Are Contributions from Chinese Physicists Undercited?

Jinzhong Guo^{1,2}, Xiaoling Liu¹, Liying Yang², Jinshan Wu^{3†}

¹School of Information Management, Xinjiang University of Finance and Economics, Urumchi 830012, China

²National Science Library, Chinese Academy of Sciences, Beijing 100190, China

³School of Systems Science, Beijing Normal University, Beijing 100875, China

Citation: Jinzhong Guo, Xiaoling Liu, Liying Yang, Jinshan Wu (2019). Are contributions from Chinese physicists undercited? Journal of Data and Information Science, 4(4), 84–95

DOI: 10.2478/jdis-2019-0022

Received: Sep. 23, 2019

Revised: Oct. 7, 2019

Accepted: Oct. 26, 2019

Abstract

Purpose: In this work, we want to examine whether or not there are some scientific fields to which contributions from Chinese scholars have been under or over cited.

Design/methodology/approach: We do so by comparing the number of received citations and the IOF of publications in each scientific field from each country. The IOF is calculated from applying the modified closed system input–output analysis (MCSIOA) to the citation network. MCSIOA is a PageRank-like algorithm which means here that citations from the more influential subfields are weighted more towards the IOF.

Findings: About 40% of subfields in physics in China are undercited, meaning that their net influence ranks are higher (better) than the direct rank, while about 75% of subfields in the USA and German are undercited.

Research limitations: Only APS data is analyzed in this work. The expected citation influence is assumed to be represented by the IOF, and this can be wrong.

Practical implications: MCSIOA provides a measure of net influences and according to that measure. Overall, Chinese physicists' publications are more likely overcited rather than being undercited.

Originality/value: The issue of under or over cited has been analyzed in this work using MCSIOA.

Keywords Input-Output Analysis; Scientific impact; Citation networks

1 Introduction

Evaluating scientific contribution from each country to each discipline has been a trending topic in Scientometrics. In practice, for science policymakers, it is also important to have proper recognition of their own country's academic position,

especially for the fast-developing countries like China. Often in this kind of study, researchers count the number of publications or received citations of the countries and in the fields of interest and rank them accordingly (King, 2004; Meho & Yang, 2007; Moed & Halevi, 2015).

However, a simple counting of publications and citations might underestimate or overestimate the contribution of a country. How the papers are cited, say by important mile-stone like papers or by insignificant homework-like papers, should make a large difference. In fact, that is exactly the idea behind PageRank algorithm (Brin & Page, 1998) and Leontief Input-Output Analysis (LIOA) of Economics (Leontief, 1941; Miller & Blair, 2009): citation or more generally a linkage from a more influential node, which can be papers, web pages, economic sectors, and scientific fields, should weight more than that from an insignificant node. In the LIOA, economists are interested in the same question. Given an input-output table between economic sectors, which sector is more influential than others. For precisely this purpose, Loentief proposed the LIOA, which regard the economic sectors and the input-output relation among them as an open system, while taking final demanders and labor input as the external sector (Leontief, 1941). In Shen et al. (2016), we extended LIOA into a Modified closed system input-output analysis (MCSIOA) to make the analysis to be applicable to an input-output table of any flow between any nodes beyond economic sectors and to even closed systems.

LIOA starts from the direct input-output coefficient matrix B , where B_j^i means how many units of a product i is needed to produce one unit of product j . Then the idea of LIOA is that for each unit of product j from the final demanders, denoted as Y^j , the economic system needs to produce first Y^j , and then also the raw materials need to produce Y^j , thus BY^j , and then also the raw materials to produce BY^j , thus $BBY^j = B^2Y^j$ and so on. Overall, one arrives at the famous Leontief inverse input-output matrix $X = Y^j + BY^j + B^2Y^j + \dots = (1 - B)^{-1}Y^j \triangleq LY^j$. Depending on the structure of L , some sectors Y^j might lead to a large X , even when it is needed by the final demanders only a small amount. In terms of scientific influence, this is like to say that if a field A cites heavily a field B and the field B also cites heavily a field C, then the field A should be considered strongly influenced by the field C.

