Online Appendix for "Geography, Ties, and Knowledge Flows: Evidence from Citations in Mathematics"

Keith Head* Yao Amber Li† Asier Minondo‡

August 21, 2018

A Construction of variables and samples

A list of the journals included in the database along, with the year of earliest article from that journal can be found at the following URL: `http://yaoli.people.ust.hk/HLM_Annex1.pdf`

A.1 Sources of data

Web of Science (previously called Thomson Reuters' ISI Web of Knowledge.)

We use the WOS to record citations (the dependent variable), the author lists to obtain coauthorship links, and to find the affiliations of authors. The affiliations allow us to construct ties variables from career histories and to measure geographic proximity. The WOS provides a record per each article published in the journals covered in the database. The record provides data on the title of the article, the journal in which it was published, the year of publication, the authors, the affiliation of the authors, and the cited articles.

From WOS we select all 255 journals included in the category "Mathematics" in 2009. Our database covers all the articles published in these journals in the period 1975–2009. However, for a large number of journals abstracting and indexing of articles started later than 1975. With

*Sauder School of Business, University of British Columbia, CEPR Research Fellow, Centre for Economic Performance (International Affiliate). keith.head@sauder.ubc.ca.

†Department of Economics and Faculty Associate of the Institute for Emerging Market Studies (IEMS), Hong Kong University of Science and Technology, Research Affiliate of the China Research and Policy Group at University of Western Ontario. yaoli@ust.hk

‡Deusto Business School, University of Deusto, Research affiliate of Instituto Complutense de Estudios Internacionales. aminondo@deusto.es

these limitations, the database contains information about 339,613 articles. A shortcoming of WOS is that it does not provide the affiliation for a substantial number of authors. The WOS provides affiliations for 69% of the author-article combinations. Following procedures described in A.3 we raise the fraction of affiliation identifications to 84%.

Academic genealogy data

The second main database used by this paper is the Mathematics Genealogy Project (MGP). The MGP records the doctoral degrees awarded in mathematics since the 14th century. The MGP provides the university and year in which each degree recipient completed their Ph.D., as well as the names of their doctoral advisors. We merged this data set with the citing authors and cited authors in our database. The MGP is not an exhaustive list of all mathematicians but we were able to match the records by author for around 44% of records.

Table A.1: MGP vs Non-MGP authors

Author	Career duration	#Institutions	USA(%)	#Coauthors	Productivity
MGP	5.8	2.0	31.5	4.0	2.3
	(6.6)	(1.3)	(46.5)	(5.2)	(7.5)
Non-MGP	5.1	1.9	22.7	3.7	1.9
	(6.1)	(1.3)	(41.9)	(5.1)	(7.0)

Note: Career duration is the difference between the last year and the first year in which the author appears in the database. USA reports the percentage of authors affiliated to a US university. Productivity is computed dividing the total citations received by the author by her career duration. Standard deviations in parentheses.

The MGP data are central to the analysis conducted here because they permit the construction of detailed educational ties that are pre-determined at the time the authors' careers begin. However, a natural concern is that these authors were *selected* for inclusion in the data set based on special characteristics. Table A.1 compares MGP authors with other authors on several relevant dimensions. The MGP authors have longer careers: the period over which they publish averages eight months more than non-MGP authors. Both types of authors work at two institutions on average and have four co-authors. The MGP authors receive on average 0.4 more citations per year but there is huge variation in productivity within both groups. In sum, MGP authors tend to be more active and prominent but the between-group differences seem small relative to intra-group variation. The most salient difference is that US-based mathematicians seem over-represented in the MGP. To the extent that mathematicians at US departments have different citation patterns, this will be more heavily weighted in the MGP sample. We address this in our empirical analysis by estimating distinct geography and ties effects for US-residents.

Mathematics subject classification data

We used Zentralblatt MATH (zbMATH) to obtain the Mathematics Subject Classification

(MSC) for the articles in our sample.[1] The MSC is a 5-digit classification scheme maintained by Mathematical Reviews and zbMATH which is used to categorize items in mathematics (broadly defined). We focus on the 3-digit codes (two numerical and one letter), of which there are 422 in the year 2000 revision. We also use 5-digit codes, which gives extra detail (2175 fields). An example of a 3-digit code is 15A, "basic linear algebra." Within that "inequalities involving eigenvalues and eigenvectors" is a 5-digit code. The drawback of using the 5-digit codes is a massive reduction in the estimating sample (which we explain in section B of this Appendix).

Geographic data

We consider three geography variables, distance, borders, and language difference. Each variable is expressed such that a large value indicates greater separation. The national border dummy takes the value of 1 if none of the authors of the citing papers are based in the same country as any of the cited authors. The language dummy is based on the official language of the country hosting each authors' institution, which need not be the native language of the author in question.

We extracted the latitude and longitude information for all top 1000 institutions from Google Maps, enabling construction of distances between each institution pair. We code the distance of authors at the same institution as zero. Much of the prior work uses coarse measures of location such as residing in the same metropolitan area. Even Belenzon and Schankerman (2013), who measure intercity distances, cannot calculate decay in citation propensities *within* cities. For example, within the Boston metro area, the distance between Harvard and MIT is only 3km but the distance of MIT to Brandeis University is 14km. This permits us to estimate the profile of information decay non-parametrically over fine and broad scales.

