Collaborating with People Like Me: Ethnic Coauthorship within the United States

Collaborating With People Like Me: Ethnic Co-authorship within the US

Richard B. Freeman, Harvard University and NBER

Wei Huang, Harvard University

Abstract

This study examines the ethnic identity of authors in over 2.5 million scientific papers written by US-based authors from 1985 to 2008, a period in which the frequency of English and European names among authors fell relative to the frequency of names from China and other developing countries. We find that persons of similar ethnicity co-author together more frequently than predicted by their proportion among authors. Using a measure of homophily for individual papers, we find that greater homophily is associated with publication in lower impact journals and with fewer citations, even holding fixed the authors' previous publishing performance. By contrast, papers with authors in more locations and with longer reference lists get published in higher impact journals and receive more citations than others. These findings suggest that diversity in inputs by author ethnicity, location, and references leads to greater contributions to science as measured by impact factors and citations.

The globalization of science has changed the ethnic and national origin of US-based scientists and engineers (Freeman 2006). 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 research assistants during their PhD studies and as post-doctoral workers afterward (Bound et al. 2009; Franzoni et al. 2012; Stephan 2012). 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 (Borjas and Doran 2013).

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 (WoS) with US addresses had "English" names, 13% had European names while 30% had names of other ethnic groups. 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 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; and whether homophily in collaborations is associated with more or less valuable scientific work.

To determine the extent of homophily in scientific collaborations, we use names to identify the ethnic identify of the co-authors of 2.57 million papers with US addresses in the Thomson-Reuters Web of Science data base. To assess the scientific contribution of papers with differing ethnic composition, we examine the impact factors of the journals in which the papers appear and the numbers of forward citations to the paper. Despite extensive studies of co-authorship patterns among scientists (Barabasi et al 2002; Newman 2001a, 2001b, Jones et al. 2008), to our knowledge this is the first study of homophily in scientific collaborations. 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 ethnic distribution of all authors of scientific papers or of authors in the specific discipline in which the paper fits.

2.Homophily is associated with publication in a lower impact factor journal and with fewer citations of papers.

3. Researchers with weaker previous publications records are especially likely to write papers with persons of the same ethnicity, but this accounts for only part of the reduced impact factor and citations associated with higher homophily among co-authors.

Section one documents the existence of homophily in the ethnic composition of co-authorship for US-based papers and develops an Homophily Index at the level of

papers for ensuing analysis. Section two estimates the relation between the Homophily Index of a paper and the impact factor of the journal of publication and numbers of future citations, conditional on other characteristics of papers and authors. Section three examines the previous publication record of researchers who are more/less likely to write papers like themselves and examines whether the previous publication record accounts for the negative relation between homophily and impact factors and citations. We conclude by placing the estimated effect of homophily on impact factors and citations in the context of the other factors associated with those outcomes of scientific work.

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, using the Thomson-Reuters Web of Science database from 1985 to 2008, we created a file of papers 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 thus potentially form a collaborative project. Limiting the sample to papers written solely in the US allows us to construct a probabilistic model of the distribution of ethnic co-authorship absent homophily that would be difficult to develop for foreign collaborations.

Second, using William Kerr's name-ethnicity matching program, we assigned an ethnic identity to authors. Kerr's program combines information on the distribution of names by ethnicity and the metropolitan statistical areas in which individuals live to

determine their likely ethnicity (Kerr 2008, Kerr and Lincoln 2010). The identification hinges on the fact that last names such as Kim are especially likely to represent Koreans while names like Zhang are likely to be Chinese, and so on. Because persons of a particular ethnicity live disproportionately in some areas, area 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' surnames, initials of first names and in later years first names and addresses. On the notion that first authors and last authors have greatest responsibility for the paper, we limited our analysis to papers in which we identified the ethnicity of first and last authors. Thus our sample has ethnic identification for both authors in two-author papers, and for the first and last author in other papers, but sometimes lack ethnic identification for some intermediate authors in papers. We match names with ethnicity at a rate of 86%. The rate of match increased 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. Appendix table A1 gives the numbers of papers in our sample after the matching process. In total we had 2.634 million papers, of which 2.299 million were the co-authored papers on which we focus. Of those 1.505 million were papers with two to four authors which constitutes the main sample on which

we report results; they constitute 65% of all co-authored papers in our data set. Appendix Table A2 gives the mean and standard errors of key variables for those papers. We also analyzed papers with five to ten authors, which gave results much like those for two to four authored papers and report some of those results. Table A3 gives the means and standard errors for the summary statistics of five-authored to ten-authored papers. We examine papers separately by numbers of authors to allow for possible different interactions among co-authors as the number of authors changes.

