NBER WORKING PAPER SERIES

COLLABORATING WITH PEOPLE LIKE ME: ETHNIC CO-AUTHORSHIP WITHIN THE US

Richard B. Freeman
Wei Huang

Working Paper 19905
http://www.nber.org/papers/w19905

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
February 2014

Comments are appreciated and can be sent to freeman@nber.org. We especially thank The Sloan Foundation for support of the NBER Science and Engineering Project, and The Cheung Yan Family Fund to Support Chinese Studies and Students in Economics. We thank William Kerr for his name matching program, and two referees and seminar participants at the October 25, 2012 NBER Conference on High-Skill Immigration for very helpful comments. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Collaborating With People Like Me: Ethnic co-authorship within the US
Richard B. Freeman and Wei Huang
NBER Working Paper No. 19905
February 2014
JEL No. J01,J1,J15

ABSTRACT

This study examines the ethnic identify of the authors of over 1.5 million scientific papers written solely in the US from 1985 to 2008. In this period the proportion of US-based authors with English and European names fell while the proportion of US-based authors with names from China and other developing countries increased. The evidence shows that persons of similar ethnicity co- author together more frequently than can be explained by chance given their proportions in the population of authors. This homophily in research collaborations is associated with weaker scientific contributions. Researchers with weaker past publication records are more likely to write with members of ethnicity than other researchers. Papers with greater homophily tend to be published in lower impact journals and to receive fewer citations than others, even holding fixed the previous publishing performance of the authors. Going beyond ethnic homophily, we find that papers with more authors in more locations and with longer lists of references tend to be published in relatively high impact journals and to receive more citations than other papers. These findings and those on homophily suggest that diversity in inputs into papers leads to greater contributions to science, as measured by impact factors and citations.

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

Wei Huang
Harvard University and NBER
1050 Massachusetts Avenue
Cambridge, MA 02138
huangw@nber.org

The globalization of science has changed the ethnic and national origin of US-based scientists and engineers. From the mid-1970s to the 2000s the foreign-born proportion of science and engineering PhDs granted by US universities roughly doubled, increasing the supply of foreign-born persons to US-based science as student research assistants during their PhD studies and as post-doctoral workers afterward.[1] Expansion of doctorate science and engineering education worldwide increased the supply of potential non-US educated immigrant scientists and engineers to US-based science as well.[2]

These developments substantially changed the ethnic composition of the scientists and engineers who produce scientific papers in the US. In 1985 about 57% of authors on papers in the Web of Science with US addresses had "English" names, 13% had European names while 30% had names of other ethnic groups.[3] The proportion of authors with English names dropped below 50% in 1994 and continued falling to 46% in 2008. By contrast, the proportion of Chinese named authors increased substantially, as did the proportion of authors with names associated with Indian, Hispanic/Filipino, Russian, and Korean ethnicity. In 2008 14% of the names on papers written in the US had Chinese names and 8% had Indian/Hindi/South Asian names.

Given the increasingly collaborative nature of science (Wuchty, et al 2007), it is natural to ask whether or not newly emergent groups of primarily foreign-born researchers work disproportionately with persons of their ethnicity, producing homophily in co-authorship similar to that found in many other areas of human and animal behavior[4]; and whether homophily in collaborations is associated with more or less valuable scientific work.

This study seeks to answer these two questions. To determine the extent of homophily in scientific collaborations, we examine the ethnic identify of the co-authors of over 1.5 million papers with US addresses in the Thomson-Reuters Web of Science (WOS) data base. To assess the scientific contribution of papers with differing ethnic composition, we examine the impact factors of

[1] The share of US science and engineering PhDs going to persons without US citizenship or permanent residence from 17% in 1977 to 33% in 2009.The 1977 figure is calculated from NSF Science and Engineering Indicators,1993, appendix table 2-28 http://www.nsf.gov/statistics/seind93/chap2/doc/02app93.htm. The 2009 figure is calculated from NSF Science and Engineering Indicators, 2012, table 2-28: http://www.nsf.gov/statistics/seind12/appendix.htm

[2] The largest expansion in the past 20-30 years has been in China, which made huge investments in doctorate training to recover from the Maoist destruction of higher education and now surpasses the US in PhD production.

[3] As described shortly on the basis of a name-ethnicity program developed by William Kerr. European excludes Russian or Hispanic/Filipino names.

[4] 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 Miller McPherson, Lynn Smith-Lovin, and James M Cook Birds of a feather: Homophily in Social Networks " Annu. Rev. Sociol. 2001. 27:415–44. For an insightful analysis of the potential payoff from homophily 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"

the journals in which the papers appear and the numbers of citations of the paper. Despite extensive studies of co-authorship patterns among scientists (Barabasi, et al 2002; Newman, 2001a. 2001b, Jones, et al 2008), this is to our knowledge the first study of homophily in scientific collaborations and its relation to the measured contribution of research. We find:

1. Substantial homophily among research teams, with co-authors more likely to be of the same ethnicity than would occur by chance given the distribution of ethnicity among all authors of scientific papers.
2. That researchers with weak previous publications records are especially likely to write papers with persons of the same ethnicity
3. That homophily is associated with publication in a lower impact factor journal and fewer of citations of papers, even holding fixed the previous publishing performance of authors.

Section one documents the existence of substantial homophily in the ethnic composition of co-authorship for US-based papers and develops an index of homophily for ensuing empirical analysis. Section two examines the past publication experience of the authors of papers of papers written by teams of differing ethnic backgrounds. Section three assesses the relation between the extent of homophily among authors of paper, as measured by our index, and several other factors on the impact factor of the journal of publication and numbers of citations garnered by the paper. We conclude with brief comments on the implications of our findings on the productivity of scientific collaborations more broadly.

1. Ethnic composition of US-based authors and homophily of research teams

To measure the ethnic composition of US-based researchers, we undertook a two-step procedure.

First, we used the Thomson-Reuters's Web of Science[5] (WOS) database for the years 1985 to 2008 to create a file of papers for co-authors in which all authors had US addresses. We limited the sample to US-based authors so that authors could meet at seminars, conferences, or other scientific events in the country and connect to collaborate on a project. Limiting the sample to papers written solely in the US allows us to construct a probabilistic model of the distribution of co-authorship

[5] The Thomson-Reuters Web of Science provides data on the articles published in 12,000 plus scientific journals and one of the two major sources for bibliometric material on scientific publications, citations, and related information. http://thomsonreuters.com/products_services/science/science_products/a-z/web_of_science/

among ethnic groups absent homophily that would difficult to develop for foreign collaborations. To focus on collaborations in which preferences for working with persons like oneself may affect the selection of scientific teams we focus on papers with 2-4 authors. These constitute 65% of all co-authored papers in our data set. But we have also analyzed papers with five to ten authors and found similar results to those in the main body of the paper (see Appendix B).

Second, we used William Kerr's name-ethnicity matching program, which combines information on the distribution of names by ethnicity and on the metropolitan statistical areas in which individuals live to determine their likely ethnicity, to assign an ethnic identity to WOS authors.[6] The identification hinges on the fact that last names such as Kim are more likely to represent Koreans than any other group while names like Zhang are likely to be Chinese, and so on. Because persons of a particular ethnicity live disproportionately in some MSAs, MSA information helps distinguish ethnicity as well. We divide ethnicity 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).

The WOS provides authors' complete surnames, initials of first names[7] and addresses to match names to ethnicity. On the notion that first authors and last authors have greatest responsibility for the paper, we limited our data set to papers in which we identified the ethnicity of first and last authors. This means that our sample has ethnic identification for both authors in two-author papers, for the first and last author in other papers, but lacks ethnic identification for some intermediate authors in papers with three or more authors. We match names with ethnicity at a rate of 86%, with the rate of match increasing over time, in part because in later years the WoS has more first names, which allows the matching program to more accurately identify ethnicity than initials.[8]

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. The "not identified" group is middle positioned authors whose ethnicity we could not identify. The biggest change in the ethnic distribution of authors is the near tripling in the frequency of Chinese names, which increased steadily from 4.79 percent in 1985 to 14.45 percent in 2006 and then dropped slightly in 2007 and

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

[7] First names are available for all 2008 papers and in small numbers for papers in 2006 and 2007. We use first names when they are available.

[8] We identify both authors in 2-authored papers at 73.0%; 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%. This is lower than matching rate obtained when the data provide both given names and surnames rather than initials.

2008. The proportion of names from other developing country backgrounds such as Indian/Hindi/South Asian, Hispanic/Filipino, and Vietnamese also increased, as did the proportion of Russian and Korean names. By contrast, the proportion of English names decreased from 56.56 percent in 1985 to 45.56 percent in 2008, while the proportion of European names decreased from 13.47 percent to 11.18 percent.

