Collaborating With People Like Me:

Ethnic Co-authorship within the US

Richard B. Freeman, Harvard and NBER

Wei Huang, Harvard and NBER

NBER Conference on High-Skill Immigration

October 25, 2012

First draft for comment only

The globalization of science has changed the ethnic and national origin of scientists and engineers in the US. The influx of international students has raised the share of the foreign-born among science and engineering new PhDs. Over half of post-doctorate workers in science labs come from overseas. Expansion of higher education worldwide has increased the supply of non-US educated scientific researchers, which contributes to the flow of immigrant scientists and engineers to the US. Political shocks such as the collapse of the Soviet Union have led to sudden influxes of scientists and engineers to the US job market.[1]

This paper examines the pattern of ethnic collaborations among US-based researchers and the relation between ethnic collaborations and the impact factor of the journals in which collaborators publish their papers. We measure ethnicity by the names of coauthors on scientific papers. Our analysis of names shows a huge increase in the share of US-based scientific authors from developing economies, particularly China, and a corresponding decrease in the share with traditional English names. We find that researchers are more likely to work with people of the same ethnicity than would arise by chance – homophily in co-authorships.[2]

The pattern of homophily in scientific publications could reflect the productivity advantages of such arrangements – the greater ease of communication and trust associated with similar language and culture.[3] It could also reflect the greater probability of meeting persons of the same ethnicity due to

[1] John Bound, Sarah Turner, Patrick Walsh, "Internationalization of U.S. Doctorate Education" in Science and Engineering Careers in the United States: An Analysis of Markets and Employment (2009), Richard B. Freeman and Daniel L. Goroff, editors (p. 59 – 97). George J. Borjas, Kirk B. Doran "The Collapse of the Soviet Union and the Productivity of American Mathematicians" NBER Working Paper No. 17800 February 2012

[2] -Homophily refers to the "birds of a feather flock together" pattern in which people of similar backgrounds congregate together. Such behavior is found in many areas of social life: marriage, residence, business partnerships, seating arrangements in university dining halls, and so on. See Deepak Hegde, New York University, and Justin Tumlinson, Ifo Institute at the University of Munich, "Can Birds of a Feather Fly Together? Evidence For the Economic Payoffs of Ethnic Homophily" for an analysis of the economics of homophily in capital investments.

[3] Tanyildiz, (2008) shows that students from a given country are more likely to enroll in universities with faculty from their native country, are more likely to be in labs labs that are directed by foreign-born faculty are more likely to be populated by students from the same country of origin than are labs that are directed by native (U.S. born) faculty. and with the number of existing students from their country at a the university, and attributes this both Tanyildiz, Zeynep Esra, "The Effects of Networks on Institution Selection by Foreign Doctoral Students in the U.S." (2008). Public Management and Policy Dissertations. Paper 25. http://digitalarchive.gsu.edu/pmap_diss/25

ethnic groups congregating together in other social spheres[4] or it could reflect tastes/preferences for working with persons like oneself, which are likely to reduce scientific productivity by adding non-productivity factors to the formation of scientific teams. Our evidence suggests that homophily is associated with less valuable scientific work, as measured by the impact factor of the journals in which a paper appears, compared to the work by teams that have a more diverse ethnic mix. We also find, however, that ethnicity has effectively no impact on the persistence of collaborations, which depend largely on the impact factor of the journal which published the co-authored paper.

This paper is organized as follows. The next section describes the data used and how the ethnic composition of US-based researchers changes over time. Section 2 introduces the homophily measures and provide evidence for homophily in co-authorship. Section 3 shows the relation between authors' characteristics and homophily. Section 4 presents the how homophily is associated with papers' impact factor. Section 5 shows the role of homophily in team persistence. Finally, Section 6 concludes.

1. The changing ethnic composition of US-based researchers

To measure the ethnic composition of researchers in the US, we undertook a two-step procedure. First, we created a database from the Thomson Reuters Web of Science of papers with two, three and four authors with addresses in the US at the time of the paper for 1990 to 2003. We limited our sample to US-based authors so that the persons could encounter each other at scientific seminars and conferences or other scientific events in the country - the first step in deciding to collaborate on a project. We limited the sample to authors of 2-4 papers so that each author was likely to have decided to collaborate rather than to have participated as part of a huge team in which preference for collaborating with persons like themselves might not enter the decision. Over half of all co-authored

[4] The standard model of job search model provides a valuable framework for analyzing this issue. The model shows that the appropriate strategy for job search over a distribution of jobs differing in their level of pay (or other characteristics) is to form a reservation wage and to accept the first job above the reservation wages.

papers have between 2 and 4 authors.[5] Our sample contains 930,188 papers. The data set provides authors' complete surnames, initials of first names, authors' addresses6. The total number of the papers increased by 79 percent from 47,640 in 1990 to 84,159 in 2003. When we extend the analysis through 2008, we will be able to use first names for the most recent years to improve our measures.

Second we applied a program developed by Bill Kerr that matches the names of persons to their likely ethnicity. This program uses names and MSAs to determine the likely ethnicity of authors. Names such as Kim are far more likely to represent Korean people than any other, while names like Zhang are likely to be Chinese. Because persons of a particular ethnicity are more likely to live in some MSAs than others, MSA information helps distinguish ethnicity among people as well. Ethnicity is divided into nine categories: Chinese (CHN), Anglo-Saxon/English (ENG), European (EUR), Indian/Hindi/South Asian (HIN), Hispanic/Filipino (HIS), Japanese (JAP), Korean (KOR), Russian (RUS) and Vietnamese (VNM)7. We matched names to ethnicity at a rate of 74%.[8] This is lower than that in usual matching with both given names and surnames (about 95%). [9]

Table 1 presents the distribution of authors in two- three- and four- author papers by ethnicity in our data set. The sum of statistics in a row equals to one. From 1990 to 2003, the proportion of Chinese authors nearly doubles from 7.2 to 14.1 percent while the proportion of Anglo-Saxon/English names falls from 52.9 percent to 43.5 percent, and percentage of European names drops from 13.5 percent to 11.9 percent. The proportions of other ethnic groups also increased. If we take the ethnicity groups

[5] Richard Freeman, Ina Ganguli, and Raviv Goroff-Muriciano "International Collaboration in Research" (The Changing Frontier: Rethinking Science and Innovation Policy, Pre-Conference, NBER, Oct 26, 2012) find that 55% of coauthored papers in nano-technology, 58% of coauthored papers in particle physics and 52% of coauthored papers in biotechnology and applied microbiology have 2, 3, or 4 authors.

[6] Since some authors may have multiple addresses and different authors may have different ones, we are not able to identify the address for a specific author in a paper because the addresses of the authors are pooled together.

