NBER WORKING PAPER SERIES

THE CONTRIBUTION OF CHINESE DIASPORA RESEARCHERS TO SCIENTIFIC PUBLICATIONS AND CHINA'S “GREAT LEAP FORWARD” IN GLOBAL SCIENCE

Qingnan Xie
Richard B. Freeman

Working Paper 27169
http://www.nber.org/papers/w27169

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2020

We gratefully thank Dr. Xi Hu from University of Oxford for her support in collecting data from Scopus. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

The Contribution of Chinese Diaspora Researchers to Scientific Publications and China's “Great Leap Forward” in Global Science
Qingnan Xie and Richard B. Freeman
NBER Working Paper No. 27169
May 2020
JEL No. F1,I2,J2,J3,J5,O3

ABSTRACT

China-born scientists and engineers who conduct their research outside China, the diaspora researchers of our title, contributed to global science through the exceptional quantity and quality of their scientific work and through distinctive connections to China-based researchers and research. Analysis of the Scopus database of English language scientific journal articles shows that Chinese diaspora research publications are a substantial and growing proportion of global scientific publications, receive an above average number of citations per article, and are published at above average rates in high Scopus CiteScore journals. In addition, diaspora researchers helped China advance to the forefront of science through collaboration on papers with China-based researchers and through the citation network linking China-based research to research outside the country.

Qingnan Xie
Nanjing University of Science and Technology
200 Xiaolingwei Street
Xuanwu District, Nanjing, Jiangsu 210094
China
2362626753@qq.com

Richard B. Freeman
NBER
1050 Massachusetts Avenue
Cambridge, MA 02138
freeman@nber.org

China's “great leap forward' in science and engineering in the first two decades of the 21st century moved the country to the forefront of global scientific research. In 2018 China became the global leader in scientific publications with 25% more papers than the US.[1] From 2000 to 2015, the average number of 3-year forward citations to a China-addressed English language journal article jumped from about one quarter of the world average to just above the world average. The increase in papers and citations per paper combined to raise China's share of citations in the Scopus data base from a negligible 0.7% in 2000 to 17.5% in 2015.[2]

In this study we examine the contribution to scientific literature of a group of *China-born* researchers that has received little attention in China's “great leap forward” in science – *diaspora researchers* who conduct part or all of their research outside China. We identify these researchers by their having Chinese first and last names, and having one or more journal articles where their address is outside China.

If Chinese diaspora researchers were a small group of scientists doing average quality work or if they had a similar connection to China as their non-Chinese named peers, there would be little gain from differentiating their work from that of others. But our data show that far from being modest contributors to science, Chinese diaspora researchers have produced a large and growing body of papers that has gained exceptional numbers of citations and publications in journals with high Scopus CiteScores.[3] Diaspora researchers top lists of most productive scientists, publish extensively in *Nature* and *Science*, and have stronger network links to China-addressed researchers in co-authorship and citations than other scientists and engineers publishing papers with addresses outside of China.

This paper documents these claims. Section 1 defines Chinese diaspora research and measures its contribution to global scientific publications. Section 2 examines the citations and Scopus CiteScores of diaspora papers and performance of diaspora authors. Section 3 describes the collaboration and citation links between diaspora and China-based research. Section 4 concludes.

Section 1: Chinese Diaspora Papers and Researchers

Researchers publish contributions to knowledge in scholarly papers that contain their name, institutional affiliation, and address. Analyses of research contributions focus on the number of published papers, citations received and/or the prestige of the journals of publication. We use author's name and country address in the Scopus database to identify the Chinese diaspora researchers on whom we focus. We use a Chinese last name to identify someone as ethnic Chinese[4] and a Chinese first name to differentiate those likely to be mainland born from those born elsewhere. Our measure identifies Qing Yang as coming from China and David Yang as coming from outside China, and labels a US addressed paper by Qing Yang as a diaspora paper and one by David Yang as a US paper.[5] We define a paper with all non-China addresses as an NC paper and define an NC paper with at least one diaspora (D) author as an *NCD paper* and define an NC paper with no Chinese named author as an *NCN paper.* We further differentiate a paper with China-addressed (C) and non-China addressed (NC) authors as a China Joint paper (CJ), and label CJ papers in which one or more Chinese named author has an NC address as a *CJD paper* and those without a Chinese named author at an NC address a*s CJN papers.*

To measure the "diasporaness" of a paper, we use the proportion of Chinese first and last-named authors with non-China addresses on the paper as indicating membership in the set of CJD diaspora papers. A three authored paper with all authors having Chinese first and second names and addresses outside China would have 100% membership in the diaspora set, while a paper with two Chinese named authors with outside China addresses would have 67% membership, and so on.

Turning to authors, we define a *Chinese diaspora author* as a mainland born researcher who has written at least one paper with an address outside mainland China and examine the proportion of papers in which they have a non-China address. Those who publish entirely at NC addresses are 100% diaspora. Those with a lengthy publication record where all or most papers were at a non-China address are likely to be permanent emigrants. Persons with, say 20% NC addresses on their papers, would be 20% diaspora and would likely be temporary diaspora researchers who wrote those papers during overseas research visits or study. Degree of membership in the diaspora set of authors will vary over time as an author changes the country in which they work. Young researchers who study or do post-docs outside China and write NC papers would have high membership in the NC set early in their careers, which would fall if they returned to permanent jobs in China.

Diaspora papers, 2018

Table 1 records the number of English language journal articles with China addresses and names and the estimated numbers that meet our diaspora definition in the 2018 Scopus database. It also shows those numbers relative to all Scopus journal articles and to China addressed articles.[6] We use Scopus because it is the largest bibliometry of scientific journals, with more journals and wider coverage of China-published English and Chinese language journals than its Web of Science rival. We focus on natural and physical sciences, including engineering and mathematics, as providing a less politicized area of study than social sciences or humanities.

We limit analysis to articles written in English because articles in Chinese (and other non-English languages) face a language barrier that makes them less visible to much of the research community. English language journals constitute about 88% of Scopus journal articles. By contrast, only 4.8% of 2018 articles were in Chinese.[7] Indicative[8] of the language barrier between Chinese and English language publications, articles published in 2015 with only China addresses averaged 1.4 citations if they were written in Chinese, compared to 9.1 citations if they were written in English.[9] Ninety-seven percent of citations to the Chinese language articles came from China-addressed papers.[10]

Line 1 records the number of journal articles with China Only (CO) addresses in the 2018 Scopus data base. These articles constituted 16.8% of all journal publications in that year – a massive increase over the 2.4% of articles that had China Only addresses in 2000.[11] By definition, none of these articles are diaspora articles, though their authors could be diaspora authors by virtue of previously publishing non-China addressed articles.

Line 2 records the number of articles with only non-China (NC) addresses. They constitute a bit over 3/4ths of the 2018 scientific literature. We counted the number of Chinese last-named authors in the NC only addressed articles from the full Scopus data, but had to estimate the proportion of those authors with Chinese first names to meet our diaspora criterion. To do this, we randomly sampled 2,000 NC papers published in 2018 (see Appendix A), and hand checked the first names of all Chinese last-named authors. We found that 79.7% had a Chinese first name, which implied that there were 152,255 non-China addressed diaspora papers (NCD) in 2018. As we had only 324 papers with at least one Chinese last name in the 2018 NC sample, we drew a larger sample of 2,000 NC papers with at least one Chinese last-named author and obtained a virtually identical estimate of 79.9% with Chinese first names.[12] These papers were 9.5% of Scopus articles and were 46.8% as numerous as China addressed articles.

Line 3 gives the number of China Joint (CJ) papers. Those with a first and last-named Chinese author at a non-China address fit our diaspora definition.[13] To estimate the proportion in this situation, we randomly sampled 2,000 CJ papers in 2018, as described in Appendix A, and counted the number of first and last-named Chinese authors. Line 3a gives the estimated number of CJ papers with at least one Chinese last-named author at a non-China address. Line 3b gives the estimated number of CJ papers in which the Chinese last-named author had a Chinese first name and thus were CJD papers.

Line 4a sums the relevant numbers in lines 2 and 3 to obtain the total number of papers with at least one Chinese named author at an NC address. Line 4b estimates the more limited number with first and last-named Chinese authors. The 220,974 papers that met this definition were 13.8% of all Scopus articles and 67.9% as numerous as China-addressed papers.

Taking line 4b's number of diaspora papers and line 1's number of China only (CO) addressed papers gives us a broad measure of China's “presence” in the scientific literature. Figure 1 shows that in 2018, China had a presence on nearly one in three papers – a tripling of its presence in 2000 when one in ten papers had a China name or address with, strikingly, a majority being diaspora papers.

Counting papers with Chinese names and addresses as part of China's contribution to science comparable to China-addressed papers arguably exaggerates the size of diaspora research. The number of diaspora authors on a non-China addressed papers can range from a single person amid many non-Chinese co-authors to all the authors on a paper. The number of China addresses on a China Joint paper can similarly vary from one among many to all but one while the number of Chinese names could range from one name to all names.

To adjust for the “diasporness” of diaspora papers, we calculated the percentage of authors with a Chinese first and last name on the NCD papers in our 2,000 sample of NC papers and come up with an estimate of 37.5%. Multiplying the line 2b number by 37.5% gives the fractional count of NCD papers in line 5a. Similarly, we estimated the percentage of first and last Chinese named authors on CJD papers in our 2,000 sample of CJ papers at 27.6% of the names on those papers. Multiplying the line 3b number by 27.6% gives the fractional count of CJD papers in line 5b. The sum of the fractional counts of NCD papers with non-China address and of diaspora papers on international collaborations in line 5c shows that diaspora papers were 4.7% of all 2018 papers – a proportion which exceeds the fractional count of papers by address for all countries in save for China, the US, and India.[14]

Finally, note that even our fractional counts of the Chinese diaspora share of articles are crude estimates of the country's *contribution* to global publications. They are crude because they *assume* that addresses and names count equally in the contribution. If address better reflected contribution than the Chinese name of an author, more of the diaspora share should go to the addressed country than to China, and conversely if Chinese origin better reflected contribution.[15] Absent a counterfactual model of what would happen to scientific research and publications in China and outside China if Chinese diaspora researchers could not work overseas,[16] our calculations can be viewed as drawing attention to the crediting problem for diaspora researchers that addressed-based measures ignore entirely. They are also crude because they ignore the potential impact of the paper, which we address in Section 2.