Using publications to track author locations over time, we calculate distances (and other measures of geographic separation) at the time the *citing* article is written. Past work using patents calculated distances between inventors using the cited inventors' addresses in the year the *cited* patent was obtained. For example, suppose paper i is being written in 2005. It may be more likely to cite paper d, written in 1980 at a very distant institution, if the authors of paper d had by 2005 moved closer to the authors of paper i, thus increasing their likelihood of interacting around the time paper i is written. Thus, our *contemporaneous* distance measure more precisely captures the geographic separation when the true knowledge flow occurs, i.e., when the new knowledge is created rather than at the time that the prior knowledge was created.

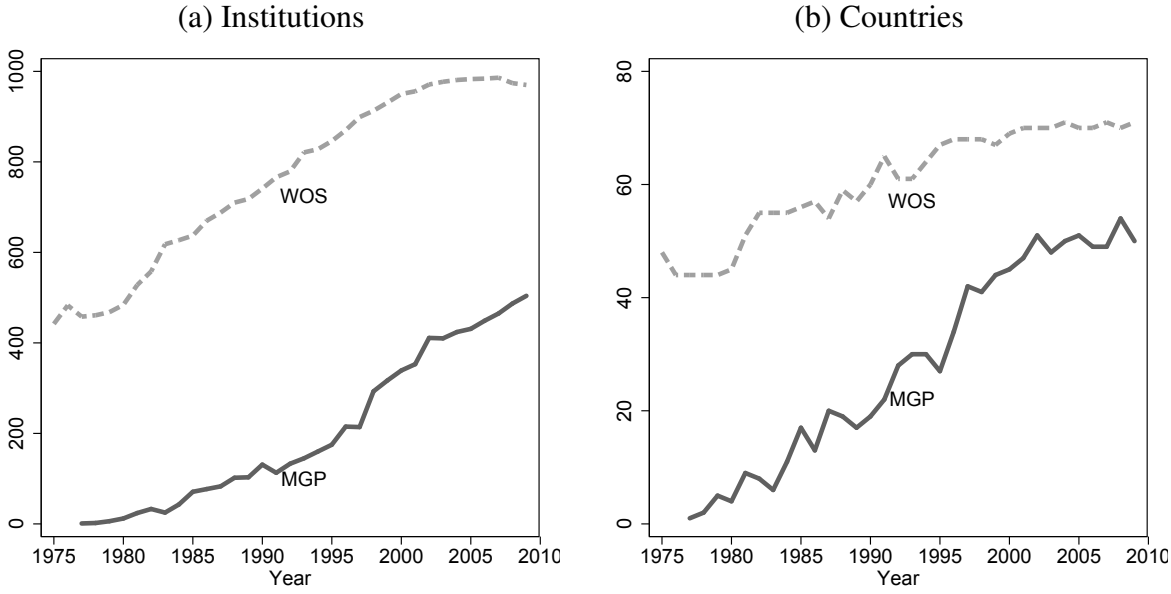
There is an important caveat regarding our contemporaneous distances. Location of each mathematician is revealed from their affiliations only in the years when they publish an article. Not surprisingly, there were many gaps in affiliation histories. As described in A.3, we fill these gaps through interpolation and extrapolation, assuming that moves occur in the midpoint

[1] zbMATH describes itself as "the world's most comprehensive and longest running abstracting and reviewing service in pure and applied mathematics." `https://zbmath.org/about/`

between the periods we observe location.

There has been a notable increase in the number of articles and authors per year; moreover, the rate of increase seems to have accelerated from the early 2000s onwards. The number of articles published in 1975 was 5,830, written by 5,193 different authors. The number of articles published in 2009 was 19,699, written by 22,787 different authors. Much of this huge expansion comes from the WOS adding 195 journals to the data base between 1975 and 2009. Considering only the journals included in 1975, we find a 30% increase in the number of articles and a doubling in the number of authors.

Figure A.1: Number of institutions and countries, 1975–2009



Note: Dashed lines correspond to the count of distinct institutions (left) and countries (right) represented in the sample of math citations obtained from entries in the Web of Science (WOS) database. Solid lines count only within the subset of citations from and to authors included in the Mathematics Genealogy Project (MGP).

Meanwhile, the numbers of institutions and countries represented in the WOS citation data increase over time. Figure A.1 shows that during the period 1975–2009 the set of institutions with citing or cited author affiliations rises to nearly 1000 (some institutions disappear) and the corresponding number of countries rises to 71. The sample containing MGP information on all authors starts very small but eventually represents 504 institutions located in 50 countries. Over the whole period there are 65 citing countries and 62 cited countries with a total of 1,113 dyads with at least one citation. This number of country pairs in our analysis is unprecedented in the citations literature, which has mainly focused on cross-metropolitan area citations within the United States.[2]

[2]Peri (2005) and Singh and Marx (2013) include international citations but the challenge of determining locations for individual patentees limited Peri's sample to 18 countries, whereas

A.2 Construction of estimating sample

Table A.2: Citation Data: Web of Science (WOS)

	Citing articles	Cited articles	Realized citations
Start	339,613	1,247,171	4,915,374
Study period*	339,613	987,056	3,665,145
Math. category journals	339,613	321,447	1,788,981
Partial affiliation data	221,908	162,457	1,044,673
Full affiliation data	187,062	133,429	749,257
Excluding self-citations	168,054	108,214	562,024
Authors at top 1000 inst.	131,347	86,536	425,399
With 5-digit MSC field	69,558	68,755	268,527
MGP data all authors	13,256	12,608	29,404

Note:* 1980–2009 for citing papers and 1975–2009 for cited papers.

