Interrelations among scientific fields and their relative influences revealed by an input–output analysis

Zhesi Shen\textsuperscript{a}, Liying Yang\textsuperscript{b}, Jiansuo Pei\textsuperscript{c}, Menghui Li\textsuperscript{d}, Chensheng Wu\textsuperscript{d}, Jianzhang Bao\textsuperscript{a}, Tian Wei\textsuperscript{a}, Zengru Di\textsuperscript{a}, Ronald Rousseau\textsuperscript{e}, Jinshan Wu\textsuperscript{a,}\textsuperscript{*}

\textsuperscript{a} School of Systems Science, Beijing Normal University, Beijing 100875, PR China
\textsuperscript{b} National Science Library, Chinese Academy of Sciences, Beijing 100190, PR China
\textsuperscript{c} School of International Trade and Economics, University of International Business and Economics, Beijing 100029, PR China
\textsuperscript{d} Beijing Institute of Science and Technology Intelligence, Beijing 100044, PR China
\textsuperscript{e} KU Leuven, Department of Mathematics, 3000 Leuven, Belgium

1. Introduction

Science funding agencies and science policymakers often have to decide on which science or technology fields to prioritize for a period of time for an efficient management of scientific resources and activities. To answer this question, the funding agencies need to assess the (future) relative importance of all scientific fields. Furthermore, once the target, i.e. the prioritized field, is chosen, it becomes an important consideration to find other fields which support the target field, as these be supported too.

These two questions are relevant not only to policymakers and committees in such agencies, but also to individual scientists, academic committees and university departments. Of course, one can apply peer review, relying on the opinions, feelings and visions of individual experts. However, with the rise of the era of big data, a natural question is whether technical analyses using large collections of published patents and research articles can help answer such questions.

The question of the relative importance of and influences between scientific fields has not yet been answered completely, admitting that investigating connections between scientific fields and technological sectors is one of the areas of investigation.
in the field of informetrics (JST, 2015; Narin, Hamilton, & Olivastro, 1997). In JST (2015), the Japan Science and Technology Agency (JST) was interested in knowing, for a given sector of patents, which scientific fields have been the primary sources of published information. The simple approach used in JST (2015) is to calculate how journal articles cited in a specific sector of patents are distributed across all scientific fields. In Narin et al. (1997), the authors were more focused on how the patterns of citations between patents and scientific publications changed due to national origin and over time. Such analyses based on directly counting the number of articles, patents and citations, are referred to as direct analyses. In this simple, direct statistical approach, an indirect contribution from scientific fields to sectors of patents is missing: if there is one sector of patents $T_a$, which heavily relies on one scientific field $S_i$, which in turn makes use of concepts and techniques from another scientific field $S_j$, then it is clear that even if there are no direct citations from $T_a$ to $S_j$, $S_j$ is a major contributor to $T_a$. These connections are referred as indirect connections. They are the main topic of this investigation.

This idea of considering direct as well as indirect relations, though straightforward, cannot be underestimated. Results of such approaches are sometimes described as network effects (West & Vihlена, 2014). In Fig. 1A, we provide an example of a citation relationship between scientific fields in which indirect connections (between node 1 and node 4 or node 1 and node 3) could in principle play a more important role than direct ones, due to the lack of a direct connection between nodes 1 and 4 and a weak connection between nodes 1 and 3. While network science researchers, including those from social network analysis, have often used this perspective (Barabási, 2015), the network perspective is not yet a commonplace in informetrics. This remark does not imply that informetricians have not valued the network perspective (Otte & Rousseau, 2002; West & Vihlена, 2014). Indeed, the network effect is the key idea behind Google’s PageRank algorithm (Brin & Page, 1998) and its scientific predecessor, the Pinski-Narin influence methodology (Franceschet, 2011; Pinski & Narin, 1976). The PageRank algorithm has been used to measure the relative importance of journals (Bergstrom, West, & Wiseman, 2008) and articles (Chen, Xie, Maslov, & Redner, 2007; Ma, Guan, & Zhao, 2008). However, the PageRank algorithm focuses mainly on ranking the nodes in a network, not on interrelations among the nodes. We consider here both the tasks of ranking as well as describing interrelations.

Now that our work has been placed in its proper context, we first note that we will focus on scientific fields instead of journals and articles. Therefore, we may naively adopt the PageRank algorithm or equivalently the Pinski-Narin influence methodology for our study, by classifying publications into scientific fields.
However, our interest goes beyond a measure of relative importance. We also want to know which fields support or are supported by a given field. Therefore, we consider the Leontief Input–Output Analysis (LIOA) in economics (Leontief, 1941; Miller & Blair, 2009). LIOA is a method of answering similar questions about economic sectors. In fact, the similarity between the ideas and motivations behind LIOA and PageRank has previously been described by Franceschet (2011). In LIOA, one starts with a direct input–output matrix $B$, where $b_{ij}$ represents the number (or monetary value) of product $i$ required for producing one product $j$. Sector $N$, the last sector, is reserved for final consumers, so $b_{iN}$ refers to the number (or value) of products from sector $i$ used per final consumer. This sector is also called final demands. Two typical questions in LIOA are the following: First, what happens if the final demand increases? How will the total output of the other sectors change to match an increment in the demand for certain products? Second, which economic sector is the most important for the whole economy? What are the effects of removing one sector, e.g. sector $i$, from the economy, on each of the other sectors? The former is usually discussed in terms of the Leontief inverse (Miller & Blair, 2009), a solution to a specific linear equation while the latter is often discussed in terms of the so-called Hypothetical Extraction Method (HEM) (Temurshoev, 2010). Roughly speaking, in HEM people compare various quantities calculated in the complete LIOA and in the LIOA without sector $i$. In this way, if there is a large change in one of the quantities, e.g. sector $j$’s output, sector $i$ is regarded as important for and especially influential on sector $j$.

Because these two questions concerning the relative importance of industrial sectors and their interrelations, such as the effect of changes in the output of product $i$ on product $j$, are very close to what we are interested in, we use the ideas of LIOA for the present study. To do so, we need to define an input–output matrix $B$ based on the citation relationships between scientific fields. Entries in $B$ could be, for example, the ratio between the number of citations from field $j$ to field $i$ and the total number of citations received by field $j$. In a sense, this ratio stands for the number of citations of papers in $i$ required for producing a citation in $j$. This provides a close parallel between LIOA and the problem we intend to study. In principle, other definitions of $B$ can be considered. For example, one can use numbers of papers instead of numbers of citations. One can also define $B$ to be something like the PageRank matrix, i.e. the ratio between the number of citations from field $j$ to field $i$ and the total number of citations received by field $i$. In a sense, this means how all citations received by field $i$ are initiated from papers in field $j$. Of course, different definitions study different aspect of citation phenomena. Since here our questions of interest are close to the question studied in LIOA, we define our matrix $B$ in the same way as the one used in LIOA. The PageRank $B$ will be more appropriate when studying popularity (including both direct and indirect effect) of fields. Other definitions and their meanings could be the topic of further investigations.