Diaspora papers among diaspora researchers

To what extent do diaspora authors write all or most of their papers at a relatively permanent non-China address as opposed to writing just a few papers outside China?

To answer this, we calculated the distribution of papers written inside and outside China for a sample of authors with at least one diaspora paper in 2018. Figure 2 shows that 47.5% of the authors wrote **all** of their papers at a non-China address; while 62.7% wrote over 90% outside China; and 72.1% wrote over 80% outside China. Sampling from 2018 under-represents persons who wrote few diaspora papers and published them in years other than 2018 but the concentration of papers among those primarily writing outside China documents the existence of a permanent or relatively permanent diaspora research group. Given visa requirements for residence in a country, these researchers likely have citizenship/permanent residency in the country where they work.

In the 2010s the diaspora research community was boosted by the huge number of Chinese students earning PhDs and other post-graduate degrees in advanced countries and by the substantial number of Chinese PhDs working as overseas post-docs.[17] Social Security data for the US, which hosts the largest number of Chinese international students, show that 84% of Chinese students who gained US science and engineering PhDs in 2007-2009 and 85% of those who earned US PhDs in 2003-2004 were working in the country in 2013 (Finn and Pennington, 2018: tables 6 and 7). In addition, the US employed many China-educated PhDs as post-docs, all of whom would have US addresses on papers.

Section 2: The Quality of Chinese Diaspora Research

If Chinese diaspora research was close to the global average in research quality the Table 1 numbers would capture its main contribution to science. If the research was above (below) average quality, the Table 1 numbers would understate (overstate) its contribution. To assess the relative quality of diaspora research, we compared 3 year forward citations and the CiteScores of the journals which published diaspora papers with other papers. We chose the 3 year period to provide a reasonable indicator of the likely position of papers in citation distributions over time.[18] We analyze both citations **and** CiteScores because while the two are highly correlated (r = 0.50 in our data),[19] the decision processes for them differ enough to allow for substantial within-journal variation in citations.[20]

On the citation side, individual researchers decide whether or not to cite a paper based on their view of its quality and relevance to their research, among other factors. Articles in high impact journals, in fields with more researchers, and on hot topics gain more attention, raising their chances of being cited. Articles by persons with large research networks in a field, determined by the number of persons similar to them in, say, gender, graduating institution, or national origins (Schubert and Glänzel, 2006; Yan and Ding, 2012; Maliniak at el., 2013; Freeman and Huang, 2015) also gain more citations. And people cite themselves and famous persons ("Matthew Effect") more than others.

The CiteScore of the journal that publishes a paper depends on the quality of the paper and on decisions by the papers' authors on the journal to which they submit the paper and on journal editors' decisions to accept or reject a submission. Since more prestigious journals have low acceptance rates, researchers' submission decision will balance the lower likelihood of getting an acceptance against the potential advantages in prestige and future citations. Journals editors will decide whether to accept or reject a paper based on the reviews of the experts they asked to review it and on the editors' assessment of how journal readers and the scientific community may respond to it. The idiosyncratic factors that affect the CiteScores almost certainly differ from those that impact its future citations.

Panel A of Table 2 presents the mean and median of citations and the mean of citations for the upper decile of the citation distribution for diaspora papers and comparison papers. We give the median and the mean citation in the upper 10% because the distribution of citations is heavy tailed (with power law or related shape) so that means and standard errors do not fully capture its properties. The Table gives statistics for NCD and CJD papers and for NCN, CJN, and CO papers without diaspora authors.

The Table shows that diaspora papers, written at non-China addresses (NCD) or with collaboration from China-addressed authors (CJD), gained roughly **twice** the citations of the NCN papers and CO papers that make up the bulk of scientific publications. The diaspora advantage is larger in mean citations than in median citations, indicating a heavier tail to the diaspora distribution (also evinced in the higher mean of the upper 10% of papers for diaspora papers). NCD papers lead all others in citations in means but CJD papers top them in medians. CJ papers without diaspora authors obtain more citations than CO papers and NC papers without diaspora authors but fall short of the citations received by diaspora papers.

The high number of citations of diaspora research can also be seen in very different statistics on individual authors in lists of top scientists. In 2011 Clarivate Analytics named the "Top 100 Materials Scientists" based on 2000-2010 citations in the Thomson Reuters Web of Science data. Table 3 shows that 5 of the top 10 had Chinese first and last names and worked in the US and thus fit our definition of diaspora authors. All remarkably had undergraduate training from the same Chinese university, which suggests their Chinese education played a strong role in their success".[21] In the entire list, 12 of the top 100 material scientists were diaspora scientists. Clarivate also reported "Top Scientists" in Chemistry, where 3 of the top 10 and 10 of the top 100 were Chinese diaspora researchers.

CiteScores

The CiteScores in panel B of Table 2 relate to citations relative to publications in the 3 *previous* years' publications, which gives them a different time period and metric than the 3 year forward citations. Our CiteScores are from Scopus' 2017 tabulation and thus depend on citations from 2014-2016 while our forward citations are for 2016 to 2018. Consistent with the citation data, diaspora papers gain higher CiteScores than papers with all non-China addresses and no diaspora authors and higher CiteScores than papers with all China addresses.[22]

Finally, to exemplify the growing success of diaspora papers in getting into top scientific journals, we examined their share of papers in *Nature* and *Science* in 2000 and 2018. Table 5 shows that in 2000 *Nature* and *Science* published virtually no papers with China Only addresses and relatively few joint China-other country collaborative papers. But 16.4% of Nature papers and 18.1% of *Science* papers were NCD papers – far in excess of the diaspora share of papers. Between 2000 and 2018 when the CO share of all scientific articles increased massively, the CO share of *Nature* and *Science* articles increased just marginally. The big increase in China's presence in *Nature* and *Science* was in diaspora articles, written entirely at non-China addresses (which reached about one-fourth of papers in *Nature* and *Science*) or in smaller numbers, as CJD papers. In 2018 30.3% of papers in Nature and 35.0% in Science had a diaspora author.

As Chinese authors and addresses are only part of the authors and addresses on diaspora papers, we have fraction counted the diaspora contribution, which necessarily reduces the credit given to China. Even so, in 2018 diaspora papers had a larger share of *Nature* (3.4%) and *Science* (3.9%) articles than the far more numerous China Only papers (0.9% and 2.6%, respectively). Diaspora researchers were in the forefront of quality research from Chinese-named scientists.

Regression estimates of the diaspora quality effect

The Table 2 differences in citations and CiteScores between diaspora papers and other papers could be due to differences in the attributes of papers beyond addresses and names, such as their field of study, number of authors, or other factors associated with citations or publication in more prestigious journals (Börner et al., 2010; Abramo and D’Angelo, 2015). To see whether the Table 2 patterns hold up in analyses that contrast papers identical in measurable determinants of citations and CiteScores, we regressed citations and CiteScores on dummy variables that distinguish 21 scientific fields and on the numbers of authors on a paper as well as on dummy variables for diaspora and other name-address groups relative to NCN papers as the deleted name-address group.

The regressions in Table 5 show that while field and number of authors substantially impact citations and CiteScores, their inclusion in the analysis leaves standing the finding that diaspora papers score higher on these measures than papers written without a diaspora presence. The column 1 regression shows that NCD and CJD papers gain about as many extra citations versus NCN papers and versus China Only papers in the Table 2 mean differences. The column 2 regression shows that addition of the CiteScore of the journal of publication has a huge impact on the estimated effects of diaspora groups reducing the coefficients on NCD and CJD papers by 70-80%. The implication is that much of their advantage comes through publication in higher impact journals. The column 3 regression for CiteScores shows indeed showed diaspora research papers get into higher CiteScore journals than papers with CO addresses or NC addresses with no diaspora authorship.

Given the non-normal distribution of citations and the number of papers with zero citations in the three-year period, the linear regressions do not fully capture the relation between citations and CiteScores and independent variables. Accordingly Appendix B explores four other specifications: (1) a log form regression where one citation is added to the citations for each observation to keep 0 citation papers in the data set; (2) a log regression limited to positive citation observations with a separate equation that estimates the impact of the groups of papers on the probability of positive citations; and (3) a regression in which the citation and CiteScores were scaled into a 0-1 interval by dividing each observation of a variable by its maximum value in the data set; and (4) a power law that regresses the Ln of citations on the Ln rank of citations.[23] All of these forms confirm that diaspora papers produce more citations and higher CiteScores than non-diaspora papers. Whether CJD papers or NCD papers top the citation list varies with specification.

The higher citations and CiteScores associated with diaspora research suggest that the Section 1 estimate of the diaspora contribution to the scientific literature based on numbers of papers understates that contribution. To assess the extent of understatement, we adjusted the numbers of papers for citations using our table 5 regression estimates of the difference in citations between diaspora papers and an NCN paper. Given the sizable regression coefficients on CND and JCD papers, the estimated diaspora contribution increases from 13.8% of papers to 24.9% of papers “citation adjusted”.[24]

Section 3. Impact of Diaspora Research on China-Addressed Science

To what extent does the existence of highly productive China-born researchers working outside China impact research in China?

There are two competing views on how emigration of highly qualified workers affects countries. Traditional brain-drain literature views emigration as a loss that weakens the ability of the source country to upgrade its productive capacity and catch up with economically advanced countries (Docquier and Rapoport, 2012). Emigrants and the countries to which they move benefit at the expense of the countries they left. By contrast, the “ethnic network view” analyzes emigrants to more advanced countries as a positive channel of communication and knowledge that allows the source country to access advances in science and technology more rapidly than otherwise (Kerr, 2008).