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 consists of middle positioned authors whose ethnicity we could not identify. The biggest change in the ethnic distribution of names is the near tripling in the frequency of Chinese names, which increased from 4.79 percent in 1985 to 14.45 percent in 2006 and then dropped slightly in 2007 and 2008. The proportion of names associated with ethnic origins in other developing countries 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 dropped from 56.56 percent in 1985 to 45.56 percent in 2008, while the proportion of European names dropped from 13.47 percent to 11.18 percent.

The distributions in the table do not distinguish between American-born persons and foreign-born persons of the same ethnicity. For the fastest growing group, persons

with Chinese names, the increase is driven by increased numbers of researchers born overseas. 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 US-born Chinese are more Anglicized than the names of China-born Chinese, 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 of Chinese names, this implies that 85 percent of the increased number of Chinese named authors in the US were born in China.

1.1 Measuring homophily in the co-author population

To determine 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 observed population of authors (vide table 1). If 20% of authors in the population of 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 for papers with 2-4 authors, 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 and funded the work and who set the overall direction of the research. Intermediate positions reflect the activity of other contributors of varying importance in the project. Panel A shows that in the 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 first authors and 60.2 percent of 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 dominance of that group among senior scientists.

Our test for homophily in co-authorship compares the observed ethnic distribution of the authors on papers to the counter-factual distribution that would arise from random draws of co-authors from the pool of authors by position. Rather than giving the full distribution of ethnicity, the table records the proportion of papers in which all authors are of a given ethnicity and treats other ethnic groups as “other”. 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 with all authors of 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. The differences 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 the proportion of 3 and 4 authored papers in which those two authors had the same ethnicity and found that this proportion (not reported in the table) also exceeded that produced by chance. We conclude that *homophily is substantive among co-authors of scientific papers.*

To see the extent to which the observed homophily reflects the concentration of persons with a given ethnicity among scientific disciplines or residence in the same region of the country, we developed a regression model that modifies the random proportion by geographic location and field. In this analysis someone residing in, say California, where many Chinese reside, would be more likely to have a Chinese co-author than someone in Houston. Someone in a specialty with many Chinese specialists 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. As a further check, we compared the actual distribution of co-authors by ethnicity with the expected distribution based on the probability model applied to each of the twelve fields in which WoS

classifies papers for every year in our data set and also found strong evidence for homophily.

The comparisons of the observed pattern of co-author ethnicity with the pattern that would arise by chance document that homophily is a feature of scientific collaborations but do not identify the structure of preferences or behavior that produces the homophily. Researchers could be disproportionately writing with people like themselves because persons in each group prefer to work with persons of their ethnicity; or because persons in one group prefer to work with persons of their ethnicity while persons in other groups have no such affinity; or from different rates of preference for homophily among the groups. Since every author is a co-author of someone in the data set it is not possible 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 would show more persons working with their own group than the random model. But the same observed distribution could arise if persons in group B preferred working with persons like themselves while those in group A did not care; if both groups

cared equally; and so on. Sophisticated modeling might give some insight into which group evinces greater preferences for working with people like themselves but are no substitute for direct information about preferences . To the extent that preferences toward working with one's own group vary *within* ethnic groups, moreover, models based on average preferences will miss the role of heterogeneity of preferences in determining observed homophily.

1.2 The Homophily of a Paper

With many ethnic groups and many authors on papers, measuring homophily as a dichotomous "all persons of the same kind" variable does not adequately capture the phenomenon. A paper with three authors of one ethnicity and one author of another ethnicity exhibits more homophily than a paper with two authors of the same ethnicity and two authors of different ethnicity, while the simple dichotomy puts them all under the same non-homophily category. The probabilistic framework underlying table 2 directs attention at a further complication: dependence of papers written by persons of the same ethnicity due to the proportion of that ethnicity in the population. A paper with all authors from a relatively small group is more reflective of homophily than a paper with all authors from a relatively large group. Having a paper with all Korean-named authors, for instance, is stronger evidence of homophily than a paper with all English-named authors. Indeed, a four-authored paper with 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. Analysis of homophily at the level of papers must take account both of the ethnic similarity among authors and the ethnic distribution of the underlying distribution of the population of authors.