MCSIOA (Shen et al., 2016) follows the same spirit but works on closed systems where there is not a natural external sector like the final demanders in LIOA. Therefore, instead of a matrix inverse, which requires an external sector, MCSIOA uses the largest matrix eigenvalue and the corresponding eigenvector (called the largest eigen pair for simplicity), which is applicable to closed systems and also takes into account higher-order effects of the matrix. To see why the largest eigen pair includes higher-order effects, one can use the power method calculation of it: starting from a random initial vector X_0 , the iterative multiplication $X_n = BX_{n-1}$ leads

to eventually the largest eigenvalue pair. we also refer the readers to Shen et al. (2016) for further details, which will also be briefly explained in the section of Data and Method.

In this work, we will apply MCSIOA to an input-output table between the subfields of physics from major countries. We will regard the overall influence of each sector calculated by MCSIOA as the net influence and use the number of citations as the direct influence of each sector, and then compare the net and the direct influence. When a country is ranked higher according to the net influence than the one according to the direct influence, then we say publications from that country have been undercited. We say publications from a country are overcited if the net influence rank is lower than the rank according to the direct citation counts.

We are fully aware that the above approach is just one possible way to measure the net influence. Simply counting of citations is a limited influence measure, but it is easy to understand and easy to implement. The use of MCSIOA, on the other hand, requires further justification. However, from the above explanation of its spirit and also from its success in economics application and the previous Scientometric study, we think it provides a reasonable measure of the net influence by taking not only the direct but also the indirect connections into consideration.

The rest of this manuscript is organized as follow, the data and method are in section 2. In section 3, we illustrate the main results. Conclusions and discussions will be in section 4.

2 Data and method

The data we use in this work is provided to us by the American Physical Society (APS) and it includes all papers published in APS journals between 1977 and 2013. There are a total of 404,496 papers and 6,039,964 citations. All papers have been classified according to the Physics and Astronomy Classification Scheme (PACS) codes, which is a classification system of subfields of physics. We chose the first and third-level PACS code for our analysis. There are $F = 1,281$ subfields in total.

Totally 165 countries and regions are identified from 337,768 authors' address. For non-USA addresses, the last part of the address string is usually a country name. For that, we match this last part to a list of countries. For the USA addresses, there is often not a country name in the address strings, we then match the state names to a list of states in the USA. In very rare cases, we find a match of the last part of the address string in both lists and in those cases, we check each of them manually. We use the full count when assigning papers to countries. As illustrated in Figure 1, for each citation between a citing paper A and a cited paper B, we identify the corresponding fields f^A, f^B and country c^A, c^B , which can be more than one, and then add this citation to the citation count from $f^{A \times c^A}$ to $f^{B \times c^B}$,

Figure 1. (a) paper B, in field 75.10 (also 75.30, 75.40, and 75.50) and from Japan, is cited by paper A, in field 75.10 (also 75.30) and from the USA, German and Japan. (b) Citations (from A to B and from A to C) are converted into a citation network among the countries $\times$ subfields of physics.

$$x_{f_{A \times C}^A}^{f_{B \times C}^B} = x_{f_{A \times C}^A}^{f_{B \times C}^B} + 1 \quad (1)$$

We then keep the countries/regions with more than 1,000 publications in our record and group other countries and call it “others”. In the end, there are $C = 45$ countries/regions in our record.

Once we have the input-output matrix $x = (x_j^i)_{(C,F) \times (C,F)}$, where each element x_j^i representing the number of citations from j to i , we define the direct input-output coefficient matrix F

$$F_j^i = \frac{x_j^i}{\sum_k x_i^k} \quad (2)$$

and perform the MCSIOA that is the net influence of sector j is (Shen et al., 2016).

$$S_{IO}^j = 1 - \lambda_{MAX}^{-j} \quad (3)$$

where λ_{MAX}^{-j} are the largest eigenvalue of $F^{(-j)}$, which is the matrix after removing the j th row and column of F . It is called IOF in (Shen et al., 2016). We then rank all sectors according to respectively their total number of received citations $X^j = \sum_k x_k^j$ and their IOF S_{IO}^j and compare the two ranks to determine sector j is overcited or undercited.