The evidence in this section shows that Chinese diaspora research best fits the ethnic network view. We compare the diaspora proportion of nodes linking Chinese research to research outside China to the proportion that we would expect if diaspora researchers had the same ties to China as other scientists working outside China. The comparisons show overwhelmingly that diaspora researchers are more linked to China than other non-Chinese named researchers. Diaspora researchers collaborate more with China-addressed researchers and their papers cite China-addressed papers more, and are cited more by China-addressed papers compared to non-China born researchers working outside China.

International co-authorship

We assess the likelihood that diaspora researchers collaborate more with China-addressed authors than other non-China addressed researchers in two steps.

First, using Table 1 statistics on papers we compare the share of CJ papers that have a diaspora author with the share that would have a diaspora author if non-China collaborations came randomly from all NC papers. Under the null hypothesis that diaspora scientists do not differ in their connection to China than other non-China addressed scientists, the diaspora share of non-China addressed authors in joint China-other country collaborations would approximate the diaspora share of authors on NC papers. Line 1 of Table 1 shows that 12.3% (=152,255 /1,233,660) of NC papers had at least one diaspora author, which would yield a similar proportion of papers with at least one diaspora authors in CJ papers. Instead, the statistics in lines 3a and 3b of Table 1 show that CJ papers with at least one diaspora author make up 69.2% (= 68,719/ 99,316) of CJ papers – a 5.8-fold difference.

Second, counting the number of authors with Chinese first and last names and the number of all authors on NC addressed papers, we estimate that 5.7% of authors on NC papers were born in China (smaller than the diaspora share of NC papers). Taking 5.7% as the likelihood that a random draw of a co-author from the pool of NC addressed authors would have both a first and last Chinese name, we estimated the proportion of diaspora authors on papers with given numbers of non-China addressed co-authors under random selection and compared it to the observed percentage of Chinese first and last named NC addressed co-authors, Figure 3 gives the results for papers with 1-5 non-China addressed co-authors. For CJ papers with just one NC addressed author, 44.3% of CJ papers had a Chinese named co-author with an NC address compared to the 5.7% expected from the random draw. The larger the number of NC addressed co-authors on a paper, the greater the gap between the observed and random distributions. Given the small diaspora proportion of NC authors, virtually no paper with 3 or more non-China co-authors should have 3 or more diaspora co-authors while in fact one in three papers had 3 or more diaspora co-authors.

Finally, looking at CJ papers from 2000 to 2018, Figure 4 shows an upward trend in the Chinese presence on joint papers in terms of both the Chinese share of addresses and diaspora share of non-China addressed authors. The proportion of China addresses, with addresses counted separately for each author, increased steadily from 43.7% in 2000 to 58.6% in 2018. The proportion of NC authors who had Chinese first and last names also increased steadily from 25.5% in 2000 to 34.9% in 2018. Dividing credit for the work of the diaspora authors equally between their names (credit to China) and their address (credit to NC), the two trends raise China's share of credit for the papers from 51% in 2000 to 64.0% in 2018.[25] By this metric CJ papers have become more of a collaboration between China born researchers working outside China and China-based researchers than a 50-50 split between China and non-China based research.

Importing and exporting knowledge through citations

The flow of citations from China-addressed papers to papers written outside China measures the extent to which China-based research “imports” knowledge from the rest of the world, or alternatively the extent to which the rest of world “exports” knowledge to China. Conversely, the flow of citations from non-China addressed papers to papers written in China measures the import of knowledge from China or, alternatively, China's export of knowledge to the rest of the world.

To see if diaspora research has a special role in transmitting non-China based research to China addressed papers, we contrasted the three year forward citations received by NCD and NCN articles published in 2015 from CO articles. Table 6 shows that NCD articles received 2.3 future citations from CO papers while the NCN articles received just 0.9 future citations – a differential of 2.6 to 1.0, which suggests that China addressed researchers import more knowledge from papers with diaspora authors than from papers without those authors due to the greater connection between the diaspora researchers to China. There is, however, an alternative possible explanation for the differential. This is that the high quality of diaspora research articles shown in Section 2 produced the greater number of citations from CO papers to NCD papers than to NCN papers.

We use the citation statistics in the “from NCN” column in Table 6 to adjust for the quality differential. This column shows that, consistent with NCD papers being high quality, NCN papers also cite NCD papers more than other NCN papers. But the rate at which NCN papers give more citations to NCD papers compared to NCN papers is1.6, which is much lower than the rate at which CO papers give more citations to NCD than to NCN papers. To the extent that the NCN differential citing of NCD papers reflects their differential quality, we can use the 2.6 ratio of citations from CO papers relative to the 1.6 ratio for NCN papers – 1.56 – as reflecting the greater connection between diaspora and China-based researchers on citations. By this interpretation, papers with only China addresses are 56% more likely to cite an NCD paper than an NCN paper cites an NCD paper because of the ethnic link between diaspora authors on the NCD papers and China addressed authors.

In Table 7 we apply the same logic to assess the flow of citations from non-China addressed papers to China Only addressed papers. If diaspora researchers are more connected to Chinese research, we would expect them to cite China Only addressed papers more frequently than authors on a non-China addressed paper without a diaspora researcher. The data in line 1 appears at first blush to reject the hypothesis that diaspora papers were more likely to cite CO papers. It shows that China Only addressed papers averaged 2.1 three year future citations from NCN papers compared to 0.6 three year future citations from NCD papers. But diaspora papers made up just 11.9% of NC papers in the 2016-2018 period in which we computed citations,[26] which implies a huge differential in citations due solely to the sizes of the citing populations. To assess the differential *preference* of NCD and NCN papers for citing CO work, we scaled the citations to reflect the different sized citing populations and obtained the average number of CO citations per NCD paper and average number of CO citations per NCN paper in columns 3 and 4.[27] The differential here shows a huge differential of 2.11 to 1.0 in NCD citations of China only papers compared to NCN citations.

Line 2 shows a similar greater diaspora link among papers with both China and non-China addresses but with no Chinese-named authors at a non-China address. NCD papers give relatively more citations per paper to these papers than do NCN papers, by a ratio of 1.48 to 1.00. This is our smallest estimate of NCD/NCN differential diaspora effect on citing papers with a China address.

Line 3 shows that NCD papers cite CJ papers with a diaspora author by a 2.90 to 1.00 ratio over NCN papers. This is our biggest estimate of a diaspora effect on citing papers with a China address. The likely reason is that CJD papers, as the overlap papers between CJ and diaspora papers, reflect the strongest network connections/homophily among those authors.

Finally, recognizing that the univariate analysis in Tables 6 and 7 leaves open the possibility that the link between diaspora papers and China-addressed papers could reflect factors associated with the papers beyond the diaspora connection, we estimated the impact of diaspora papers on citations to and from China-only addressed papers in a regression model that conditons on papers being in the same field of study and having the same number of authors. The results summarized in Appendix C show that while field and number of authors impact citations, they do not substantially change the finding that diaspora papers are more likely to cite and to be cited by China Only addressed papers.

Section 4. Conclusion

Standard assessments of country contributions to scientific publications credit a country for papers in which an author's address is in that country, regardless of the origin of the scientist. Using the names of Chinese scientists to differentiate those likely to have been born in China from those likely to have been born elsewhere, our evidence shows that despite being relatively few in number, Chinese diaspora researchers accounted for a substantial (fraction counted) proportion of journal articles; had a presence on many more articles; gained above average citations and publication in top journals; disproportionately co-authored with researchers in China; and were a key node in the flow of citations between China and the rest of the world.

Since countries govern borders, the development of the Chinese diaspora research community required supportive or permissive migration policies by China as the source country and by destination countries. In contrast to the former Soviet Union which discouraged scientists from engaging with scientists in other countries and viewed emigrant scientists as traitors, China's government began to sponsor and support overseas education and scholarly research visits in the early 1980s, following the end of the Cultural Revolution and Deng Xiaoping's rise to become the paramount leader of China.[28] It continued to support international students and research trips even as many Chinese students and scholars chose to seek permanent residence and citizenship overseas. As tensions over trade, refugee crises and COVID-19 pandemic fears have led to more local nationalist orientation, the diaspora contribution deserves attention as a success of globalization that spread knowledge and talent widely and spurred the growth of global scientific publications even as publications slowed in such traditional scientific leaders as the US and Japan.

Our research raises questions about the diaspora experience and transmission of scientific knowledge across country lines that can illuminate the ways science progresses in a global world. How much do extended stays overseas impact career trajectories in the form of future collaborations or research topics compared to attending or presenting at a single conference (Chai and Freeman, 2019)? Does working or studying overseas matter more for persons from developing countries or from more advanced countries? Is overseas exposure or diversity of backgrounds more useful in some disciplines than in others? To what extent, if at all, does the return of diaspora scientists to their country of origin alter research in that country? Do diaspora scientists from other countries contribute "above their weight" to science as those from China have done and if not, why not? Finally, while we treated the co-authorship network and citation networks as separable processes, analysis of the interactions between these networks[29] and possible synergies and trade-offs between knowledge gained by collaborations compared to reading/citing published studies might suggest ways to go beyond our analogy of citations as "exports" and "imports" in the transfer of knowledge to greater insight into the development of comparative advantage in national expertise in different scientific specialties.

INDEX OF ACRONYMS:

C: China addressed

CJ: China Joint, International collaboration (Papers with at least one Non-China address and with at least one China address)

CJD: Chinese diaspora paper = CJ papers with at least one D author.

CJN: CJ papers without D author

CO: China only addressed papers

D author: Diaspora author; Author with Chinese first and last names and a Non-China address.

NC: Paper with all non-China addresses

NCD: Papers with all non-China addresses and at least one D author

NCN: Papers with all non-China addresses without a D author

References

Abramo, G. and C. A. D’Angelo (2015), ‘The Relationship Between the Number of Authors of a Publication, its Citations and the Impact Factor of the Publishing Journal: Evidence from Italy,’ *Journal of Informetrics*, 9(4), 746–761.