The distributions in the table do not distinguish between American-born persons of an ethnicity and foreign-born persons of the same ethnicity. For the fastest growing group, persons with Chinese names, the increase is driven largely by increased numbers of researchers born overseas rather than by increased numbers of US-born Chinese. We determine this by exploiting the fact that persons born in China are more likely to have initials with the letters Z, Y, Q and X than are persons born in the US. In our data set 0.3 percent of English names have Z, Y, Q, X first initials compared to 24.2 percent of Chinese names. Assuming that the first names of the US-born Chinese are more Anglicized than the names of Chinese born in China,[9] we estimate that 70.2 percent of Chinese named authors in 1985 and 79.1 percent of Chinese named authors in 2008 were born in China. Given the growth rate of Chinese named authors in our data, this implies that 85 percent of the increased number of Chinese named authors in the US were born in China.

1.1 Measuring homophily overall and at the level of papers

To determine the extent of homophily among co-authors we compare the observed ethnic distribution of names on papers to the ethnic distribution that would arise if co-authorship resulted from random draws from an urn with the distribution of names in the actual population of authors (vide table 1). If 20% of authors in the population of author names had a given ethnicity, our null hypothesis would be that 4% (= 0.20^2) of two authored papers would have authors of that ethnicity and that 0.8% (= 0.20^3) of three authored papers would all have that ethnicity, and so on.

The results of this analysis, summarized in Table 2, provide strong evidence of homophily in scientific teams. Columns 1–4 refine the table 1 distribution by differentiating authors' ethnicity 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 whose laboratory housed the work and who raised the funds and set the overall direction of the research. Intermediate positions reflect the activity of other contributors of varying importance in the project. Panel A

[9] For example a US born "Wang" might be named Richard whereas someone born in China might be named Xia .

shows that in two-author paper sample, 16.6 percent of the first authors and 9.2 percent of second ones have Chinese names; while 49.8 percent of the first authors and 60.2 percent of the second authors have English names. The higher proportion of Chinese names among first authors reflects the entry of young Chinese researchers into US research, while the high proportion of English names among second authors reflects the dominant role of native-born graduates from US universities among senior scientists.

Our test for homophily in co-authorship compares the observed ethnic distribution of the authors on papers to the counter-factual ethnic distribution based on random draws of co-authors from the pool of authors by position. Rather than examining full distributions of ethnicity, the table focuses on the proportion of papers in which all authors are of a given ethnicity for each of the ethnic groups. Column 5 records the expected proportion of papers based on an ethnicity's proportion of first authors, second authors, third authors, and fourth authors. The 1.52% for Chinese-named authors in two-author papers is the multiplicand of 16.6% in column 1 and 9.15% in column 2. Column (6) shows the actual proportion of papers on which all authors have the same ethnicity.

Comparing column 6's realized proportion of authors of the same ethnicity with column 5's expected proportions that authors would be the same ethnicity, we see that the realized proportions are uniformly greater. The absolute differences between the random and realized proportions in column 7 are statistically significant by the t-statistic of difference in means, and are largest for the largest groups. The ratios of the realized to random probabilities in column 8 are larger for smaller groups. Given the likely greater role of first and last authors in the research, we also calculated but do not report in the proportion of 3 and 4 authored papers in which those two authors had the same ethnicity and found that this proportion also exceeded that produced by chance.[10]

We conclude that homophily is substantive among co-authors of scientific papers.[11]

To see the extent to which the high level of homophily reflects the decisions of persons with a given ethnicity to choose the same scientific fields, live in the same region of the country, or have distinct interests in topics relevant to their country of ethnicity, we developed a regression model that used geographic location and field to modify the random proportions. 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; someone in scientific specialties with many Chinese specialists

[10] The statistics look very similar to those authors on two-author papers. Available on request from authors.

[11] We also examined homophily conditional on an author's position in the paper, for instance taking as given the ethnicity of a first author and estimating if the second author was exceptionally likely to be of the same ethnicity, and then taking as given the ethnicity of the second author and estimating if the first author was exceptionally likely to be of the same ethnicity. The conditional probabilities also show considerable homophily.

would be more likely to have a Chinese co-author, and so on. The results of this counter-factual give similar results of homophily to those in the table.[12]

Table 2 documents that homophily as an important feature of scientific collaborations and shows differences in the extent of homophily among groups. But it does not tell us about the structure of preferences that produced the patterns of co-authorship. Homophily could result from persons in each group preferring to work with persons of their ethnicity; or it could result from persons in one group preferring to work with persons of their ethnicity while persons in other groups have no such affinity; or from different preferences for homophily among the groups. Since every author is a co-author of someone in our data it is impossible to identify whose preferences lie behind the observed pattern. To illustrate this point, consider the random distribution of authors from two ethnic groups in two-authored papers. If 50% of authors came from group A and 50% came from group B, the random distribution would have ½ of authors writing with persons of their own group (¼ all A co-authorship and ¼ all B co-authorship) and ½ writing with someone from the other group. If persons in group A had an affinity for working with people like themselves while persons in group B did not care with whom they worked, the distributions for both groups would show more persons working with their own group than the random model. But the same observed distribution could have arisen if persons in group A did not care with whom they worked and those in group B preferred working with persons like themselves. Sophisticated modeling might yield some insight into the differential magnitude of preferences for working with persons of a similar ethnicity[13] but direct information about the preferences of members of groups would almost certainly be more illuminating. To the extent that preferences regarding working with members of one's own group vary within ethnic groups, models based on average preferences will be approximations to reality at best.[14]

Finally, we build on the probabilistic framework underlying table 2 to develop a measure of homophily at the level of the individual paper that we use in ensuing analysis of papers by authors

[12] Results available from authors on request.

[13] The existence of three or more groups can help identify the magnitude of preferences for working with persons of a similar ethnicity. Assume that the preference is for one's own group. Then 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 for 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.

[14] As in the economic theory of discrimination, the realized distribution of outcomes will depend on the distribution of preferences in different groups and the costs of searching to find persons fitting those preferences.

with different mixes of ethnicity. The probability framework makes it clear that the simple dichotomous measure of homophily between "all persons of the same kind" and its complement does not adequately capture the behavior that produces homophily. The reason is that the divergence of a distribution from the random distribution for a given ethnicity depends on the proportion of that ethnicity in the overall population. A paper with all authors from a group from a group that makes up a small proportion of the population will be more reflective of homophily as opposed to random selection of authors than a paper with all authors from a group that make up larger proportions of the population. In our case, having a paper with all English-named authors will occur by chance with greater probability than a paper with all Korean-named authors and thus be less reflective of homophily. Going further, the probabilistic framework shows that a paper with more than one ethnicity could also be more reflective of homophily than a paper with a single ethnicity for its authors. A four-authored paper with say three authors of a small ethnic group and one English author could deviate more from the random distribution than if all four authors were from the larger English-named group.

Viewing homophily as a deviation in the ethnic distribution of authors from the likely distribution that would arise from chance we devised a *homophily index* of the degree of homophily among authors of a paper and use the index to measure the homophily of the paper. Ignoring for simplicity the ordering of authors, our index is a function of the numbers of authors of a given ethnicity compared to the number that we would expect on the basis of the group's proportion of all authors, using a square functional form. Let N(i) = the number of authors of ethnic group i on a paper with T authors so that $\Sigma N(i) = T$, and let p(i) be the proportion of authors of ethnicity i in the population of authors. Then our Homophily Index for a paper with T authors is:

$$(1)\ \Sigma\,[N(i) - p(i)N(i)]^2/N^2 \;=\; \Sigma\, N(i)^2\ [1 - p(i)]^2/N^2$$

where $N(i) - p(i)N(i)$ is the difference between the number of authors on the paper and the number we would expect from its share of all authors; and where N^2 is a scaling factor that puts papers with different numbers of authors onto a similar scale. The index increases as the number of authors of a single group increases ($N(i)^2$) and at a greater rate for groups with small p ($[1 - p(i)]^2$).

To see how the homophily index works consider a 2 authored paper with persons from two groups. Group A has 80% of the population of authors and group B has the remaining 20%. A paper written by 2 persons in group A would get the score $2^2\ [1 - 0.8]^2/2^2 = 0.36$ whereas a paper with two authors from group B would get the score $2^2\ [1 - 0.2]^2/2^2 = 0.96$ and a paper written by an author from each group would be $[1 - 0.2]^2 + (1 - 0.8)^2/2^2 = 0.33$. The papers with the authors from only the

A or B group have a higher score on the index than the papers written by authors from the two groups but the homophily score from the minority B population far exceeds that for the majority A group, reflecting its smaller probability of occurring by chance. By construction the index gives papers with less likely ethnic combinations greater homophily scores. Since the index allows for multiple groups and any number of authors and depends on the deviation of the observed ethnic distribution on a paper from chance, we use it rather than the dichotomous "all authors of the same ethnicity" measure of homophily in the remainder of this paper. It can be generalized to take account the position of authors' on papers as well by using different p(i)s for different positions, per the table 2 calculations.[15]

2. Characteristics of Authors on Papers with Greater/Lesser Homophily

Given the homophily index, what are the characteristics of papers with more or less homophily? In this section we address this question by examining the extent to which the past publication performance of authors and the number of addresses on the paper are associated with more or less homophily using a least squares regression model. We take the homophily index of each paper and regress it on the numbers of previous papers and the average impact factor[16] of the journals which published those papers for the first and last authors of each paper, and a variety of covariates to compare like with like.