The idea behind the definition of S_{IO}^j can be seen from the following two facts. Firstly, the largest eigenvalue of F takes into account both direct and indirect connections in F . Secondly, the largest eigenvalue of the original F matrix is 1, and it can be regarded as the production efficiency of F , meaning that all the input when supplied according to the right combination, the corresponding eigenvector. Therefore, the largest eigenvalue $F^{(-j)}$ also captures both direct and indirect

connections, and it means the percentage of production efficiency of the system after the sector j removed, due to which some unmatched supplies will be wasted. Therefore, if the sector j is well-connected to the rest of the system and rest of the system deeply relies on sector j , then the percentage of production efficiency of the rest of the system will be low. Our previous study (Shen et al., 2016) have shown that indeed influential sectors do lead to large S_{IO}^j .

The difference between net influence and the direct influence and the direct influence means that indirect citation does not follow a similar pattern of direct citations. For example, if most of the citations are from the same sector (country $\times$ field), then the dissipation power of the sector is lower, and then the net influence will also be lower. Or the net influence is lower when most of the citations come from low impact papers. On the other hand, if most citations from other sectors and from high impact papers, the net influence will be higher.

3 Results

3.1 The direct input-output flow among countries/regions

We first illustrate the input-output flow among countries/regions on a world map (Csomos, 2018). The nodes indicate countries/regions and the links represent the citations of scientific papers within APS. As shown in Figure 2, each link is color-coded. The red (green) part corresponds to the number of received citations (citing references). The thickness of the line corresponds to the number of citations, the thicker, the larger. For each line, the node starting with the red (green) line has more (less) received citations than the number of citing references. For example, the edge between the USA and Europe is red near the USA and green near Europe, and this means that the USA received more citations from Europe than the other way around. This line is also quite thick, indicating there are a lot of citations on this line. Furthermore, the ratio between the length of red and green parts is set to be the ratio between the received citation of the USA and the received citation of Europe. In this way, we code a lot of information on this world map. We can see that the edge between Europe and Japan is red near Japan while the edge between Europe and China is red near Europe. On each node c , we also plot the number of received citations and the calculated IOF, S_{IO}^c .

We can see from this world map in Figure 2 that the USA is a source of knowledge to all other countries and regions since the lines starting from the USA are all red. On the contrary, China is more like a sink, or a consumer of knowledge since almost all the lines starting from China are green. The Europe as a whole can be regarded as a proxy where edges are mixed with red and green colors. Japan can also be seen as a proxy (Zhang et al., 2013).

Figure 2. The direct citations are shown on a world map. For each node c , we show the number of received citations and the calculated IOF, S_{io}^c . On each edge e_j^i on the world map, we code with the thickness of the line the value of both x_i^j and x_j^i : x_i^j is the line near i and x_j^i is the line near j . Each edge is also color-coded: the line starting from i is red when $x_i^j > x_j^i$ and green otherwise.

In principle, we can draw such a world map for each subfield, but then by looking at each subfield separately, we will be missing the citations among subfields. Therefore, next, we apply MCSIOA to the input-output table of countries $\times$ subfields to take into account of those cross-field citations and also to provide a measure of net influence of each subfield in each country so that we can compare the direct and the net influence to answer the equation raised in the introduction that which field in which country is undercited or overcited.

Based on this input-output matrix $x = (x_i^j)_{(C \cdot F) \times (C \cdot F)}$ and the MCSIOA analysis (Shen et al., 2016), we calculate the IOF of each of the 972 subfields of physics in each of the 45 countries and regions. We then rank all the countries $\times$ subfields together and compare these two ranks.

3.2 The net influence (IOF) ranks of countries $\times$ subfields

From Figure 3, we see that many USA subfields are above the diagonal line. Thus, they are undercited. Or we say that according to their net influence, there should be more citations to these USA subfields. Meanwhile, a lot of Chinese subfields are below the diagonal line. Thus, they are overcited. To see how many subfields in each country are above or under the diagonal line, we count the percentage of undercited fields and also calculate the relative and absolute ranking difference, and show them in.