Abrishami, A. and S. Aliakbary (2019), ‘Predicting Citation Counts Based on Deep Neural Network Learning Techniques,’ *Journal of Informetrics*, 13(2), 485–499.

Biscaro, C. and C. Giupponi (2014), ‘Co-Authorship and Bibliographic Coupling Network Effects on Citations,’ *PloS one*, 9(6), e99502.

Börner, K., N. Contractor, H. J. Falk-Krzesinski, S. M. Fiore, K. L. Hall, J. Keyton, B. Spring, D. Stokols, W. Trochim and B. Uzzi (2010), ‘A Multi-Level Systems Perspective for the Science of Team Science,’ *Science Translational Medicine*, 2(49), 49cm245–49cm24.

Bornmann, L., L. Leydesdorff and J. Wang (2014), ‘How to Improve the Prediction Based on Citation Impact Percentiles for Years Shortly After the Publication Date?’ *Journal of Informetrics*, 8(1), 175–180.

Brzezinski, M. (2015), ‘Power Laws in Citation Distributions: Evidence from Scopus,’ *Scientometrics*, 103(1), 213–228.

Callaway, E. (2016), ‘Beat It, Impact Factor! Publishing Elite Turns Against Controversial Metric,’ *Nature News*, 535(7611), 21.

Chai, S. and R. B. Freeman (2019), ‘Temporary Colocation and Collaborative Discovery: Who Confers at Conferences,’ *Strategic Management Journal*, 40(13), 2138–2164.

Chen, X. (2009), ‘Review on China's Policies for Chinese Students Going Abroad Over Last Three Decade,’ *Journal of Xuzhou Normal University*, 35(4), 1–8 (In Chinese language).

Ding, Y. (2011), ‘Scientific Collaboration and Endorsement: Network Analysis of Co-authorship and Citation Networks,’ *Journal of Informetrics*, 5(1), 187–203.

Docquier, F. and H. Rapoport (2012). ‘Globalization, Brain Drain, and Development,’ *Journal of Economic Literature*, 50(3), 681–730.

Finn, M. and L. Pennington (2018), ‘Stay Rates of Foreign Doctorate Recipients from US Universities, 2013,’ [Electronic resource]. Oak Ridge Institute for Science and Education:

United States. Available at: https://orise.orau.gov/stem/reports/stay-rates-foreign-doctorate-recipients-2013.pdf (Accessed: 4, 2020).

Freeman, R. B. and W. Huang (2015), ‘Collaborating With people Like Me: Ethnic co-authorship within the United States,’ *Journal of Labor Economics*, 33(1), 289–318.

Gabaix, X. and R. Ibragimov (2011), ‘Rank − 1/2: A simple way to improve the OLS estimation of tail exponents,’ *Journal of Business & Economic Statistics*, 29(1), 24–39.

Golosovsky, M. (2017), ‘Power-Law Citation Distributions Are Not Scale-Free,’ *Physical Review E*, 96(3), 032306.

Gupta, H. M., J. R. Campanha and R. A. Pesce (2005), ‘Power-Law Distributions for the Citation Index of Scientific Publications and Scientists,’ *Brazilian Journal of Physics*, 35(4A), 981–986.

Kerr, W. R. (2008), ‘Ethnic Scientific Communities and International Technology Diffusion,’ *The Review of Economics and Statistics*, 90(3), 518–537.

Larivière, V., V. Kiermer, C. J. MacCallum, M. McNutt, M. Patterson, B. Pulverer, S. Swaminathan, S. Taylor and S. Curry (2016), ‘A Simple Proposal for the Publication of Journal Citation Distributions,’ *BioRxiv*, No. 062109.

Maliniak, D., R. Powers and B. F. Walter (2013), ‘The Gender Citation Gap in International Relations,’ *International Organization*, 67(4): 889–922.

Miao D., Z. Wei, Y. Bai, M. Long and X. Chen (2009), ‘Memorabilia in the 60 Years of China’s Oversea Education,’ *World Education Information*, 10, 35–40 (In Chinese).

National Science Board, National Science Foundation (2020), *Science and Engineering Indicators 2020: The State of U.S. Science and Engineering*. NSB-2020-1. Alexandria, VA. Available at https://ncses.nsf.gov/pubs/nsb20201/.

National Science Board, National Science Foundation. 2020. *Research and Development: U.S. Trends and International Comparisons. Science and Engineering Indicators 2020. NSB-2020-3*. Alexandria, VA. Available at https://ncses.nsf.gov/ pubs/nsb20203/.

Schubert, A. and W. Glänzel (2006), ‘Cross-National Preference in Co-Authorship, references and citations,’ Scientometrics, 69(2): 409–428.

Scopus Database. Available at: https://www.scopus.com.

Xie, Q., and R. B. Freeman (2019). ‘Bigger Than You Thought: China's Contribution to Scientific Publications and Its Impact on the Global Economy,’ *China & World Economy*, 27(1), 1–27.

Yan, E. and Y. Ding (2012), ‘Scholarly Network Similarities: How Bibliographic Coupling Networks, Citation Networks, Cocitation Networks, Topical Networks, Co-authorship Networks, and Coword Networks Relate to Each Other,’ *Journal of the American Society for Information Science and Technology*, 63(7), 1313–1326.

Table 1 – Numbers of Journal Articles by Address and Names of Authors and Numbers Relative to World and China Papers, 2018

Definition and Source	Number	Number Relative to	
All Journal Articles (n= 1,602, 030)		All articles	China addressed papers*
1. Papers with China Only address (CO)	269,054	16.8%	82.7%
2. Papers with Only Non-China address (NC)	1,233,660	77.0%	379.3%
a) NC Papers with at least one Chinese *last-named* author	191,040	11.9%	58.7%
b) NC Diaspora Papers, estimated from 2,000 NC papers (NCD)	152,255	9.5%	46.8%
3. Papers with at least one C and one NC address (CJ)	99,316	6.2%	30.5%
a) CJ papers with at least one Chinese last name at NC address, estimated from 2,000 CJ papers	83,908	5.2%	25.8%
b) CJ Diaspora papers (CJD), based on % papers with at least one Chinese first & last-named authors at NC address in 2,000 CJ sample	68,719	4.3%	21.1%
4. Papers with Chinese names and Non-China Addresses			
a) NC Papers with at least one Chinese *last*-named author, 2a+ 3a	274,948	17.2%	84.50%
b) NC papers with at least Chinese *first and last*-named author, 2b +3b	220,974	13.8%	67.9%
5. Diaspora Papers Fractional Counts by Chinese Diaspora Proportion of Authors			
a) Fractional Count NC Diaspora Papers, based on 37.5% share of China names on papers from 2,000 NC sample x line 2b	57,093	3.6%	17.6%
b) Fractional Count CJD papers based on 27.6% estimated Chinese names on NC address from 2,000 CJ sample x line 3b	18,951	1.2%	5.8%
c) Fractional Count of all Diaspora Papers (5a + 5b)	76,044	4.7%	23.4%

*China number of papers fractionated by giving China a proportion of each CJ paper dependent on % of authors with China address, with China credited for authors with a C and one or more NC addresses, proportion to China's share of addresses.

Source: Scopus English language journal articles in science, mathematics and engineering. This excludes papers in social sciences; arts and humanities; psychology; business, management and accounting; economics, econometrics and finance; decision sciences, and undefined. Statistics from all Scopus data with estimates from sample of papers described in Appendix A.

Table 2. Average 3 year forward citations and CiteScores of papers published in 2015

Panel A: Citations			
	Mean	Median	Mean for top decile of group
1 NCD – NC (Non-China Only) papers with one or more China named authors	**18.3**	**8.0**	**103.9**
2 **CJD – CJ papers with diaspora author (CJD)**	**17.5**	**10.0**	**85.5**
3 CJN – CJ papers without diaspora author	12.4	7.0	51.2
4 CO – China Only addressed papers	9.1	5.0	37.4
5 NCN – NC papers with no China named author	8.5	5.0	34.3
Panel B: CiteScores			
1 NCD – NC Papers with one or more China named authors	5.0	4.1	14.4
2 **CJD – CJ papers with diaspora author (CJD)**	4.9	4.1	13.6
3 CJN – CJ papers without diaspora author	4.2	3.4	11.0
4 CO – China Only addressed papers	3.1	2.7	8.3
5 NCN – NC papers with no China named author	3.4	2.7	9.3

Note: The standard errors for the means in citations are 0.3, 0.7, 1.0, 0.9, 2.1, 0.3
The standard errors for means of CiteScores are 0.1, 0.1, 0.1, 0.2, 0.2, 0.2
The CiteScore values are assigned to papers based on the CiteScore of the journals in which they appeared. Scopus does not assign a CiteScore to new or inactive journals so observations on those journals are excluded at the CiteScore calculation. We use the 2017 version CiteScore list issued by Scopus, Downloaded at 25 May 2018.

Source: All measures are based on 2000 yearly CO, CJ and NC samples, see Appendix A for details.

Table 3: Top Ten Material Scientists, 2000-10, Ranked by Total Citations

Rank	Name	Current Employer	Bachelor's degree if had China education.	Citations	Papers
1	**Peidong Yang**	Univ Calif Berkeley	University of Science and Technology of China	13,900	36
2	**Younan Xia**	Washington Univ, St. Louis	University of Science and Technology of China	11,936	83
3	**Yiying Xu**	Ohio State	University of Science and Technology of China	9,590	74
4	N. Serdar Sarificitci	Johnnes Kepler Univ, Linz		6,444	74
5	**Yadong Yin**	Univ Calif Riverside	University of Science and Technology of China	6,387	32
6	Alan Heeger	Univ Calif Santa Barbara		5,788	49
7	Frank Caruso	Melbourne		5,589	
8	Michael Huang	National Tsing Hua University, Taiwan		5439	34
9	**Yugang Sun**	Argonne Nat'l Lab	University of Science and Technology of China	**5,231**	**37**
10	Galen Stuckey	Univ Calif Santa Barbara		5,095	72

Note: Our ranking is based on total citations, whereas the Clarivate ranking is based on the ratio of citations to papers, which causes some differences between their statistics and ours. Diaspora researchers are in bold.