We expect that researchers with fewer papers and publications in less prestigious journals will be more likely to work with co-ethnics than authors with better publication records primarily because researchers with weaker publication records are likely to have a smaller network of research connections from which to draw collaborators than researchers who publish more articles and who publish in higher impact journals. The less productive may find it easier to work with persons they know for reasons of homophily in other parts of their lives than to tap into persons of different ethnicity.

Determining the past publishing record of authors is, however, difficult in the Web of Science. The problem is that with information only on the first initial and last name of authors, there

[15] The homophily index is of course correlated with the dichotomous measure. Using the dichotomous measure we obtained qualitatively similar results to those in the text. Calculations with the 0-1 dichotomization of papers available from the authors on request.

[16] The impact factor of a Web of Science journal in a year is the average number of citations to articles in the journal in the preceding two years. See Thomson-Reuters, Introducing the Impact factor, http://thomsonreuters.com/products_services/science/academic/impact_factor/

will invariably be errors in which two or more people with the same same initial and last name would appear to be the same person. This risks attributing more papers to researchers with common names than to those with uncommon names. The J. Kim who is first author on a given paper may have not written any earlier papers but his namesake J. Kim may have done so, producing measurement error if we attach the second J. Kim's papers to the first J. Kim. Measurement error of this sort will bias downward estimated effects of past paper performance in equations that relate the homophily of papers to the characteristics of authors.

We have addressed this problem in two ways.

First, we examined the distribution of names in the WOS papers with US addresses and sought to distinguish authors with the same name by using the fields of the journals in which their papers appeared. Appendix table A1 gives the statistics regarding the distinct names in our data set. Our sample contains over 2.57 million papers and over 7.4 million names. Dividing names by papers, we have an average of 2.88 authors per paper. But many of the names in the data set are the same. The number of distinct names is 1,303,224, which implies that on average a name appears 5.69 times. There are 569,618 names (43.7% of all the distinct names) that appear once, so there is no problem of confusing their work with that of anyone else, but they have no track record of research with which to judge their productivity prior to the paper in our data. The remaining 733,606 names appear more than once. The disambiguation problem is to differentiate which of the multiple appearances of the names reflect the same person writing more than one paper and which reflect different persons with the same name writing some of those papers. Assuming that people with the same names writing in different fields are in fact different (J. Kim who publishes on physics is different than J. Kim who publishes on biochemistry) and that those with the same name in the same field are the same we differentiate names into separate people by using the 11 major fields of science that WOS sorts papers. This yields 1,390,470 names-field that appear more than once -- nearly twice as many potentially distinct authors as the 733,606 different author names with more than a single publication. By construction, differentiating names by field produces more individual authors. We could go further to differentiate names by narrow subfields but this creates the danger of failing to attribute papers to a given author whose research crosses narrow disciplinary lines. Instead of further dividing names by field, we conducted robustness checks on our findings by eliminating names with “large” numbers of papers for varying definitions of large. We obtained results similar to those in the text tables, so that we can rule out the possibility that our results are driven by large numbers of papers due to two or more similarly named authors whom we have failed to distinguish.

Our second method for dealing with disambiguation of names problem is to replicate our calculations on the life sciences subset of the WoS that overlaps with Pub Med papers using Torvik and Smallheiser's (2009) computer program disambiguation of Pub Med names.[17] We will present results from the disambiguated Pub Med sample as well as for the larger WoS sample.

We also examine the relation between the homophily of a paper and the geographic locale of paper authors. Here there are no disambiguation issues because the WoS reports the addresses of all authors on its data base, though not until 2008 does it link the addresses to specific authors so we cannot tell how many authors on a multi-authored paper with two addresses worked at one of the addresses rather than the other or if someone worked at both.

Should we expect greater homophily among authors in the same locale or across locales? On the one side, students of the same ethnicity often work in the same labs for professors of their ethnicity, which suggests that papers with the same address would evince greater homophily.[18] To the extent that authors collaborate across geographic areas to combine scientific expertise and/or care less about ethnicity of someone with whom they do not interact regularly, we would also expect greater homophily on papers where authors are at the same address. But geographic closeness may substitute for ethnic closeness in connecting researchers. Researchers may be more likely to meet persons of different ethnicity at their university based on geographic propinquity than to meet someone of a different ethnicity far away. This would produce less homophily on papers with fewer addresses.[19] Absent direct measures of how collaborators met,[20] it is impossible to distinguish these factors so we turn to the statistics to assess the net direction of effects.

We measure geographic proximity by information on the paper of whether co-authors have the same address, two addresses, or three or more addresses. Since the name disambiguation problem occurs only with measures of author's past publications, which could potentially distort estimates of the relation between the homophily of a paper and addresses (and other paper-specific measures such as numbers of references) we estimate equations for the relation between the

[17] There are other samples from which one can obtain disambiguated names: Cite See (Huang, 2013) and for other data sets (Ferreira, et al 2011). Lai, et al, (2011) disambiguation of names on patents is not helpful to us.

[18] Tanyildiz, (2008) shows that students from a given country are more likely to enroll in universities with faculty from their native country and which already have many students from their country and are likely to work in labs populated by students from the same country of origin under the direction of foreign-born faculty. 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

[19] For analysis of spatial and social effects on knowledge flows in patents, see Agrawal, Kapur, and McHale, 2008)

[20] See Freeman, Ganguli, and Murciano-Goroff (2014) for some information on how co-authors meet in international and other collaborations.

homophily index and addresses with and without including researchers' past publication experience.

Table 3 records ordinary least squares estimates of the relation between our homophily index and the characteristics of the paper, authors' locations, and authors' previous publishing records. The analysis treats papers with 2, 3, and 4 authors separately. Because first and last authors are presumptively the most important, we focus on their publishing record. Estimates show that including the publishing record of intermediate authors does not greatly affect the table 3 results for the first and last authors. Note that the regressions include a large group of dummy variable covariates, as specified in the table notes: year of publication; 180 subfields based on the journal of publication; state or area of the US; and dummy variables for each of the nine ethnic groups. This is to better isolate the independent effects of the authors' publishing record and location on homophily.

The odd-numbered columns record coefficients for our full sample of papers. The key estimated coefficients are on the dummy variables for the number of previous papers, where the reference group are persons with no previous papers. The results for two-author papers show that authors with a larger number of previous papers were less likely to co-author a paper with someone of the same ethnicity than persons who had no previous papers. The columns for three and four authored papers are similar, though there is greater variability in the magnitude and significance of estimated coefficients on the number of papers. We conclude that researchers who have written more papers in the past are less likely to write with persons of the same ethnicity.

The even-numbered columns cover the smaller sample of cases in which authors had previous publications, which changes the reference group on numbers of papers from zero papers to 1-5 papers. The estimated coefficients on the average impact factor of previous papers show that the impact factor of the previous papers is also negatively associated with homophily, with a markedly larger coefficient for the last author. This suggests that the presumptively senior author plays a particularly important role in determining the composition of the papers' team, possibly taking prime responsibility for forming the collaboration. In an analysis not reported in the table, we also examined the effect of numbers of papers and impact factors for intermediate authors and found that they were not substantively related to homophily.[21]

Turning to the geographic dimension of collaborations, as reflected in the number of addresses on a paper, the table 2 estimates show a substantial difference in the relation to homophily

[21] 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.

between two-authored papers and three and four-authored papers.[22] Among two-authored papers, the estimated coefficient on having two addresses is associated with a positive though statistically insignificant link to homophily. Among three and fourth-authored papers, the number of different addresses is strongly associated with smaller levels of homophily. Since about 70% of the papers in the sample have 3 or 4 authors and the coefficient on addresses are more precisely estimated in those samples, pooling the data to form a single sample would yield a negative coefficient on the number of addresses variables. The smaller homophily among the three and four authored papers where authors having different addresses could reflect the reaching out of labs for expertise from another lab regardless of the ethnic ties of researchers and/or the greater importance of homophily preferences for people close by with whom the researchers connect with regularly.

3. Relation between homophily and impact factors and citations

Do researchers working with persons of the same ethnicity write papers that have greater or lesser scientific impact than researchers working with persons with different ethnicity?

To the degree that working with people of one's own group makes communication easier, homophily should raise the productivity of the research team. People from the same group can communicate in similarly accented English or switch to their native tongue if the English does not work. But to the degree that co-authoring reflects tastes/preferences for working with persons like themselves at the expense of complementary research skills or knowledge, homophily should be associated with reduced productivity, per standard analysis of discriminatory preferences. Similarly, if preferences aside, persons of the same ethnicity have been through similar educational experiences or think more alike for whatever reason, working together may reduce the diversity of perspectives that can produce more interesting scientific results.

