{(0)}.$$

We then define our measure of within-group impact of country t, by $\mu_t^{(0)} = \mu_{0:t|0}$; explicitly, we estimate this value for a specific paper by

$$\mu_t^{(0)} = \frac{P_t^{(0)}}{P_0^{(0)}} = \frac{n_t^{(0)}/n_0^{(0)}}{\alpha_t/\alpha^{(0)}}, \tag{8}$$

and report the average of this value over all papers in the superset. For the important special case in which class 0 is the entire universe, then $\alpha_0 = 1$ and $n_0^{(0)} = n$, the total number of citations in the paper, and this reduces to:

$$\mu_t^{(0)} = f_t^{(0)}/\alpha_t. \tag{8a}$$

More generally, we can estimate $\mu_t^{(0)}$ in terms of fractions of citations of a category:

$$\mu_t^{(0)} = \frac{f_t^{(0)}/f_0^{(0)}}{\alpha_t/\alpha^{(0)}}.$$

This measure suggests the so called normalized or relative impact measure, used by Bibliometricians to measure a nation's impact on the world's research. But bibliometricians normally calculate the ratio of the percentage of references (in all papers published world-wide) to a particular country's papers and that country's share of papers in the total database, whereas we measure the value for a single paper, and then take its average over the world's research output.

5.3.2. *Alternative contrast sets*

We note that in Eq. (8a), the size of $\mu_t^{(0)}$ is constrained by the value of α_t. This reflects actual constraints imposed by the relative sizes of the sets 0 and t,[4] and is the result of using the full superset as the contrast set. This consequence, which may be disturbing to some people, can be corrected by using alternative, more restrictive, contrast sets.

We first generalize the previous measure to define a *relative* within-group impact measure, denoted by $\mu_{t|t'}^{(0)}$:

$$\mu_{t|t'}^{(0)} = \frac{\mu_t^{(0)}}{\mu_{t'}^{(0)}} = \frac{f_t^{(0)}/f_{t'}^{(0)}}{\alpha_t/\alpha_{t'}}. \tag{9}$$

This suggests, as our second possibility, a relative measure that contrasts the impact, on the superset, of t with the impact of all members of the *superset* other than t. If the superset is the whole universe of citations itself, the contrast set is just $\bar{t}$. More generally we will denote this set by $\bar{\bar{t}}$. We thus define the alternative measure $\mu_t'^{(0)}$, in terms of Eq. (9), by,

$$\mu_t'^{(0)} = \mu_{t|\bar{\bar{t}}}^{(0)} = \frac{P_t^{(0)}}{P_{\bar{\bar{t}}}^{(0)}} = \frac{f_t^{(0)}/f_{\bar{\bar{t}}}^{(0)}}{\alpha_t/\alpha_{\bar{\bar{t}}}}. \tag{10}$$

Finally, we note that if the superset is the total universe of items, the notation simplifies, since $f_{\bar{\bar{t}}}^{(0)} = 1 - f_t^{(0)}$, and $\alpha_{\bar{\bar{t}}} = 1 - \alpha_t$:

$$\mu_t'^{(0)} = \frac{P_t^{(0)}}{1 - P_t^{(0)}} = \frac{f_t^{(0)}/(1 - f_t^{(0)})}{\alpha_t/(1 - \alpha_t)}. \tag{10a}$$

5.3.3. *Excluding self-citation*

All the earlier within-group measures are influenced by the impact of self-citation: among the papers citing papers in t were source papers also in t. It is possible, within our framework, to exclude these. We conclude by noting a couple of possibilities for assessing a nation's impact on the global research community where the impact of self-citation is excluded.

In constructing the measure, we may have a specific contrast set, t', in mind. If so, we can offer a relative measure of the impacts of t and t' on all members of the global research community, other than countries t and t' themselves. By this restriction on the source set, self-citation is excluded. Such a measure allows us to explore the relative impact of t and t' on the rest of the "world". If we define T as the union of t and t', a measure satisfying our demand is defined by $\mu_{\bar{\bar{T}}} : t|t'$, where the double bar allows us to consider collectivities other than the whole world as our superset.