Source: Tabulated from *Clarivate Science Watch*, ‘Top 100 Materials Scientists’ http://archive.sciencewatch.com/dr/sci/misc/Top100MatSci2000-10/

Table 4. Chinese Diaspora Papers in *Nature* and *Science*, 2000 and 2018

	2000	2018	2000	2018
	Nature		Science	
Proportion of papers				
China Only addressed papers (CO)	0.3%	0.9%	0.2%	2.6%
Papers without China address but with at least one China named authors (NCD)	**16.4%**	**24.6%**	**18.1%**	**27.0%**
China Joint papers with diaspora authors (CJD)	**0.2%**	**5.7%**	**0.2%**	**8.0%**
China Joint papers without diaspora authors (CJN)	0.2%	3.4%	0.5%	2.1%
Papers without China address or China named author (NCN)	82.8%	65.3%	80.9%	60.3%
Papers with either China address or China name (CO+NCD+CJD+CJN)	17.2%	34.7%	19.1%	39.7%
Proportion of papers, fractional counts by addresses and names				
China Only addressed authors (CO authors)	0.3%	0.9%	0.2%	2.6%
Papers without China address but with at least one China named authors (NCD)	**2.5%**	**3.4%**	**3.1%**	**3.9%**
China Joint papers with diaspora authors (CJD)	**0.1%**	**1.7%**	**0.1%**	**3.2%**
China Joint papers without diaspora authors (CJN)	0.1%	1.5%	0.2%	0.8%
Papers without China address or Chinese named author (NCN)	97.0%	92.5%	96.4%	89.4%
Papers with either China address or Chinese name (CO+NCD+CJD+CJN)	3.0%	7.5%	3.6%	10.6%

Source: Tabulated from every edition of *Nature* and *Science* in the specified year.

Table 5: Regression Estimates and Standard Errors Relating 3 Year Forward Citations and CiteScores of 2015 Papers to Groups of Paper Authors, with Field Variables and Number of Authors

	Citations	Citations	CiteScore
NCD (Diaspora Papers in NC addressed group	9.44 (0.000)	3.54 (0.004)	1.42 (0.000)
CJD (Diaspora Papers in CJ group)	8.55 (0.000)	2.01 (0.016)	1.58 (0.000)
CJN (Papers without Diaspora authors in CJ)	3.88 (0.004)	-0.02 (0.989)	0.94 (0.000)
CO (China Only papers)	1.24 (0.144)	1.86 (0.013)	-0.15 (0.131)
NCN (Papers with no China address and no diaspora authors)	-	-	-
CiteScore	-	4.15 (0.000)	-
Other Factors			
21 Field	yes	yes	yes
#Authors	0.27 (0.000)	0.14 (0.000)	0.03 (0.000)
Adj R-squared	0.0634	0.2753	0.2293
NOB	5318	5318	5318

Note: NCD is the dummy variable of NCD papers; CJD is the dummy variable of CJD papers; CJN is the dummy variable of CJN papers; CO is the dummy variable of CO papers; NCN is our benchmark group. CiteScore value is assigned to a paper based on the 2017 CiteScore value of the journal it published on. The 21 fields are: Multidisciplinary; Agricultural and Biological Sciences; Biochemistry, Genetics and Molecular Biology; Chemical Engineering; Chemistry; Computer Science; Earth and Planetary Sciences; Energy; Engineering; Environmental Science; Immunology and Microbiology; Materials Science; Mathematics; Medicine; Neuroscience; Nursing; Pharmacology, Toxicology and Pharmaceutics; Physics and Astronomy; Veterinary; Dentistry; Health Professions.

Source: Tabulated from a sample of 2,000 CO papers, a sample of 2,000 CJ papers, and a sample of 2,000 NC papers published in 2015. Observations without valid address or name information are omitted, papers are also omitted if the journals they published on haven't been assigned a 2017 version of CiteScores by Scopus, mainly because those journals are newly established. The number of observations for each group are NCD: 364; CJD: 1269; CJN: 401; CO: 1838; NCN: 1446.

Table 6. Three Year Forward Citations to Non-China Addressed Papers published in 2015 from China-Addressed and Non-China Addressed Papers, by presence of diaspora author

Papers published in 2015	Three year forward Citations		Col.1/Col.2
	From CO	From NCN	
NCD papers	2.3	10.5	*0.22*
NCN Papers	0.9	6.4	*0.14*
Row 1/ Row 2	*2.56*	*1.64*	***1.56***
Differential of CO citation of NCD Papers to NCN citation of NCD papers is 1.56			

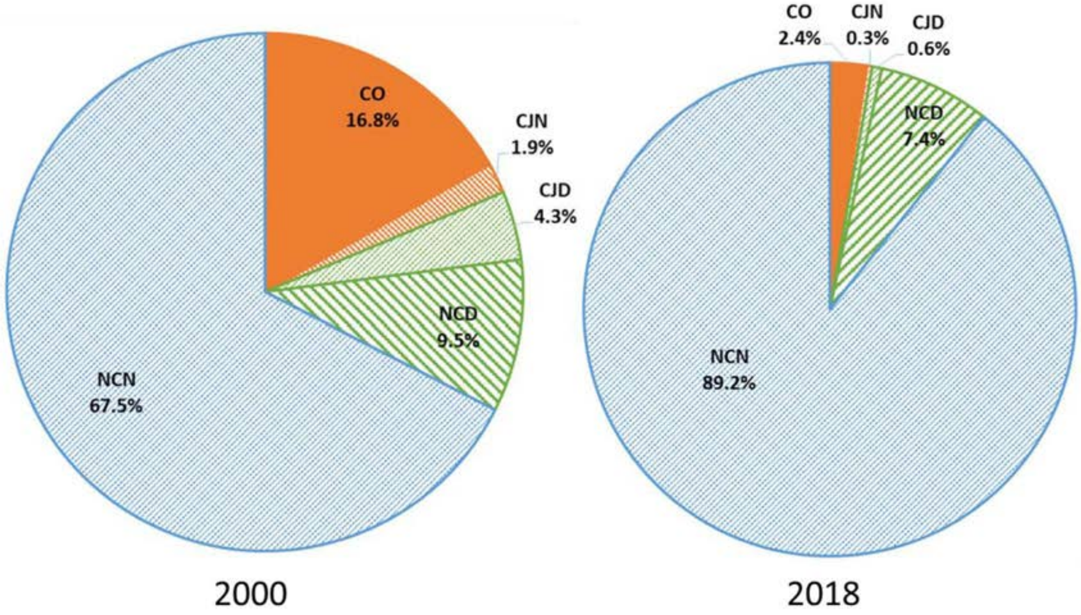
Note: Citations counts are 3-year forward citation counts. Citations to NCD and NCN papers estimated from 2,000 NC papers, described in appendix A.

Table 7. Three Year Forward Citations to China Addressed Papers published in 2015 from Non-China Addressed Papers, by presence of diaspora author

Papers published in 2015	Three Year Forward Citations Received per 2015 Published Paper by type of 2016-2018 citing paper		Three Year Citations Given to 2015 papers per Citing Paper, by type of 2016-2018 citing		**Ratio (Col.3/Col.4)**
	(1)	(2)	(3)	(4)	
	NCD citing	NCN citing	NCD citing	NCN citing	
CO	0.6	2.1	*0.28*	*0.13*	***2.11***
CJN	0.9	4.5	*0.0391*	*0.0265*	***1.48***
CJD	2.2	5.6	*0.25*	*0.09*	***2.90***

Note: Citations counts are 3-year forward citation counts. Column 1 and 2 citations to CO papers are estimated from sample of 2,000 CO papers. Citations to CJD and CJN papers are estimated from sample of 2,000 CJ papers. Columns 3 and 4 multiply the citations per 2015 receiving paper by the number of 2015 papers published in each group, from Table 1, and divided by the estimated number of NCD an NCN papers published in 2016, 2017, and 2018 in a three-step procedure. First, we obtained from the Scopus website using query strings the total number of NC papers which were: 2016: 1,213,200; 2017: 1,215,647; 2018: 1,233,660. Second, we counted the number of NC papers with at least one China last named author in 2016: 185,799; 2017: 187,903; and 2018: 191,040. Third, we estimated ratios of NCD papers to NC papers with at least one China first named author based on a sample of 2,000 NC papers in 2016, 2017, and 2018 described in appendix A with estimates of 2016: 74.5%; 2017: 78.1%; 2018: 79.6%. This produced an estimated number of NCD papers is 138,234 in 2016; 146,376 in 2017; and 152,255 in 2018. The number of NCN papers are #NC papers minus #NCD papers, which is 1,074,966 in 2016; 1,069,271 in 2017; and 1,081,405 in 2018.