We use a two step procedure to differentiate homophily of a paper due to behavior from the homophily that would result from the size of ethnic groups in the population of authors. First, we create an Homophily Index that measures the ethnic concentration of authors on a paper regardless of whether it comes from homophilous behavior or the concentration of an ethnicity in the population. Our Index is analogous to the Herfindahl-Hirschman Index of concentration of production among firms that specialists in industrial organization use to measure market structure. Let N = number of identified ethnic groups and s_i be the share of the ith group in the authors of a paper and let A be the number of authors on a paper. Then we define the Homophily Index for a given paper as the sum of the squares of the shares of each group among the authors of the paper:

1) $H = \sum s_i^2$, where the summation is from i=1 to N

If all of the authors on a paper have the same ethnicity, H is 1.0, which is the maximum value of homophily. If the paper has authors of different ethnicity, H takes different discrete values for papers depending on the number of ethnic groups and number of authors on a paper. When the number of ethnic groups is greater than or equal to the number of authors (as in our analysis), the minimum value of H for a paper is 1/N. For two-authored papers, H has 2 values: the maximum 1.0 and 1/2 (each author of

different ethnicity). For a three-authored paper, H has three values: the maximum 1.0, a minimum of $1/3^{rd}$ (each author of a different ethnicity), and an intermediate value of 5/9ths (2 authors with same ethnicity and one with different ethnicity). For a four-authored paper H has five values: maximum 1.0, a minimum of 1/4th, and intermediate values of 6/16ths (one group has 2 authors and the remaining two authors come from different groups), 10/16ths (three authors from one group and one from a different group) and 1/2 (authors evenly divided between two groups). With the nine ethnic groups we identify and the residual unidentified ethnicity group, the H index follows this pattern through the ten-authored papers that are the maximum we consider.

The nature of our data produces a slight deviation in the H-index from that just described. Because persons with a given surname can be of differing ethnicity, the ethnicity identification program does not produce pure dichotomous measures of ethnic status. Instead, it gives a probability to the ethnicity of a person – for instance Kim might be associated with 99.0% Korean and 1% with some other ethnicity. We use the full information in the data so that a "dummy variable" for Kim being Korean takes the value 0.99 rather than 1.00 and so on. The probabilities in most names are sufficiently close to 1 for one ethnicity and 0 for all others that the distribution of the Homophily Indexes for papers follows closely the pattern based on 0/1 measures described above.

While the Homophily index captures the concentration of authors by ethnicity, it does not adjust for the differing frequency of ethnic groups in the population of authors. The index gives a paper with all Korean names the same 1.0 H-measure as it gives a paper with all-English names, although, as noted, the all-Korean paper is a rarer event and thus reflective of greater ethnic affinity than the all-English paper. To take account of this in ensuing statistical calculations, we include ethnic "dummy variables" for every author on a paper, where by dummies we mean the probabilities that a given name is of a

given ethnicity. A regression of H on the ethnic "dummy variables" will give relative larger coefficients on the dummies for groups that have a larger representation of authors compared to any base reference group. To illustrate the calculation and show how it captures the impact of population size on the index, consider a world of authors from two ethnic groups – a large English-named group and a small Korean-named group. The homophily index is 1 when either English-named authors write with English-named authors and when Korean authors write with Korean authors, and have the 1/2 otherwise. With more English-named authors in the population, random association will produce more values of 1 for English-named authors than for Korean-named authors. If population size is the sole reason for homophily, the regression of H on a dummy variable for the first author having an English-name and for the second author having an English-name will each obtain positive coefficients. If we then regress an outcome of the paper – says the impact factor of the journal of publication – on the homophily index and dummy variables for ethnicity of authors, the homophily index captures the effect of homophily above and beyond that due to the size of ethnic groups, which are reflected in the dummy variables. Put differently a regression of the outcome on homophily and the ethnic dummy variables of authors gives the coefficient on homophily that one would obtain from taking the residual of the regression of the outcome on the ethnic dummy variables on the residual of the regression of the homophily index on the ethnic dummies – the effect of homophily on outcomes independent of the ethnic groups on the paper.

To make sure that our findings do not hinge critically on the way we measured homophily and adjusted for its dependence on ethnic group size, we have experimented with other metrics and find patterns analogous to those in the text.

2. Relation between homophily of a paper and impact factors and citations

Does working with persons of the same ethnicity produce papers that have greater or lesser scientific impact than working with persons of different ethnicity?