However, as we will show furtheron, this approach is not as straightforward as it may seem. New concepts and techniques are required to make LIOA applicable to study the informetric problems that we are interested in. The key difference is that LIOA is performed on an open system, but the system of scientific fields is a closed system. There is not a natural external sector paralleling the final demand sector in economics unless, perhaps, if one includes patents. This would be a further step requiring more data than what we have at the moment. Thus, we need an input–output analysis method for closed systems. Furthermore, the number of citations is not a conserved quantity in the production of scientific works: the total number of citations received by a field is usually not the same as the number of citations initiated from the field.

Fortunately, as we illustrate later, eigenvalues, which are the basis of our definition of Input–Output Factor (IOF), and eigenvectors, which are the basis of our definition of Input–Output Influence (IOI), are the key concepts we need for our closed-system input–output analysis. This relates our method to the PageRank algorithm or, equivalently, the Pinski–Narin influence methodology. Therefore, the method developed in this study – an extension of LIOA for a closed system – can also be regarded as an extension of the PageRank algorithm that makes it applicable to influences among the nodes in a network with an input–output relation. It also extends the PageRank algorithm in that it makes it possible to answer the second question – the interrelation among nodes, not only the ranking problem. A comparison between our closed-system input–output analysis and the PageRank algorithm is provided in Section 5.

Aside from the methodological contributions toward answering the two questions we raised in the beginning, we find that, although overall our IOF is strongly correlated with the number of citations or publications, there are outliers in the correlation plots between the IOF and the number of citations or publications. Those outliers have either much stronger (i.e., Statistical Physics) or much weaker (i.e., Relativity), influences on other fields when compared with the number of citations or publications in them. It seems to us that these outliers are intuitively understandable and plausible. Similar meaningful outliers have been identified in relational studies, in which influences on and from individual fields are considered. For example, we found that 03 (Quantum Mechanics) is closely related to 37 (Mechanical control of atoms) while there are relatively few direct citation between the two. This demonstrates that our network-based analysis can go beyond studies based on direct statistics using the number of citations or publications.

We present the main idea and the formulae in the next section. After that, in Sections 3 and 4, we use a closed-system analysis to investigate relationships between the subfields of physics using records from the American Physical Society (APS) of published journals articles and discuss the validity of the information revealed by our analysis. In Section 5 we compare the proposed method against the PageRank algorithm. A more general discussion of the validity of our closed-system input–output analysis can be found in Section 6. Discussions of some technical issues of our method are reported in Appendix A.
2. Modified closed system input–output analysis (MCSIOA): the core idea

We will first summarize the open-system LIOA in economics and then modify it to make it applicable to closed systems. In fact, the first input–output model (Leontief, 1941) that Leontief proposed was a closed-system model and only later he and the vast majority of his followers turned to an open-system analysis. Let us assume that the whole economy has \( N \) sectors and each sector is a component such as Agriculture, Mining, Textiles, etc. Starting from a matrix \( x = (x_i^j)_{N \times N} \) representing the number or monetary value of all products of sector \( i \) that are required for producing the products of sector \( j \), one defines a matrix of direct input–output coefficients

\[
b_j^i = \frac{x_j^i}{x_i^i},
\]

where \( x_i^i = \sum j x_j^i \). With these elements \( b_j^i \), we obtain

\[
X' = \sum_j b_j^i X_j = B X,
\]

meaning that \( X \) is an eigenvector of matrix \( B \) with eigenvalue 1, the largest eigenvalue of matrix \( B \). For simplicity, we call the eigenvector corresponding to the largest eigenvalue the largest eigenvector.

If we separate the final demand sector, say sector \( N \), from the other sectors of an economy, and denote it as \( x_N^N = y' \), we have

\[
X' = \sum_{j=1}^{N-1} b_j^i X_j + y' \Rightarrow X^{(-N)} = (I - B^{(-N)})^{-1} Y^{(-N)},
\]

where \( X^{(-N)} (Y^{(-N)}) \) is what remains of vector \( X (Y) \) after its \( N \)th element is removed and, similarly, \( B^{(-N)} \) is what remains of matrix \( B \) after its \( N \)th row and \( N \)th column are removed. The inverse matrix is known as the Leontief inverse, and is denoted as \( L = (1 - B^{(-N)})^{-1} \). \( L \) is also called the full input–output coefficient matrix because it takes into account not only the direct coefficients but also the indirect ones. This can be observed even more clearly if we rewrite \( L \) as follows:

\[
\Delta^{(-N)}X = L \Delta^{(-N)}Y = \sum_n (B^{(-N)})^n \Delta^{(-N)}Y,
\]

assuming \( \Delta^{(-N)}Y \) – a change in the final demand – is known. Intuitively, this means that to meet the demand \( \Delta^{(-N)}Y \), as a response, \( \Delta^{(-N)}X \) first needs to produce at least \( \Delta^{(-N)}Y \) products; to do so, it needs to have enough products to make \( \Delta^{(-N)}Y \) and thus \( (B^{(-N)}) \Delta^{(-N)}Y \), and so on. This is the basic idea of the open-system Leontief input–output analysis. In order to use Eq. (4) to find \( \Delta^{(-N)}X \), in principle, one needs \( \Delta^{(-N)}Y \).

In addition to the question of the system’s response to a change in the final demand, LIOA can be applied to measuring the relative importance of sectors and the influences among them. This is called the Hypothetical Extraction Technique (HEM) (Temuruho, 2010). The basic idea is that for a given \( \Delta^{(-N)-j}Y \) (removing further the \( j \)th element from \( \Delta^{(-N)}Y \), one can define

\[
\Delta^{(-N)-j}X = L^{(-j)} \Delta^{(-N)-j}Y = (1 - B^{(-N)-j})^{-1} \Delta^{(-N)-j}Y,
\]

where \( B^{(-N)-j} \) is what remains of matrix \( B \) after both the \( j \)th and the \( N \)th \((j \neq N)\) row and column are removed. One then compares \( \Delta^{(-N)}X \) with \( \Delta^{(-N)-j}X \). If they are quite different (or, specifically, the \( k \)th element differs), then the \( j \)th sector is essential to the economy (to the \( k \)th sector). One may say that the importance of sector \( j \) to the economy and to each other sector is concealed in the difference between \( L \) and \( L^{(-j)} \).

Due to the difference in the time scales of producing next-generation labor and manufacturing other products, it is plausible to separate the sector of final consumers from the other industrial sectors. However, in principle the sector of final consumers is an intrinsic ‘manufacturing’ sector of the economy because it provides labor and accepts products. Let us now turn to the closed-system approach to input–output analysis, in which it is neither necessary nor possible to treat one sector as external to the system.