[4] An extreme example makes this clear. We examine the typical case in which the superset is the total universe, so $\alpha^{(0)} = 1$. First suppose that the number of items cited is fixed at n. Then we first conclude that $P_0^{(0)} = n/N$. Now allow $P_t^{(0)}$ to take its *maximum* value – this requires that all citations are to items in set t, that is, $n_t^{(0)}$ *also* equals n – it cannot take a value greater than the total number of items cited. This implies $P_t^{(0)} = (1/\alpha_t)(n/N)$. Thus the intrinsic constraints of the problem impose the condition that $P_t^{(0)} \leqslant (1/\alpha_t)P_0^{(0)}$, which is reflected in the bound indicated in Eq. (8a).

To give an explicit example: Let us say our "world" consists of the set of countries $\{A,B,C,D,E\}$, and we are interested on the relative influence of A and B on the rest of the world – that is, the relative impact of A and B on $\{C,D,E\}$. We compute this measure as $\mu_{\{C,D,E\}:A|B}$, that is, $s=\{C,D,E\}$; $t=A$; $t'=B$.

If an explicit contrast set is not apparent, we can use the "rest of the world". We thus conclude by suggesting, without further comment, an alternative measure, constructed in the same spirit. We here simply look at how often the rest of the world cites t. This is directly measured by $\mu_{\bar{\bar{t}}:t|\bar{\bar{t}}}$.

Appendix A. Notation

The generality of our measures is matched by a corresponding complexity of notation. We here summarize the notation used for reference.

N: Number of items in universe of citable items.
$n_{s:t}$: Number of citations in a "typical" paper in s to papers in t. We will use n to denote both the value for a specific paper in s, and its population average – relying on the context to make the distinction clear. We will let n denote the total number of citations in a source paper. On occasion, when s is a fixed, predetermined set, we will represent s as a superscript.
α_t: Each citable paper is associated with one (or more) sets, according to national designation(s); α_t is the fraction of potentially cited papers classified as being associated with the target set t.
$\bar{t}$: It will be useful below, given a class indexed by t, to have a notation for the set of items for which the index value t does not apply. We use the bar to indicate this. For example, if t refers to a nation, then the set $\bar{t}$ denotes all papers for which that national designation does not apply: $\bar{t}$ is just the complement of t. On occasion, when the complement is taken relative to a given subset of the universe, a double bar will be used.
$P_{s:t}$: Probability that a paper in the source set s cite (or in some other way select) a specified item from t.We can simplify the notation for special cases: The parameter P_0 of Bookstein and Yitzahki (1999, 1999a), representing the probability that s will cite an item of the same category, is now denoted by $P_{s:s}$; more succinctly (since the second s is understood), we will denote this by P_s. On occasion, when s is a fixed, predetermined set, we will represent s as a superscript.
$f_{s:t|t'}$: Empirically estimated proportion of choices in a specific paper by a member of the source set s of items in t, when the choice is restricted to be only of items from nations t or t'. We will slightly overburden the notation by using it for the average of this value over all members of s: we hope the distinction will be clear from the context. We similarly define $f_{s:t'|t}$, as the fraction of items cited by s that belong to nation t', given the same restriction. If the sets t and t' are disjoint, then clearly, $f_{s:t'|t}=1-f_{s:t|t'}$, since each cited item can be assigned to only one of the two classes. On occasion, when s is a fixed, predetermined set, we will represent s as a superscript.
Following the precedent of P, we can simplify the notation for special cases. Very often, the most appropriate choice for t' (the contrast set) will be $\bar{t}$, all citable items not classed in t. This occurs frequently enough to justify the simpler notation, $f_{s:t}$, suppressing the $\bar{t}$ in $f_{s:t|\bar{t}}$. This is the fraction of *all* citations by s to papers put in class t. The fraction of self-citation, denoted by F in Bookstein and Yitzahki (1999, 1999a), is now denoted by $f_{s:s}$, or more simply, by f_s. It is useful to note that $f_{s\bar{t}}=1-f_{s:t}$, the fraction of items cited by s that do not belong to t.
$\mu_{s:t|t'}$: The μ family of parameters will denote the preferences, defined below, of a member of s for members of t, when the only options are t or t'. Continuing the pattern established above, when $t'=\bar{t}$, we denote the simpler preference measure by $\mu_{s:t}$: the preference by s to papers in class t over all papers not so classified. And when self-preference is intended, we use μ_s. Occasionally, when we want to emphasize the role of s, we will place it as a superscript.