To see which effect, if any, dominates, we relate two measures of the scientific contribution of papers to the homophily index – the impact factor of the journal which published the paper and the number of citations the paper received as of the last year of our sample, 2008. Impact factors are available upon publication. They are an imperfect measure of the quality of a journal and of any

[22] Appendix table A, which summarizes the statistics in our sample, shows that the mode for co-authored papers is a single address: 69% of the authors of two-authored papers, 52% of the authors of three-authored papers, and 40% of the authors of four-authored papers report a single address. Twenty-seven percent of the two authored papers report two addresses while 5% report three addresses, due to some authors having two addresses. Similarly, thirty-three percent of papers with three and four authors report two addresses, and 16% and 27% respectively report three or more addresses.

paper in it (European Association of Science Editors, 2007) but provide some indication of how editors and reviewers from prestigious journals judged a paper. Citations presumably are a more accurate indication of the paper's scientific merit as they reflect the "wisdom of crowds (of knowledgeable scientists)" rather than the views of a few though they also are subject to problems (International Mathematical Union, 2008). A mediocre paper in a large field where the norm is to cite many papers may receive more citations than a path-breaking paper in a small field where the norm is to cite few papers. Our inclusion of dummy variables for 180 subfields serves as a control for this problem.

Because journal impact factors are based on the citations of the papers published in recent years, impact factors and the citations of articles in a journal are highly correlated. But there is a wide dispersion of citations within journals. Some papers in lower impact journals invariably gain more citations than the vast majority of papers in high impact journals. Lozano, Larivière, and Yves Gingras (2012) show that since 1990, the relation between IFs and paper citations has been weakening, potentially because Internet search engines make it easier to find relevant articles in less widely circulated journals. In any case, given that impact factors and citations are imperfect measures of the scientific value of papers, we use both in the analysis.

To estimate the effect of homophily on the impact factor associated with a paper we regressed the impact factor on the homophily index, the number of previous papers and the average impact factor of previous first and last author publications for papers with two, three, and four author, taken separately. In addition, we examine the relations between impact factors and the number of references in the papers, and the number of addresses on the paper. Number of references provides a measure of the breadth of the paper and its use and possible contribution to a range of scientific activity.[23] As in table 3, we include dummy variables for the year of publication, subfield, state of the location of the paper, and author ethnicity.

Table 4 reports the estimated coefficients and standard errors from this analysis for two-author, three-author, and four-author papers, respectively. The columns labeled "full sample" cover all papers while those labeled "papers with authors having previous papers" are limited to those where we identified earlier papers and used the journal in which they appeared to calculate average impact factors. The calculations for the full samples give highly significant negative coefficients for the homophily index while those for the sample of papers with authors having previous publications

[23] It also may reflect pressures by editors to cite their journal as part of the acceptance process (Willhite and Fong, 2012) or decisions of authors to include references to papers from potential referees. Assuming that these factors affect the references of many papers in a similar way, this will create measurement error in the true use of past work in a paper.

give weaker negative coefficients. In each case inclusion of measures for authors' previous number of papers and the impact of those papers reduces the estimated coefficient on homophily substantially, indicating that a large part of the homophily effect is due to the weaker publication records of the researchers associated with homophily. In the three-author and four-author papers the estimated coefficients on the homophily index are sufficiently reduced to lose their statistical significance by standard criterion. In all the calculations, moreover, the past publication performance of the last author have larger and statistically more significant links to the impact factor of the journal than the past publication performance of the first author, which suggests that the last author has a particularly important role in getting a paper into a higher impact journal, possibly by contributing to a higher quality paper or possibly by having better connections within the scientific journal world.24

As for the paper-related variables, the number of references in a paper has a large significant relation to the impact factor of the journal of publication. To the extent that the number of references indicates the body of knowledge that went into a paper and thus the breadth of its scientific contribution, this pays off in the form of greater likelihood of publication in high impact journals. The positive estimated coefficients on the number of addresses of authors can also be interpreted as reflecting the breadth of knowledge that went into the paper. Researchers working in different universities or research centers are likely to bring a wider range of ideas, perspectives, and materials to the analysis than those working in the same lab. From this perspective, the negative effect of homophily on the impact factor suggests that papers written by persons of the same ethnicity may reflect a narrower research perspective than that provided by a more diverse set of authors.

Table 5 turns to our second measure of the quality of a paper – the number of citations it garners. Because the distributions of papers by citations has a peculiar shape, with at one extreme on the order of 20% to 30% of papers obtaining no citations while at the other extreme papers with many citations follow a power law distribution (Redner, 2005; Gupta et al, 2005), we transform the number of citations of a paper into a percentile distribution based on citations to papers in the year of publication and use the 0-100 measure as our indicator of citations. The subfield dummy variables in the list of covariates allow for different levels of the citations among fields.25 As with the calculations for the impact factors, our estimates treat two samples: the full sample of papers, and the

[24] These alternative explanations can be tested by examining future citations of papers that were "boosted" into higher impact journals by characteristics of the final author, that go beyond the scope of our analysis.

[25] The percentile thresholds and number of citations of papers in the percentiles in our data are available on request.

sample of papers for which we identify previous papers for authors.

The estimated coefficients on the homophily index and the number and addresses and references in columns 1 to 3 tell a clear story. Homophily is significantly negatively associated with numbers of citations for two-author, three-author, and four-author papers while addresses and references are positively associated with citations. The estimated coefficients in columns 4, 6, and 8 for papers for the sample of authors with previous papers yields smaller and less significant but still negative coefficients on the homophily index. But the addition of the average impact factor of the previous papers of the first and last authors and the numbers of papers they have written reduces the estimated coefficient on the homophily measure to insignificant positive or negative in columns 5, 7, and 9. The implication is that papers written by persons of the same ethnicity are less cited than papers written by persons with differing ethnicities because the authors of papers with greater homophily have weaker past publication records. There are a various possible reasons for weaker publication records, ranging from weak educational or post-doctoral experiences, choice of research topic within their subfield, heavier teaching loads, to hysteresis in their publication trajectory due to bad luck at the outset. A life cycle analysis of the publications of individual researchers could illuminate these and other possible reasons for the differences in research outputs.

Inclusion of past publication records in table 5 also diminishes the coefficient on number of addresses, though it has a much smaller effect of the estimated coefficient for number of references. Finally, the estimated relation between the number and impact factor of authors' prior papers to the citation percentile differs between the first author and the last author. The estimated regression coefficients show that the last and presumably senior author's publishing record has a larger effect on citations than the first and presumably junior author's past record, though the difference in coefficients is smaller for citations than it was for impact factors in table 4.

The weakened negative relation between homophily and the impact factor and citations of a paper with addition of measures of the past publications record of the papers first and last author suggests that homophily may have its largest adverse effect on persons with no previous papers. To examine that possibility we interacted our homophily index with dummy variables for the past number of papers of the last author: no publications, 1-5, 6-10 and 10 or more. Figure 1 shows the results of this analysis in terms of the estimated interaction coefficients; Appendix table A3 gives the details of the regression analysis. The coefficients in the upper panel for impact factors support the posited relation. For papers with two , three, and four authors, the coefficients are significant negative on papers with a last author who has no previous publications and decline as the number of

publications increases. But the estimated coefficients in the bottom panel for citations show greater variability, particularly between papers whose authors have no other publications and papers whose authors have 1-5 publications. With the small number of authors on the papers, the variability may reflect the publication records of those authors as well, which we have not investigated.

To what extent, if at all, do our findings generalize to papers written with more authors? To answer this question, we replicated the analyses in tables 4 and 5 on the link between homophily and the impact factor and citations for papers with five, six, seven, eight, nine, and ten authors. Appendix B gives the results of that analysis. Table B1 summarizes the statistical properties of the papers in the analysis while table B2 records the estimated regression coefficients on the homophily index, number of references, and number of addresses. The estimates show a similar pattern to that in the text tables: negative relations between impact factors and citation percentile and the homophily index and positive relations between the impact factor and citation percentile and the number of references and addresses.

As another check on the robustness of our results, we replicated the analysis in tables 4 and 5 for papers in Pub Med for the mid-1980s to 2008 period. This restricts our sample to life science papers but allows us to use the disambiguated names from the Torvik and Smalheiser (2008) algorithm for differentiating same-named people. Given the dependence of the negative relation between impact factors and citations and homophily on measures of the past publications of authors, and the importance of name disambiguation in creating those measures, it is important to see whether or not our results hold up with names data disambiguated using more subtle techniques than ours.

Table 6 presents the results of our Pub Med analysis based on the Torvik-Smaheiser disambiguated names data in terms of the estimated coefficients from impact factor and citation percentiles on the homophily index and other measures of the attributes of papers and authors. The table differentiates between the full sample of papers, in this case in PubMed, and the sample of papers in which the authors had previous publications, based on the disambiguated names. The estimated coefficients on the homophily index in the impact factor regressions are negative, though with weaker statistical significance than the table 4 estimates for the larger Web of Science papers, The estimated coefficients on the homophily index in the citation percentile regressions are also noticeably smaller in the PubMed data than in the comparable table 5 analysis, presumably because of some differences between the life sciences and other sciences that goes beyond counting papers,

impact factors, and citations. Addition of the past publications record of the first and last author further weakens the negative relation between homophily and outcomes, with the impact factor and numbers of papers of the last author having stronger positive links to both outcomes than the impact factor and number of papers of the first author, mimicking the results in table 4 and table 5. By contrast, while the estimated positive relation between the number of addresses and number of references and the impact factor and citation percentile weaken with the addition of the information on authors' publication record, they continue to have a substantial positive link to the outcome variables, again as in table 4 and table 5.