[7] More details can be found in William R. Kerr and William F. Lincoln, "The Supply Side of Innovation: H-1B Visa Reforms and US Ethnic Invention," The Journal of Labor Economics 28:3 (July 2010), 473-508. and William R. Kerr, "Ethnic Scientific Communities and International Technology Diffusion," The Review of Economics and Statistics, 90:3 (August 2008), 518-537.

[8] We identify both authors in 2-authored papers at73.0%; identify at least two of the three-author papers in three-authored papers at 73.1% and identify 3 or four authors of four-author papers at 74%.

[9] In the small number of cases where we have multiple addresses we use the average ethnicity distribution of those in the different geographical positions.

exclusive of Anglo-Saxon, European, Russian, and Japanese as representing names of persons with developing country backgrounds, the percentage of persons from those backgrounds increases from 16.6% in 1990 to 28.8% in 2003. The percentage of authors whose ethnicity the program cannot identify is stable in the time span, ranging from 12.0 to 12.5.

The distributions in the table do not distinguish between American-born persons of a given ethnicity and foreign-born persons nor have any information on the citizenship status of the foreign born. Given the huge number of persons from China earning science and engineering PhDs in the US and their propensity to remain in the US for many years, the huge increase in Chinese names in table 1 is likely driven largely by persons born overseas rather than by US-born Chinese. Assuming that the first names of the Chinese born in the US are more Anglicized than of Chinese born in China, we can assess the importance of international students and immigrants on the increased share of scientific authors with Chinese names. Names given persons born in China are far more likely to have initials with the letters Z, Y, Q and X persons than persons born in the US, whose first names are often Anglicized. For example someone born in US might be named Richard Wang (217 people listed on white pages http://names.whitepages.com/Richard/Wang) whereas someone born in China might be named Xia Wang (58 people in white pages with this name http://names.whitepages.com/Xia/Wang. The name-ethnicity program shows that 0.3 percent of English names have Z, Y, Q, X first initials compared to 24.2 percent of Chinese names. Using the full distribution of initials of English names and Chinese names and assuming that US-born Chinese use English first names, we estimate that the vast bulk of the increase in Chinese names is associated with the upward trend in the number of Chinese born researchers working in the US – international collaborations within the same country that measures of international collaborations have traditionally ignored. [10]

[10] We can compare the names of persons with China addresses with that of the Chinese in the US to get a potentially better measure of the differentiation of place of birth on the initials, but the results will almost surely give a similar pattern.

2. Measuring homophily

To determine the extent of homophily we compare the observed distribution of co-authors by the ethnicity of their names to a predicted distribution of co-authors by name that would arise if co-authorship was determined by random draws from an urn that contained names with the distribution of names in table 1.

Columns 1 – 4 of table 2 record the actual ethnicity distribution of authors differentiated by the position of the authors in the paper. In most scientific fields, the first-author is the junior person who did the most work on the paper while the last author is the senior person in whose laboratory the work was conducted. Intermediate positions reflect contributions of researchers who contributed in other ways. Some journals require the paper to give the ways in which the various authors contributed, which we do not treat in this study. The data in columns 1 and 2 shows that in two-author paper sample, 17.6 percent of the first authors and 9.8 percent of second ones are Chinese, while 48.3 percent of the first authors and 59.3 of the second authors are Anglo-Saxon/English. This presumably reflects the entry of young Chinese researchers into US research compared to Anglo-Saxon/English researchers who tend to be older senior investigators. Column 5 gives the realized probability of the authors having the same ethnicity regardless of position in the paper.

The key to measuring homophily in co-authorship is to develop a counter-factual distribution of authors that assumes that they have no preference for collaborating with people like them. Column 6 records the probability that the authors belong to the same ethnicity if they randomly chose their coauthors from the pool of authors by ethnicity in table 1. The probability of authors having the same ethnicity is just the product that persons of their ethnicity would be first authors, second authors, and where relevant third and fourth authors.

If authors are more likely to coauthor with same-ethnicity persons, the statistics in column 5

should be larger than those in column 6.[11] The results give strong evidence for the homophily phenomenon in co-authorship: researchers are more likely to coauthor with the others who have the same ethnicity. With the large samples of authors and papers in our data set, the differences are statistically significant in all cases save for the four-authored papers for Vietnamese named persons, where the sample size is too small. The absolute differences tend to be largest for the largest groups whereas relative differences tend to be larger for smaller groups. Appendix table B uses the separate distributions of ethnicity for authors by position on the paper to examine homophily conditional on an author's position in the paper. It contrasts the probability that a first author from a given ethnicity has a second author of the same ethnicity and the converse that a second author of a given ethnicity has a first author of the same ethnicity. The conditional probabilities also show considerable homophily.

As noted at the outset, a diverse set of factors can induce people of similar ethnicity to work together. To control for some of these factors and go beyond the random distribution in table 2, we developed a set of random models that differentiated persons by geographic location and field. In this analysis someone residing in, say San Francisco, where many Chinese reside, would be more likely to have a Chinese co-author than someone in Houston, and someone in scientific specialties with many Chinese specialists would be more likely to have a Chinese co-author, and so on. Appendix Table C summarizes the results of this more refined counter-factual and finds again strong evidence of homophily.

Because the decisions of all co-authors to collaborate is necessary for a collaboration, the homophily found in table 2 and appendix tables B and C could result from persons in each of the groups preferring to work with persons of their ethnicity or from persons in only one of the groups preferring to work with persons of their ethnicity. Without additional information or assumption, the general equilibrium aspects of co-authorship make it impossible to determine the impetus behind the

[11] Except for the last row because there is no Vietnamese authors coauthoring together in reality due to very small population.

observed pattern. To illustrate the issue, consider the random distribution that would arise in a two-authored paper from two ethnic groups. If 50% of authors were in group A and 50% in group B, the random distribution would predict that ½ of authors would write with persons from their own group (¼ all A co-authorship and ¼ all B co-authorship) and ½ would collaborate with someone from the other group. If group A cared about ethnicity but group B did not, the distributions for both A and B would show more persons in both groups working with their own group than in the random model. Similarly if group B cared about ethnicity but group A did not or if both groups cared about ethnicity. Finding that each group in the table co-authors with members of their own group more than in the random model does not identify whether that reflects their preferences or that of another group.

The existence of more than two ethnic groups can help identify the magnitude of preferences for working with persons of a similar ethnicity. Assuming that the only force at work is preference for one's own group, the magnitude of deviations from the random pattern can identify the roles of differences in preferences for working with one's own group in creating the overall pattern of homophily among groups. If one third of authors are in each of three groups, A, B, and C and the only group that prefers to work with itself is A, the deviation from the random pattern will be largest of A as the secondary effects will be divided between B and C. With groups of different sizes, there is a comparable but more complex computation.