The linear equation technique is clearly no longer applicable to our closed-system input–output analysis, but we may study the largest non-negative eigenvector of \( B \) and \( B^{(-j)} \) as long as those matrices have such an eigenvector. Ideally, we would also like to expect that such a largest non-negative eigenvector is unique for a given matrix \( B \) or \( B^{(-j)} \). However, in principle this is not necessarily true although this is almost always the case in the following empirical analysis. We introduce a robust analysis by adding a perturbative term to matrices \( B \) and \( B^{(-j)} \) to make the values all positive just as is used in the PageRank algorithm. Details are provided in Appendix A. For simplicity of notation, we still call those perturbed positive matrices \( B \) and \( B^{(-j)} \), of which each has a unique all positive largest eigenvector.

We then consider the difference between the eigenvalues and eigenvectors of \( B^{(-j)} \) and \( B \). This relies on another interpretation of Eq. (2): the vector \( X \) can be regarded as the specific combination of products that, when supplied to the economy, results in one hundred percent of the input becoming the output, i.e., the economy operates at full efficiency because the
corresponding eigenvalue is 1 and it is the maximum eigenvalue. Similarly, the maximum eigenvalue and the corresponding eigenvector of $B^{(-)}$ are associated with the highest efficiency and the corresponding combination of products for the economy without sector $j$. Imagine the case in which sector $j$ has hardly any connections to other sectors, i.e., the values in the $j$th row and/or column are very small compared with other elements of $B$. Denoting the largest eigenvalue of matrix $B^{(-)}$ by $\lambda^{(-)}$, then, $\lambda^{(-)}$ will be very close to 1. Otherwise, when elements in the $j$th row and column are relatively large, $\lambda^{(-)}$ will be much smaller than 1. The fact that all eigenvalues of the matrix $B^{(-N)}$ (and also all $B^{(-)}$) must be less than or equal to 1 in magnitude is shown in Appendix A.

Therefore, we propose using the IOF (input–output factor) defined by

$$S_{ji}^{(I)} = 1 - \lambda^{(-)}$$

(6)

to measure the relative importance of sector $j$. This answers the first question we raised in this paper.

Let us now attempt to provide an answer to the second question. Intuitively, the influence of sector $j$ on each of the other sectors is concealed in the difference between $X$ and $|\lambda^{(-)}|$, which are respectively, the largest eigenvector of $B$ and $B^{(-)}$. Thus, we propose the following quantity, which we call IOInfluence (IOI), to provide a comparison between $X$ and $|\lambda^{(-)}|$,

$$\Delta'_k = \left( \frac{k[X] - \lambda^{(-)} (\sum_{l \neq k} X_l)}{(k[X])} \right),$$

(7)

where $|\lambda^{(-)}|$ is the largest eigenvector of matrix $B^{(-)}$ (the L1-norm of $|\lambda^{(-)}|$ is 1), and $|k|$ is the column vector with all zeros except for the $k$th element. In a sense this eigenvector represents the best combination of products when sector $j$ is removed from the economy. The total output of the new system without sector $j$ intuitively should be $\lambda^{(-)}$ times the original total output, thus the term $\lambda^{(-)} (\sum_{l \neq k} X_l)$. Note that this definition of $\Delta'_k$ is based on intuition and has not been fully justified.

Eqs. (6) and (7) are the two core formulae in this paper. All of the calculations in the following sections are based on these two formulae. Within the general framework of the closed-system input–output analysis sketched above, we will now answer the two central questions raised at the beginning of this article.

3. MCSIOA applied to relationships between subfields in physics: the results

The above closed-system input–output analysis is now applied to the study of relative importance of and influences among scientific fields. We consider subfields of physics as a case study.

3.1. Construction of the closed input–output system

We use data regarding all papers published in APS (American Physical Society) journals between 1980 and 2013. A total of 390,191 papers have Physics and Astronomy Classification Scheme (PACS) codes. PACS is a classification system of subfields in physics consisting of 6-digit 4 to 5-level codes. We will, however, use only the first 3 levels. There are 10 (resp. 78 and 937) PACS codes at level 1 (resp. level 2 and level 3). APS papers come with several author–defined PACS codes. The rich information encoded in such a classification system has been discussed in e.g. Wei et al. (2013).

To establish the input–output system of subfields, we regard each PACS code as a sector. A citation received by a papers in one sector (PACS code) from a paper in another sector (PACS code) is modeled as an input from sector $i$ to sector $j$. We then count the papers and citations within the APS data. For example, if one paper $p$ published in sector $j$ cites a paper $q$ published in sector $i$, there is a link from $i$ to $j$. Each paper may have multiple PACS codes. For instance, if in a time window $t$, a paper $p$ having $P_p$ PACS codes, one of which is $j$, and cites $C_p$ papers, one of which is $q$, which has $P_q$ PACS codes one of which is $i$, then the contribution toward the input–output relation from $i$ to $j$ due to the citation from paper $p$ to paper $q$ is

$$X'_{ij}(p \rightarrow q) = \frac{1}{P_p P_q C_p}.$$

(8)

In this study we use a time window of five years. We provide an example of the weighted network in Fig. 1, where a citation, as in Fig. 1A, from Paper A to Paper B is converted into a network, as in Fig. 1B, and a matrix representing the weighted network, as in Fig. 1C, following Eq. (8). Input–output networks/matrices $X_{ij}^{(N)}$ of PACS codes can be established at various levels in this way. In LIOA in economics, $X_i = X_i^{(1)}$; the total input to an economic sector equals the total output from that sector. Here it is not necessarily true that the citation count from the field is the same as the citation count to the field. Luckily for us, we do not need this to be the case for the analysis to work.

One thing we should note about this extension from analysis of economics to informetrics is that the matrix elements $b_{ij}$, which stand for how much input one needs from $i$ to $j$ for a unit of product in $j$, in informetrics are not strictly respected by developments of scientific fields while in economics they are better defined as constants. Of course, in both cases, it is assumed that the investigation is carried out in a short time window so time is not really a problem. The problem lies rather in production: in economics, it is natural that producing every new unit of product $j$ needs the same input $b_{ij}$ from $i$, while a new paper in field $j$ only roughly, in an average sense, needs $b_{ij}$ from $i$. However, when the field size is much more than a few of papers, we believe that the assumption of ‘following the old paths’ – how papers in one field on average cite papers from
all fields – is still reasonable. Furthermore, in this work, we focus more on a descriptive study than a predictive study, i.e. we consider the direct and indirect influences of each field by removing the fields and see how such hypothetical extraction propagates in the already formed network. A similar assumption is used in the EigenFactor (Bergstrom et al., 2008), which applied the PageRank algorithm to evaluate relative importance of journals.