The Web of Science data we extracted begins with 339,613 citing articles that yield a set of nearly five million citations to over a million distinct articles. Table A.2 shows how our sample declines to the much smaller sets (the last two rows) that we use in regressions. The first cut we make is to limit the period of *cited* articles to the period 1975–2009. Absence of pre-1975 WOS data papers reduces the set of cited articles by 21%. The WOS only identifies the first author of the cited articles. To identify the institutional affiliation of the first author, and the identity and affiliations of any coauthors, we matched the cited articles with our original database providing more complete information on the citing authors. As our database is restricted to the 255 journals included in Mathematics category, we can only identify the authors and coauthors of the cited articles belonging to this set. Only one third of the cited papers (containing about half the citations) were published in the pure math journals included in our database.[3] Inability to obtain complete affiliation information for the citing authors and the cited authors reduces the number of realized citations by 58% (0.75 million compared to 1.8 million). We then remove all self-citations, that is all article pairs where any of the citing authors has the same zbMATH author code as any of the cited authors.[4] This subtracts a surprisingly high one quarter of the realized citations.

There are 11,383 different affiliations for the citing authors and 7,722 different affiliations for the cited authors. To keep the set of required geographic information manageable, we select the 1000 affiliations with the highest number of citing articles. The top 1000 affiliations

Singh and Marx (2013) limit their sample to cited patents with US-resident inventors.

[3]The lost citations include books, book chapters, and proceedings. We also lose citations due to spelling discrepancies.

[4]A.4 describes how we identified and removed self-citations.

account for 76% of the realized citations observations (after all previous cleaning steps). Failure to obtain a subject classification from Zentralblatt MATH further shrinks the sample of realized citations by 37%.[5]

Applying the filters described above leaves us with 269 thousand realized citations to use in our initial estimations that omit educational histories. The biggest decline in realized citations occurs when we require MGP data to be available on all authors. The 89% reduction in realized citations in the last row of Table A.2 raises concerns that the new sample might not be representative.

Table A.3: Comparison of means in the Web of Science (WOS) and Mathematics Genealogy Project (MGP) samples

	Only realized citations		Only control citations	
mean of variables	WOS (1)	MGP (2)	WOS (3)	MGP (4)
Different institution (Distance > 0)	0.922	0.917	0.984	0.987
ln Distance | Distance > 0	7.099	6.990	7.800	7.741
Different country	0.637	0.634	0.749	0.758
Different language	0.500	0.476	0.600	0.578
Co-authors	0.099	0.090	0.019	0.014
Coincided past	0.085	0.088	0.030	0.027
Worked same place	0.049	0.048	0.029	0.030
Observations	268,527	29,404	268,527	412,388

Note: Realized citations are article pairs in which i cites d. Control Citations are articles matched to i by citing year and 3-digit field that did *not* cite d.

Table A.3 displays the differences between the characteristics of realized and control citations. In line with our expectations, we see that realized citations are more likely to be at the same university, same country, and from countries that use the same official language. Citing authors reside on average half the distance to the nearest cited author of non-citing (control) authors.[6] In terms of ties, citing authors are many times more likely to coauthor with the (realized) cited authors. They are also more than twice as likely to have worked at the same university either at the same or different times. Since all these variables are correlated we estimate regressions to determine the partial relationships.

Comparing columns (1) and (2) and columns (3) and (4) of Table A.3 we see that the average characteristics of the WOS and MGP samples are very similar. Imposing the criteria that all

[5]We match the Zentralblatt MATH and the WOS databases using the title of the article.

[6]The calculation is $\exp(6.990 - 7.741) = 0.47$ for the MGP sample and $\exp(7.099 - 7.800) = 0.50$ for the WOS.

citing and cited authors have MGP data leaves a much smaller sample of realized citations but it does not seem to change the average values of the geography and ties variables in a systematic way. The number of observations in column (4) is much higher than column (3) because the WOS sample only contains one control per case in column (1) whereas there are on average 14 controls per case in the MGP sample.

A.3 Affiliation identification and histories

There are 536,454 author-article combinations included in our database, of which 31% lack affiliations We recover affiliation information for many of these authors by applying the procedures developed by Tang and Walsh (2010), as implemented in Agrawal et al. (2013). For each record without author's affiliation we check whether there is another record with the same author name (full surname and name or full surname and initials) with an affiliation. We assign this latter affiliation to the missing record as long as both articles cite, at least, two articles that are not highly cited. The low citation benchmark is set at less than 50 citations. This increases the author-article combinations with affiliation information for some authors from 69% to 80%. Of those, 84% have affiliations for all authors.

We impute affiliation information for years in which an author does not publish by using his or her affiliation before or after those years. Our algorithm uses, iteratively, the closest information relative to the information gap. For example, suppose that author A published an article in 1990 when she was affiliated to MIT, and then published her next article in 1994 when she was affiliated to Princeton. In this example, we have holes in the affiliation history of this mathematician from 1991 to 1993. In the first iteration, the algorithm will fill the 1991 hole with information from 1990 (the closest available year), and the 1993 hole with information from 1994. After the first iteration we will still have a hole for the year 1992. We apply the second iteration to the algorithm. In this case, the author will have a double affiliation for the year 1992, because she has two different affiliations in the closest years (1991 and 1993).

A.4 Self-citation

To identify self-citations, we developed a unique author code that combines data from WOS, MGP and zbMATH databases (see below). MGP and zbMATH provide the name and surname of the authors, plus a unique author identification code. WOS only provides the surname and initials of the author. As zbMATH identifies the author at the article level, for those articles included in the zbMATH database, we were able to match WOS authors with zbMATH author codes. The personnel at zbMATH also provided us with a correspondence between zbMATH author codes and MGP author codes. For the rest of authors, we assigned a zbMATH author code if there was only one author code for a surname+initials combination. For the remaining

cases, we created a unique author code. To be conservative, we consider a self-citation if any of the citing authors has the same zbMATH code as any of the cited authors; and when any citing author has the same surname and initials of any of the cited authors.