There are factors that might lead homophily to be associated with higher quality research. To the degree that working with persons like oneself 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 there are factors in the opposite direction as well. To the degree that co-authoring with persons like oneself reflects tastes/preferences for socializing 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 think more alike for whatever reason, working together may reduce the diversity of perspectives that can produce novel scientific results.

To see which effect, if any, dominates, we examine the relation between the Homophily Index and two measures of the scientific contribution of papers – the impact factor of the journal which published the paper and the number of citations the paper

received as of 2008, the last year of our sample. Impact factors have the advantage of being available upon publication and remaining fixed over time. But they are an imperfect measure of the quality of a paper (European Association of Science Editors, 2007). They show whether the paper was judged worth publication in a more prestigious journal by presumably tough editors and reviewers but not whether any reader of the journal found it useful or not. For this reason, numbers of citations are arguably a more accurate indication of the paper's scientific merit. They also have the virtue of reflecting the "wisdom of crowds (of knowledgeable scientists)" rather than the views of a few. But they also are subject to problems (International Mathematical Union, 2008). One problem is that citations vary with the number of scientists working in a field and with field specific norms for citing others' work. We include of dummy variables for 180 sub fields to control for this problem. Another problem is that citations likely depend on the size of a scientists' network: a paper by a senior researcher with many students and collaborators may gain more citations than a comparable paper by a new researcher with fewer connections. The fact that citations follow a distinctive non-normal distribution, with 20% to 30% of papers obtaining no citations over an extended period and with highly cited papers following a power law distribution (Redner, 2005; Gupta et al, 2005) creates statistical problems. To deal with this, we transform the number of citations of a paper into a percentile distribution based on citations in the year of publication and use the 0-100 percentile as our measure of citations.

Since journal impact factors are calculated from the citations of the papers in a journal in the previous two years, impact factors and number of citations of articles are highly correlated. But even so the two measures have enough independent variation to justify treating them separately. Within the same journal, citations of papers varies widely. Across journals some papers in lower impact journals invariably gain more citations than most papers in high impact journals. Lozano, Larivière, and Yves Gingras (2012) show that since 1990 the relation between impact factors and paper citations has weakened, possibly because Internet search engines make it easier to find relevant articles in less widely circulated journals.

We estimate the effect of homophily (H) on the impact factor of a paper (I) and the citation percentile of the paper (CP) using the following linear regression model:

(2) I or CP = a H + b Vector of "Dummies" for Ethnicity of Authors + c Number of Addresses + d Number of References + e (Vector of dummy variables for year of publication, subfield, state, and interactions between publication year and subfield) + residual

The vector of ethnicity dummies for paper authors picks up the effect of differences in the size of ethnic groups on the Homophily index. It is a key covariate to identify the impact of homophily behavior on outcomes exclusive of the size of ethnic groups as well as of any ethnic-specific pattern in outcomes.

We introduce the number of addresses on the paper as an indicator of the likely

diversity of ideas in the research. 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, which suggests that the coefficient on number of addresses will be positive. The measure of addresses has one problem, however. WoS reports the addresses of all authors on its data base, but before 2008 it did not link addresses to specific authors so we cannot tell how many authors on a multi-authored paper worked at any given address: on a 4 authored paper with two addresses, all authors but one could be at one or other of the addresses and or the authors could split evenly between the addresses.

The number of references in the paper indicates the breadth of knowledge on which the paper presumptively relied and thus measures "the shoulders" of previous researchers the paper built on. We expect it to obtain a positive coefficient on impact factors and citations. But number of references is an imperfect measure of the breadth of information in the paper, as authors put in references for other reasons as well.

The subfield dummy variables in the list of covariates allow for different levels of impact factors and citations among fields.

Table 3 summarizes our estimates of the equation for impact factors (panel A) and citation percentile (panel B). Columns 1-3 with the heading *main sample* report results for the 2-4 author papers that we focus on in our analysis. Columns (5) to (9) give the results for five to ten author papers that make up the bulk of the remaining co-authored papers in the WoS. The negative significant coefficients on the Homophily index for Impact factors and Citation percentiles in nearly all of the calculations shows that greater homophily is associated with publication in lower impact journals and fewer citations,

which presumptively implies that those papers are of lower quality than other papers. By contrast, the number of addresses and references in a paper have significant estimated positive effects on the impact factor of the journal of publication and number of citations. To the extent that the estimated impact of homophily reflects a narrower research perspective among persons of the same ethnic group and that the number of addresses and references reflect a wider research perspective among authors in different locations and/or who rely on a larger base of previous work, both sets of results can be interpreted as reflecting the value of diversity in the inputs going into a paper on its contribution.