The relative and absolute ranking difference, which sometimes are also called ranking mobility (Dagostino & Dardanoni, 2009), is respectively defined as

$$M_c = \sum_f R_{c,f}^{(d)} - R_{c,f}^{(n)} \quad (4)$$

$$|M|_c = \sum_f |R_{c,f}^{(d)} - R_{c,f}^{(n)}| \quad (5)$$

where $R_{c,f}^{(d)}$ and $R_{c,f}^{(n)}$ are respectively the direct citation count rank and the net influence rank of the subfield f of the country c . For an overall undercited country, M_c will be larger than zero. $|M_c|$ shows how large is the difference between the direct rank and the net rank.

From Figure 4(a), we see that physicists from the USA have made a contribution to most subfields (972 out of 1,281) and also the percentage of undercited subfields are high $74\% = \frac{724}{972}$. Similar situations are found for German (844 out of 1,281, and $75\% = \frac{629}{844}$), France (806 out of 1,281, and $75\% = \frac{589}{806}$), and British (757 out of 1,281, $75\% = \frac{547}{757}$). For China, the coverage is 696 out of 1,281, and the percentage of the undercited fields is $40\% = \frac{492}{696}$. Both coverage percentages and undercited

Figure 3. The net influence rank of countries $\times$ subfields. Each country is represented by its flag. The ones with higher net influence rank than the direct rank are above the diagonal line, thus undercited, while they are under the diagonal line when their net influence ranks are lower, thus overcited.

percentages are much lower than those of the countries mentioned above. Furthermore, in Figure 4(b), we look into the relative and absolute ranking differences of each country. M_c provides more detailed information than the undercited percentage. We found that again, the USA, German, France have high M_c while China, Iran, and Korean are the countries with the lowest M_c , indicating that overall publications from those countries are overcited.

For readers who are interested in knowing what are the undercited or overcited fields for each country, we provide a list of top 10 undercited or overcited subfields of four of the countries, including the USA, German, Japan, and China in Table 1.

Figure 4. (a) Percentage of undercited fields of each country are plotted in a figure of the number of undercited fields v.s. the number of contributed fields. More information than just the percentage of undercited fields can be seen from (b) the relative and absolute ranking difference between the net and the direct rank of each country c.

Table 1. Top 10 undercited or overcited subfields of USA and China.

USA		China	
UnderCited		UnderCited	
1	14.17 Properties of specific particles	12.25 Models for gravitational interactions	
2	91.67 Geochemistry	96.40 Cosmic rays	
3	25.38 Properties of specific particles	11.40 Currents and their properties	
4	66.46 Quantum tunneling of defects	42.66 Physiological optics	
5	26.10 Nuclear astrophysics	31.90 Other topics in the theory of the electronic structure of atoms and molecules	
6	23.70 Heavy-particle decay	93.30 Information related to geographical regions	
7	24.87 Surrogate reactions	51.35 Mechanical properties; compressibility	
8	91.80 Geochronology	34.60 Scattering in highly excited states	
9	95.80 Astronomical catalogs, atlases, sky surveys, databases, retrieval systems, archives, etc.	47.37 Hydrodynamic aspects of superfluidity	
10	42.27 Wave optics	25.43 Antiproton-induced reactions	
OverCited		OverCited	
1	68.25 Phase transitions in liquid thin films	43.80 Bioacoustics	
2	68.48 Solid-gas/vacuum interfaces: types of surfaces	91.65 Geophysical aspects of geology, mineralogy, and petrology	
3	53.35 Other topics in physics of plasmas and electric discharges	82.75 Molecular sieves, zeolites, clathrates, and other complex solids	
4	07.58 Infrared, submillimeter wave, microwave and radiowave instruments and equipment	28.50 Fission reactor types	
5	07.90 Other topics in instruments, apparatus, and components common to several branches of physics and astronomy	87.56 Radiation therapy equipment	
6	62.90 Other topics in mechanical and acoustical properties of condensed matter	35.80 Atomic and molecular measurement	
7	85.15 Electronic and magnetic devices; microelectronics	46.15 Computational methods in continuum mechanics	
8	68.40 Chemisorption/physisorption: adsorbates on surfaces	29.17 Electrostatic, collective, and linear accelerators	
9	01.50 Educational aids	85.40 Microelectronics: LSI, VLSI, ULSI; integrated circuit fabrication technology	
10	25.50 Photonic nuclear reactions	84.30 Electronic circuits