4. Conclusion

Our analysis has shown that homophily is a substantive phenomenon in co-authorship in scientific papers and thus in the make-up of the research teams that produce the papers and papers; that research teams with greater homophily produce papers with lower impact factors and fewer citations than other papers, with most of the negative relation attributable to the weaker prior publication performance of the authors of papers with great homophily. To illuminate the pattern of homophily and its link to scientific outcomes requires analysis of the decisions of researchers with different publication trajectories to collaborate with others and the factors that produce stronger and weaker publications trajectories. Here, the evidence that the attributes of last authors are more important in explaining the data than the attributes of first authors, while that of intermediate authors is usually negligible, suggests that the easiest path toward a theory of collaboration may through treating the last author as the initiator or entrepreneur for the collaboration as opposed to viewing the collaboration as a partnership among equals.

Going beyond homophily, our analysis has also found that two variables that reflect diversity of authors and the knowledge they use in a paper – the number of addresses and the number of references are strongly associated with publishing in a higher impact journal and gaining more citations. A reasonable interpretation of this pattern and that for homophily is that greater diversity and breadth of knowledge of a research team contributes to the quality of the scientific papers that the team produces. This hypothesis requires text-mining and latent semantic analysis of the content of the papers written by teams with differing degrees of homophily, different numbers of addresses, and different numbers and possibly types of references to see if in fact the novelty of papers varies with those factors. Our analysis and findings will hopefully spark further work that to illuminate not only homophily in science but the factors that produce more impactful and valuable science from

research teams, be they of the same ethnicity or of multiple ethnicities.

References

Agrawal, A., Kapur, D., and McHale, J. 2008. How Do Spatial and Social Proximity Influence Knowledge Flows? Evidence from Patent Data, Journal of Urban Economics, 64: 258-269.

Barabasi, A., Jeong, H., Neda, Z., et al. (2002). Evolution of the social network of scientific collaborations. Physica A: Statistical Mechanics and its Applications, 311(3–4), 590–614.

Breschi, S. and Francesco L., 2009. Mobility of skilled workers and co-invention networks: an anatomy of localized knowledge flows. Journal of Economic Geography 9(4):439-468.

The European Association of Science Editors "The EASE Statement on Inappropriate Use of Impact Factors" 2007. Available at: http://www.ease.org.uk/sites/default/files/ease_statement_ifs_final.pdf

Ferreira, Anderson, Marcos André Gonçalves, and Alberto H. F. Laender "A Brief Survey of Automatic Methods for Author Name Disambiguation" SIGMOD Record, June 2012 (Vol. 41, No. 2) 15

Freeman, Richard B. Ina Ganguli, Raviv Murciano-Goroff January 2014 Why and Wherefore of Increased Scientific Collaboration NBER WP 19819

Gupta, Hari, Jose R. Campanha, and Rosana A. G. Pesce " Power-Law Distributions for the Citation Index ofScientific Publications and Scientists" Brazilian Journal of Physics, vol. 35, no. 4A, December, 2005

Hegde, Deepak and Tumlinson, Justin, Can Birds of a Feather Fly Together? The Payoffs of Ethnic Proximity in US Venture Capital (October 5, 2011). Available at SSRN: http://ssrn.com/abstract=1939587 or http://dx.doi.org/10.2139/ssrn.1939587

Huang, Jiang, Seyda Ertekin, C. Lee Giles "Fast Author Name Disambiguation in CiteSeer" . Available at http://web.mit.edu/seyda/www/Papers/IST-TR_DisambiguationCiteseer.pdf

International Mathematical Union, 2008, Joint IMU/ICIAM/IMS–Committee on Quantitative Assessment of Research Citation Statistics, Available at: http://www.mathunion.org/fileadmin/IMU/Report/CitationStatistics.pdf

Jones, Ben, Stefan Wuchty, and Brian Uzzi. "Multi-university Research Teams: Shifting Impact, Geography, and Stratification in Science," Science, Fall 2008

Kerr, William 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;

Kerr, William R. "Ethnic Scientific Communities and International Technology Diffusion," The Review of Economics and Statistics, 90:3 (August 2008), 518-537. diss/25

Lai Ronald, Alexander D'Amour, Amy Yu, Ye Sun, Vetle Torvik, Lee Fleming Disambiguation and Co-authorship Networks of the U.S. Patent Inventor Database, June 9 2011 Available at : funginstitute.berkeley.edu/sites/default/files/Disambiguation%20and%20Co-authorship%20Networks%20of%20the%20U.S.%20Patent%20Inventor%20Database.pdf

Lozano, George A., Vincent Larivière, Yves Gingras "The weakening relationship between the impact factor and papers' citations in the digital age" Journal of the American Society for Information Science and Technology Volume 63, Issue 11, pages 2140–2145, November 2012

Miller McPherson, Lynn Smith-Lovin, and James M Cook Birds of a feather: Homophily in Social Networks " Annu. Rev. Sociol. 2001. 27:415–44

Newman, M. (2001a). The structure of scientific collaboration networks. PNAS, 98(2), 404–409.

Newman, M. (2001b). Scientific collaboration networks. I. Network construction and fundamental results. Physical Review E, 64(1), 16131.

Redner S (2005) Citation Statistics from 110 Years of Physical Review. Physics Today 58: 49–54.

Strotmann, Andreas and Dangzhi Zhao "Author name disambiguation: What difference does it make in author-based citation analysis?" Journal of the American Society for Information Science and Technology Volume 63, Issue 9, pages 1820–1833, September 2012

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_

Torvik, V. and M. Weeber, D. Swanson, N. Smalheiser, 2005. A Probabilistic Similarity Metric for Medline Records: A Model for Author Name Disambiguation. Journal of the American Society for Information Science and Technology, 56(2):140–158.

Torvik, V. and N. Smalheiser, 2009. Author Name Disambiguation in MEDLINE. ACM Transactions on Knowledge Discovery from Data, Vol. 3., No. 3, Article 11.

Velden, Theresa Asif-ul Haque, and Carl Lagoze "A New Approach to Analyzing Patterns of Collaboration in Co-authorship Networks" Scientometrics October 2010, Volume 85, Issue 1, pp 219-242

Wilhite,Allen W. and , Eric A. Fong* " Coercive Citation in Academic Publishing" Science 3 February 2012: Vol. 335 no. 6068 pp. 542-543 DOI: 10.1126/science.1212540

Wuchty S, Jones BF, Uzzi B. (2007). The increasing dominance of teams in production of

knowledge, Science, 316(5827):1036-9.

Table 1: Number of papers and distribution of authors by ethnicity, for papers with 2 to 4 authors, US addresses only

Year	Number of papers	Distribution of authors by ethnicity (%)									
		Anglo-Saxon/ English (ENG)	Chinese (CHN)	European (EUR)	Indian/Hindi/ South Asian (HIN)	Hispanic/ Filipino (HIS)	Japanese (JAP)	Russian (RUS)	Korean (KOR)	Vietnamese (VNM)	Not identified
1985	43270	56.56	4.79	13.47	4.23	2.87	2.45	0.71	2.05	0.14	17.15
1986	43790	56.07	4.81	13.35	4.26	3.08	2.31	0.78	2.04	0.16	17.53
1987	44571	55.66	5.27	13.27	4.37	2.99	2.39	0.84	1.98	0.15	17.40
1988	46615	54.83	5.80	13.35	4.42	3.14	2.32	0.92	1.95	0.16	17.32
1989	48218	54.23	6.35	13.14	4.49	3.25	2.32	1.00	2.00	0.17	17.12
1990	49896	53.32	6.98	12.94	4.76	3.33	2.29	1.07	2.09	0.17	17.24
1991	52462	52.34	7.73	12.92	4.82	3.41	2.36	1.14	2.07	0.18	17.15
1992	53134	51.47	8.50	12.67	5.23	3.32	2.28	1.12	2.09	0.18	17.27
1993	53344	50.64	9.55	12.43	5.15	3.43	2.27	1.19	2.28	0.22	16.95
1994	53596	49.56	10.16	12.31	5.43	3.43	2.24	1.31	2.29	0.22	17.02
1995	55886	48.97	10.62	12.16	5.49	3.58	2.13	1.35	2.57	0.25	16.89
1996	62576	48.55	10.91	12.03	5.67	3.64	2.02	1.33	2.55	0.23	16.97
1997	64092	48.37	11.23	11.96	5.74	3.64	1.93	1.38	2.63	0.28	16.80
1998	80914	49.15	11.33	11.97	6.01	3.81	1.50	1.32	2.74	0.32	16.09
1999	78320	48.61	11.48	11.86	6.09	3.96	1.52	1.38	2.86	0.32	16.15
2000	77946	48.22	11.78	11.77	6.04	3.99	1.50	1.49	2.94	0.36	15.96
2001	75443	47.56	12.20	11.73	6.17	4.00	1.54	1.59	3.11	0.33	15.93
2002	74852	46.99	12.49	11.45	6.45	4.02	1.57	1.67	3.18	0.37	15.82
2003	77973	46.20	13.12	11.35	6.80	4.14	1.45	1.69	3.14	0.34	15.74
2004	79872	45.43	13.63	11.13	7.07	4.29	1.47	1.83	3.18	0.35	15.45
2005	82377	44.71	14.16	11.03	7.37	4.28	1.36	1.87	3.20	0.37	15.47
2006	86177	45.21	14.45	11.14	7.64	4.77	1.35	1.89	3.41	0.33	13.21
2007	85018	45.65	14.28	11.21	7.93	5.03	1.33	1.90	3.42	0.36	11.97
2008	77959	45.56	14.16	11.18	7.98	5.05	1.30	1.93	3.41	0.35	12.14
Total	1548301	48.97	10.92	11.99	6.06	3.89	1.79	1.44	2.75	0.28	15.85