To check the robustness of our results, we also estimated equation 2 on 2-4 authored papers in Pub Med, which many analysts use in scientometric work because Torvik and Smalheiser (2009) developed a sophisticated algorithm for differentiating same-named people. While name disambiguation is not an issue in equation (2) it is critical in the analysis in section 3, which estimates the part of the observed homophily-outcome relation that is due to differences in the past publications of authors in our sample with differing levels of homophily. Table 4 gives the estimated coefficients of impact factor and citation percentile on homophily, number of addresses, and number of references in 2-4 authored papers in the PubMed data set. The results are quite similar to those in table 3: negative coefficients on homophily and positive coefficients on addresses and references.

3. Probing the Negative Homophily-Paper Outcome Relation

The negative relation between the homophily of a paper and the impact of its journal of publication and the number of citations to the paper could be due to two possible factors. The first is selectivity of researchers. Researchers of lesser scientific prominence, measured say by their prior publication performance, may disproportionately write with persons like themselves and produce papers with publication and citation outcomes consistent with their past performance. The second reason is that interactions among persons of similar ethnicity produce less novel research than interactions among persons from more diverse backgrounds. In this section we examine the extent to which the weaker outcomes of papers with greater homophily in tables 3 and 4 are attributable to the presumably weaker publication performance of their authors.

This analysis faces one major problem: disambiguation of names in determining the past publication record of authors in the WoS. Until 2008 the WoS reported the first

initial and last name of authors, which creates potential errors in attributing papers to researchers with common names and first initials. The J. Kim who is first author on a given paper may have not written any earlier papers but his or her namesake J. Kim may have done so, producing measurement error if we attach the second J. Kim's papers to the first J. Kim. Appendix Table A4 gives the statistics regarding the distinct names in our WoS data set. Our sample contains over 2.57 million papers and over 7.4 million names, of which 1,303,22, are distinct names, implying that on average a name appears 5.69 times. Of those, 569,618 names (nearly 44 percent of all distinct names) appear once and thus cannot be used to obtain a research track record with which to judge productivity prior to the paper in our data. To differentiate which of the 733,606 names that appear more than once reflect the same person and which reflect different persons, we assume that people with the same names writing in the same scientific field are the same while those with papers in different fields are another person. In our disambiguation the J. Kim who publishes on physics is different than the J. Kim who publishes on biochemistry. We use the twelve major fields for which WoS sorts papers to differentiate same named authors. The resultant calculation yields 1,390,470 names-field that appear more than once – nearly twice as many distinct authors as the number of author names with more than a single publication. By construction, differentiating names by field produces more individual authors.

We could further differentiate names by narrow sub fields but this risks failing to attribute papers to an author whose research crosses narrow disciplinary lines. Instead of further dividing names by field, we undertook robustness checks on our findings. We eliminated all names with "large" numbers of papers for varying definitions of large and obtained results similar to those in the text table. We also do separate calculations on the life sciences subset of the WoS that overlaps with Pub Med papers. As noted, the advantage of Pub Med is that Torvik and Smallheiser (2009) have developed a disambiguation algorithm of Pub Med names that is more sophisticated than ours. Thus, PubMed offers a more accurate test of relations between outcomes and past author publications at the cost of a smaller sample limited to the life sciences.

Our analysis proceeds in two steps. First, we assess the relation between the prior

publications of the first and last authors and the homophily index – a necessary prerequisite for those publications to influence the estimated impact of the index on paper outcomes. Second, we add measures of prior publications to the equation (2) regressions and compare the estimated coefficient on homophily in equations which include the past publication records to those in tables 3 and 4 which do not.

We focus on the publication records of the first and last authors since in most fields these are the two most important authors: the first author is often the younger person who does the bulk of the laboratory work and the last author is often the senior professor who initiated the project and in whose laboratory it took place. A preliminary analysis found little relation between the past publication records of intermediate authors and paper outcomes, supporting this decision.

We examine two measures of authors' prior publication records: number of papers written and the impact factors of the journal in which those papers appeared, which can be viewed crudely as quantity and quality of past work. We formed three dummy variables from these measures to reflect how the the authors on a given paper compared to other first and last authors: whether the number of the previous papers was (1) above the median number of papers for authors in their position; or (2) below the median number of papers for authors in their position; and (3) whether the average impact factor of previous publications was above the median impact factor for authors in their position. The reference group for these measures are persons who wrote no previous papers and

had by definition no impact factors for previous papers.