3. Researcher characteristics and homophily

To see the characteristics that cause some researchers to work with persons of their ethnicity while others work with persons of other ethnic backgrounds, we estimated a linear probability model that relates dichotomous variables measuring whether or not a paper was written by authors of the same ethnicity or with other ethnic patterns to the characteristics of authors, in particular their previous publishing record. We measure authors' previous publishing record by the average impact factor and

the number of previous papers. We also include dummy variables for the number of different addresses of authors in the paper. Finally, in each calculation we include dummy variables for the state in which the last author is located, dummy variables for each of the ethnic groups in Table 2, dummy variables for the field of the paper as determined by the WOS category for the journal which published it, and dummy variables for the year of publication. The state and field variables are designed to eliminate potential factors that might affect the probability of co-authorship independent of preference for collaborating with people of the same ethnicity.

Table 3 presents the estimated coefficients and their standard errors for papers with two, three and four authors separately. The last row of the table gives the mean values of the dependent variables. The most striking result is that the characteristics of the last author, who is presumably the principal investigator or most senior person, has the greatest impact on homophily. The estimated coefficient on the average impact factor of the previous papers of last authors is negatively associated with homophily in all of the columns: principal investigators who publish in more visible journals are more likely to connect with co-authors outside of their ethnic group. By contrast, the coefficients on first author's previous impact factor are generally insignificant (the exception is for four-author papers) and the coefficients on intermediate position author' impact factors are unrelated to homophily.12 The results for number of previous papers tells a similar story: the estimated coefficient on the dummy variables for numbers of papers are negatively correlated with homophily measures for all authors, with again the last author's numbers having the largest effect. If you write more papers you are less likely to coauthor with same ethnicity researchers. Taken together the results on impact factors and numbers of papers imply that researchers with better publishing track records are less likely to co-author with persons of the same ethnicity.

[12] We identify authors by surnames and initials of first names, so there may be some name disambiguation problems here that we will examine further, but this should just add measurement error to the analysis and is unlikely to affect the pattern by the position of authors on the paper. The probability of having other authors with same identifiers should be random across the authors in different positions. To help with the disambiguation we used the field of the journal publication as an additional identifier and obtained similar results.

The majority of co-authored papers in our sample are located at the same institution: 86% of the authors of two-author papers, 77% of the authors of three-author papers, and 71% of the authors of four-author papers report a single address. On the two author papers, 13% report two addresses, while the remaining 1% report three addresses, due to some authors having two addresses. Similarly on the three and four author papers, most have two addresses. Thus, the key coefficients in table 3 on numbers of addresses are those for two addresses compared to the reference group of a single address. Our initial expectation was that researchers would find it easier to collaborate with authors of different ethnicity if they were located at the same place, so that papers with authors having two addresses would more likely be written by persons of the same ethnicity. This is consistent with the result for two-author papers where the coefficient on having two addresses increases the likelihood that authors would have the same ethnicity, but is not consistent with the results for three-author and four-author papers, for reasons we do not understand. Since the WOS does not link authors with their addresses but simply lists all addresses in a separate field, the SCOPUS data base, which records the address of each author separately, may provide greater insight into this pattern.

4. Characteristics of the paper and homophily

Do researchers working with persons of the same ethnicity write papers that have greater or lesser impact than researchers with different ethnicity? The table 3 finding that homophily is associated with last authors' having lower average impact factors on previous papers suggests that the impact factor of the current paper will be negatively related to same ethnicity collaborations because of the characteristics of the last author. To isolate the effect of homophily from the characteristics of authors on the productivity of the paper, we regressed the impact factor of the current paper on measures of the ethnicity of authors and on measures of the characteristics of the authors, including the impact factors of previous papers. We also include measures of the period of time during which

persons have collaborated, defined as the time span between the first co-authorship and the final one in our sample and a variable for the year of the first co-authorship because the more recent it is, the less likely that we will observe another co-authorship. There are two reasons to expect longer collaborations to be more productive than shorter collaborations. Authors may gain person-specific collaborative experience. And authors will maintain successful collaborations for longer periods,

Table 4 reports the estimated coefficients and standard errors from this analysis. The principal result is that the dummy variables for authors with the same ethnicity have negative effects on the impact factors of publications even with the inclusion of measures for authors' previous impact factor and number of papers13. This result is robust across different measures of homophily in authorship. The regressions further show that the length of time that people have been co-authors is positively related with co-authorship period. Conditional on the length of co-authorship, the relation between times of co-authorship and impact factor varies in across the author-number samples. In two-author paper, the impact factor is higher when the authors coauthor twice or more than twice. In the three-author paper sample, the estimated coefficient on the impact factor of the papers written by the authors who co-authored twice is positive but the coefficient is negative for those who co-authored three or more times. For the four-author paper, we find that the number of co-authorship times is negatively associated with impact factor.

Finally, in all the calculations, the previous impact factors of the co-authors are positively associated with the current impact factor, with last authors' impact factor having the largest and most significant coefficient. The dummy variables for the number of the papers that authors have written are also positively associated with current paper's impact factor and the coefficients are larger for the dummy variable for ten and above papers than for one to ten papers except for the first author. More experienced authors who have published in higher impact journals in the past are more likely to publish

[13] We add these as control variables because we have learned from Table 3 that the junior researchers are more likely to coauthor with same-ethnicity ones. The coefficients of homophily measures are larger if we exclude the previous impact factor and number of papers.

the current paper in higher impact journals as well.

The impact factor of the journal in which a paper is published is far from an ideal measure of the productivity of the collaboration. A better measure might be citations to a paper, which are positively correlated with impact factor, but which require a longer time period to obtain. In addition, the impact factor downgrades the quantity of papers. A collaboration the produces two papers and places them in lower impact journals may advance science as much or more as a collaboration that produces one paper and places it in a higher impact journal. Finally, same-ethnicity researchers may be doing important research relevant to their home country. For example, Chinese researchers working on the atmosphere quality in Beijing would more likely to publish in a Chinese oriented journal than in a higher impact factor US or British journal, whereas an American or English researcher doing precisely the same research on the atmosphere quality in London or Washington DC would publish in a higher impact British or American journal.

These issues notwithstanding, the evidence in table 4 indicates that same-ethnicity collaborations have lower scientific productivity at least in terms of the impact factors of their papers than do multi-ethnic collaborations.

5. Persistence of teams

Another way to assess the success of a team collaboration is whether the team continues to work together and produce papers in the future. Members of a team judge presumably judge the likely science that they can produce working together compared to what each might produce individually. If one coauthor sees a better opportunity working with someone beyond their most recent co-author, they presumably will end the current collaboration. There is no obvious way for the author who views the co-authorship as desirable to him or her to reallocate credit to maintain the collaboration. A scientific collaboration differs in this way from business partnerships or marriage where the partners can

reallocate profit shares or other benefits in favor of the person with better outside option and maintain the relation.