Appendix B. Statistical commentary

In the text of this paper, our focus has been on developing measures that made sense. To do this, we relied on simple probabilistic models. Since our interest was not in statistical analysis, we were relaxed in making distinctions that would be appropriate in a more formal discussion. In this appendix, we point to a few of these fine points.

One simplification was to simplify our notation a bit by not distinguishing values for a single paper and population values (more formally, between random variables and expected values). Thus, we wrote $n_{s:t}$ to denote both the random variable indicating the number of citations in a paper drawn at random from a source set s to papers in target set t; and also to some average of this value taken over all papers in s. This simplified our notation and discussion; we hope that the context allowed us to do this without confusing the reader.

In this spirit we also did not discuss the effect of taking averages. For example, we evaluated $n_{s:t}/n_{s:t'}$ for a single paper in s, and spoke of using an average over papers in s for our measures. But should we take the average values of $n_{s:t}$ and $n_{s:t'}$, and then take their ratio? Or should we compute the ratio of these quantities for each source paper, and take the average of these ratios? Our inclination is that, so long as we are consistent, either can be used for our heuristic measure without producing misleading results.

We try to justify this by a somewhat more precise analysis that should, at least, make clear the issues involved. Our question is, is there a significant difference between $E(n_t/n_{t'})$ and $\bar{n}_t/\bar{n}_{t'}$, where E is the expectation operator, and $\bar{n}$ is the expected value of the n's. (We suppress the s index, which is shared for all n's.) But, we can rewrite the ratio, using a straightforward algebraic identity,

$$\frac{n_t}{n_{t'}} = \frac{\bar{n}_t}{\bar{n}_{t'}} \frac{1 + (n_t - \bar{n}_t)/\bar{n}_t}{1 + (n_{t'} - \bar{n}_{t'})/\bar{n}_{t'}}.$$

The denominator can now be expanded as a geometric expansion to yield,

$$\begin{aligned} \frac{n_t}{n_{t'}} &= \frac{\bar{n}_t}{\bar{n}_{t'}}\left(1 + \left(\frac{n_t - \bar{n}_t}{\bar{n}_t}\right)\right)\left(1 - \left(\frac{n_{t'} - \bar{n}_{t'}}{\bar{n}_{t'}}\right) + \left(\frac{n_{t'} - \bar{n}_{t'}}{\bar{n}_{t'}}\right)^2 - \dots\right) \\ &= \frac{\bar{n}_t}{\bar{n}_{t'}}\left(1 + \left(\frac{n_t - \bar{n}_t}{\bar{n}_t}\right) - \left(\frac{n_{t'} - \bar{n}_{t'}}{\bar{n}_{t'}}\right) + \left(\frac{n_{t'} - \bar{n}_{t'}}{\bar{n}_{t'}}\right)^2 - \left(\frac{n_t - \bar{n}_t}{\bar{n}_t}\right)\left(\frac{n_{t'} - \bar{n}_{t'}}{\bar{n}_{t'}}\right) + \cdots\right) \end{aligned} \tag{11}$$

We can now take the expectation of both sides. If we note that the linear terms give zero contribution (because of the definition of the expected value), we conclude,

$$E\left(\frac{n_t}{n_{t'}}\right) \approx \frac{\bar{n}_t}{\bar{n}_{t'}}\left(1 + (\sigma_{t'}/\bar{n}_{t'})^2 - \rho_{tt'}\frac{\sigma_t}{\bar{n}_t}\frac{\sigma_{t'}}{\bar{n}_{t'}}\right) = \frac{\bar{n}_t}{\bar{n}_{t'}}\left(1 + \frac{\sigma_{t'}}{\bar{n}_{t'}}\left(\frac{\sigma_{t'}}{\bar{n}_{t'}} - \rho_{tt'}\frac{\sigma_t}{\bar{n}_t}\right)\right), \tag{12}$$

where we have neglected higher order terms. The above expansions are valid, provided the variability, as measured by the standard deviation, is less than the expected value. If so, then if the standard deviations are small compared to the expected values, the error in using $\bar{n}_t/\bar{n}_{t'}$ instead of $E(n_t/n_{t'})$ will be correspondingly small. But also, if the measures are used only for comparisons, then the results will be consistent provided the standard deviations, relative to the expected values, do not vary very much from variable to variable.

This result simplifies if $t\prime = \bar{t}$, for then it is easy to confirms that $\sigma_{t'} = \sigma_t$, and $\rho_{t\bar{t}} = -1$, both following from the definitions of standard deviation and correlation coefficient, since $n_{t'} = n - n_t$.