NOTES: Sample limited to papers with only US addresses. Because most of the sample contains initials rather than first names as well surnames, the match rate is 84.15%, below rate of matching in patent data when both first and last names are available. For two-author papers, we keep those papers in which both authors are identified; in three- and four- author paper sample, we keep those with at least two authors identified, so that the "not identified" occur only in papers with more than two authors.

Table 2: Percentage of authors by ethnicity and position, and comparison between realized and random collaborating patterns

Ethnicity	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Authors' ethnicity distribution by position (%)				Probability of all authors same ethnicity (%)			Ratio
	First	Second	Third	Fourth	Random	Realized	Difference (6) - (5)	(6)/(5)
Panel A: Two-author paper								
CHN	16.6	9.15			1.522	4.157	2.636	2.73
ENG	49.8	60.2			29.99	33.56	3.57	1.12
EUR	12.8	14.7			1.870	2.274	0.404	1.22
HIN	7.71	6.53			0.504	1.605	1.102	3.19
HIS	4.57	3.76			0.172	0.429	0.257	2.50
JAP	2.24	1.31			0.029	0.270	0.241	9.23
KOR	2.39	1.02			0.024	0.135	0.111	5.58
RUS	3.55	3.15			0.112	0.397	0.285	3.55
VNM	0.35	0.23			0.001	0.009	0.008	11.1
Panel B: Three-author paper								
CHN	13.7	10.25	6.63		0.093	1.243	1.149	13.3
ENG	42.8	48.9	54.0		11.30	13.34	2.044	1.18
EUR	11.0	11.6	12.9		0.164	0.221	0.057	1.35
HIN	6.83	5.86	5.08		0.020	0.314	0.294	15.4
HIS	4.15	3.81	3.28		0.005	0.074	0.069	14.3
JAP	2.21	1.64	1.17		0.000	0.087	0.086	206
KOR	1.93	1.31	0.78		0.000	0.023	0.023	118
RUS	2.79	2.67	2.49		0.002	0.031	0.029	16.6
VNM	0.32	0.29	0.18		0.000	0.000	0.000	208
Panel C: Four-author paper								
CHN	12.5	10.5	8.33	5.76	0.006	0.526	0.520	83.6
ENG	41.2	45.5	49.0	52.4	4.823	6.969	2.147	1.45
EUR	10.7	10.8	11.5	12.5	0.017	0.032	0.015	1.93
HIN	6.30	5.42	4.76	4.40	0.001	0.076	0.075	106
HIS	4.04	3.97	3.64	3.25	0.000	0.034	0.034	178
JAP	2.35	1.90	1.58	1.11	0.000	0.047	0.047	6003
KOR	1.70	1.35	1.09	0.70	0.000	0.005	0.005	2785
RUS	2.56	2.42	2.35	2.30	0.000	0.007	0.007	201
VNM	0.30	0.31	0.27	0.16	0.000	0.000	0.000	-

NOTES: Papers include only those where the name-ethnicity program identifies the ethnicities of first and last authors. The differences between columns 6 and 7 are significant except for the fourth authored VNM row, where there are too few observations.

Table 3: Estimated Coefficients and Standard Errors for Regression of Homophily Index of a Paper on first and last author's prior papers and number of addresses on paper for two, three, and four-author papers

	(1)	(2)	(3)	(4)	(5)	(6)
	Two-author papers		Three-author papers		Four-author papers	
VARIABLES	Homophily Index					
Number of FIRST author's prior papers						
1 - 5	-0.021		-0.032		-0.079***	
	(0.038)		(0.026)		(0.025)	
6 - 10	-0.165***	-0.118*	-0.020	-0.020	-0.070*	-0.015
	(0.060)	(0.064)	(0.042)	(0.046)	(0.039)	(0.042)
10 +	-0.486***	-0.488***	-0.056	-0.126***	-0.105***	-0.064
	(0.055)	(0.058)	(0.040)	(0.043)	(0.037)	(0.040)
Number of LAST author's prior papers						
1 - 5	-0.162***		-0.101***		-0.130***	
	(0.050)		(0.033)		(0.031)	
6 - 10	-0.176***	-0.007	-0.113***	0.046	-0.143***	0.085*
	(0.060)	(0.065)	(0.041)	(0.047)	(0.038)	(0.043)
10 +	-0.328***	-0.041	-0.072**	0.117***	-0.077**	0.133***
	(0.054)	(0.055)	(0.036)	(0.039)	(0.034)	(0.036)
Average impact factor of prior papers						
First author		-0.022*		-0.019**		-0.027***
		(0.012)		(0.009)		(0.008)
Last author		-0.097***		-0.062***		-0.075***
		(0.015)		(0.011)		(0.010)
Number of addresses	0.052	0.034	-0.164***	-0.164***	-0.161***	-0.145***
	(0.041)	(0.054)	(0.019)	(0.026)	(0.013)	(0.018)
Observations	478,349	273,229	569,015	315,035	457,667	257,736
R-squared	0.605	0.617	0.405	0.438	0.348	0.393
More addresses than authors	Yes	Yes	Yes	Yes	Yes	Yes
Publish year (# of years: 24)	Yes	Yes	Yes	Yes	Yes	Yes
Subfield (# of subfields: 180)	Yes	Yes	Yes	Yes	Yes	Yes
Subfield * Publish year	Yes	Yes	Yes	Yes	Yes	Yes
State (# of states: 53)	Yes	Yes	Yes	Yes	Yes	Yes
Author ethnicity (Each: 9)	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. Covariates include all authors' ethnicities and dummies for states, publish year, subfields, and interations between publish year and subfields. In the 2, 4 and 6 columns, only those papers whose first and last author have previous publications are kept.

Table 4: Estimated Coefficients and Standard Errors for Regression of Impact Factor of Journal of Publication on on Homophily Index, Attributes of Papers, and Previous Articles of Co-Authors

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Full sample			Papers with Authors having previous papers					
Number of authors	Two	Three	Four	Two		Three		Four	
Variable	Dependent variable: Impact factor (IF)								
Homophily index	-0.161***	-0.183***	-0.364***	-0.167***	-0.092***	-0.135***	-0.056	-0.250***	-0.094
	(0.027)	(0.033)	(0.048)	(0.035)	(0.033)	(0.042)	(0.040)	(0.063)	(0.060)
Number of addresses	0.068***	0.046***	0.048***	0.068***	0.047***	0.042***	0.031***	0.048***	0.037***
	(0.009)	(0.005)	(0.005)	(0.012)	(0.012)	(0.007)	(0.007)	(0.007)	(0.006)
Number of references	0.008***	0.012***	0.014***	0.007***	0.005***	0.011***	0.009***	0.014***	0.010***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Ave. IF of prior articles									
First author					0.109***		0.121***		0.128***
					(0.005)		(0.005)		(0.006)
Last author					0.294***		0.297***		0.293***
					(0.007)		(0.007)		(0.007)
Num. of first author's prior articles: Reference group is 1-5 papers									
6 - 10					0.011		-0.014		-0.012
					(0.011)		(0.010)		(0.012)
10 +					0.026**		-0.020**		-0.015
					(0.010)		(0.009)		(0.011)
Num. of last author's prior articles: Reference group is 1-5 papers									
6 - 10					0.064***		0.070***		0.054***
					(0.011)		(0.011)		(0.013)
10 +					0.136***		0.121***		0.128***
					(0.010)		(0.009)		(0.011)
Observations	478,349	569,015	457,667	273,229	273,229	315,035	315,035	257,736	257,736
R-squared	0.523	0.553	0.545	0.535	0.570	0.551	0.590	0.539	0.575
More addresses than authors	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Publish year (# of years: 24)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subfield (# of subfields: 180)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subfield * Publish year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State (# of states: 53)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Author ethnicity (Each:9)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

NOTE: Standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1. Covariates include all authors' ethnicities and dummies for states, publish year, subfields, and interations between publish year and subfields.