To see whether collaborations of co-ethnic teams are more or less likely to persist over time, we examine whether persons in our sample continue collaboration after the first time they publish a paper together. We regressed a dummy variable for whether authors who collaborate together once collaborate a second time on measures of same ethnicity, the characteristics of the authors, and the impact factor of the initial collaboration.

Table 5 records the estimated coefficients and standard errors in this analysis for two-authored papers. On average 16.2% of initial collaborations produce an additional collaboration through the period covered by our data. Column 1 shows that conditional on the current paper impact factor, same ethnicity has no discernible relation to the continuance of a co-authorship. This result holds if we add measures of the impact factor of the authors' earlier papers, which enter the regression with substantial negative coefficients. This pattern suggests that authors may choose to continue a co-authorship if the paper they wrote together gets into a higher impact journal than their previous papers. We modeled this by using past impact factors to predict the likely impact factor of a collaboration and then entered the predicted impact factor and the difference between the actual impact factor and the predicted impact factor into the equation for whether or not the collaboration persisted. The coefficients in column 3 on the predicted impact factor and the difference between the actual and predicted impact factor are both positive. The implication is that a co-authorship is likely to continue if authors anticipate a successful outcome based on their track record of publications and if the current publication beats the expectation.

The pattern for three-author papers and four-author papers (not reported in the table) is more complex. In three-author papers there is little relation between homophily and the continuance of co-

authorship.[14] But four-author papers show a different pattern that we do not yet understand.[15]

6. Conclusion

Our analysis of the authors of scientific papers shows significant homophily in scientific collaborations and suggests that the pattern is associated with lower scientific productivity as measured by the impact factor of the journals in two ways. The last author on papers written with persons of the same ethnicity publish in lower impact journals than the last author on papers written with persons of different ethnicity. The impact factor of papers written by persons with the same ethnicity tend to be lower than the impact factors of papers written by persons with different ethnicity. By contrast, co-ethnicity has no impact on continuance of two-authored papers, where the dominant factor in determining continuance is the impact factor of the paper from the collaboration.

[14] The co-ethnicity of the first and last author and all authors in three-author papers has no impact on the continuance.
[15] One way to get greater insight into the role of ethnicity and other factors in decisions to pursue or end a scientific collaboration would be to follow up the authors of the many co-authored papers that did not produce an additional co-authorship: did the authors of papers written with persons of the same ethnicity disproportionately choose persons of the same ethnicity in any follow-up research or did they follow the pattern in the random distribution model? We leave that analysis for future work.

Table 1: The distribution of authors by ethnicity

Year	Number of Papers	Distribution of authors by ethnicity									
		Chinese (CHN)	Anglo-Saxon/ English (ENG)	European (EUR)	Indian/Hindi/ South Asian (HIN)	Hispanic/ Filipino (HIS)	Japanese (JAP)	Korean (KOR)	Russian (RUS)	Vietnamese (VNM)	Others (OTHER)
1990	47640	0.072	0.529	0.135	0.048	0.033	0.026	0.011	0.020	0.002	0.124
1991	50478	0.078	0.519	0.135	0.048	0.034	0.027	0.011	0.020	0.002	0.125
1992	51567	0.086	0.510	0.133	0.052	0.033	0.028	0.012	0.021	0.002	0.124
1993	52001	0.096	0.500	0.130	0.051	0.034	0.027	0.012	0.023	0.002	0.123
1994	52405	0.101	0.490	0.129	0.054	0.035	0.027	0.013	0.023	0.002	0.125
1995	54877	0.106	0.484	0.128	0.055	0.036	0.026	0.014	0.025	0.002	0.123
1996	61732	0.110	0.478	0.127	0.056	0.037	0.025	0.014	0.025	0.002	0.125
1997	64166	0.113	0.474	0.127	0.056	0.037	0.024	0.015	0.027	0.003	0.124
1998	84463	0.116	0.473	0.127	0.057	0.038	0.023	0.015	0.028	0.003	0.120
1999	82210	0.119	0.466	0.126	0.057	0.040	0.023	0.016	0.030	0.003	0.121
2000	82788	0.123	0.460	0.125	0.056	0.040	0.023	0.017	0.030	0.003	0.122
2001	80859	0.129	0.451	0.123	0.058	0.040	0.024	0.019	0.032	0.003	0.121
2002	80843	0.132	0.444	0.121	0.060	0.041	0.024	0.020	0.033	0.003	0.122
2003	84159	0.141	0.435	0.119	0.063	0.042	0.023	0.021	0.032	0.003	0.121
Total	930188	0.112	0.474	0.127	0.056	0.038	0.025	0.016	0.027	0.003	0.123

NOTES: Only US papers are kept. The "Others" are those names not identified. Because we use surnames to match only, the match rate is 74%, which is lower than ususal match rates when both first and last names are available. For two-author papers, we keep those papers in which both authors are identified; in three-author sample, we keep those with at least two authors identified; in four-author papers, we only keep those with at least three identified. The summary statistics by number of authors can be found in Appendix Table A.

Table 2: Homophily Results

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Ethnicity	Authors' ethnicity distribution by position (%)				Probability of authors same ethnicity (%)		
	First	Second	Third	Fourth	Random	Realized	Difference (7) - (6)
Panel A: Two-author paper							
CHN	17.60	9.80			1.73	4.67	2.94
ENG	48.30	59.30			28.68	32.33	3.65
EUR	13.30	15.00			1.99	2.50	0.51
HIN	7.30	6.30			0.46	1.57	1.11
HIS	4.30	3.40			0.15	0.44	0.29
JAP	2.60	1.70			0.04	0.42	0.37
KOR	2.60	1.10			0.03	0.18	0.15
RUS	3.60	3.10			0.11	0.47	0.36
VNM	0.30	0.20			0.00	0.01	0.01
Panel B: Three-author paper							
CHN	14.20	10.70	6.90		0.10	1.51	1.41
ENG	39.60	45.40	50.10		9.03	10.63	1.61
EUR	11.40	11.90	13.00		0.17	0.25	0.07
HIN	6.20	5.40	4.70		0.02	0.29	0.28
HIS	4.10	3.70	3.10		0.00	0.10	0.10
JAP	2.80	2.20	1.60		0.00	0.19	0.19
KOR	2.10	1.50	0.90		0.00	0.06	0.06
RUS	2.80	2.70	2.40		0.00	0.06	0.06
VNM	0.30	0.30	0.20		0.00	0.00	0.00
Panel C: Four-author paper							
CHN	14.10	12.10	9.60	6.70	0.01	0.95	0.94
ENG	40.90	45.10	48.40	52.00	4.63	6.33	1.70
EUR	11.80	12.00	12.70	13.60	0.02	0.06	0.03
HIN	6.00	5.20	4.60	4.30	0.00	0.10	0.10
HIS	4.40	4.20	3.80	3.30	0.00	0.07	0.07
JAP	3.70	3.10	2.80	2.10	0.00	0.21	0.21
KOR	2.00	1.60	1.40	0.90	0.00	0.03	0.03
RUS	2.80	2.50	2.40	2.40	0.00	0.02	0.02
VNM	0.30	0.30	0.30	0.20	0.00	0.00	0.00

NOTE: The differences between columns 6 and 7 are significant except for the final row.