A.5 Article-level aggregation of geography and ties

Mathematics has traditionally been characterized by more sole authorship than other fields. The average number of authors in mathematics has risen over time[7] but remains just 1.88 in 2009. In contrast, the average number of authors in evolutionary biology articles was 4 in 2005 (Agrawal et al., 2013), 3.75 in biomedical research (1961–2000), and 2.5 in physics (1991–2000,) 2.22 in computer science (1991–2000) (Newman, 2004), and 2.19 in economics (2011) (Hamermesh, 2013).

For multiple-author article pairs, a method for aggregating geography and ties of coauthors must be selected. For example, suppose paper i has authors A and B, whereas the authors of paper d are C and D. Then there are four combinations (A-C, A-D, B-C, B-D) of primitive **G** and **L** variables (e.g. distance between A's and C's respective institutions or whether A was C's Ph.D. advisor). There are two obvious ways to aggregate and both have been employed in prior papers. The min/max approach (used by Singh (2005) in defining past collaboration between citing and cited inventor teams) implicitly assumes perfect information flow between coauthors. Thus, it takes the *minimal* value of each measure of geographic separation (since separation is hypothesized to reduce flows). For example, the distance between article i and article d is defined as the minimum distance between the institutions to which citing authors are located and the institutions to which cited authors are located. For connections, which are hypothesized to increase flows, we use the maximal value between the author pairs. Thus the advisor citing indicator would "turn on" if *either* A or B was the Ph.D. advisor of either C or D. The min/max approach may be thought of as making the most optimistic assumption about flows of information between members of the same author team: if one knows about a paper, then all do.

A natural alternative is to average across the sets of bilateral relationships. The averaging approach implicitly assumes that knowledge transfer within teams is imperfect. More linkages therefore increase information flow. Under averaging, advisor citing would take a value of 1 only if A advised C and D and so did B. In other cases it would take fractional values. We use min/max as our main specification because we find the binary ties and geography variables are easier to interpret. Table D.9 shows that the averaging approach yields similar results for geography variables but stronger coefficients for ties.

[7] Agrawal et al. (2016) show that Soviet-rich fields of math have seen disproportionately large increases in coauthorship, suggesting that the integration of Soviet mathematicians has increased the gains from collaboration by shifting out the knowledge frontier.

B How controls for relevance affect estimates

The fixed effects in our baseline results control for the 3-digit subject field of the citing paper. The goal is to neutralize the issue of paper relevance so as to estimate the impact of geographic separation and ties on *awareness*. Table B.1 shows how the results vary as we tighten the criteria for the subject component of the fixed effect (and the corresponding set of control observations). The purpose is to see whether the effects of geography and ties are stable. To trim down the number of effects to be compared across specifications, we average the coefficients of all fourteen tie indicators. The table is organized such that the first column removes matching based on subject altogether and instead considers a randomly selected article published in the same year as the case observation. Not needing MSC data, the number of realized citations rises to 47,670. We add up to 25 random controls per case, with an average of 24.5. This number was chosen to approximately match the sample size of column (2), where the control set comprises all other papers published in the same journal and the same year as the citing paper. Column (3) reproduces column (5) from Table 1.

Table B.1: Sensitivity of results to alternative controls for article relevance

	(1)	(2)	(3)	(4)	(5)	(6)
Control group:	nil	journal	MSC-3d		MSC-5d	keyword
Distance > 0	-0.840*	-0.782*	-0.571*	-0.589*	-0.367*	-0.529*
	(0.062)	(0.059)	(0.073)	(0.073)	(0.091)	(0.163)
ln Dist $\mid$ Dist > 0	-0.045*	-0.030*	-0.037*	-0.034*	-0.033*	-0.047*
	(0.007)	(0.007)	(0.008)	(0.008)	(0.010)	(0.017)
Different country	-0.035	-0.041	-0.090*	-0.098*	-0.086†	-0.098
	(0.027)	(0.027)	(0.031)	(0.031)	(0.041)	(0.068)
Different language	-0.014	0.026	-0.025	-0.020	0.007	-0.127†
	(0.023)	(0.023)	(0.026)	(0.026)	(0.035)	(0.054)
Average effect of ties	1.639*	1.114*	0.585*	0.570*	0.379*	0.419*
	(0.048)	(0.037)	(0.033)	(0.031)	(0.034)	(0.069)
Cocitation				3.277*	2.151*	1.704*
				(0.057)	(0.077)	(0.197)
Observations	1215286	1135825	441792	441792	75926	22680
pseudo-R^2	0.181	0.144	0.091	0.127	0.097	0.114

Notes: Average effect of ties refer to the mean effect of 14 (3 WOS and 11 MGP) ties. Significance: *: 1%, †: 5%, $^{\sim}$: 10%. Robust standard errors clustered by cited article in parentheses.

The results shown in specification (1) of Table B.1 make it clear that the use of subject fixed effects and corresponding control observations is a crucially important element of the method. With random controls, the average coefficient on ties rises from 0.585 to 1.64. This means

that the presence of a linkage goes from multiplying the odds of citation by 1.80 up to 5.15. This is a statistical confirmation of what introspection would already have made obvious: our connections are influenced by common topics of interest. Column (2) finds that an intermediate form of matching, forcing the control to come from the same journal as the case, leads to intermediate results for ties (implying multiplication of citation odds by three).