Table 5: Estimated Coefficients and Standard Errors for the Factors That Affect Citations

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Full sample			Papers with Authors having previous papers					
Number of authors	Two	Three	Four	Two		Three		Four	
Variable	Dependent variable: Citation percentile (0-100)								
Homophily index	-0.732**	-1.818***	-3.009***	-0.536	-0.047	-0.995*	-0.619	-1.863**	-1.137
	(0.367)	(0.428)	(0.577)	(0.497)	(0.493)	(0.575)	(0.571)	(0.770)	(0.765)
Number of addresses	0.881***	0.404***	0.369***	0.769***	0.575***	0.250***	0.155*	0.268***	0.195***
	(0.106)	(0.062)	(0.053)	(0.136)	(0.135)	(0.083)	(0.083)	(0.072)	(0.071)
Number of references	0.408***	0.431***	0.422***	0.384***	0.370***	0.403***	0.388***	0.406***	0.387***
	(0.004)	(0.004)	(0.009)	(0.004)	(0.004)	(0.005)	(0.005)	(0.004)	(0.004)
Ave. IF of prior articles									
First author					0.641***		0.679***		0.737***
					(0.032)		(0.028)		(0.029)
Last author					1.408***		1.449***		1.408***
					(0.040)		(0.036)		(0.037)
Num. of first author's prior articles: Reference group is 1-5 papers									
6 - 10					0.978***		0.905***		0.856***
					(0.152)		(0.140)		(0.154)
10 +					1.814***		1.411***		1.828***
					(0.139)		(0.129)		(0.143)
Num. of last author's prior articles: Reference group is 1-5 papers									
6 - 10					1.518***		1.215***		1.132***
					(0.155)		(0.143)		(0.161)
10 +					3.374***		2.913***		2.783***
					(0.132)		(0.120)		(0.133)
Observations	478,349	569,015	457,667	273,229	273,229	315,035	315,035	257,736	257,736
R-squared	0.351	0.337	0.326	0.360	0.369	0.341	0.351	0.325	0.336
More addresses than authors	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Publish year (# of years: 24)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subfield (# of subfields: 180)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subfield X Publish year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State (# of states: 53)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Author ethnicity (Each:9)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

NOTE: Standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1. Covariates include all authors' ethnicities and dummies for states, publish year, subfields, and interations between publish year and subfields.

Table 6: Estimated Coefficients and Standard Errors for Relation of Homophily and Other Factors to Impact Factor and Citations, Pub-med data

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Full sample in PUBMED						**Papers with Authors having prior papers in PUBMED**					
Variable	Impact factor			Citation percentile			Impact factor			Citation percentile		
Number of authors	Two	Three	Four	Two	Three	Four	Two	Three	Four	Two	Three	Four
Homophily index	-0.161***	-0.083	-0.229***	-0.673	0.332	-1.164*	-0.095	0.028	-0.053	-0.997	1.281	-0.131
	(0.049)	(0.057)	(0.070)	(0.469)	(0.576)	(0.706)	(0.069)	(0.079)	(0.100)	(0.645)	(0.796)	(0.992)
Number of addresses	0.059***	0.042***	0.041***	0.922***	0.378***	0.324***	0.026	0.025**	0.011***	0.588***	0.219**	0.359***
	(0.015)	(0.008)	(0.006)	(0.139)	(0.075)	(0.060)	(0.021)	(0.011)	(0.001)	(0.184)	(0.105)	(0.005)
Number of references	0.006***	0.013***	0.015***	0.349***	0.399***	0.398***	0.002***	0.008***	0.024***	0.311***	0.347***	0.150*
	(0.000)	(0.000)	(0.001)	(0.005)	(0.003)	(0.012)	(0.001)	(0.000)	(0.009)	(0.005)	(0.005)	(0.084)
Ave. IF of prior articles												
First author							0.115***	0.116***	0.117***	0.575***	0.604***	0.601***
							(0.007)	(0.006)	(0.006)	(0.034)	(0.028)	(0.029)
Last author							0.251***	0.258***	0.257***	1.200***	1.228***	1.289***
							(0.008)	(0.007)	(0.007)	(0.043)	(0.038)	(0.038)
Num. of first author's prior articles: Reference group is 1-5 papers												
6 - 10							-0.038	-0.076***	-0.126***	1.190***	1.427***	1.378***
							(0.028)	(0.020)	(0.020)	(0.257)	(0.208)	(0.210)
10 +							-0.024	-0.146***	-0.172***	3.463***	2.375***	2.275***
							(0.032)	(0.022)	(0.022)	(0.277)	(0.234)	(0.238)
Num. of last author's prior articles: Reference gorup is 1-5 papers												
6 - 10							0.122***	0.055***	0.119***	1.549***	1.698***	1.774***
							(0.025)	(0.018)	(0.019)	(0.235)	(0.191)	(0.197)
10 +							0.169***	0.160***	0.170***	3.391***	3.315***	3.297***
							(0.023)	(0.017)	(0.017)	(0.207)	(0.170)	(0.174)
Observations	206,879	306,228	291,019	206,879	306,228	291,019	107,648	154,226	147,007	107,648	154,226	147,007
R-squared	0.491	0.526	0.531	0.311	0.293	0.286	0.506	0.546	0.551	0.326	0.307	0.305
More addresses than authors	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Publish year (# of years: 24)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subfield (# of subfields: 180)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subfield * Publish year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State (# of states: 53)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Author ethnicity (Each:9)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

NOTE: Standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1. Covariates include all authors' ethnicities and dummies for states, publish year, subfields, and interations between publish year and subfields.

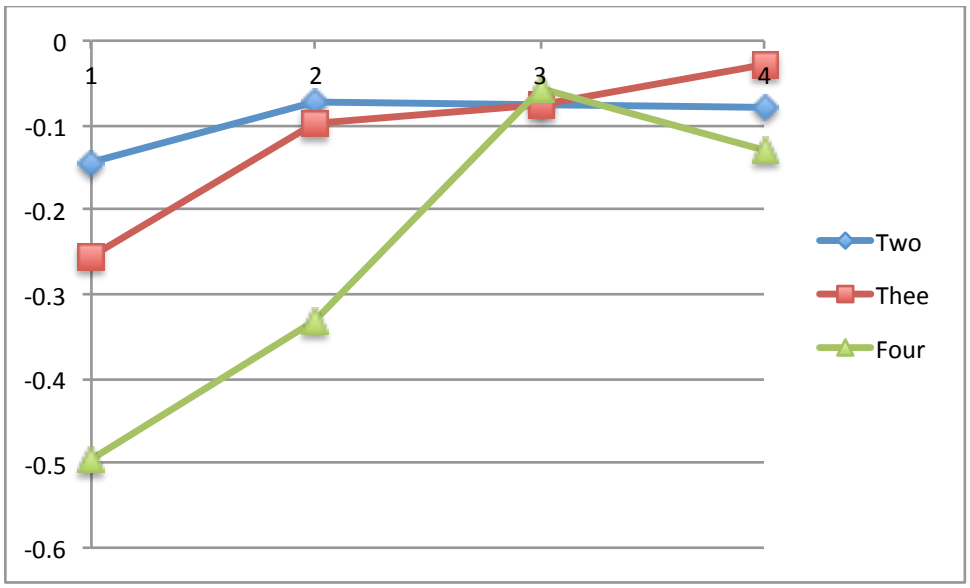

Figure 1a. Coefficients for Impact factor regressions

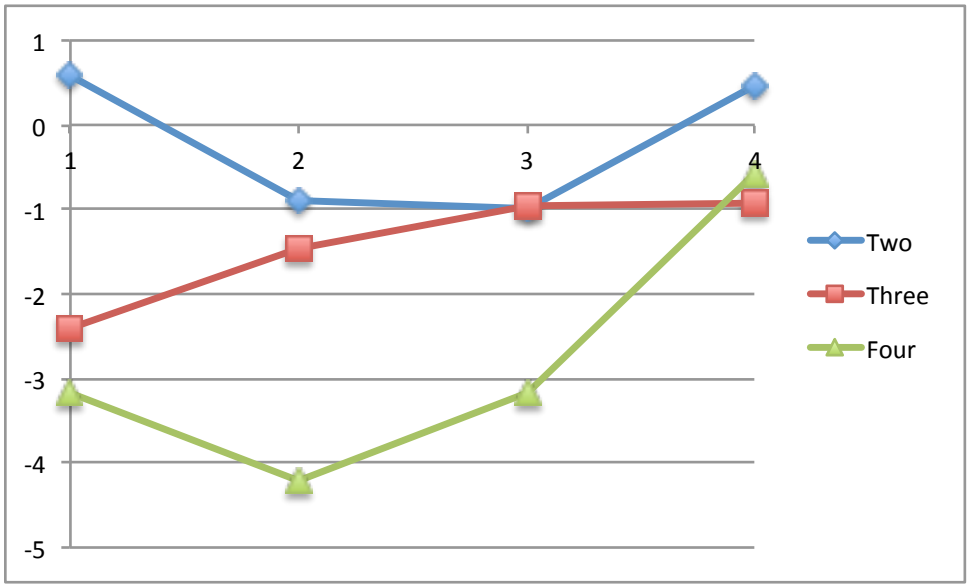

Figure 1b. Coefficients for Citations regressions

Figure 1. Coefficients on Interactions between homophily index and categorical dummies for number of previous papers