Table 3: Homophily and characteristics of coauthorship

VARIABLES	(1) Two-author paper	(2)	(3)	(4)	(5)	(6)	(7)
		Three-author paper			Four-author paper		
	Authors same ethnicity	Authors same ethnicity	Over half authors (2+) same ethnicity	First and last author same ethnicity	Authors same ethnicity	Over half authors (3+) same ethnicity	First and last author same ethnicity
Mean of dependent variables	0.43	0.13	0.61	0.28	0.08	0.37	0.30
Average impact factor of previuos papers							
First author	0.0531	-0.0398	-0.0437	0.0218	-0.0581**	-0.149***	-0.00936
	(0.0437)	(0.0251)	(0.0354)	(0.0311)	(0.0256)	(0.0410)	(0.0386)
Last author	-0.330***	-0.142***	-0.353***	-0.149***	-0.118***	-0.464***	-0.233***
	(0.0414)	(0.0247)	(0.0347)	(0.0305)	(0.0247)	(0.0396)	(0.0373)
Second author		0.0169	0.0257	-0.0139	-0.0281	-0.0625	-0.0273
		(0.0251)	(0.0353)	(0.0310)	(0.0255)	(0.0409)	(0.0385)
Third author					-0.0393	0.0277	0.0565
					(0.0250)	(0.0401)	(0.0378)
Number of first author's previous papers							
None (Reference)							
One to ten	-0.403**	-0.374***	-0.395***	-0.503***	-0.529***	-1.171***	-0.714***
	(0.190)	(0.108)	(0.152)	(0.134)	(0.111)	(0.178)	(0.168)
Ten and above	-1.450***	-0.757***	-0.273	-0.721***	-0.952***	-1.160***	-0.972***
	(0.243)	(0.146)	(0.205)	(0.180)	(0.150)	(0.241)	(0.227)
Number of last author's previous papers							
None (Reference)							
One to ten	-1.486***	-0.701***	-1.611***	-1.073***	-0.934***	-2.045***	-1.462***
	(0.239)	(0.134)	(0.189)	(0.166)	(0.139)	(0.222)	(0.209)
Ten and above	-2.408***	-0.892***	-2.222***	-2.028***	-1.056***	-2.497***	-2.544***
	(0.259)	(0.147)	(0.207)	(0.182)	(0.153)	(0.245)	(0.230)
Number of second author's previous papers							
None (Reference)							
One to ten		-0.601***	-1.113***	-0.570***	-0.506***	-1.366***	-0.392**
		(0.114)	(0.160)	(0.141)	(0.113)	(0.181)	(0.171)
Ten and above		-1.126***	-1.136***	-0.364**	-0.840***	-1.863***	-0.415*
		(0.139)	(0.196)	(0.172)	(0.146)	(0.234)	(0.220)
Number of third author's previous papers							
None (Reference)							
One to ten					-0.299**	-1.144***	-0.424**
					(0.116)	(0.187)	(0.176)
Ten and above					-0.588***	-1.456***	0.274
					(0.142)	(0.227)	(0.214)
Multiple addresses							
One (Reference)							
Two	0.715***	-0.717***	-1.121***	-0.641***	-0.895***	-2.229***	-1.248***
	(0.228)	(0.114)	(0.161)	(0.141)	(0.112)	(0.180)	(0.170)
Three	1.721	-0.0807	-0.112	-0.293	-0.622***	-1.331***	-0.527*
	(1.136)	(0.250)	(0.352)	(0.310)	(0.208)	(0.333)	(0.313)
Four and above	3.845	1.508	-0.778	0.402	-0.479	0.997	0.161
	(3.413)	(1.072)	(1.510)	(1.327)	(0.447)	(0.716)	(0.675)
Constant	43.56***	27.73***	72.93***	41.74***	22.30***	45.10***	43.17***
	(0.892)	(0.544)	(0.767)	(0.674)	(0.567)	(0.909)	(0.856)
Observations	283,749	388,753	388,753	388,753	257,686	257,686	257,686
R-squared	0.336	0.330	0.357	0.419	0.242	0.403	0.413
State dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes
All authors' ethnicity	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Publish year dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Field dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes

NOTE: Standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1. The coefficients are interpreted as percentage because all dependent variables are multiplied by 100.

Table 4: Relation of ethnic and other characteristics of coauthorship to impact factor