The fourth, fifth, and sixth specifications impose tighter controls for relevance. Column (4) begins with a new proxy for topic similarity, cocitation. Reasoning that two articles that have been cited together in *other* papers are likely to deal with related topics, we add a co-citation dummy set equal to one if there exists a paper j that cites both i and d (and set to zero if the papers have never appeared jointly in the reference sections of the papers in our sample). We find this proxy for similarity in topic massively increases citation probability (factor of 26) and inclusion of the cocitation dummy lowers the estimated network effects. However, the reduction is minor (2%) and the network effects remain strong and statistically significant.

Column (5) of Table B.1 changes the data set by imposing that the control observation must be a paper in the same 5-digit field as the case. At the same time the triad fixed effect is modified to depend on the 5-digit citing subject. The cost of tighter matching is that we now find far fewer control observations—the sample falls by 83% to 75,926 observations. The coefficients on ties decline but the effects remain large (increasing citation odds by 46% on average) and precisely estimated.

The final estimation of Table B.1 specifies the triad and control observations based on the criteria of common "keywords." This presents an even stronger cut in the availability of controls than the 5-digit fields. The same-keywords sample has 95% fewer observations than the same 3-digit sample and 70% fewer than the same 5-digit sample. This possibly non-random attrition seems unacceptably high. The average standard error for network effects and distance effects almost doubles. The average coefficient on ties actually rises slightly when using the keywords control, suggesting that finer controls would not wipe out the estimated effects of ties. Indeed, an unavoidable trade-off emerges between tighter matching restrictions and sample size. If we defined the subject of the citing article sufficiently narrowly, there would be no other potential citing papers for a given cited paper. We view the 3-digit controls as hitting the "sweet spot" between controlling adequately for relevance and retaining a full set of comparison non-citing articles.[8]

Table B.2 removes the ties indicators, but is otherwise identical to Table B.1. Failure to control for ties dramatically magnifies the estimated impact of the geography variables. Gener-

[8]The trade-off between fineness of comparisons and sample attrition recalls the debate between Thompson and Fox-Kean (2005) and Henderson et al. (2005). The former argued that using more detailed (6-digit) technology classes for the control sample eliminates localization of patent citations. The counterargument was that such fine controls cause excessive non-random reductions in the sample. Using a novel method, Murata et al. (2014) show that distance matters even for 6-digit controls.

Table B.2: Sensitivity of results to alternative controls for article relevance (excluding ties)

	(1)	(2)	(3)	(4)	(5)	(6)
Control group:	nil	journal	MSC-3d		MSC-5d	keyword
Distance > 0	-1.846*	-1.663*	-1.243*	-1.254*	-0.914*	-1.043*
	(0.051)	(0.051)	(0.065)	(0.065)	(0.083)	(0.141)
ln Dist $\mid$ Dist > 0	-0.082*	-0.066*	-0.068*	-0.066*	-0.058*	-0.080*
	(0.006)	(0.006)	(0.008)	(0.008)	(0.010)	(0.017)
Different country	-0.239*	-0.213*	-0.232*	-0.236*	-0.193*	-0.266*
	(0.026)	(0.026)	(0.031)	(0.031)	(0.040)	(0.065)
Different language	-0.115*	-0.052†	-0.082*	-0.074*	-0.039	-0.159*
	(0.022)	(0.022)	(0.026)	(0.026)	(0.034)	(0.052)
Cocitation				3.339*	2.203*	1.670*
				(0.055)	(0.074)	(0.192)
Observations	1215286	1135825	441792	441792	75926	22680
pseudo-R^2	0.045	0.037	0.033	0.073	0.055	0.056

Notes: Significance: *: 1%, †: 5%, $^{\sim}$: 10%. Robust standard errors clustered by cited article in parentheses.

ally speaking they are twice as large, regardless of which fixed effect for relevance is employed. Thus we see that this key result from the baseline estimates is very robust.

C Conference Data

We draw data from the American Mathematical Society Annual Meetings over the 1990–2009 period. This conference is also known as the Joint Mathematics Meetings, since it is organized jointly with the Mathematical Association of America. It gathers the largest number of mathematicians in America, and is considered the most important annual conference in mathematics.[9]

For each annual meeting, we extract the information contained in the full program web page.[10] It provides the name of the presenter, the title of the paper, and the session. The full program also identifies the special sessions' organizers. On average, 1459 scholars participate in the conference every year as presenters or session organizers, and 1037 papers are presented.

[9]Worldwide, the most important meeting is the International Congress of Mathematics, organized by the International Mathematical Union, which takes places every four years. The winners of the Fields Medal are announced in this congress. Since the Joint Mathematics Meetings takes place every year, and its web page provides more information about papers and presenters, we chose this latter meeting to maximize observations.

[10]It can be accessed from `http://www.ams.org/meetings/national/national_past.html`

We merge the conference participation database with our citations sample using the name of the scholar and the title of the paper as links. First, we analyze whether geographical barriers impede participating in a conference. We pool the observations and estimate a Logit model with year fixed effects. As shown in Table C.1-column 1, a larger distance, being located in a different city and in a country whose official language is not English reduce the likelihood of attending the conference. In contrast, a scholar affiliated to a Canadian university has a higher likelihood of attending the conference. In column 2 we control for participant fixed effects. All coefficients, except for different country, keep their sign, although distance is the only coefficient which remains statistically significant. In columns 3 and 4, we estimate a linear probability model. As expected, the value of the coefficients is much lower. However, results are qualitative similar.