Appendix A: Characteristics of the Data in Our Analysis

Table A1: Numbers of papers and names in data set of US-address papers from WOS, 1985-2008

Number of papers	2,570,999
Number of author names in total	7,415,643
Number of *Different* author names	1,303,224
Names appearing only once	569,618
Nmaes appearing multiple times	733,606
Fields	11
Number of individuals	1,960,088
Number of individuals of those appearing once	569,618
Number of individuals of those appearing more	1,390,470

Table A2: Mean and standard deviations of statistics for papers in the two, three, and four authored sample in text analysis

	(1) Two-author articles	(2) Three-author articles	(3) Four-author articles
Same ethnicity (All authors)	0.43	0.15	0.08
	(0.49)	(0.36)	(0.26)
Homophily index	0.30	0.27	0.23
	(0.17)	(0.11)	(0.08)
Impact factor	2.32	2.59	2.89
	(2.90)	(3.00)	(3.21)
Citation percentile	45.71	46.25	46.62
	(31.72)	(31.32)	(31.06)
Num. of references	29.67	29.62	29.93
	(20.07)	(18.67)	(18.00)
Distribution of Number of addresses			
One	0.69	0.52	0.41
	(0.46)	(0.50)	(0.49)
Two	0.26	0.33	0.33
	(0.44)	(0.47)	(0.47)
Three and above	0.05	0.16	0.26
	(0.22)	(0.36)	(0.44)
Average impact factor of previous papers			
First author	2.49	2.65	2.71
	(2.26)	(2.25)	(2.20)
Last author	2.54	2.71	2.86
	(2.23)	(2.14)	(2.08)
Second author		2.66	2.72
		(2.24)	(2.20)
Third author			2.74
			(2.17)
Distribution of Number of Authors' prior papers since 1985			
First author			
0	0.33	0.34	0.35
	(0.47)	(0.48)	(0.48)
1-5	0.40	0.41	0.40
	(0.49)	(0.49)	(0.49)
6-10	0.11	0.10	0.10
	(0.31)	(0.30)	(0.31)
10+	0.16	0.15	0.15
	(0.37)	(0.35)	(0.35)
Last author			
0	0.18	0.19	0.20
	(0.38)	(0.39)	(0.40)

1-5	0.31	0.30	0.28
	(0.46)	(0.46)	(0.45)
6-10	0.15	0.15	0.14
	(0.36)	(0.36)	(0.35)
10+	0.36	0.37	0.37
	(0.48)	(0.48)	(0.48)
Second author			
0		0.30	0.34
		(0.46)	(0.47)
1-5		0.36	0.37
		(0.48)	(0.48)
6-10		0.13	0.11
		(0.33)	(0.32)
10+		0.21	0.18
		(0.41)	(0.38)
Third author			
0			0.32
			(0.47)
1-5			0.34
			(0.47)
6-10			0.12
			(0.32)
10+			0.22
			(0.42)
Number of observations	478349	569015	469703

Table A3: Estimated Coefficients and Standard Errors for Interactions of Homophily of Paper and Previous Paper of Last Authors

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Two-authored articles			Three-authored articles			Four-authored articles		
VARIABLES	IF	Citation percentile		IF	Citation percentile		IF	Citation percentile	
Homophily * No previous paper	-0.144***	0.587	1.003*	-0.257***	-2.410***	-1.602*	-0.497***	-3.161**	-1.666
	(0.047)	(0.620)	(0.598)	(0.072)	(0.874)	(0.834)	(0.102)	(1.281)	(1.208)
Homophily * Num. Papers 1 - 5	-0.072**	-0.900*	-0.693	-0.097*	-1.476**	-1.172*	-0.330***	-4.221***	-3.227***
	(0.036)	(0.501)	(0.488)	(0.051)	(0.684)	(0.663)	(0.085)	(0.983)	(0.944)
Homophily * Num. Papers 6 - 10	-0.076*	-0.998	-0.779	-0.076	-0.967	-0.729	-0.056	-3.166**	-2.997**
	(0.045)	(0.652)	(0.634)	(0.063)	(0.907)	(0.883)	(0.112)	(1.265)	(1.225)
Homophily * Num. Papers 10 +	-0.080**	0.475	0.706	-0.027	-0.937	-0.852	-0.130**	-0.574	-0.182
	(0.033)	(0.488)	(0.476)	(0.041)	(0.591)	(0.574)	(0.058)	(0.791)	(0.767)
Observations	478,349	478,349	478,349	569,015	569,015	569,015	457,667	457,667	457,667
R-squared	0.557	0.363	0.394	0.590	0.349	0.386	0.581	0.339	0.379
Impact factor	No	No	Yes	No	No	Yes	No	No	Yes
Previous publication records	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Publish year (# of years: 24)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subfield (# of subfields: 180)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subfield X Publish year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State (# of states: 53)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Author ethnicity (Each:9)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

NOTE: Standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1. Covariates include all authors' ethnicities and dummies for states, publish year, subfields, and interations between publish year and subfields.

Appendix B: Analysis of Papers with 5 to 10 co-authors

Table B1: Means and standard errors for Summary statistics fors with over four authors

	(1)	(2)	(3)	(4)	(5)	(6)
Number of authors	Five	Six	Seven	Eight	Nine	Ten
Homophily index	0.21	0.19	0.18	0.17	0.16	0.15
	(0.08)	(0.07)	(0.06)	(0.06)	(0.06)	(0.05)
Citation percentile	46.75	46.79	46.84	46.83	46.97	46.90
	(30.98)	(30.95)	(30.92)	(30.94)	(30.91)	(30.98)
Impact factor	3.19	3.47	3.77	4.04	4.35	4.57
	(3.46)	(3.66)	(3.95)	(4.20)	(4.55)	(4.68)
Num. of addresses	2.27	2.57	2.87	3.16	3.47	3.79
	(1.24)	(1.44)	(1.61)	(1.81)	(1.99)	(2.24)
Num. of references	30.89	31.63	32.35	32.99	33.46	33.58
	(17.45)	(17.40)	(17.38)	(17.58)	(17.41)	(17.36)
Observations	322502	207854	123586	74468	44350	27466

Standard deviations in parentheses

Table B2: Estimated Coefficients and Standard Errors for Factors That Affect Impact Factor and Citations

	(1) Five-authored	(2) Six-authored	(3) Seven-authored	(4) Eight-authored	(5) Nine-authored	(6) Ten-authored
Panel A: Dependent variable is Impact factor						
Homophily index	-0.554***	-0.548***	-0.478***	-1.564***	-1.100***	-0.871
	(0.070)	(0.112)	(0.176)	(0.268)	(0.421)	(0.620)
Number of references	0.015***	0.015***	0.016***	0.012***	0.016***	0.012***
	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)	(0.002)
Number of addresses	0.045***	0.056***	0.066***	0.088***	0.070***	0.065***
	(0.005)	(0.006)	(0.007)	(0.009)	(0.013)	(0.016)
Observations	319,351	206,204	122,735	74,004	44,065	27,328
R-squared	0.548	0.541	0.537	0.541	0.558	0.569
Panel B: Dependent variable is citation percentile						
Homophily index	-5.703***	-6.563***	-9.342***	-10.918***	-8.128**	-16.278***
	(0.764)	(1.137)	(1.639)	(2.314)	(3.323)	(4.734)
Number of references	0.401***	0.376***	0.348***	0.303***	0.316***	0.273***
	(0.003)	(0.004)	(0.005)	(0.007)	(0.010)	(0.012)
Number of addresses	0.289***	0.305***	0.481***	0.569***	0.393***	0.356***
	(0.052)	(0.057)	(0.068)	(0.082)	(0.099)	(0.122)
Observations	319,363	206,212	122,739	74,006	44,066	27,328
R-squared	0.330	0.332	0.345	0.366	0.400	0.425
More addresses than authors	Yes	Yes	Yes	Yes	Yes	Yes
Publish year (# of years: 24)	Yes	Yes	Yes	Yes	Yes	Yes
Subfield (# of subfields: 180)	Yes	Yes	Yes	Yes	Yes	Yes
Subfield * Publish year	Yes	Yes	Yes	Yes	Yes	Yes
State (# of states: 53)	Yes	Yes	Yes	Yes	Yes	Yes
Author ethnicity (Each: 9)	Yes	Yes	Yes	Yes	Yes	Yes

NOTE: Standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1. Covariates include all authors' ethnicities and dummies for states, publish year, subfields, and interations between publish year and subfields.