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Two-author paper	Three-author paper			Four-author paper		
VARIABLES	Impact factor						
Mean of dependent variables	2.24	2.52			2.86		
Authors same ethnicity	-0.0618***	-0.0808***			-0.127***		
	(0.0104)	(0.0130)			(0.0209)		
Over half authors same ethnicity			-0.0759***			-0.114***	
			(0.00921)			(0.0130)	
First and last authors same ethnicity				-0.0674***			-0.0492***
				(0.0105)			(0.0138)
Coauthorship period	0.0157***	0.0548***	0.0550***	0.0548***	0.0489***	0.0493***	0.0491***
	(0.00358)	(0.00551)	(0.00551)	(0.00551)	(0.0122)	(0.0122)	(0.0122)
Same authors co-authored twice	0.161***	0.0304**	0.0309***	0.0304**	-0.0418**	-0.0421**	-0.0424**
	(0.0118)	(0.0119)	(0.0119)	(0.0119)	(0.0199)	(0.0199)	(0.0199)
Same authors co-authored 3+ times	0.148***	-0.0367**	-0.0369**	-0.0373**	-0.127***	-0.128***	-0.128***
	(0.0152)	(0.0179)	(0.0179)	(0.0179)	(0.0341)	(0.0341)	(0.0341)
Average impact factor of previuos papers							
First author	0.148***	0.139***	0.139***	0.139***	0.124***	0.124***	0.124***
	(0.00243)	(0.00203)	(0.00203)	(0.00203)	(0.00271)	(0.00271)	(0.00271)
Last author	0.407***	0.387***	0.387***	0.387***	0.366***	0.366***	0.366***
	(0.00230)	(0.00200)	(0.00200)	(0.00200)	(0.00262)	(0.00262)	(0.00262)
Second author		0.154***	0.154***	0.154***	0.116***	0.116***	0.116***
		(0.00203)	(0.00203)	(0.00203)	(0.00270)	(0.00270)	(0.00270)
Third author					0.174***	0.174***	0.174***
					(0.00265)	(0.00265)	(0.00265)
Number of first author's previous papers							
None (Reference)							
One to ten	0.107***	0.109***	0.108***	0.109***	0.0721***	0.0714***	0.0724***
	(0.0108)	(0.00885)	(0.00885)	(0.00885)	(0.0118)	(0.0118)	(0.0118)
Ten and above	0.0825***	0.0888***	0.0891***	0.0890***	0.0512***	0.0510***	0.0519***
	(0.0138)	(0.0119)	(0.0119)	(0.0119)	(0.0160)	(0.0160)	(0.0160)
Number of last author's previous papers							
None (Reference)							
One to ten	0.528***	0.543***	0.542***	0.543***	0.541***	0.540***	0.542***
	(0.0134)	(0.0109)	(0.0109)	(0.0109)	(0.0147)	(0.0147)	(0.0147)
Ten and above	0.698***	0.713***	0.712***	0.713***	0.724***	0.723***	0.724***
	(0.0148)	(0.0120)	(0.0120)	(0.0120)	(0.0162)	(0.0162)	(0.0162)
Number of second author's previous papers							
None (Reference)							
One to ten		0.0821***	0.0817***	0.0822***	0.0218*	0.0209*	0.0223*
		(0.00927)	(0.00927)	(0.00927)	(0.0120)	(0.0120)	(0.0120)
Ten and above		0.143***	0.143***	0.144***	0.0415***	0.0404***	0.0424***
		(0.0114)	(0.0114)	(0.0114)	(0.0155)	(0.0155)	(0.0155)
Number of third author's previous papers							
None (Reference)							
One to ten					0.156***	0.155***	0.156***
					(0.0124)	(0.0124)	(0.0124)
Ten and above					0.246***	0.245***	0.246***
					(0.0151)	(0.0151)	(0.0151)
Multiple addresses							
One (Reference)							
Two	-0.0856***	-0.0418***	-0.0420***	-0.0416***	0.0496***	0.0482***	0.0501***
	(0.0127)	(0.00924)	(0.00924)	(0.00924)	(0.0119)	(0.0119)	(0.0119)
Three	-0.0413	-0.0762***	-0.0762***	-0.0763***	-0.0243	-0.0250	-0.0238
	(0.0630)	(0.0202)	(0.0202)	(0.0202)	(0.0220)	(0.0220)	(0.0220)
Four and above	0.00467	0.0322	0.0304	0.0313	-0.0547	-0.0530	-0.0541
	(0.189)	(0.0867)	(0.0867)	(0.0867)	(0.0473)	(0.0473)	(0.0473)
Constant	-0.377***	-0.552***	-0.519***	-0.546***	-0.645***	-0.622***	-0.652***
	(0.0500)	(0.0446)	(0.0450)	(0.0447)	(0.0615)	(0.0616)	(0.0616)
Observations	283,749	388,753	388,753	388,753	257,686	257,686	257,686
R-squared	0.445	0.482	0.482	0.482	0.473	0.473	0.473
State dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes
All authors' ethnicity	Yes	Yes	Yes	Yes	Yes	Yes	Yes
The year started to co-author dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Publish year dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Field dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes

NOTE: Standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1.

Table 5: Continue co-authorship, same ethnicity and impact factor (Two-author paper)

	(1)	(2)	(3)
	Two-author paper		
VARIABLES	Continue co-authorship after the first time (Yes = 1, and multiplied by 100)		
Mean of dependent variables		16.21	
All authors same ethnicity	-0.0779	0.0677	-0.0182
	(0.196)	(0.196)	(0.197)
Current paper's impact factor	0.710***	0.624***	
	(0.0332)	(0.0346)	
Predicted impact factor			1.275***
			(0.0878)
Impact factor Diff (Realized - Predicted)			0.624***
			(0.0350)
Average impact factor of previuos papers			
First author		-0.204***	
		(0.0445)	
Last author		-0.128***	
		(0.0430)	
Number of first author's papers			
Zero (Reference)			
One - Ten		-0.421**	
		(0.187)	
Ten and above		-0.532**	
		(0.243)	
Number of last author's papers			
Zero (Reference)			
One - Ten		6.083***	
		(0.196)	
Ten and above		11.40***	
		(0.233)	
Multiple addresses			
One (Reference)			
Two	-3.837***	-3.786***	-3.730***
	(0.206)	(0.206)	(0.206)
Three	-4.346***	-4.328***	-4.229***
	(0.943)	(0.950)	(0.943)
Four and above	-6.423***	-6.038**	-6.235**
	(2.422)	(2.398)	(2.422)
Constant	25.11***	24.57***	24.68***
	(0.925)	(0.928)	(0.928)
Observations	217,838	217,838	217,838
R-squared	0.040	0.049	0.040
Ethnicity variables	Yes	Yes	Yes
State dummies	Yes	Yes	Yes
Field dummies	Yes	Yes	Yes
Publish year dummies	Yes	Yes	Yes

NOTE: Standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1. The observations of further co-authorship are dropped. Previous impact factor is the average of the impact factor that the author's papers before the co-authorship. We use same variables and regressions in Table 4 to predict the impact factor. The coefficients are interpreted as percentage because all dependent variables are multiplied by 100.