We call for domain experts to examine more closely this table and even the results of net and direct ranks of all major countries in each subfield. We will be happy to provide the corresponding data.

4 Conclusion and discussion

In this work, using the IOF calculated from the general input-output analysis (Shen et al., 2016) as a measure of net influence and the direct citation counts as a measure of the direct influence, we discuss the question of whether or not publications from physicists from a country, especially China, have been undercited. We find that 75% percent of German subfields are undercited and 74% for the USA, while China has 40% percent undercited subfields. We also provide a list of such highly undercited or overcited subfields for each country.

Our discovery that there are more overcited subfields than undercited subfields for China implies that often the paper citing papers from Chinese physicists are often with low influence themselves. We have not looked into whether or not it is indeed the case. Doing so will require a paper-level application of the PageRank algorithm or input-output analysis. This will be the topic of future studies.

The data we analyzed in this work is only on physics and only from the APS journals. The method is applicable and should be applied to other disciplines or even all the disciplines together. After all, evaluating and recognizing properly scientific contributions of our own countries properly can be meaningful not only to science policymakers and educators but also to individual researchers and even citizens.

The definition of overciting or under-citing in this work is based on the comparison between direct citation ranks and the IOF rank. We admit that this is not the only way to define overciting or under-citing.

Author contributions

Jinshan Wu (jinshanw@bnu.edu.cn), Liying Yang (yangly@mail.las.ac.cn) and Jinzhong Guo (guojinzhong123@163.com) designed this study. Guo and Xiaoling Liu (liuxiaoling.xmu@163.com) performed the analysis. All participated in writing up the manuscript.

References

- Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN systems*, 30(1–7), 107–117.
- Csomos, G. (2018). A spatial scientometric analysis of the publication output of cities worldwide. *Journal of Informetrics*, 12(2), 547–566.
- Dagostino, M., & Dardanoni, V. (2009). The measurement of rank mobility. *Journal of Economic Theory*, 144(4), 1783–1803.

- King, D.A. (2004). The scientific impact of nations. *Nature*, 430(6997), 311–316. URL: <http://www.nature.com/articles/430311a>. doi:10.1038/430311a.
- Leontief, W. (1941). *The structure of the American economy, 1919–1929* Harvard University Press. Cambridge (new, enlarged edition, Oxford University Press, New York, 1951).
- Meho, L.I., & Yang, K. (2007). Impact of data sources on citation counts and rankings of list faculty: Web of Science versus Scopus and google scholar. *Journal of the American Society for Information Science and Technology*, 58(13), 2105–2125.
- Miller, R.E., & Blair, P.D. (2009). *Input-output analysis: Foundations and extensions*. Cambridge University Press.
- Moed, H.F., & Halevi, G. (2015). Multidimensional assessment of scholarly research impact. *Journal of the Association for Information Science and Technology*, 66(10), 1988–2002. doi:10.1002/asi.23314.
- Shen, Z., Yang, L., Pei, J., Li, M., Wu, C., Bao, J., Wei, T., Di, Z., Rousseau, R., & Wu, J. (2016). Interrelations among scientific fields and their relative influences revealed by an inputoutput analysis. *Journal of Informetrics*, 10(1), 82–97. doi:10.1016/j.joi.2015.11.002.
- Zhang, Q., Perra, N., Gonçalves, B., Ciulla, F., & Vespignani, A. (2013). Characterizing scientific production and consumption in physics. *Scientific Reports*, 3(1), 1640.

This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivs License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).