In general, n is variable, and we would have to consider its impact on our approximation. This adds extra complexity without extra insight. For simplicity, we assume n is constant. Thus,

$$E\left(\frac{n_t}{\bar{n}_{\bar{t}}}\right) \approx \frac{\bar{n}_t}{\bar{n}_{\bar{t}}}\left(1 + \frac{\sigma_t^2}{n - \bar{n}_t}\left(\frac{1}{n - \bar{n}_t} + \frac{1}{\bar{n}_t}\right)\right) = \frac{\bar{n}_t}{\bar{n}_{\bar{t}}}\left(1 + \frac{n}{\bar{n}_t}\left(\frac{\sigma_t}{n - \bar{n}_t}\right)^2\right),$$

which again approaches $\bar{n}_t/\bar{n}_{\bar{t}}$ for small σ's, specifically, if $\sigma_t << n - \bar{n}_t$.

Appendix C. Global vs local statistics

The text of this paper was based on the assumption that for every paper in the source set being considered (the set indexed by s), it is easy to determine the number of target documents in set t that were cited, and then to take the average over s: this was the value used to compute the μ-measure of the strength of linkage between s and t. But it may be easier to determine the global statistics, giving the total number of items in t cited by the documents in s. We now offer a rough translation between this global statistic and the local value used in this paper.

To keep our notation simple, let us eliminate from consideration all source papers except for those in *s*, and all target papers except those in *t*. This permits us to not carry the index values explicitly. Suppose that our global statistics are:

- T: Total number of items (in *t*) available for citation.
- T_c: The overall number of these that are in fact cited.
- S: Number of documents in the source set.
- n: The *effective* value for the number of items from the target set cited in any specific source document in *s*. Ideally, this would be the average, over documents in *s*, of target items cited. By "effective" we mean that this will be a value computed below, which we will then use as if it were constant over each paper in *s*; we would then use this value to compute μ.

As a first step in determining the relationship between local and global values, we develop a model in which every item in *t* has the same probability of being cited, and in which each paper in *s* picks at random the *n* items from *t* it cites; in this model, we are using *n* as the constant number of items from *t* cited across the papers in *s*. We will relieve these assumptions somewhat below. With this assumption, the relation is straightforward. Consider a specific target document. The probability that it be cited by a given source document that selects *n* items at *random* is n/T, and $1 - n/T$ the probability that it not be selected. Thus, the probability that it not be selected by any of the *S* source documents being evaluated, each choosing *n* items at random, denoted by $\overline{P}$, is given by

$$\overline{P} = (1 - n/T)^S = (1 - n/T)^{(T/n)(Sn/T)}.$$

We now note that T/S is the value we would get if all of the *T* citable items were evenly split over the *S* source documents under study. We shall denote T/S by n_0 and use it as the unit in terms of which we measure *n*. With this notational simplification, since n/T is very small, we can approximate $(1 - n/T)^{T/n}$ by *e*, and $\overline{P}$ by

$$\overline{P} = \mathrm{e}^{-n/n_0}.$$

This implies that the expected value of the number of target items not cited is $T\mathrm{e}^{-n/n_0}$; and finally, that the expected global number of target items cited, denoted above by T_c, is given by,

$$T_c = T(1 - \mathrm{e}^{-n/n_0}).$$

This gives us our basic relation between the global values T_c and *n*. If it is *n* that we are interested in, for use in evaluating a value of μ, we easily solve to conclude,

$$n = -n_0 \ln(1 - T_c/T).$$

Below we will examine how close this value is to the desired average we ideally need to compute μ-measures.

If T_c/T is small, a more easily computed approximation could be used, since we can use the Taylor series to expand the logarithm, and take the first couple of terms:

$$n \approx n_0 \frac{T_c}{T}\left(1 + \frac{1}{2}\frac{T_c}{T}\right),$$

or, alternatively,

$$n \approx \frac{T_c}{S}\left(1 + \frac{1}{2}\frac{T_c}{T}\right). \qquad (13)$$

We can use T_c/S as a 0th order approximation to the relationship between *n* and the observed global value T_c: it is the value *n* would take if there were no overlap among the sets of items cited by different documents. The second term gives us a first order correction that recognizes the impact of overlaps. If T_c/T, the overall fraction of documents cited, is small, this level of approximation might suffice.