Table 5 gives the estimated coefficients and standard errors for the regression relation between the homophily index (multiplied by 100 to make the coefficients easier to read) and these measures of past publication and of one additional variable – the number of addresses on the paper, whose coefficients indicate whether homophily is more or less likely when co-authors are in differing locations. Columns 1-3 give the results for the WoS sample. Columns 4-6 give the results for the PubMed sample. The estimated coefficients vary by number of papers and between the WoS and PubMed samples but the pattern of results are clear. The impact factors of previous papers for the last author and for the first author are negatively related to homophily: the index obtains significant negative coefficients in five of the six columns for the last author's publications and in four of the six columns for the first author's publications. By contrast, the number of papers written by authors shows a weaker and less consistent sign pattern across the numbers of papers and the data sets. These differences could reflect differences in the way teams form of different sizes or between the bio-medical sciences and others. Absent a formal model and study of the decisions to form teams, we simply note the correlation patterns, as they will affect the impact of adding these measures to analysis of the homophily-outcome relation.

Turning to the number of addresses, it is unclear a priori whether we should expect homophily to be larger or smaller among authors in the same locale or across

locales. On the larger 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. If authors care more about the ethnicity of geographically close collaborators with whom they interact often than about the ethnicity of distant collaborators, we would also expect greater homophily on papers with 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 than at some distant location. This would produce less homophily on papers with fewer addresses (Agrawal et al. 2008). Absent direct measures of how collaborators actually met and decided to work on project together (see Freeman et al. 2014 for survey evidence on how authors meet), the most we can do is to estimate the net direction of numbers of addresses on homophily. The estimates in table 5 on number of addresses shows strong negative effects on homophily for three-author and four-author papers in the WoS and PubMed but differing effects for the two-author papers between the WoS and PubMed samples.

Overall, the strong negative relation between the average impact factors of first and last author's paper and homophily and the varying weaker effects of numbers of papers on homophily, suggest that, as long as the publication performance of scientists shows some persistence over time, adding these measures as explanatory factors to the earlier regressions of paper outcomes on homophily will weaken the estimated adverse

effects of homophily. The regressions in tables 6 and 7 show that this is indeed the case for impact factors and citations in both the WoS and the PubMed data sets, but that homophily still has negative effects on the two outcome measures.

Columns 1-3 of Table 6 summarize the results of the regressions of impact factors on homophily and other key explanatory variables in the WoS data set with the new measures of the publishing record of first and last authors. The estimated negative coefficients on homophily are smaller than those in table 3 by about 15-20%. The measure for the average impact factor has a huge effect in these regressions, particularly for last authors, whose impact factor measure obtains a regression coefficient roughly twice that for first authors (0.608 vs 0.305 in the two-authored papers column 1 regression), which suggests that the last author has a greater role in getting a paper into a higher impact journal than the first author. The huge strong relation between average past impact factors and the current impact factor may reflect the persistence of authors publishing in the same or similar journals, possibly because of the subject matter of the research or its quality. The results on numbers of previous papers are more mixed, with positive estimated coefficients for last author's papers on the impact factor but negative coefficients on first-authors papers. Perhaps the quantity and quality of papers are substitutes for first authors, with the more able researchers moving to first authorship quickly while the less able obtain many middle author positions before they become first authors. The only way to unpack the observed pattern is to follow the life cycle trajectory

of beginning researchers, which goes beyond this paper.

Columns 4-6 of table 6 give comparable results when the dependent variable is the number of citations received by the paper through 2008. Addition of the publishing record measures reduces the negative coefficients on homophily for three-author and four-author papers while increasing the positive effect for two-author papers. The regression coefficients on the past publication variables for both first and last authors are strongly positive but are still considerably larger for last authors than for first authors. Whichever measure of past publication record we look at, last authors have a larger effect on the impact factor of publication and citations of a paper than first authors.

The results from the analogous analysis of the Pub Med data set summarized in Table 7 tell a similar story. The addition of authors' past publishing records reduces but does not eliminate the negative effects attributed to the homophily index on impact factors and citations (and slightly increasing the insignificant positive coefficient among three-author papers in column 5 compared to the comparable regression in table 4). The estimated effects of last authors' publication record on both impact factors and citations are considerably larger than those for first authors' publications.

Finally, note that in both tables addition of the authors past publication record reduces but does not eliminate the estimated positive relation between the number of addresses and number of references and the impact factor and citation percentile outcome measures .