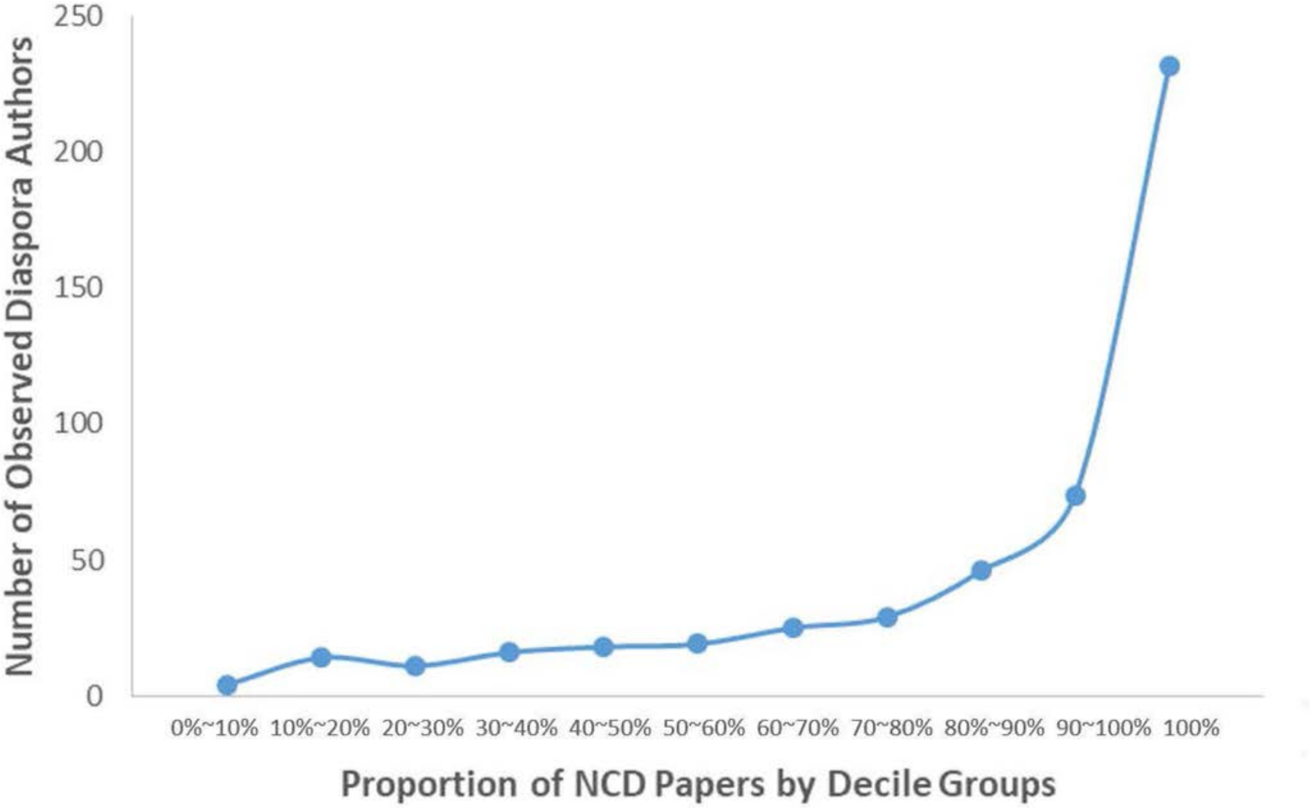
Figure 1: China's Presence in Global Scientific Publications, 2000 and 2018



Notes: The 2018 Figure is drawn based on numbers from Table 1. For the 2000 Figure, the measure of CO is accurate number from Scopus data base, the measure of CJD and CJN are estimated based on a sample of 2,000 CJ papers published in 2000, and the measure of NCD and NCN are estimated based on a sample of 2,000 NC papers published in 2000.

Source: English journal articles in Scopus which are published in 2000 and 2018, samples of CJ papers and NC papers are described in Appendix A.

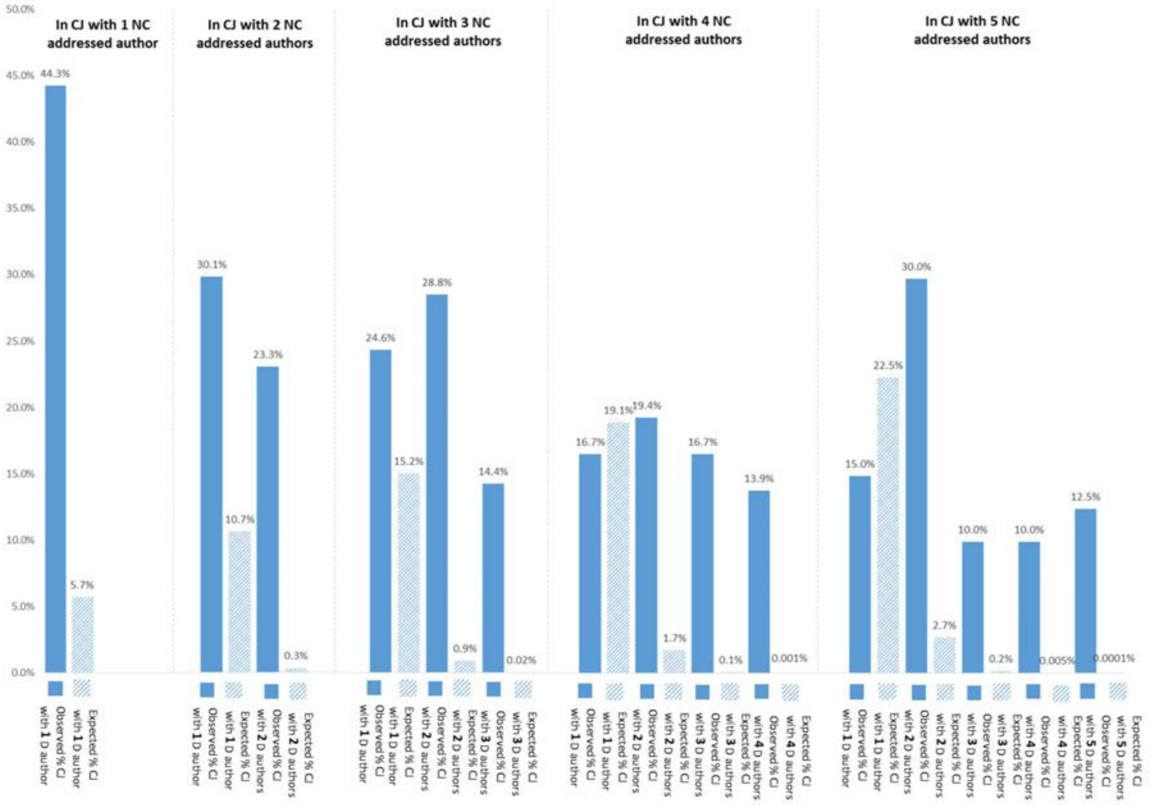
Figure 2: Proportion of Diaspora Authors by Percentage of Lifetime Papers at Non-China Address.



Note: Decile groups defined as having more than or equal to the starting value and less than the ending value – i.e. 0% ~10% means an author has more than or equal to the 0% and less than 10%; 10% ~20% means an author has more than or equal to the 10% and less than 20%.

Source: Based on the lifetime publication of 488 diaspora authors on the sampled 2000 NC papers published in 2018. Only English journal articles are counted when calculating the number of lifetime publication of an author.

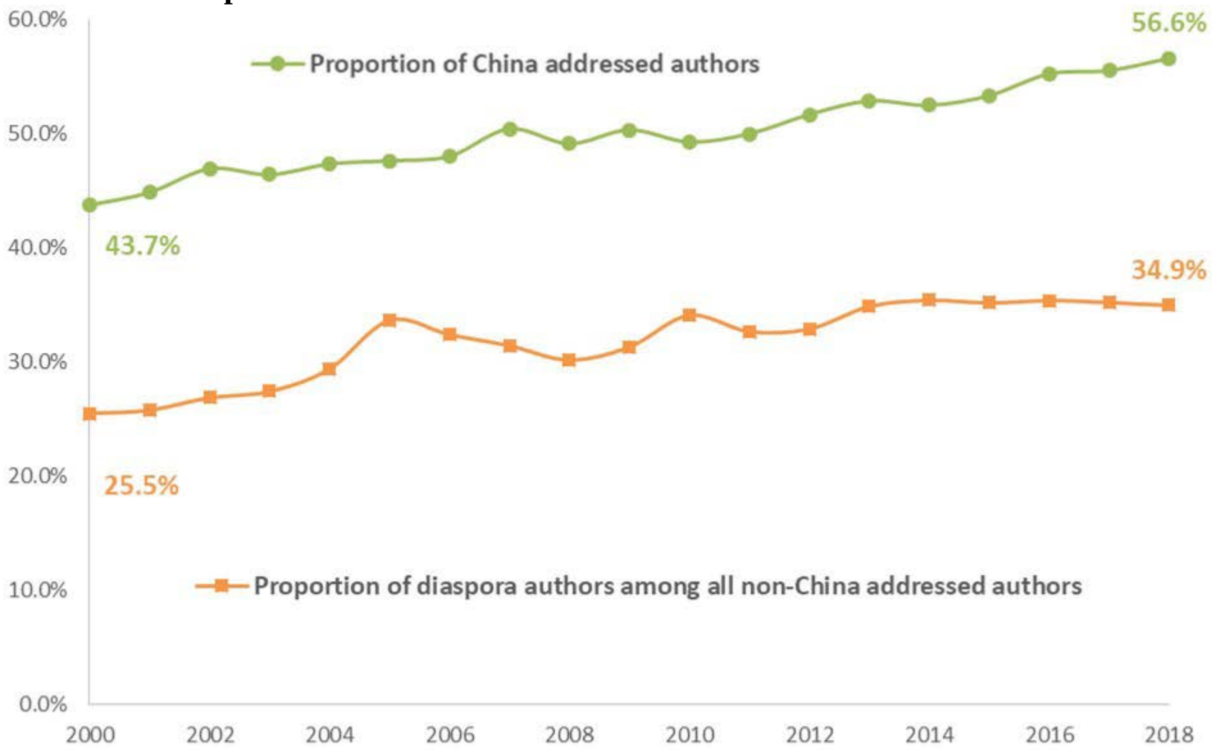
Figure 3: Distribution of Number of Diaspora Authors on China Joint international Collaborative Papers published in 2018, with one to five NC addressed authors.



Note: Actual distribution of numbers on papers with different numbers of NC addressed authors. Expected based on percent of diaspora authors among all NC authors of 5.7%. In cases where one author has 2 or more papers in our sample, the author is counted each time as an author.

Source: Sample of 2,000 NC and CJ papers as described in Appendix A

Figure 4. Proportion of China addressed (C) authors among all authors listed and proportion of diaspora authors among all non-China addressed authors in China Joint Collaborative Papers



Notes: Credit for Authors with more than one address in China and a foreign address prorated by giving ½ to each of the two country addresses. Two authors with the same address count as two addresses. We count the proportion of China address and proportion of diaspora authors among all non-China addressed authors for each paper, then calculate the mean of proportions for each paper in our data set.