3.2. The relative importance of subfields and its evolution

With the set of input–output networks/matrices \( (X_i^j)_{j \in N, i \in N} \) and matrices \( B \) of PACS codes for different time periods, we first discuss the relative importance of subfields and how this evolves.

First, we examine the correlation between the relative importance, as measured by the IOF, and by the number of times each subfield is cited. In Fig. 2A, we compare the IOF rankings of PACS codes with the rankings obtained from the total number of citations received by all papers with corresponding PACS codes. As shown in the figure, although the two rankings are correlated, there are some outliers: some fields, such as 05 and 02, have relatively higher IOF rankings (smaller \( y \) values, toward the top in the figure) whereas others, such as 04 and 98, have higher citation rankings (smaller \( x \) values, toward the right in the figure).

PACS 05 is the field of “Statistical physics, thermodynamics and nonlinear dynamical systems” (Statistical physics for short). From the correlations for 2009 – 2013 shown in Fig. 2, we see that 05 has a large influence on other fields of physics relative to the number of citations it received, and this has been the case for this field for the past few decades (see Fig. 3). This means that not only were papers in Statistical physics (05) cited directly by many papers in other fields, but that 05 plays an important indirect role: many other influential papers cited those papers who directly cited papers in 05 and so on. This picture of the importance of Statistical physics is consistent with our own intuition that, in recent years, concepts, models and methods from statistical physics have been extensively used in other scientific fields.

A similar be it slightly different behavior can be observed for PACS 02, “Mathematical methods in physics”. It has a relative low IOF ranking and total number of citations. However, considering its low number of citations, its IOF score is outstanding. This means that the total number of citations received directly by this field is not very high, but its indirect effect makes this field more important than suggested by the number of received citations.

PACS 04 and 98 are among the fields that have higher citation rankings than their IOF rankings. This result does not imply that those fields are less important: it just means that they have smaller influence on other fields. It is understandable that each of these fields is more like a closed field of their own. Many physicists may not need to know much about stellar systems (98) to conduct their research.

We performed a similar comparison between the citation rankings and publication rankings of the subfields. Fig. 2B shows that these rankings are better correlated than the previous pair of rankings, so that, generally speaking, the outliers in Fig. 2B stand out less. Consider, for example, the subfields 04 and 05 in the two figures: they are quite different in Fig. 2A while they are both on the diagonal line in Fig. 2B. We want to emphasize that by including indirect connections, IOF rankings provides more insights and valuable information than citation rankings and publication ranking (at least in this case) because the latter only consider direct connections.
Fig. 3. IOFs are plotted against citation counts for all subfields during each year from the year 1991 to 2011. Evolutions of sub fields 05, 03, 04, 32, 61, 68, 74, 78, 82, 98 are highlighted with their 2011 positions marked on this figure. Full data and a list of the top 20 at all levels can be found in Section 4. An animated version is provided in Supplementary materials (animated.gif).

Fig. 4. Multi-level pie charts for the years 1991, 2001 and 2011 showing the composition of major branches of physics within each of the four quartiles. Colors represent the level-1 PACS codes, which are for the major branches of physics here. From the years, in the core region of the first quartile influential subfields, General Physics (00) and Interdisciplinary (80) have taken a larger part while Atomic and Molecular Physics has been shrinking. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

There are other outliers in the correlation figure, but we focused on some fields with which we have personal knowledge. The complete data set is provided in Supplementary materials for further examination. The results on parallel studies on level-1 and level-3 subfields are reported in the next Section 4.

The same plot can be used to reveal the time evolution of relative importance of the subfields. In Fig. 3, we plot the IOF values (z-score, as normalized according to mean and standard deviation of all IOF values of each year), instead of the rankings, of the IOF and citation counts of all subfields between 1996 and 2011. We intend to use IOF purely for ranking the subfields. In fact, all that we know about the eigenvalues is that their range is [0, 1] and they can be ordered, but there is no operation such as summation defined over the set of eigenvalues. However, we do find that the z-score plot gives us more information on how far the distributions of IOF values are from the normal distribution: IOF values are more clustered in the negative region while there are less positive values, especially large positive values, which makes those values thus the corresponding subfields stand out more. The trajectories of some subfields (05, 03, 04, 32, 61, 68, 74, 78, 82, 98) are highlighted. The following two facts were interesting and surprising to us. First, for a very long time (before the year 2007) 05 (Statistical physics) had a higher IOF than 03 (Quantum Mechanics), and second, that several subfields of 60 (Condensed matter I) and 70 (Condensed matter II) have decreasing IOFs even in cases of increasing citation counts. For example, the citation count of 74 (Superconductivity) increases while its IOF decreases. Thus, a simple citation count of 74 (Superconductivity) will not find that the influence of this subfield to others has not been improving during the years. See Section 4 for the figures of subfields at other levels.

To look more closely into the finer structure, we analyzed the relative importance of the level-3 subfields and then plotted the level-3 subfields according to their level-1 classifications. Result of this analysis are presented using multi-layer pie charts in Fig. 4. Each level of the charts from inner to outer layers represents ordered quartiles of subfields (from the most to the least influential one). In each layer, we use different colors to represent the level-1 PACS codes of the subfields.
Here level-1 PACS codes are regarded as the major branches of physics. In this way, we can see how each region is composed from major branches of physics and how this composition changed over time.

The pie chart for 1991 shows that the top quartile consisted mostly of Condensed Matter (PACS 60 and 70), General Physics (PACS 00) and Elementary Particles and Fields (PACS 01), with small contributions from Atomic and Molecular Physics (PACS 30), Electromagnetism, Optics, Classical Mechanics (PACS 40) and Interdisciplinary Physics (PACS 80). When the 2001 and 2011 pie charts are compared with the 1991 pie chart, we see that General Physics and Interdisciplinary Physics become larger parts of the core region while Atomic and Molecular Physics shrank. We can also look at the change in the distributions of a particular color between all four quartile sets. For example, Interdisciplinary Physics moves steadily toward the center, the more influential level, whereas large portions of Geophysics, Astronomy and Astrophysics (PACS 90) migrate from the second to the third quartile.

3.3. Influences among the subfields

For a given subfield $j$, we calculate $\Delta^j_k$. This describes how much the number of citations received by the subfield $k$ changes, directly and indirectly, if subfield $j$ is removed from the field of physics. Subfield $k$ relies strongly on subfield $j$ when $\Delta^j_k < 0$ and subfield $k$ can be regarded as a substitute for subfield $j$ when $\Delta^j_k \gg 0$.