Table C.1: The effect of geographical barriers on the probability of attending a conference, 1990–2009 (pooled data)

	(1)	(2)	(3)	(4)
Different city	-0.618*	-0.158	-0.043*	-0.022*
	(0.216)	(0.167)	(0.015)	(0.009)
ln Distance	-0.050†	-0.136*	-0.002$^{\sim}$	-0.004*
	(0.025)	(0.016)	(0.001)	(0.000)
Different country	-1.331*	0.041	-0.028*	0.001
	(0.063)	(0.074)	(0.002)	(0.002)
Participant from Canada	0.611*	0.001	0.008*	-0.000
	(0.080)	(0.145)	(0.002)	(0.004)
Different language	-0.060	-0.047	-0.001	-0.001
	(0.074)	(0.111)	(0.001)	(0.002)
N. obs.	667399	97867	667399	667399
Participant FE	No	Yes	No	Yes
Model	Logit	Logit	LPM	LPM

Note: *, †, $^{\sim}$ statistically significant at 1%. 5% and 10% respectively. In specifications (1) and (3) standard errors clustered by the location of the conference and the location of the institution in which the conference participant is affiliated. In specifications (2) and (4) standard errors clustered by participant.

Second, we analyze whether coinciding at a conference raises the likelihood of citation. To test this hypothesis, we build four new tie variables:

1. Some citing and cited authors coincided at a conference before the citation.
2. Some citing and cited authors coincided at a conference and session before the citation.
3. Some citing and cited author coincided at a conference where the cited paper was presented before the citation.

4. Some citing and cited authors coincided at a conference where the citing paper was presented before the citation.

Table C.2 presents the absolute and mean values for these variables. We report data for the realized and the control citations. All the probabilities are very low. For example, the probability that some citing and cited authors have coincided at a conference before the citation is 0.03, and the probability that some citing and cited author coincided at a session in the same conference before the citation is 0.0041. Few citing or cited papers included in our citations' database were presented at the Joints Mathematics Meetings. For all variables, the probabilities are larger for realized than for control citations, suggesting a positive correlation between coinciding at a conference and citation.

Table C.2: New conference-participation tie variables. Realized vs. Control citations

Variable	Total		Average	
	Realized	Control	Realized	Control
Citations	29,404	412,388		
Coincided at a conference	918	9,895	0.0312	0.0240
Coincided at a conference and session	121	770	0.0041	0.0019
Coincided at a conference where the cited paper was presented	10	22	0.0003	0.0001
Coincided at a conference where the citing paper was presented	15	158	0.0005	0.0004

Source: Authors' own calculations, based on Joint Meetings full programs and the citations database.

Table C.3 presents the estimates of the baseline regression including the four new conference variables. Since conferences provide an opportunity to share information about research, we expect all conference coefficients to be positive. As expected, both in the Logit and LPM estimations, we find a positive and statistically significant effect for coinciding at the same session, and coinciding at a conference where the cited paper was presented. Coinciding at a conference is not precisely estimated, even when the citing paper was presented in it.

D Subsamples and other robustness checks

This subsection reports the findings of additional robustness checks. Our baseline table in subsection 4.1 first shows estimates for all the authors in the WOS before restricting the sample to papers where all the authors have MGP data. Table D.1 splits the WOS sample used in our baseline estimates columns (1) and (2) into MGP (11%) and non-MGP (89%) subsets in order to provide an additional check for selection bias. The coefficients on geography and career ties in the MGP sample have confidence intervals (CI) that are wide enough to include the non-MGP coefficients in every case except distance> 0 which lies just outside the CI. These results provide some assurance that the MGP sample does not suffer from selection bias.

Table C.3: Baseline regression with conference variables

	(1) Logit	(2) LPM
Coincided conference	-0.032	-0.003
	(0.065)	(0.004)
Coincided conference+session	0.423*	0.032†
	(0.151)	(0.013)
Coincided conference cited paper presented	2.122†	0.205†
	(0.835)	(0.081)
Coincided conference citing paper presented	0.083	0.012
	(0.307)	(0.021)
4 Geography variables	YES	YES
14 Ties variables	YES	YES
pseudo-R2 or R2	0.091	0.058

Notes: Robust standard errors clustered by cited article in parentheses.
$^{\sim}p < 0.1, ^{\dagger}p < 0.05, ^{\star}p < 0.01$

Table D.2 adds an indicator for No Shared Association to the set of geographic barriers employed in Table 1. The idea is to test whether continental conference blocs might be an important omitted variable in our specification of the geography variables. There are four major continental mathematics associations: the African Mathematical Union, the European Mathematical Society, the South East Asian Mathematical Society, and the Latin American Society. We code two papers as sharing an association if (1) any member of the citing team is located in an institution in the same continental (or bi-national) association as any member of the cited team, or (2) any citing author is in the same country as any cited author and that country has a *national* association. No Shared Association enters significantly only in specifications that lack full controls for distance and ties. In those cases it enters with a positive sign, which is unexpected since the variable is coded (like the other geography indicators) in the form of a barrier. The inclusion of No Shared Association reduces the Different Country and Different language effects but by less than a standard error in each case.

Table D.3 shows the results obtained by re-estimating our logit regressions using the linear probability model (LPM), employed in some studies including Belenzon and Schankerman (2013). While the magnitudes of logit coefficients are much larger, the results are very similar in other dimensions.[11] All 51 coefficients in this table have the same sign as the corresponding coefficient in Table 1. Significance levels are the same for 47 coefficients. In general, an effect

[11]The smaller size of LPM coefficients follows from the fact that they estimate marginal effects on probabilities rather than log odds. With logit on one explanatory variable, x_i, the probability of a positive outcome is $p_i = (1 + \exp[-\beta x_i])^{-1}$. Differentiating by x_i, we see that $b_{\text{lpm}} \approx (1/N)\sum_i p_i(1-p_i)\beta \approx \bar{p}(1-\bar{p})\beta$. In our data $\bar{p}(1-\bar{p}) = 0.06$ so we expect logit coefficients to be about 17 times higher than LPM coefficients. The log distance effect in Table 1 is 18.5 times larger than the one in Table D.3.