Source: Sample of 2,000 CJ English journal articles in Scopus database from 38,000 CJ papers in 2000~2018. Most omitted articles lack valid address; some lack name. See Appendix A

[1] National Science Board, (2020) *Science & Engineering Indicators*, NSB-2020-6 TABLE 5A-1 S&E articles in all fields, for 15 largest producing regions, countries, or economies: 2008 and 2018.

[2] Three-year citations include citations from the same publication year. National Science Board, *Science and Engineering Indicators* (2020) show a similar pattern in its publication output: *US Trends and International Comparisons* (2020), Figure SA-9, with an increase in all year citations for China relative to the world from 0.33 in 1996 to 1.12 in 2016.

[3] To measure the impact of journals Scopus uses CiteScore, the number of citations to the journal in one year to documents published in the three previous years, divided by the number of documents indexed in Scopus in the same three years. See https://service.elsevier.com/app/answers/detail/a_id/14880/supporthub/scopus/. The Web of Science measure of the number of citations received by a journal is its impact factor.

[4] We identify Chinese last-named authors using a list of common Chinese last names from the Chinese Ministry of Public Security household registrations: http://www.mps.gov.cn/, accessed June 26, 2017. This list has last names for 84.8% of the population, leaving out uncommon last names, including some non-Han minority names. We distinguish Mainland Chinese names from other Chinese language speaking areas, by analyzing differences in the Chinese Pinyin spelling (Pinyin is the official romanization system for Standard Chinese in mainland China). We check authors' first names by hand.

[5] This method limits mislabeling country of birth to immigrants who changed their first name to fit the country to which they moved and to native born persons whose parents gave them a mainland first name.

[6] The sum of China Only (CO) addressed articles and fractional counts of articles with both Chinese and non-China addresses proportionate to China's share of the article's addresses.

[7] Scopus contained 350 active Chinese language journals that made it the 2nd largest language in the bibliometry in 2018.

[8] Indicative because differences in research quality almost surely account for a large proportion of the difference in citations.

[9] The average 3-year-citation to all non-English language papers published in 2015 is 1.3, while the average 3-year-citation to all English language papers published in 2015 is 9.5, so the citation pattern goes beyond the Chinese language.

[10] Papers published in 2015 with NC addresses gave just 0.05% of references to Chinese language papers compared to a 6.4% share for China addressed English language articles. These statistics ignore citations to papers from the many Chinese language scientific publications in the China National Knowledge Inventory database (https://en.wikipedia.org/wiki/CNKI) not included in Scopus (Xie and Freeman, 2019).

[11] We obtained the number of CO English language papers published in 2000 in the science and engineering fields from the Scopus online database and divided it by the number of all English journal articles in those fields in 2000.

[12] To check this statistic, we obtained an independent sample of 2,000 papers all with at least one Chinese last-named author. In this sample 79.9% of the papers had at least one China first named author, giving us an estimated number of 152,641.

[13] We prorated the address share of credit for an author with addresses in China and another country on the extent of a paper being diaspora by giving ½ to each of the two country addresses. For example, we credited a paper with one Chinese named author and with both a Chinese and a US address 3/4th to China and 1/4ths to non-China. In an n-authored paper, this gives 3/4n to China and 1/4n to non-China. Because non-Chinese named researchers with China addresses are a negligible part of China addressed papers, we ignored them but their contribution could be divided similarly by names and addresses.

[14] National Science Board (2020) Table 5A-1 shows India with 5.3%, Germany with 4.1% and Japan with 3.9% of papers in 2018. We obtain the same ranking in our compilation of English journal articles in Scopus.

[15] If we used number of citations as the metric of value of a paper, the measure of contribution would be the relative impact on citations of having a Chinese-named author or a China-addressed author.

[16] Such an analysis would have to estimate the likely lower productivity of diaspora scientists if they were in China rather than working outside China and the likely lower productivity of their potential replacements on projects outside China. The impact would likely depend on the size of the counterfactual change in flows.

[17] Ministry of Education of the People's Republic of China data show that from 1978 to 2018, 5.9 million students studied abroad, with 1.5 million students outside China at the end of 2018. Among those 4.3 million students who completed their studies, 3.6 million students returned to China:

http://www.moe.gov.cn/jyb_xwfb/s5147/201909/t20190927_401309.html; http://www.moe.gov.cn/jyb_xwfb/s5147/201909/t20190924_400587.html : accessed on March 2020. Over the same period Institute of International Education data show that China became global leader in international students. https://www.iie.org/research-and-insights/open-doors/data/international-students

[18] Three year forward citation in a sample of 5989 papers published in 2000 had correlations of 0.97, 0.89, and 0.68 to their citations 5, 7, and 10 years in the future. An extensive literature examines ways to predict later citations from early citations and other attributes of papers (Bornmann et al., 2014; Abrishami and Aliakbary, 2019).

[19] This is for a sample of 5,540 papers with valid cite scores published in 2018.

[20] Larivière et al. (2016) analyze the distribution of citations within journals. Callaway (2016) notes that variation in the within-journal variation could potentially be explicable by attributes of the journals, including cite scores.

[21] Requiring a Chinese first name meant that Michael Huang, from Taiwan and educated in part in the US, is not counted.

[22] As cite scores are highly correlated over time, the results should be similar with modestly different year coverage. The correlation for the cite score of Scopus journals is 0.93 between 2017 and 2015, and is 0.87 between 2017 and 2011.

[23] Gupta et al. (2005); Brzezinski (2015); Golosovsky (2017) discuss power laws of citations in Scopus and elsewhere.

[24] The number of diaspora papers was 220,974 (line 5 of Table 1) and the number of all English journal articles was 1,602,030. Using the Table 5 column 1 regression coefficients that estimated the relation between citations and types of papers conditional that papers are in the same fields and have the same number of author, we estimated that diaspora share of citations adjusted for relative number of citations as (NCD+CJD) = [(9.44+7.92)*152,255 + (8.55+7.92)*68,719] / [(9.44+7.92)*152,255 + (8.55+7.92)*68,719 + (3.88+7.92)*30,597 + (1.24+7.92)*269,054 + 7.92*1,081,405] = 24.9%. 7.92 is the mean of NCN deleted group in the regression.

[25] This procedure credits China with 43.7% + (0.5) (0.255) 56.3% = 50.9% in 2000 and with 58.6% + 0.5 (0.349) 41.4% = 64.0% in 2018.

[26] Computed using data in source of Table 7.

[27] Specifically, we multiplied the estimated average number of citations that CO papers received from the NCD and NCN papers in columns (1) and (2) by the total number of CO papers to obtain the total number of citations from NCD and fromo NCN paperts to CO papers. We then divided these numbers by the total number of NCD and NCN papers respectively, to obtain the numbers in columns (3) and (4).

[28] The Tiananmen Incident in April 1976 was followed by the arrest of the Gang of 4 in Oct 1976, and Deng Xiaoping's becoming leader in 1978. In 1977 China restored the college entrance examination system, which had been interrupted for ten years due to the Cultural Revolution and universities began to operate “normally”. The Minister of Education made a goal to send 3000 more Chinese students to study overseas in 1978 and succeeded in sending 4252 government-sponsored Chinese students, including 3006 visiting scholars, 537 graduate students, and 649 undergraduate students. In 1981, the state allowed self-supporting overseas education. After China’s accession to the WTO in 2001, the number of Chinese students/researchers going abroad boomed. (Chen, 2009; Miao et al., 2009).

[29] Ding (2011); Yan and Ding (2012); and Biscaro and Giupponi (2014) examine connections between networks.

Appendix A. The data set of sampled papers.

There are two ways to use data from Scopus in analysis. The first method is to download a file that contains bibliographic data on of papers from the Scopus online website https://www.scopus.com using the Scopus query string (https://service.elsevier.com/app/answers/detail/a_id/11365/c/10545/supporthub/scopus/). The second is to make requests to the server of Elsevier and get the response content through its API (Application programming interface). Downloading files from the first channel does not provide the first names of researchers that we need to differentiate mainland-born persons from citizens or permanent residences born in other countries that meets our definition of diaspora researchers. It also does not give sufficiently detailed data to determine the position of diaspora researchers in the citation network of papers. It records the number of citations a paper receives but little about the citing papers. It also does not report the address or name of authors of the papers in the reference part of a paper.

To extract evidence on those aspects of papers, we undertook a two-part analysis.

First, we randomly selected samples of 2000 articles from the Scopus English journal articles with valid address or name information that are the focus of our study. The query string in Scopus allows 2,000 papers to be downloaded in any query. It reports up to 100 pages of data for each query, with each page containing from 20 to 200 items. We specify the result page to show 100 items per page. To draw the random samples, we generated 20 random numbers between 1~100 from the random function in Excel and used the numbers to select 20 pages with papers for our sample. The 100 papers in each of the 20 pages gives us a sample of 2,000 papers out of the 10,000 items in the query. The downloaded files contain the author name and address information and other bibliographic information – the title of paper, the publication year, and the ISSN number of the journal etc. But they don’t report the first names of authors nor which publication in Scopus cites the selected papers.

Second, using the paper identifier in the downloaded files, we added the desired information to the samples through Elsevier API. We find information on the first names of authors and the papers that cited the paper using the unique identifier assigned to papers in Scopus – eid (see: https://dev.elsevier.com/guides/ScopusSearchViews.htm) and added the first names and the author and address information of the citers of the selected samples via the API portal provided by Elsevier (see: https://dev.elsevier.com/api_docs.html). To get the address and name information of the references in papers in our sample, we accessed the metadata of papers to get the eid code of the references indexed in Scopus through the Elsevier API. We then obtained the detailed address and name information of those cited papers using their eids also through Elsevier API.