In Fig. 5 we use two specific subfields – 98 (Stellar systems) and 03 (Quantum Mechanics) – in the time interval 2004–2008 as examples. We see that there is a large difference between the influential sets, according to IOI (IO Influence) and citation counts for subfield 03, while the difference is smaller for subfield 98. It is also important to note that, according to Fig. 5A, the top 10 fields with the greatest influence on 98 are generally in astronomy, relativity, stars, etc., which makes intuitive sense. This observation supports our intuitive definition of $\Delta^j_k$. From Fig. 5B, we see that, if, for example, one wants to boost the development of 03, then it might be necessary to increase funding for 37 (Mechanical control of atoms, etc.) and 39 (Instrumentation and techniques for atomic and molecular physics, later partially merged into 37), which are not in the top five fields cited from 03. A complete map of all the physics subfields at all levels is provided in the next section (Section 4).

4. Additional results

In the previous section, we demonstrated that some valuable information can be extracted from the input–output table of subfields with our MCSIOA method by showing firstly that outliers in the correlation plot of IOF and number of citations received of subfields provide insightful information and those outliers are more visible in the plot of IOF vs. number of publications than in plot of number of citations vs. number of publications, and secondly, for some given subfields, IOI reveals other subfields which are closely related to the fields but not quite visible from simply direct citations among the two fields. These are the main results of the current work, which intends to mainly illustrate the validity and usefulness of our MCSIOA method.

In this section, we try to provide some additional and more detailed results on the system to which we apply our method, i.e. this section is more about the system while the previous sections were more about the method.
better gases) ranks. (B) Correlation between citations ranks and the IOF ranks of level-3 subfields is plotted in this figure. Fields 67.85 (Ultracold gases, trapped gases) and 78.67 (Optical properties of low-dimensional, mesoscopic, and nanoscale materials and structures) have better IOF ranks than their citation ranks while fields 71.45 (Collective effects) and 98.80 (Cosmology) have better citation ranks than their IOF ranks.

4.1. Tables and figures for influences of subfields of physics

In Fig. 2 of the main text, we reported correlations between ranks of level-2 subfields based on the IOF indicator and the number of citations. Here we provide in Fig. 6 the same correlation plots on level-1 and level-3 subfields of physics.

In Fig. 6 (A) Correlation between citations ranks and the IOF ranks of level-1 subfields is plotted in this figure. PACS codes such as 00 (General Physics) have better IOF ranks than their citation ranks while citation ranks of PACS codes such as 10 (The Physics of Elementary Particles and Fields) are better than their IOF ranks. (B) Correlation between citations ranks and the IOF ranks of level-3 subfields is plotted in this figure. Fields 67.85 (Ultracold gases, trapped gases) and 78.67 (Optical properties of low-dimensional, mesoscopic, and nanoscale materials and structures) have better IOF ranks than their citation ranks while fields 71.45 (Collective effects) and 98.80 (Cosmology) have better citation ranks than their IOF ranks.

4.2. Figures for influences among subfields of physics

In Fig. 6, we choose two specific subfields at level-2, 98 and 03, and present subfields that are closely related to these two subfields. Here we report influences among all the level-2 subfields and also among all level-1 and level-3 subfields. We use a heatmap for this purpose: the size of each circle represents number of citations from the column subfield to the row.
Table 1
Top 20 most influential level-2 subfields at the year of 2011. Full data can be downloaded at subfield_list_level2.txt from Supplementary materials.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Pacs code</th>
<th>Subfield</th>
<th>IOF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>03</td>
<td>Quantum mechanics, field theories, and special relativity</td>
<td>0.0334</td>
</tr>
<tr>
<td>2</td>
<td>05</td>
<td>Statistical physics, thermodynamics, and nonlinear dynamical systems</td>
<td>0.0323</td>
</tr>
<tr>
<td>3</td>
<td>42</td>
<td>Optics</td>
<td>0.0267</td>
</tr>
<tr>
<td>4</td>
<td>71</td>
<td>Electronic structure of bulk materials</td>
<td>0.0261</td>
</tr>
<tr>
<td>5</td>
<td>73</td>
<td>Electronic structure and electrical properties of surfaces, interfaces, thin films, and low-dimensional structures</td>
<td>0.0245</td>
</tr>
<tr>
<td>6</td>
<td>75</td>
<td>Magnetic properties and materials</td>
<td>0.0233</td>
</tr>
<tr>
<td>7</td>
<td>78</td>
<td>Optical properties, condensed-matter spectroscopy and other interactions of radiation and particles with condensed matter</td>
<td>0.0179</td>
</tr>
<tr>
<td>8</td>
<td>61</td>
<td>Structure of solids and liquids; crystallography</td>
<td>0.0174</td>
</tr>
<tr>
<td>9</td>
<td>68</td>
<td>Surfaces and interfaces; thin films and nanosystems (structure and nonelectronic properties)</td>
<td>0.0141(3)</td>
</tr>
<tr>
<td>10</td>
<td>74</td>
<td>Superconductivity</td>
<td>0.0140(7)</td>
</tr>
<tr>
<td>11</td>
<td>72</td>
<td>Electronic transport in condensed matter</td>
<td>0.0134</td>
</tr>
<tr>
<td>12</td>
<td>64</td>
<td>Equations of state, phase equilibria, and phase transitions</td>
<td>0.0132</td>
</tr>
<tr>
<td>13</td>
<td>11</td>
<td>General theory of fields and particles</td>
<td>0.0124</td>
</tr>
<tr>
<td>14</td>
<td>47</td>
<td>Fluid dynamics</td>
<td>0.0122</td>
</tr>
<tr>
<td>15</td>
<td>87</td>
<td>Biological and medical physics</td>
<td>0.0121</td>
</tr>
<tr>
<td>16</td>
<td>02</td>
<td>Mathematical methods in physics</td>
<td>0.0118</td>
</tr>
<tr>
<td>17</td>
<td>12</td>
<td>Specific theories and interaction models; particle systematics</td>
<td>0.0113</td>
</tr>
<tr>
<td>18</td>
<td>81</td>
<td>Materials science</td>
<td>0.0105</td>
</tr>
<tr>
<td>19</td>
<td>82</td>
<td>Physical chemistry and chemical physics</td>
<td>0.0097</td>
</tr>
<tr>
<td>20</td>
<td>67</td>
<td>Quantum fluids and solids</td>
<td>0.0095</td>
</tr>
</tbody>
</table>