Table D.1: MGP vs. Non-MGP

	(1) MGP	(2) Non-MGP	(3) MGP	(4) Non-MGP
Distance > 0	-1.209*	-0.983*	-1.146*	-0.911*
	(0.092)	(0.031)	(0.093)	(0.032)
ln Distance	-0.069*	-0.074*	-0.051*	-0.052*
	(0.009)	(0.004)	(0.010)	(0.004)
Different country	-0.166*	-0.202*	-0.097†	-0.145*
	(0.037)	(0.015)	(0.038)	(0.015)
Different language	-0.089*	-0.106*	-0.065†	-0.066*
	(0.032)	(0.012)	(0.032)	(0.012)
Co-authors			1.799*	1.662*
			(0.071)	(0.022)
Coincided past			0.788*	0.704*
			(0.052)	(0.020)
Worked same place			0.530*	0.475*
			(0.056)	(0.021)
Observations	58802	478252	58802	478252
pseudo-R^2	0.049	0.044	0.093	0.085

Notes: Robust standard errors clustered by cited article in parentheses.
Significance: *: 1%, †: 5%, $^{\sim}$: 10%.

that is stronger in the logit (e.g. advisor cited vs grandparent cited) is also stronger in the LPM. Some relative magnitudes are nearly the same: The distance effect in column (5) is 54% of that in column (3) in the logit and 50% in the LPM.

Table D.4 shows the robustness of the interaction effects to changes in the sample, specification, and the method for constructing the three proxies for awareness gaps. In each case we provide the interaction with log distance, the average of the three geographic barrier indicators (different university, different country, and different language), and the average of the 13 ties interactions.

The first robustness check is to estimate the interactions using just the career ties which are available in the WOS sample. The point is to ensure that the interactions are not driven by some feature of the MGP sample. The second specification is a linear probability model (LPM). Since the LPM estimates differences in absolute risk of citation, the coefficients are expected to be much smaller. The third specification sums all 14 ties (including grandparent citing) and interacts them with the information proxies rather than averaging the interacting

Table D.2: Baseline Results with No-Shared-Association

	(1)	(2)	(3)	(4)	(5)	(6)
Specification:	Triad-fixed-effects logit (TFE-Λ)					
Sample	WOS	WOS	MGP	MGP	MGP	MGP
Geography:						
Distance > 0	-0.905*	-0.862*	-1.086*		-0.562*	
	(0.033)	(0.034)	(0.070)		(0.078)	
ln Dist $\mid$ Dist > 0	-0.089*	-0.063*	-0.091*		-0.038*	
	(0.004)	(0.004)	(0.009)		(0.009)	
Different country	-0.247*	-0.175*	-0.308*	-0.268*	-0.094*	-0.077†
	(0.016)	(0.016)	(0.035)	(0.037)	(0.035)	(0.037)
Different language	-0.095*	-0.059*	-0.060†	-0.080*	-0.024	-0.037
	(0.011)	(0.012)	(0.027)	(0.028)	(0.027)	(0.028)
No shared association	0.094*	0.067*	0.143*	-0.005	0.008	-0.074
	(0.013)	(0.014)	(0.029)	(0.050)	(0.029)	(0.051)
Ties:						
Co-authors		1.672*			1.572*	1.581*
		(0.021)			(0.050)	(0.050)
Coincided past		0.710*			0.378*	0.378*
		(0.019)			(0.043)	(0.043)
Worked same place		0.476*			0.342*	0.339*
		(0.020)			(0.043)	(0.043)
Share Ph.D. (5 years)					0.463*	0.457*
					(0.067)	(0.067)
PhD siblings					0.664*	0.665*
					(0.100)	(0.100)
PhD cousins					0.365*	0.364*
					(0.082)	(0.082)
Advisor citing					1.090*	1.079*
					(0.164)	(0.164)
Advisor cited					1.376*	1.375*
					(0.102)	(0.103)
Academic grandparent citing					-0.284	-0.255
					(0.392)	(0.390)
Academic grandparent cited					1.028*	1.024*
					(0.155)	(0.154)
Academic uncle citing					0.227$^{\sim}$	0.237†
					(0.118)	(0.118)
Academic uncle cited					0.616*	0.620*
					(0.076)	(0.076)
Alma Mater citing					0.238*	0.234*
					(0.055)	(0.055)
Alma Mater cited					0.120†	0.120†
					(0.057)	(0.057)
Observations	537054	537054	441792	441792	441792	441792
pseudo-R^2	0.044	0.085	0.033	0.034	0.091	0.091

Notes: Robust standard errors clustered by cited article in parentheses. Significance: *: 1%, †: 5%, $^{\sim}$: 10%.

coefficients. Finally, the last two specifications experiment with alternative constructions of the proxies. The "Means" specification sets binary Obscure and Recent to one when the underlying variables are less than their means (6.02 cites and 10.73 years) rather their medians (3 and 9).