We can also ask, how sensitively does *n* depend on T_c. Taking derivatives we find,

$$\Delta n/n_0 = \frac{1}{1 - T_c/T}\frac{\Delta T_c}{T}.$$

That is, small changes in T_c/T, the global fraction of items cited, begins to produce disproportionately large changes in n (relative to n_0), as the value of T_c/T approaches 1.

Impact of variability: If the number of items cited from t were actually a constant value, n, this result would trivially be the average number of target items cited we seek when computing a μ-value. But, of course, we realize that different papers cite different numbers of items, and the actual items available for citation vary in the likelihood of being cited. At minimum, a more realistic analysis would have to recognize that some papers are frequently cited, others never cited. We now ask how the value of n determined above relates to the values demanded in our formulae for the μ-values, which were averages of actual n's over s. To probe this, we will examine the impact of variability.

In general, there are many components of variability. A review paper cites more papers than a report of research results, and thus has a greater probability of citing a pertinent paper; relative probabilities vary with the content being discussed; and personalities of the authors influence the likelihood of a paper being cited. Nonetheless, much of the impact of variability is revealed by examining the impact of even simple variation: We assume that each paper in s determines whether it selects a paper in t by a random process. The probability that the ith target paper is cited is given by the probability p_i, constant over source papers, and independent of other papers that might have been cited, but varying over the target population. Since p_i, is constant over the S papers constituting the source population, we estimate it by

$$p_i \approx c_i/S,$$

with c_i the number of source papers that cite the ith target paper.

We defined p_i as the probability that a given source paper cite i. Arguing very much as we did for the simple model, we see that the probability, P_i, that *at least one* paper in s select the ith target paper (that is, that the ith target paper is cited) is given by:[5]

$$P_i = 1 - (1 - p_i)^S \approx 1 - \exp(-Sp_i), \tag{14}$$

and,

$$p_i = -\frac{1}{S}\ln(1 - P_i). \tag{15}$$

To estimate the total number of papers in the target universe that are cited at least once, we make use, for accounting purposes, of an indicator random variable $\tilde{\delta}_i$, which is equal to one if the ith target paper is cited at least once among the source documents, and zero otherwise. In terms of these indicator variables, the total number of target items cited at least once is given by the random variable $\tilde{T}_c = \sum^T \tilde{\delta}_i$. The value we denoted above by T_c is its expected value, and is given by

$$T_c = E(\widetilde{T}_c) = \sum^T E(\tilde{\delta}_i) = \sum^T P_i.$$

Note that the upper case notation, P, is called for. But the sum $\sum^T P_i$ also has an interesting interpretation. We use an argument similar the one just used, but with a minor variant of the indicator random variable: we now focus on a specific source document, chosen at random, and let the ith indicator variable take the value one if the ith target item is cited in that document. Proceeding then as above, we conclude that the expected number of items from t cited in a randomly chosen source paper, which we denote by $\bar{n}$, is given by[6]

[5] The assumption that p_i is constant over the items in s was introduced for simplicity, and with no real loss of generality. Suppose that the probability that a given source paper cite a given target paper varies with the idiosyncrasies attached to the source paper. Then the kth source paper cites the ith target paper with probability p_{ik}. Then $P_i = 1 - \Pi_k^S(1 - p_{ik}) \approx 1 - \exp(-\sum_k^S p_{ik}) = 1 - \exp(-Sp_i)$, where $p_i \equiv \sum_k^S p_{ik}/S$, the *average* value p_{ik} takes over the items in s. Similarly, the expected number of times the ith target item is cited is $c_i = \sum_k p_{ik} = Sp_i$, and the newly defined p_i could be estimated by c_i/S, just as we did before. Thus, even though in fact the probabilities vary with each paper, we can proceed as if a single probability governed the process through which a target paper was cited, bearing in mind that this probability is an average over s. For conceptual simplicity, we do continue thinking of p_i as if it were constant over source items, but we should be aware that this assumption is not needed.