Table 2
Top 20 most influential level-3 subfields at the year of 2011. Full data can be downloaded at subfield_list_level3.txt from Supplementary materials.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Pacs code</th>
<th>Subfield</th>
<th>IOF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>03.65</td>
<td>Quantum mechanics</td>
<td>0.0149</td>
</tr>
<tr>
<td>2</td>
<td>42.50</td>
<td>Quantum optics</td>
<td>0.0147</td>
</tr>
<tr>
<td>3</td>
<td>03.67</td>
<td>Quantum information</td>
<td>0.0131</td>
</tr>
<tr>
<td>4</td>
<td>05.45</td>
<td>Nonlinear dynamics and chaos</td>
<td>0.0105</td>
</tr>
<tr>
<td>5</td>
<td>03.75</td>
<td>Matter waves</td>
<td>0.0098</td>
</tr>
<tr>
<td>6</td>
<td>42.65</td>
<td>Nonlinear optics</td>
<td>0.0082</td>
</tr>
<tr>
<td>7</td>
<td>71.10</td>
<td>Theories and models of many-electron systems</td>
<td>0.0078</td>
</tr>
<tr>
<td>8</td>
<td>73.20</td>
<td>Electron states at surfaces and interfaces</td>
<td>0.0077(1)</td>
</tr>
<tr>
<td>9</td>
<td>05.40</td>
<td>Fluctuation phenomena, random processes, noise, and Brownian motion</td>
<td>0.0076(9)</td>
</tr>
<tr>
<td>10</td>
<td>75.10</td>
<td>General theory and models of magnetic ordering</td>
<td>0.0073</td>
</tr>
<tr>
<td>11</td>
<td>78.67</td>
<td>Optical properties of low-dimensional, mesoscopic, and nanoscale materials and structures</td>
<td>0.0070</td>
</tr>
<tr>
<td>12</td>
<td>75.30</td>
<td>Intrinsic properties of magnetically ordered materials</td>
<td>0.0068</td>
</tr>
<tr>
<td>13</td>
<td>05.30</td>
<td>Quantum statistical mechanics</td>
<td>0.0067(0)</td>
</tr>
<tr>
<td>14</td>
<td>71.15</td>
<td>Methods of electronic structure calculations</td>
<td>0.0066(8)</td>
</tr>
<tr>
<td>15</td>
<td>75.50</td>
<td>Studies of specific magnetic materials</td>
<td>0.0062</td>
</tr>
<tr>
<td>16</td>
<td>64.70</td>
<td>Specific phase transitions</td>
<td>0.0061(2)</td>
</tr>
<tr>
<td>17</td>
<td>67.85</td>
<td>Ultrapow gases, trapped gases</td>
<td>0.0061(1)</td>
</tr>
<tr>
<td>18</td>
<td>64.60</td>
<td>General studies of phase transitions</td>
<td>0.0060(9)</td>
</tr>
<tr>
<td>19</td>
<td>89.75</td>
<td>Complex systems</td>
<td>0.0059(4)</td>
</tr>
<tr>
<td>20</td>
<td>74.25</td>
<td>Properties of superconductors</td>
<td>0.0059(2)</td>
</tr>
</tbody>
</table>

subfield while the color in the circle corresponds to our IOI from the row to the column subfield. The value of the number of citations has been renormalized with respect to the row subfield.

First, we note that it is not always the case that the order of degree of influences is the same as the order of citation counts. Second, we found that a large numbers of IOIs are positive while a few of them are negative. We interpret the positive ones, which means that when field i is removed from the whole discipline outcomes of field j decreases, to be the relying-on relation and the negative ones, which means that outcomes of field j increases when field i is removed, to be competitive or substitutive relation. A detailed examination of all those relations, which has not been done in this work,
should be interesting to policy makers and funding agencies in physics and closely related fields. Third and finally, we also observed that in level-2 and level-3 heatmaps, overall there are relatively stronger correlations among the subfields within the same categories (the diagonal block elements) than that among the subfields across categories (the off-diagonal block elements). This means that the boundaries between different categories, represented by the hierarchical structure of the PACS codes, describe the interconnections among subfields in a meaningful way (Figs. 10 and 11).

The level-3 heat map has $940 \times 940$ entries, so it is too big to show the figure in great detail. Therefore, for this map, we also provide a png file for downloading relation_heatmap_level3.png and a data file in excel format, which is accessible via

![Fig. 9. Evolutions of top 20 most influential level-3 subfields for the years 1991, 2001 and 2011. At this level of details, we can see even more clearly that overall influences of fields in 60 and 70 are stable or slightly decreasing compared to the year of 1991 while fields in 03 and 05 are increasing. Full data can be downloaded at subfield_list_level3.txt from Supplementary materials.](image)
Fig. 10. Given a level-1 column subfield, its relations to other subfields in the row are color coded, as deeper color represents stronger influence. Size of the circles is proportional to the number of citations received by the row field from the column field. Full data can be downloaded at heatmap_level1.txt from Supplementary materials. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

Fig. 11. Given a level-2 column subfield, its relations to other subfields in the row are color coded, as deeper color represents stronger influence. Size of the circles is proportional to the number of citations received by the row field from the column field. Full data can be downloaded at heatmap_level2.txt from the Supplementary materials. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

heatmap_level3.txt (also heatmap_level1.txt and heatmap_level2.txt for the level-1 and level-2 heatmaps respectively) from Supplementary materials.

5. A comparison between MCSIOA and the PageRank algorithm

Since the proposed MCSIOA and the PageRank algorithm both considered direct and indirect effects in measuring relative importance of a node in a network, it is natural to perform a comparison of the two methods conceptually and based on obtained results.
A citation starting from \( j \) and received by \( i \) represents the flow of and idea from \( i \) to \( j \), thus a PageRank transfer matrix can be defined as
\[
MF_{ij} = \frac{X_i}{X_j}
\]
representing how much percent of all citations initiated from field \( j \) goes to field \( i \). Note that there is a difference in notation between ours and the conventional notations of the PageRank algorithm of web pages.

Consider a largest right eigenvector of \( MF \), \( \langle 1 \rangle_M = \{P_1, P_2, \ldots, P_N\} \) (we know that for a probability transfer matrix, its largest eigenvalue is 1), we have
\[
P^i = \sum_j MF_{ij}p^j = \sum_j \frac{X_i}{X_j} P_j = \sum_i \frac{X_i}{X_i} P_j = \sum_i \frac{X_i}{X_j} P_j.
\]
If we define the output–input matrix \( F \) as
\[
F_{ij} = \frac{X_i}{X_j},
\]
meaning how much output from sector \( i \) goes into sector \( j \) per unit of product received by sector \( i \), then \( \bar{P} = \left( \frac{P_1}{X_1}, \frac{P_2}{X_2}, \ldots, \frac{P_N}{X_N} \right)_j = \langle 1 \rangle_F \) is the largest right eigenvector of input–output matrix \( F \),
\[
\langle 1 \rangle_M = \langle 1 \rangle_F \cdot X,
\]
where the operator “\( \cdot \)” represents an element-wise product. This means that the PageRank vector \( \langle 1 \rangle_M \) is related to, and somewhat equivalent to, the largest right eigenvector of output–input matrix \( F \).