4. Conclusion

Our analysis shows 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. It also shows that homophily is associated with papers with lower impact factors and fewer citations; and that while part of this effect is traceable to the weaker prior publication performance of the authors of papers with greater homophily, a substantive

negative relation between homophily and paper outcomes still remains. In addition, we find that 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 the pattern for homophily, addresses, and references is that greater diversity and breadth of knowledge of a research team contributes to the quality of the scientific papers that the team produces. While we have not tested this interpretation, it is testable. One approach would be to examine the terms, techniques, and references in papers. If diversity contributes to productivity by widening ideas, papers from more diverse collaborations should contain a wider range of scientific terms, use more varied equipment, procedures, or data and reference a wider range of previous work than papers from more homogeneous groups. Another approach would be to deal directly with the possibility that having co-authors of different ethnicity increases citations through network effects rather than through novel ideas by estimating the size of co-authors networks, say in terms of the number of their co-authors, and examine the linkages among the size of co-author networks, citations, and homophily of authors. The last and potentially most difficult way to illuminate homophily and its link to scientific outcomes would be to analyze the decisions of researchers to collaborate with each other. Here, the evidence that the attributes of last authors are more important than those of first authors in explaining outcomes suggests that a theory of collaboration might fruitfully begin by treating the last author as the initiator of the collaboration rather than by treating

the collaboration as a partnership among equals.

References

Agrawal, A., Kapur, D., & McHale, J. (2008). How do spatial and social proximity influence knowledge flows? Evidence from patent data. *Journal of urban economics*, *64*(2), 258-269.

Barabási, A. L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, *311*(3), 590-614.

Borjas, G. J., & Doran, K. B. (2012). The Collapse of the Soviet Union and the Productivity of American Mathematicians. *The Quarterly Journal of Economics*, *127*(3), 1143-1203.

Bound, J., Turner, S., & Walsh, P. (2009). *Internationalization of US doctorate education* (No. w14792). National Bureau of Economic Research.

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

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, A. A., Gonçalves, M. A., & Laender, A. H. (2012). A brief survey of automatic

methods for author name disambiguation. *Acm Sigmod Record*,*41*(2), 15-26.

Franzoni, C., Scellato, G., & Stephan, P. (2012). *Foreign born scientists: Mobility patterns for sixteen countries* (No. w18067). National Bureau of Economic Research.

Freeman, R. B. (2006). Does globalization of the scientific/engineering workforce threaten US economic leadership? In *Innovation Policy and the Economy, Vol 6* (pp. 123-158). MIT Press.

Freeman, R. B., Ganguli, I., & Murciano-Goroff, R. (2014). Why and wherefore of increased scientific collaboration. In *The Changing Frontier: Rethinking Science and Innovation Policy*. University of Chicago Press.

Gupta, H. M., Campanha, J. R., & Pesce, R. A. (2005). Power-law distributions for the citation index of scientific publications and scientists. *Brazilian Journal of Physics*, *35*(4A), 981-986.

Hegde, D., & Tumlinson, J. (2011). Can Birds of a Feather Fly Together?.*Work*. Available at SSRN: http://ssrn.com/abstract=1939587 or HYPERLINK "http://dx.doi.org/10.2139/ssrn.1939587"http://dx.doi.org/10.2139/ssrn.1939587

Huang, J., Ertekin, S., & Giles, C. L. (2006). Fast author name disambiguation in CiteSeer. ISI Technical Report. Available at: http://web.mit.edu/seyda/www/Papers/IST-TR_DisambiguationCiteseer.pdf 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, B. F., Wuchty, S., & Uzzi, B. (2008). Multi-university research teams: Shifting impact, geography, and stratification in science. *Science*, *322*(5905), 1259-1262.

Kerr, W. R., & Lincoln, W. F. (2010). The Supply Side of Innovation: H-1B Visa Reforms and US Ethnic Invention. Journal of Labor Economics, 28(3), 473-508.

Kerr, W. R. (2008). Ethnic scientific communities and international technology diffusion. *The Review of Economics and Statistics*, *90*(3), 518-537.

Lai, R. A. D'Amour, D. Doolin, G. Li, Y. Sun, V. Torvik, A. Yu, and L. Fleming. Disambiguation

and co-authorship networks of the U.S. patent inventor database. Research Policy, 2014.