Appendix Table A: The distribution of authors by ethnicity and number of authors

Year	Number of Papers	Distribution of authors by ethnicity									
		Chinese	Anglo-Saxon/ English	European	Indian/Hindi/South Asian	Hispanic/ Filipino	Japanese	Korean	Russian	Vietnamese	Others (Not identified)
Panel A: Two-author paper (Keep the papers in which both authors identified)											
1990	14685	0.092	0.602	0.150	0.059	0.032	0.025	0.014	0.024	0.002	
1991	15252	0.100	0.589	0.150	0.059	0.035	0.026	0.015	0.025	0.002	
1992	15657	0.112	0.579	0.148	0.064	0.033	0.024	0.014	0.024	0.002	
1993	15583	0.122	0.565	0.145	0.064	0.035	0.025	0.015	0.026	0.002	
1994	15342	0.124	0.560	0.146	0.065	0.034	0.023	0.017	0.028	0.002	
1995	16153	0.133	0.545	0.143	0.067	0.037	0.024	0.018	0.031	0.002	
1996	17755	0.135	0.540	0.145	0.071	0.038	0.021	0.017	0.031	0.002	
1997	18632	0.140	0.539	0.137	0.071	0.039	0.022	0.017	0.033	0.002	
1998	27270	0.140	0.535	0.139	0.070	0.041	0.019	0.018	0.034	0.003	
1999	26262	0.145	0.528	0.141	0.068	0.042	0.020	0.018	0.035	0.003	
2000	25751	0.148	0.523	0.140	0.069	0.040	0.018	0.020	0.038	0.003	
2001	25219	0.153	0.516	0.136	0.069	0.042	0.021	0.021	0.039	0.003	
2002	24857	0.155	0.505	0.139	0.073	0.042	0.021	0.023	0.039	0.004	
2003	25331	0.166	0.494	0.134	0.075	0.043	0.020	0.024	0.041	0.003	
Panel B: Three-author paper (Keep the papers in which at most 1 author not identified)											
1990	20310	0.067	0.503	0.129	0.046	0.032	0.025	0.010	0.019	0.001	0.168
1991	21283	0.074	0.491	0.129	0.047	0.032	0.026	0.011	0.019	0.002	0.169
1992	21858	0.081	0.482	0.127	0.050	0.032	0.026	0.011	0.021	0.002	0.168
1993	22125	0.091	0.478	0.124	0.049	0.033	0.025	0.011	0.023	0.002	0.164
1994	22107	0.098	0.463	0.122	0.054	0.033	0.024	0.013	0.022	0.002	0.169
1995	23036	0.102	0.462	0.122	0.054	0.034	0.023	0.013	0.024	0.002	0.164
1996	26352	0.105	0.455	0.121	0.054	0.035	0.022	0.014	0.024	0.002	0.168
1997	27055	0.106	0.450	0.122	0.055	0.035	0.022	0.014	0.026	0.003	0.167
1998	34815	0.109	0.451	0.120	0.054	0.036	0.020	0.015	0.027	0.003	0.165
1999	33940	0.110	0.442	0.119	0.057	0.038	0.020	0.015	0.029	0.003	0.167
2000	34375	0.115	0.437	0.119	0.055	0.040	0.020	0.016	0.028	0.003	0.167
2001	33142	0.122	0.426	0.118	0.057	0.038	0.020	0.019	0.032	0.003	0.166
2002	33188	0.124	0.423	0.114	0.058	0.039	0.021	0.019	0.032	0.003	0.166
2003	35167	0.132	0.412	0.115	0.063	0.040	0.020	0.020	0.031	0.003	0.164
Panel C: Four-author paper (Keep the papers in which at most 1 author not identified)											
1990	12645	0.065	0.518	0.134	0.044	0.034	0.029	0.010	0.019	0.002	0.144
1991	13943	0.071	0.513	0.133	0.045	0.035	0.030	0.010	0.019	0.002	0.143
1992	14052	0.077	0.505	0.132	0.047	0.035	0.031	0.011	0.019	0.002	0.141
1993	14293	0.088	0.491	0.130	0.047	0.035	0.032	0.012	0.021	0.002	0.142
1994	14956	0.092	0.483	0.130	0.050	0.036	0.033	0.012	0.022	0.002	0.139
1995	15688	0.098	0.478	0.127	0.049	0.037	0.031	0.013	0.024	0.002	0.141
1996	17625	0.102	0.474	0.125	0.049	0.039	0.031	0.013	0.024	0.003	0.141
1997	18479	0.107	0.467	0.127	0.050	0.040	0.028	0.014	0.025	0.003	0.141
1998	22378	0.110	0.461	0.126	0.051	0.039	0.028	0.015	0.026	0.003	0.141
1999	22008	0.114	0.456	0.124	0.050	0.041	0.028	0.016	0.028	0.003	0.140
2000	22662	0.118	0.450	0.122	0.051	0.041	0.029	0.017	0.028	0.004	0.141
2001	22498	0.123	0.441	0.121	0.054	0.042	0.029	0.018	0.029	0.003	0.140
2002	22798	0.128	0.434	0.119	0.055	0.043	0.028	0.019	0.030	0.003	0.139
2003	23661	0.138	0.428	0.117	0.056	0.043	0.027	0.021	0.029	0.003	0.139

NOTES: Only US papers (all the authors have a US address) are kept. The "Others" are those names not identified.

Appendix Table B: Realized probability of same ethnicity authorship compared to random model, conditional on ethnicity

(1)	(2)	(3)	(4)	(5)	(6)	(7)
Ethnicity	Conditional on the first author's ethnicity (%)			Conditional on the last author's ethnicity (%)		
	Position	Other authors' ethnicity		Position	Other authors' ethnicity	
		Realized probability	Probability if random		Realized probability	Probability if random
Panel A: Two-author paper						
CHN	2	27.00	9.80	1	47.90	17.60
ENG	2	67.40	59.30	1	54.90	48.30
EUR	2	18.80	15.00	1	16.70	13.30
HIN	2	21.70	6.30	1	25.00	7.30
HIS	2	10.50	3.40	1	13.20	4.30
JAP	2	15.80	1.70	1	25.20	2.60
KOR	2	7.40	1.10	1	18.10	2.60
RUS	2	13.70	3.10	1	16.10	3.60
VNM	2	3.30	0.20	1	5.10	0.30
Panel B: Three-author paper						
CHN	2	30.10	10.70	1	41.20	14.20
	3	20.10	6.90	2	35.70	10.70
ENG	2	51.00	45.40	1	42.90	39.60
	3	54.20	50.10	2	48.60	45.40
EUR	2	14.70	11.90	1	13.20	11.40
	3	15.10	13.00	2	13.70	11.90
HIN	2	16.80	5.40	1	19.60	6.20
	3	14.80	4.70	2	17.70	5.40
HIS	2	12.80	3.70	1	12.70	4.10
	3	9.60	3.10	2	11.80	3.70
JAP	2	23.60	2.20	1	28.30	2.80
	3	16.60	1.60	2	21.50	2.20
KOR	2	13.00	1.50	1	19.40	2.10
	3	8.40	0.90	2	16.40	1.50
RUS	2	11.50	2.70	1	10.80	2.80
	3	9.30	2.40	2	9.80	2.70
VNM	2	1.90	0.30	1	1.80	0.30
	3	1.20	0.20	2	2.00	0.30
Panel C: Four-author paper						
CHN	2	34.30	12.10	1	42.50	14.10
	3	25.50	9.60	2	36.70	12.10
	4	19.80	6.70	3	33.30	9.60
ENG	2	52.40	45.10	1	44.50	40.90
	3	53.80	48.40	2	48.50	45.10
	4	56.70	52.00	3	51.80	48.40
EUR	2	16.10	12.00	1	13.90	11.80
	3	15.70	12.70	2	13.90	12.00
	4	16.20	13.60	3	15.30	12.70
HIN	2	16.40	5.20	1	19.50	6.00
	3	13.10	4.60	2	16.30	5.20