[6] Here again, p_i can be thought of as an average: more precisely, the expected number of items cited by the kth source paper is given by $n_k = \sum_i^T p_{ik}$. Thus, $\bar{n}$, the sought for quantity, is the average of these n_k over the S source items in s: $\bar{n} = \sum_k^S n_k/S = \sum_i^T \sum_k^S p_{ik}/S = \sum_i^T p_i$, with P_i defined as an average as indicated in the preceding footnote.

$$\bar{n} = \sum^{T} p_i = -\frac{1}{S}\sum_{T} \ln(1 - P_i).$$

It could by estimated by,

$$\bar{n} = \sum^{T} c_i/S.$$

This would be the value of n that we would be seeking when evaluating μ, if this model correctly describes the citation process. We now wish to learn the relationship of the value of n determined by our preliminary model, and the value $\bar{n}$.

Recall that in our preliminary analysis, we assumed a constant value for n, the number of target items cited, resulting in a simple relationship between the observed value T_c and the value n. We can now probe the impact of variability on the relationship between the value thus determined for n and the desired $\bar{n}$.

To do this, we rewrite the expression for $\bar{n}$ as:

$$\bar{n} = n + (\bar{n} - n) \equiv n + \Delta.$$

Substituting from the results given above, we find,

$$\Delta = -\frac{T}{S}\left[\frac{\sum \ln(1 - P_i)}{T} - \ln\left(1 - \frac{T_c}{T}\right)\right].$$

Some conclusions are immediate. We first note that

$$\frac{1}{T}\sum \ln(1 - P_i) = \ln\left(\prod (1 - P_i)^{1/T}\right),$$

while,

$$\ln(1 - T_c/T) = \ln\left(\sum (1 - P_i)/T\right).$$

Thus we are comparing the logarithms of the geometric and arithmetic means of the values $1 - P_i$; since the geometric mean of a series of positive values is always less than its arithmetic mean, we can conclude immediately that Δ must be non-negative: n will tend to underestimate the value $\bar{n}$.

We can attempt to estimate the magnitude of the error by expanding the logarithm function as a Taylor series, and take the first few terms:

$$\begin{aligned}\Delta &= -\frac{T}{S}\left[\left(-\sum P_i/T - \frac{1}{2}\sum P_i^2/T + \cdots\right) + \left(T_c/T + \frac{1}{2}(T_c/T)^2 + \cdots\right)\right] \\ &\approx \frac{1}{2}\frac{T}{S}\left[\frac{\sum P_i^2}{T} - \left(\frac{\sum P_i}{T}\right)^2\right] = \frac{1}{2}\frac{T}{S} Var(P_i). \qquad (16)\end{aligned}$$

Since all the values of $P_i < 1$, we expect higher order terms to decrease rapidly. If we can make the reasonable assumption that the variance of the P_i, is much less than one, then we are free to use the simpler formula for n when computing μ-values.

Added insight can be gained by a complementary analysis based on a Taylor expansion of Eq. (14). Taking the first few terms of the expansion yields:

$$P_i = 1 - \exp(-Sp_i) \approx Sp_i - S^2 p_i^2/2,$$

so,

$$T_c = \sum^{T} P_i = S\bar{n} - S^2 \sum^{T} p_i^2/2,$$

and,

$$\bar{n} = \frac{T_c}{S}\left(1 + \frac{T}{2T_c}\frac{\sum^{T}(Sp_i)^2}{T}\right) \approx \frac{T_c}{S}\left(1 + \frac{T}{2T_c}\frac{\sum^{T} c_i^2}{T}\right).$$

T_c/S is simply the value for the number of target items cited in a source paper, if the total number of items cited were divided, without overlap, among the source papers. The correction reveals the impact of overlaps, which is governed by $\overline{c^2}$, the average of the values c_i^2. This agrees with Eq. (13) to the extent that $(T_c/T)^2$ is approximated by $\overline{c}^2$, both values being measures of squared average citation rate.

References

Bookstein, A., & Yitzahki, M. (1999). Own-language preference: a new measure of relative language self-citation. In C. A. Macias-Chapula (Ed.), *Proceedings of the Seventh Conference of the International Society for Scientometrics and Informetrics, Colima, Mexico, July 5–8* (pp. 65–74). Colima, Mexico: Universidad de Colima.

Bookstein, A., & Yitzahki, M. (1999a). Own language preference: a new measure of relative language self-citation. *Scientometrics, 46*(2), 337–348.

Bookstein, A., Moed, H., & Yitzahki, M. (2006). Measures of international collaboration in scientific literature: Part II. *Information Processing and Management*, current issue, doi:10.1016/j.ipm.2006.03.008.

Moed, H. F. (2005). *Citation analysis in research evaluation*. Dordrecht (Netherlands): Springer.