This relation implies that if instead of \( B \) (with element \( \frac{X_j}{X_i} \)), we were defining our measures based on \( F \) (with element \( \frac{X_i}{X_j} \)), we would be effectively using the PageRank algorithm. This is one connection that we should note in comparing MCSIOA and the PageRank algorithm. When studying the flow of ideas, \( F_{ij} \) shows how much per unit incoming flow to \( i \) goes further to \( j \) while \( B_{ij} \) means that for \( j \) to receive one unit of citation how much it needs from \( i \). The former is more like a forward flow, hence the notation \( F \) and the latter is more like a backward flow, thus notation \( B \). In a special case of flows of conserved quantities, such as materials/energy in LIOA instead of a not-conserved quantity – ideas – in our case, \( F = B^T \). Therefore, in a sense, the right eigenvector of \( F \) corresponds to left eigenvector of \( B \).

Based on this closely related definition, i.e. \( B^T \sim F \sim MF \), the technique that we proposed here to measure relative influences and interrelations based on MCSIOA is quite different from the PageRank score. PageRank score is defined as the largest right eigenvector of \( MF \) (thus equivalently \( F \)). However, in MCSIOA we use the largest right eigenvector of \( B \), \( \langle 1 \rangle_B \), which corresponds, in principle, to a certain left eigenvector of \( F \) (and thus \( MF \)). In fact, we use the difference between the original \( \langle 1 \rangle_B \) and \( \langle 1 \rangle_{B^T} \), the largest right vector after removing sector \( j \). Therefore, in principle, conceptually, the two definitions are not the same, but are related in a non-trivial way. The Pagerank score measures popularity including both direct and indirect effects while the MCSIOA \( S^j_{\langle 1 \rangle_B} \) measures how much field \( j \) is needed in order to sustain the flow of ideas. Plus, in MCSIOA, we also have an interrelation matrix \( A' \), which is beyond the Pagerank algorithm.

Fig. 12. Results from IOF and the PageRank algorithm show very high correlation. Moreover, we can still see more outliers in the figure of IOF vs. PageRank (a) than the figure of PageRank vs. Citation (b), and also that IOF vs. PageRank (a) is more widely scattered than PageRank vs. Citation (b).
Now let us compare the two methods based on results by applying each on the same data, namely the citations among subfields of Physics. We can see from Fig. 12 that results from the two methods are very highly correlated, but the figure on IOF vs. PageRank (a) (also IOF vs. Citation see Fig. 2(a)) is widely scattered than the one on PageRank vs. Citation (b). This means that we can see more outliers in (a) than (b). For example, 03 (Quantum Mechanics) and 05 (Statistical Physics) are well above the diagonal line and this means that their influences are measured higher by IOF than by PageRank while for 71 this is the opposite. 71 (a subfield of condensed matter) is a large field, receiving quite a lot of citations within condensed matter, which is the largest field in physics. This (large node in a largest sector and well connected within the sector) makes it stand out according to the PageRank algorithm, but not according to our IOF. From these results, we see that the IOF and the PageRank scores intend to describe different aspects of influences.

Furthermore, we want to emphasize here again that the PageRank algorithm cannot discuss interrelations among subfields, while the proposed MCSIOA can.

6. Conclusion and discussion

In this paper we developed a method of closed-system input–output analysis and used it to study influences between subfields of physics using APS publication data. We found that by including both direct and indirect connections, our closed-system input–output analysis revealed deeper relationships among subfields than could be observed by directly looking at the numbers of citations and publications. This method provides an innovative approach for answering the two questions raised at the beginning of the paper: Given a set of fields, which is more influential and thus may be supported preferentially? Given a specific priority, what other fields are necessary foundations for the targeted field and thus may need to be prioritized? Also when combined with time-series data, this method can also be used to track the development of the influences between scientific fields.

In addition, we also did a comparison between our closed-system input–output analysis and the PageRank algorithm, and this on concepts as well as results. The two methods use similar assumptions and both consider direct and indirect effects. For reasons that are not yet completely clear overall results from the two method are strongly correlated while the exact values of influences/scores of nodes are different and their definitions are based on different mathematical concepts.

Furthermore, besides ranking the nodes, which is the output from the PageRank algorithm, our closed-system input–output analysis can be applied to discuss interrelations among nodes.

In this work, we use publications in Physics and the PACS classification system as a case study. In fact, besides PACS, currently there are other classification codes, such as JEL (on economics), MESH (on life science and medical, etc.), MSC (on math). Therefore, it is possible to apply the proposed analysis to quite a few other disciplines. Furthermore, it is our opinion that every discipline can and should have a fully structured classification code system. Thus, the proposed method can be applied to much more fields beyond Physics.

The proposed method can be regarded as an extension of Leontief’s input–output analysis, thus it is generally applicable to all systems with input–output relations. Recently several innovative and insightful works applying Leontief’s input–output analysis to environmental science have been published, see for example (Feng et al., 2013). It is possible to combine our extended input–output analysis with these investigations. Our method is also a development of network analysis since it can be seen as an alternative to the PageRank algorithm. For example, a new type of influence factor of and among journals can be established based on it. With more and more data available in this era of big data, it will be interesting to see more applications of this method. We admit though that some aspects of our method are more based on intuition than on strict mathematical results, see the discussion in Appendix A.1.

Author contributions

Conceived and designed the analysis: Jinshan Wu, Jiansuo Pei and Zengru Di.
Performed the analysis: Zhesi Shen, Liying Yang, Menghui Li, Jianzhang Bao and Tian Wei.
Wrote the paper: Ronald Rousseau, Liying Yang, Jinshan Wu, Chensheng Wu and Zhesi Zhen.
Other contribution: Ronald Rousseau also made a critical contribution to a key technical issue of this work concerning the uniqueness of largest eigenvector of B.

Acknowledgements

This work was supported by NSFC under Grant No. 61374175. We thank the anonymous referees and the main editor for valuable suggestions which clearly improved the manuscript.

Appendix A. Further details on methods

Uniqueness of the largest eigenvector of $B$ (and also $B^{-1}$) is a subtle and important technical issue of the analysis proposed in this work. Here we provide some further discussions on this issue.
A.1. Uniqueness of the largest eigenvector of B and $B^{(-j)}$

The Perron–Frobenius theorem of positive matrices states that each positive matrix has a unique eigenvector containing only positive values and the corresponding eigenvalue is the maximum real-valued eigenvalue. Therefore, positive matrices enjoy all of the good properties that we expect matrices $B$ and $B^{(-j)}$ to have. However, our matrices $B$ and $B^{(-j)}$ are not positive but only non-negative matrices. The Perron–Frobenius theorem of non-negative matrices claims that each irreducible non-negative matrix has a unique eigenvector containing only positive values and the corresponding eigenvalue is the maximum real-valued eigenvalue. Note that matrix $B$ and $B^{(-j)}$ are not necessary irreducible. Due to this, the largest eigenvalue and the corresponding largest eigenvector might not be unique. Of course, it might be the case that the largest eigenvector is still all positive and unique. Thus, we performed the following additional analysis on matrix $B$ and all matrices $B^{(-j)}$.