The 2000 paper maximum sample that Scopus allowed for an inquiry gives us an adequate number of observations for generalizing to the larger population of all papers. As most of our statistics are counts that we use to compute proportions of papers in different groups, we calculate the sampling error for estimating a proportion in a random sample of 2,000. It is quite small, with a maximum value on the order of 0.006 for a true proportion of 0.50. This allows us to distinguish modest differences in shares of the magnitudes we observe with a high level of significance. As noted in the text, in the case where we had a substantially smaller sample with just 324 persons with Chinese last names in the 2018 NC sample from which to calculate the proportion with Chinese first names, we drew a much larger sample of 2,000 NC papers with at least one Chinese last-named author and obtained virtually identical estimates of the proportion with Chinese first names as in the smaller sample.

Table A-1 lists the data samples that we created. Our focus on diaspora authors meant that we sampled papers with diaspora authors more intensely than papers with all China addresses. The number of 2,000 samples for CJ papers is particularly large because we wanted to track the change over time carefully for a related project. The 2018 sample of NC papers with China last named authors was our check on the estimated proportion of China named authors who also had Chinese first names.

Table A-1

Data Sample	**Purpose**	**Years Covered**	**Total number sampled**
Papers with only non-China addresses	Obtain data on largest group of papers; find those with China first and last names.	2000, 2015, 2016-2018	2,000 in each year for total of 10,000
Papers with only non-China addresses and China last named authors	Get larger sample to estimate the proportion of NC papers with Chinese last and first named author in NC papers with Chinese last-named author	2018	2,000 in year for total of 2,000
China Joint papers with China and other country addresses	Obtain large time series sample on international collaborations	2000-2018	2,000 in each year for total of 38,000
China Only papers	Obtain data on largest group of CO addressed papers	2000, 2015, and 2018	2000 papers in each year for total of 6,000 papers

Table A-2 records the number of cited and referenced papers we developed from our samples for 2015.

Table A-2

Data Sample	**Number of papers**	**Number of papers which cite the sampled papers published in 2015**	**Number of referenced papers of sampled papers published in 2018**
Papers with only Non-China addresses	2,000	19,415	70,561
China Joint papers with China and other country addresses	2,000	32,324	80,433
China only papers	2,000	18,160	76,556

All of the codes and the computer prints for the analysis on request from the authors.

Appendix B: Alternative specifications of citation regression equation.

Table B-1 Regression Ln value of dependent variables on specified independent variables, with citations measures as 1 + actual number of citations. This keeps 0 citation observations in calculations.

	Ln(Citations)	Ln(Citations)	Ln(CiteScore)
NC Diaspora Papers (NCD)	0.36 (0.000)	0.10 (0.062)	0.30 (0.000)
CJ with Diaspora authors (CJD)	0.48 (0.000)	0.13 (0.000)	0.38 (0.000)
CJ with No Diaspora authors (CJN)	0.27 (0.000)	0.06 (0.211)	0.23 (0.000)
China Only papers (CO)	0.06 (0.115)	0.09 (0.002)	-0.04 (0.077)
NC with no Diaspora authors (NCN)	--	-	-
Ln(CiteScore)	-	0.89 (0.000)	-
Other Factors			
21 Field	yes	yes	yes
Ln(#Authors)	0.48 (0.000)	0.21 (0.000)	0.31 (0.000)
Adj R-squared	0.1856	0.4419	0.3302
NOB	5318	5314	5314

Table B-2 Regression of 0/1 dummy variable on whether citations > 0

	Citation dummy	Citation dummy
NC Diaspora Papers (NCD)	0.028 (0.081)	0.004 (0.816)
CJ with Diaspora authors (CJD)	0.084 (0.000)	0.057 (0.000)
CJ with No Diaspora authors (CJN)	0.074 (0.000)	0.0058 (0.000)
China Only papers (CO)	0.025 (0.010)	0.028 (0.004)
NC with no Diaspora authors (NCN)	-	-
CiteScore	-	0.017 (0.000)
Other Factors		
21 Field	yes	yes
#Authors	0.001 (0.003)	0.0006 (0.102)
Adj R-squared	0.0478	0.0752
NOB	5318	5318

Note: The proportion of 0 citations varies: NCD: 8.0%; CJD: 2.8%; CJN: 5.3%; CO: 11.2%; NCN: 12.6%. We dropped 4 papers with 0 CiteScore.

Table B-3 Regression of Ln Citations on Independent Variables for observations with positive citations

	Ln(Citations)	Ln(Citations)
NC Diaspora Papers (NCD)	0.38 (0.000)	0.14 (0.000)
CJ with Diaspora authors (CJD)	0.40 (0.000)	0.13 (0.000)
CJ with No Diaspora authors (CJN)	0.19 (0.001)	0.06 (0.257)
China Only papers (CO)	0.06 (0.152)	0.11 (0.001)
NC with no Diaspora authors (NCN)	-	-

Ln(CiteScore)	-	0.85 (0.000)
Other Factors		
21 Field	yes	yes
Ln(#Authors)	0.42 (0.000)	0.18 (0.000)
Adj R-squared	0.1492	0.3915
NOB	4874	4874

Table B-4 Regression Variables divided by their maximum value in the data set so that they fall between 0-1 interval.

	Scaled Citations	Scaled Citations	Scaled CiteScore
NC Diaspora Papers (NCD)	0.015 (0.000)	0.006 (0.004)	0.047 (0.000)
CJD	0.014 (0.000)	0.003 (0.016)	0.053 (0.000)
CJ with No Diaspora authors (CJN)	0.006 (0.004)	-0.00003 (0.989)	0.031 (0.000)
China Only papers (CO)	-0.002 (0.114)	0.003 (0.013)	-0.005 (0.131)
NC with no Diaspora authors (NCN)	-	-	-
Scaled CiteScore	-	0.197 (0.000)	-
Other Factors			
21 Field	yes	yes	yes
Scaled #Authors	0.159 (0.000)	0.084 (0.000)	0.382 (0.000)
Adj R-squared	0.0634	0.2753	0.2293
NOB	5318	5318	5318

Note: Citation values, CiteScore values, and author number are divided by the maximum citation value, the maximum CiteScore value, and the maximum author number value in our data set, respectively.

Table B-5 Power law regression using Ln (citation) on the Ln (rank of citations), for all 2015 papers with positive citations and for papers in the upper decile of citations,

	Ln(Citations)	Ln(Citations) Top 10% most cited observations
Ln(Rank of Citations)	-0.8788(0.000)	-0.5496(0.000)
Other Factors		
NC Diaspora Papers (NCD)	-0.0089(0.717)	0.0030(0.513)
CJ with Diaspora authors (CJD)	0.0300(0.078)	-0.0037(0.308)
CJ with No Diaspora authors (CJN)	0.0544(0.022)	-0.0025(0.647)
China Only papers (CO)	0.0343(0.024)	-0.0009(0.807)
NC with no Diaspora authors (NCN)	-	-
Ln(CiteScore)	0.1519(0.000)	0.0045(0.094)
21 Fields	yes	yes
Ln(#Authors)	0.1908(0.083)	0.0036(0.072)
Adj R-squared	0.8691	0.9977
NOB	4874	559

Note: Given the relatively small sample of papers in the upper decile, we also estimated the Ln(citations) on the ln(rank-1/2) to reduce bias, per Gabaix and Ibragimov (2011) and obtained similar results to those in the column for the upper decile sample. The lower estimated coefficient in the column for the top 10% of cited papers is markedly smaller than in the column for all papers with citations, reflecting the “fatter tail” at the upper end of the distribution that is found in many estimates of power laws.

Appendix C: Regression Analysis of Citations Between CO and NC Papers

Table C-1 Regression Estimates and Standard Errors Relating 3 Year Forward Citations from CO Papers to 2015 NC Papers, by attributes of cited paper, with Field Variables and Number of Authors

	Citations from CO	Citations from CO
NCD – Diaspora Papers in NC addressed group	2.76 (0.000)	2.16 (0.000)
NCN – NC Papers with no diaspora authors	-	-
Citations from NCN	0.17 (0.000)	0.13 (0.000)
CiteScore	-	0.72 (0.000)
21 Field	yes	yes
#Authors	0.028 (0.647)	-0.01 (0.828)
Adj R-squared	0.2005	0.2397
NOB	1710	1710

Note: Observations are English language journal articles in from Scopus data base. Citations counts are 3-year forward citation counts. Citation data of NCD papers and NCN papers are based on sample of 2,000 NC papers, described in appendix. Observations without valid address or CiteScore information are omitted.

Table C-2: Regression Estimates and Standard Errors Relating 3 Year Forward Citations from NCD Papers to 2015 China addressed Papers, with Field Variables and Number of Authors

	Citations from NCD	Citations from NCD
CJD – Diaspora Papers in CJ group	0.96 (0.000)	0.77 (0.000)
CO – China Only papers	0.076 (0.515)	0.27 (0.015)
CJN – Papers without Diaspora authors in CJ	-	-
Citations from NCN	0.25 (0.000)	0.20 (0.000)
CiteScore	-	0.29 (0.000)
21 Field	yes	yes
#Authors	0.003 (0.406)	-0.001 (0.686)
Adj R-squared	0.4338	0.5022
NOB	3498	3498

Note: Observations are English language journal articles in Scopus data base. Citations counts are 3-year forward citation counts. Citation data of CO papers are based on sample of 2,000 CO papers, and citation data of CJD papers and CJN papers are based on sample of 2,000 CJ papers, as described in appendix A. Observations without valid address or CiteScore are omitted.

The 21 fields are: Multidisciplinary; Agricultural and Biological Sciences; Biochemistry, Genetics and Molecular Biology; Chemical Engineering; Chemistry; Computer Science; Earth and Planetary

Sciences; Energy; Engineering; Environmental Science; Immunology and Microbiology; Materials Science; Mathematics; Medicine; Neuroscience; Nursing; Pharmacology, Toxicology and Pharmaceutics; Physics and Astronomy; Veterinary; Dentistry; Health Professions.