Lines, Malcolm E. (1993) *A Number for Your thoughts* Institute of Physics (Bristol and Phil)

Lozano, G. A., Larivière, V., & Gingras, Y. (2012). The weakening relationship between the impact factor and papers' citations in the digital age. *Journal of the American Society for Information Science and Technology*, *63*(11), 2140-2145.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology*, 415-444.

Newman, M. E. (2001a). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, *98*(2), 404-409.

Newman, M. E. (2001b). Scientific collaboration networks. I. Network construction and fundamental results. *Physical review E*, *64*(1), 016131.

NSF Science and Engineering Indicators, 1993

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

Stephan, Paula (2012). How Economics Shapes Science, Cambridge, MA: Harvard University Press.

Strotmann, A., & Zhao, D. (2012). Author name disambiguation: What difference does it make in author-based citation analysis?. Journal of the American Society for Information Science and Technology, 63(9), 1820-1833.

Tanyildiz, Z. E. (2008). The effects of networks on institution selection by foreign doctoral students in the US. *PhD diss., Georgia State University*. Available at: HYPERLINK "http://digitalarchive.gsu.edu/pmap"http://digitalarchive.gsu.edu/pmap_

Torvik, V. I., Weeber, M., Swanson, D. R., & Smalheiser, N. R. (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. I., & Smalheiser, N. R. (2009). Author name disambiguation in MEDLINE. *ACM Transactions on Knowledge Discovery from Data (TKDD)*,*3*(3), 11.

Velden, T., Haque, A. U., & Lagoze, C. (2010). A new approach to analyzing patterns of collaboration in co-authorship networks *Scientometrics*, *85*(1), 219-242.

Wilhite, A. W., & Fong, E. A. (2012). Coercive citation in academic publishing. *Science*, *335*(6068), 542-543.

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, *316*(5827), 1036-103

Appendix

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, Sifan Zhou for her data support 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 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

The largest expansion was in China, which raised S&E PhDs to exceed US levels (Bound et al. 2009).

As determined by a name-ethnicity program developed by William Kerr.

Homophily refers to the "birds of a feather flock together" pattern in which people of similar backgrounds congregate together. Such behavior is found throughout social life: marriage, residence, business partnerships, seating arrangements, and so on (McPherson et al. 2001). Hegde and Tumlinson (2011) analyze the potential payoff from homophily.

The 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. See http://thomsonreuters.com/products_services/science/science_products/a-z/web_of_science/

We use first names if available. They are available for all 2008 papers and a small number of papers in 2006 and 2007.

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%. Note that these statistics are consistent an 86% identification of ethnicity for individual authors since (0.86) x (0.86) =0.74

For example a US born "Wang" might be named Don whereas someone born in China might be named Xia.

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

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.

Results available from authors on request.

These fields are Multidisciplinary Sciences, Agriculture, Biology, Biomedicine, Chemistry, Clinical Medicine, Engineering, Geosciences, ICT, Material Science, Mathematics, Physics

The existence of three or more groups can help identify preferences for working

with one's own group. For simplicity let one third of authors be in groups A, B, and C. If A is the only group that prefers to work with itself, the deviation from the random pattern will be largest for A while the secondary effects will be divided between B and C.

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.

When the number of authors exceeds the number of ethnic groups the H index follows a more complicated pattern as additional authors increase the size of one or more of the ethnic groups. For example, in a three-authored paper with two ethnic groups, the minimum of the H index is 5/9 (2 authors from one group, one author from two groups) rather than 1/3 while on a four-authored paper with 2 ethnic groups, the minimum will be 1/2 rather than 1/4th. This pattern parallels the analysis of leading digits along a line connected with Benford's law. See Lines (1993)

In an earlier version of this paper, we adjusted the Homophily Index for the size of groups by changing the actual index and obtained the same pattern of results.

The impact factor of a WoS journal in a year is the average number of citations to its articles in the preceding two years. See Thomson-Reuters, http://thomsonreuters.com/products_services/science/academic/impact_factor

Authors may face pressures by editors to cite their journal as part of the acceptance process (Willhite and Fong, 2012) or decide themselves to refer to papers from potential referees. If these factors affect the references of papers in a similar way, this will create measurement error in the true reliance on past work.

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

Other studies that disambiguate the names of scientists include (Huang 2013) and (Ferreira et al 2011).

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.

In the health sciences, a natural measure for such analysis would be the medical subject heading (Mesh) terms. Latent text analysis of keywords and of the combination of words in the text could also provide metrics of novelty.

PAGE 1