	4	14.00	4.30	3	15.90	4.60
	2	16.10	4.20	1	15.70	4.40
HIS	3	13.10	3.80	2	13.50	4.20
	4	11.60	3.30	3	13.00	3.80
	2	33.00	3.10	1	36.70	3.70
JAP	3	27.40	2.80	2	29.40	3.10
	4	20.60	2.10	3	29.20	2.80
	2	16.20	1.60	1	21.30	2.00
KOR	3	12.30	1.40	2	17.30	1.60
	4	9.70	0.90	3	15.80	1.40
	2	12.60	2.50	1	10.10	2.80
RUS	3	10.10	2.40	2	8.30	2.50
	4	8.30	2.40	3	8.40	2.40
	2	1.90	0.30	1	1.70	0.30
VNM	3	1.20	0.30	2	1.00	0.30
	4	1.00	0.20	3	2.50	0.30

Appendix Table C: Ratios of the proportion of ethnicity of different groups compared to extended random model, (homophily index) for author's of given ethnicity

(1)	(2)	(3)	(4)	(5)	(6)
First or last authors' ethnicity		Homophily Indexes			
Given one's ethnicity	The other one's ethnicity	All	Two-author	Three-author	Four-author
	CHN	**1.54**	**1.51**	**1.57**	**1.57**
	ENG	0.83	0.83	0.83	0.82
	EUR	0.85	0.85	0.85	0.84
	HIN	0.91	0.93	0.90	0.89
CHN	HIS	0.80	0.81	0.79	0.79
	JAP	0.98	0.99	0.99	0.97
	KOR	1.04	1.05	1.03	1.03
	RUS	0.82	0.83	0.82	0.83
	VNM	1.22	1.20	1.23	1.26
	CHN	0.91	0.90	0.92	0.92
	ENG	**1.12**	**1.11**	**1.11**	**1.13**
	EUR	1.08	1.08	1.07	1.09
	HIN	0.96	0.95	0.96	0.98
ENG	HIS	1.03	1.04	1.02	1.02
	JAP	0.90	0.93	0.91	0.87
	KOR	0.95	0.96	0.95	0.93
	RUS	1.03	1.03	1.03	1.04
	VNM	1.02	1.01	1.03	1.02
	CHN	0.90	0.90	0.91	0.90
	ENG	1.05	1.06	1.04	1.06
	EUR	**1.15**	**1.15**	**1.14**	**1.16**
	HIN	0.97	0.95	0.96	0.99
EUR	HIS	1.05	1.05	1.06	1.04
	JAP	0.89	0.92	0.90	0.85
	KOR	0.89	0.89	0.89	0.90
	RUS	1.08	1.07	1.08	1.08
	VNM	1.01	1.01	1.01	1.03
	CHN	0.91	0.92	0.91	0.89
	ENG	0.85	0.83	0.87	0.87
	EUR	0.89	0.87	0.90	0.91
	HIN	**1.65**	**1.62**	**1.65**	**1.70**
HIN	HIS	0.92	0.94	0.90	0.93
	JAP	0.89	0.90	0.90	0.89
	KOR	0.95	0.99	0.95	0.89
	RUS	0.94	0.95	0.93	0.95
	VNM	0.99	1.00	1.00	0.98
	CHN	0.81	0.83	0.80	0.80
	ENG	0.92	0.93	0.91	0.92
	EUR	1.01	1.02	1.01	1.02
	HIN	0.94	0.97	0.91	0.96

HIS	**HIS**	**1.69**	**1.60**	**1.66**	**1.84**
	JAP	0.89	0.91	0.90	0.88
	KOR	0.82	0.86	0.81	0.79
	RUS	0.95	0.95	0.93	0.96
	VNM	0.98	0.94	1.09	0.83
	CHN	0.85	0.86	0.86	0.85
	ENG	0.69	0.71	0.70	0.67
	EUR	0.73	0.75	0.74	0.70
	HIN	0.75	0.75	0.75	0.75
JAP	HIS	0.73	0.73	0.74	0.72
	JAP	**2.76**	**2.65**	**2.79**	**2.80**
	KOR	0.84	0.90	0.81	0.81
	RUS	0.78	0.83	0.73	0.78
	VNM	0.88	0.82	0.89	0.92
	CHN	0.96	0.97	0.97	0.92
	ENG	0.71	0.73	0.72	0.69
	EUR	0.71	0.73	0.72	0.70
	HIN	0.85	0.92	0.84	0.76
KOR	HIS	0.68	0.72	0.68	0.64
	JAP	0.89	0.97	0.89	0.82
	KOR	**2.35**	**2.15**	**2.44**	**2.49**
	RUS	0.79	0.85	0.76	0.74
	VNM	1.05	0.96	0.98	1.25
	CHN	0.80	0.79	0.81	0.82
	ENG	0.86	0.84	0.87	0.88
	EUR	0.96	0.94	0.97	0.98
	HIN	0.91	0.91	0.90	0.92
RUS	HIS	0.87	0.84	0.88	0.90
	JAP	0.88	0.91	0.83	0.91
	KOR	0.84	0.86	0.81	0.85
	RUS	**1.87**	**1.89**	**1.87**	**1.81**
	VNM	1.01	1.03	1.06	0.93
	CHN	0.97	0.92	0.99	1.04
	ENG	0.67	0.62	0.71	0.73
	EUR	0.73	0.67	0.77	0.79
	HIN	0.76	0.71	0.80	0.78
VNM	HIS	0.71	0.63	0.82	0.69
	JAP	0.78	0.71	0.81	0.87
	KOR	0.89	0.79	0.84	1.15
	RUS	0.81	0.79	0.87	0.76
	VNM	**2.68**	**3.15**	**2.40**	**2.17**

NOTE: The results have been adjusted for state, field and publish year. First, conditional on the first/last author's ethnicity, calculate the probablity of last/first author's ethinicity distribution for each field, state and publish year. Second, merge the data we get with the original data, based on the first/last author's ethnicity. Third, calculate the mean of author's distribution for each ethnicity to which the first/last author belongs to, and we suppose it to be the ethinicity distribution in random case ("random values"). Fourth, based on first/last author's ethnicity, calculate the realized ethinicity distribution of last/first author. Fifth, merge this realized values to the "random" values by first and last author's ethnicity, and calculate the ratios of realized values to "random" values. Sixth, take square roots and standize the sum of the values to 9 for each ethnicity of first/last author. (The first figures showing the ratios' distribution.)