First, we check the existence and uniqueness of this largest non-negative eigenvector in our practical calculations. After removing all sectors with no output ($X = 0$) from matrix $X$ to define matrix $B$, we find that for all cases, such a largest non-negative eigenvector exists and it is unique for matrices $B$ and $B^{(-j)}$. However, although practically it is the case in our analysis, we cannot guarantee that for other systems matrices $B$ and $B^{(-j)}$ always have this property.

Second, we check for irreducibility of matrices $B$ and $B^{(-j)}$ as that is required in the Perron–Frobenius theorem of non-negative matrices. One way to do that is to examine the strong connectivity of the graph corresponding to $B$ and $B^{(-j)}$. We have done this in this work using the non-recursive Tarjan’s algorithm with Nuutila’s modifications provided in the networkX software and find that at all PACS levels the strongly connected components of $B$ and $B^{(-j)}$ cover more than 96% of all citations. At level 1, for our 5-year period analysis, the whole network $B$ is strongly connected already and all the corresponding networks of $B^{(-j)}$ are also strongly connected. At level 2, the strongly connected subgraph of $B$ and $B^{(-j)}$, denoted as $\tilde{B}$ and $\tilde{B}^{(-j)}$, keeps 99% of the citations in the whole network. At Level 3, the citation network is relative sparse, so about 100 sectors are excluded but the remaining strongly connected component keeps about 96% of the citations. These large percentages mean that even though sometimes matrices $B$ and $B^{(-j)}$ are not irreducible, they are very close to irreducible matrices.

In principle, we can always identify and then only focus on the strongly connected $\tilde{B}$ and $\tilde{B}^{(-j)}$. This procedure is, however, quite demanding. Here we suggest to use a perturbation analysis.

Third, we use a perturbation analysis to, in a sense, calculate the largest non-negative eigenvector directly from $B$ and $B^{(-j)}$ instead of from $\tilde{B}$ and $\tilde{B}^{(-j)}$. The following idea of this perturbation analysis comes from the PageRank algorithm and is quite straightforward: we want to compare the calculated largest eigenvectors of matrix $B$ and $B_\alpha = (1 - \alpha)B + \alpha E$ with $\alpha$ as a numerical value being very close to 0, where matrix $E$ is the matrix with every element being 1. According to the Perron–Frobenius theorem of non-negative matrices, because $B$ might not be irreducible, the calculated largest eigenvector of $B$ might only be one of the eigenvectors corresponding to the eigenvalues with the same maximum magnitude, while the calculated largest eigenvector of $B_\alpha$, since it is a positive matrix, is unique and corresponds to the largest eigenvalue, which is also unique. Now when we compare those two calculated eigenvectors, denoted as respectively $\lambda(B)$ and $\lambda(B_\alpha)$, it can be the case that the two vectors are rather different or that they are quite similar. Since $\lambda(B_\alpha)$ is unique but $\lambda(B)$ is not, in principle, the two vectors can be quite different even with $\alpha$ being very close to 0: there are multiple $\lambda(B)\text{s}$ and they live in a multidimensional largest eigenvector space. Even with a tiny $\alpha$ the dimension of the largest eigenvector space collapses into a one-dimensional one. This change of dimensions has a large effect unless the largest eigenvector space of $B$ is already one-dimensional. Therefore, we can find out whether the largest eigenvector space of $B$ is one-dimensional by simply looking at whether the following expression give a value numerically very close to 1 or not,

$$u = (\lambda(B)\lambda(B_\alpha))(\alpha \approx 0).$$

We also want to compare eigenvectors of $\tilde{B}$ and $B_\alpha$ since if we want to use the largest eigenvector of $B_\alpha$ then ideally we want this largest eigenvector to be close to the one from $\tilde{B}$. Thus we define

$$v = (\lambda(\tilde{B})\lambda(B_\alpha))(\alpha \approx 0).$$

Note that ideally we expect $v$ to be close to 1 and $u$ to be slightly smaller than 1. Theoretically, this holds for arbitrarily small $\alpha$ since introducing this $\alpha$ breaks the multiplicity of the largest eigenvalue in magnitude into a simple largest eigenvector. However, in numerical calculations, there is always a problem of finite accuracy so we use a simple example in Fig. 13 to estimate the sufficiently large value of $\alpha$. Our numerical calculation is performed with the Scipy and specifically using the ARPACK linear algebra package provided by the Scipy in Python. Remember that we do not want this value to be too large such that $v$ becomes too small.

The graph in Fig. 13 is not strongly connected and the corresponding adjacency matrix has multiple largest eigenvectors, i.e. $\lambda(B)$ is not unique. The calculated largest eigenvectors of $B$ ($B^{(-j)}$) and $B$ ($B^{(-j)}_\alpha$) are compared and we find that for almost all values of $\alpha$, $\lambda(B_\alpha)$ is closer to $\lambda(B)$ than $\lambda(B)$ except when $\alpha < 0.00001$. This means that even when the original matrix $B$ has multiple $\lambda(B)$s introducing this extremely small $\alpha$ makes $\lambda(B_\alpha)$ unique and very close to the unique largest eigenvector of $\lambda(B)$. From what we have observed from this example, we use $B_\alpha$ and $B^{(-j)}_\alpha$ with $\alpha = 0.00002$ instead of directly using matrix $B$ or $B^{(-j)}$ in all of our analyses presented in the main text. Again we want to emphasize that we intend to work on $\tilde{B}$ and $\tilde{B}^{(-j)}$, which is however very demanding, thus we instead turn to the much less demanding $B_\alpha$ and $B^{(-j)}_\alpha$. 


With such an extremely small $\alpha$, we regard this to be a simple technical issue having only a minor influence on the properties of the desired largest eigenvector of $\mathbf{B}$ and $\mathbf{B}^{(j)}$. Due to this replacement, the core formulae Eqs. (7) and (8) in fact need to be adjusted accordingly. However, since this $\alpha$ is extremely small and we do this for purely technical reasons, we regard the eigenvectors and eigenvalues to be those from matrix $\mathbf{B}$ and $\mathbf{B}^{(j)}$ although they are actually not.

### Appendix B. Supplementary data

Supplementary data associated with this article can be found, in the online version, at [http://dx.doi.org/10.1016/j.joi.2015.11.002](http://dx.doi.org/10.1016/j.joi.2015.11.002).

### References


Pinski, G., & Narin, F. (1976). Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. Information Processing and Management, 12(5), 297–312. [http://dx.doi.org/10.1016/0306-4573(76)90048-0](http://dx.doi.org/10.1016/0306-4573(76)90048-0)


Wei, T., Li, M., Wu, C., Yan, X.-Y., Fan, Y., Di, Z., & Wu, J. (2013). Do scientists trace hot topics? Scientific Report, 3, 2207. [http://dx.doi.org/10.1038/srep02207](http://dx.doi.org/10.1038/srep02207)