Table D.3: Baseline Results Using LPM

	(1)	(2)	(3)	(4)	(5)	(6)
Specification:	Triad-fixed-effects LPM (TFE-LPM)					
Sample	WOS	WOS	MGP	MGP	MGP	MGP
Geography:						
Distance > 0	-0.313*	-0.254*	-0.209*		-0.093*	
	(0.010)	(0.010)	(0.008)		(0.008)	
ln Dist $\mid$ Dist > 0	-0.030*	-0.021*	-0.004*		-0.002*	
	(0.001)	(0.001)	(0.000)		(0.000)	
Different country	-0.086*	-0.058*	-0.014*	-0.016*	-0.004†	-0.005*
	(0.006)	(0.006)	(0.002)	(0.002)	(0.002)	(0.002)
Different language	-0.044*	-0.029*	-0.005*	-0.005*	-0.002	-0.002
	(0.005)	(0.005)	(0.001)	(0.001)	(0.001)	(0.001)
Ties:						
Co-authors		0.509*			0.180*	0.182*
		(0.005)			(0.007)	(0.007)
Coincided past		0.235*			0.022*	0.022*
		(0.006)			(0.004)	(0.004)
Worked same place		0.182*			0.018*	0.018*
		(0.007)			(0.003)	(0.003)
Share Ph.D. (5 years)					0.061*	0.061*
					(0.008)	(0.008)
PhD siblings					0.107*	0.107*
					(0.010)	(0.010)
PhD cousins					0.022*	0.022*
					(0.006)	(0.006)
Advisor citing					0.206*	0.205*
					(0.023)	(0.023)
Advisor cited					0.275*	0.274*
					(0.014)	(0.014)
Academic grandparent citing					-0.050	-0.049
					(0.050)	(0.049)
Academic grandparent cited					0.118*	0.117*
					(0.020)	(0.020)
Academic uncle citing					0.018$^{\sim}$	0.019$^{\sim}$
					(0.011)	(0.011)
Academic uncle cited					0.047*	0.047*
					(0.007)	(0.007)
Alma Mater citing					0.028*	0.028*
					(0.006)	(0.006)
Alma Mater cited					0.008	0.008
					(0.006)	(0.006)
Observations	537054	537054	441792	441792	441792	441792
Overall R^2	0.029	0.052	0.020	0.020	0.058	0.059

Notes: Robust standard errors clustered by cited article in parentheses. Significance: *: 1%, †: 5%, $^{\sim}$: 10%.

Table D.4: Robustness of interaction coefficients between ties and obscure, recent and different-field papers.

	Interaction	Obscure	Recent	Different-field
WOS sample	ln Distance	-0.050*	-0.006	-0.019†
	Geography indicators (3)	-0.117*	-0.228*	0.032
	Ties (13)	0.170*	0.184*	0.273*
	Observations	537054	537054	275084
	Pseudo-R^2	0.086	0.086	0.094
LPM estimation	ln Distance	-0.001	-0.002*	-0.001
	Geography indicators (3)	-0.006	-0.021*	0.004
	Ties (13)	0.014†	0.029*	0.047*
	Observations	441792	441792	225768
	R^2	0.062	0.064	0.068
Sum of ties	ln Distance	-0.031	-0.029$^{\sim}$	-0.003
	Geography indicators (3)	-0.070	-0.124†	0.039
	Ties (14)	0.134*	0.121*	0.117*
	Observations	441792	441792	225768
	Pseudo-R^2	0.080	0.081	0.086
Means / 3-digit field	lndist	-0.006	-0.033†	0.013
	Geography indicators (3)	-0.087	-0.090$^{\sim}$	-0.023
	Ties (13)	0.198*	0.175*	0.179*
	Observations	441792	441792	225768
	Pseudo-R^2	0.093	0.093	0.100
Continuous measure	ln Distance	-0.029	-0.079*	0.003
(see note 2)	Geography indicators (3)	-0.134	-0.137	-0.025
	Ties (13)	0.417*	0.427*	0.124*
	Observations	441792	441792	225768
	Pseudo-R^2	0.093	0.094	0.137

Notes: 1. Robust standard errors clustered by cited article in parentheses. Significance: *: 1%, †: 5%, $^{\sim}$: 10%. 2. The continuous measure of field difference takes the value of 0, 1, 2 or 3, depending on whether field difference is at the 5, 3, or 2-digit level. This specification controls for differences in 5-digit field as a base effect (since the triadic fixed effect does not capture this). The continuous obscure and recent measures are calculated as one minus the empirical CDFs of citations and years since publication.

The continuous measures of obscurity and recentness are based on the empirical cumulative distribution functions (ECDF) of citation counts and lags. Obscure and Recent are defined as one minus the respective ECDF. This has the advantage of keeping these variables in the unit interval. Although the tie interactions for the continuous measures of recent and obscure have larger coefficients, the continuous formulations of these variables have about half the standard deviations. Interaction sizes are again similar if expressed in terms of standard deviations. Neither the mean nor the continuous reformulations have analogous transformations for the differences in fields so we implement alternative robustness measures. In the row with means, different field is defined as papers in different 3-digit fields (instead of 2-digit). In the row with continuous measures we calculate the "tree-distance" in the field classification codes. Thus, papers in the same 5-digit field have distance 0; papers in different 5-digits but same 3-digit fields are distance 1, and so forth. In this specification it is necessary to control for the tree-distance as well as its interactions with ties and geography.

The results of the investigation of the robustness of information interactions can be summarized as follows. First, the positive ties interactions retain their strong statistical significance across a variety of specifications. Second, with two explicable exceptions, the magnitudes of the ties interactions are very similar. Third, the interactions with log distance and the other geographic barrier indicators are generally negative as expected. While not uniformly negative (5 out of the 30 reported in Table D.4 are positive), the geography/distance interactions are negative when they differ significantly from zero.

Tables D.5 to D.8 break our sample into two periods. The main interest in this is that the 2005 to 2009 period accounts for the majority of the observations in the full sample. There are too many results to compare individually but the exercise of splitting the sample leads to the following conclusions. Unsurprisingly, the decline in distance effects we observe in Figure 4 also shows up in the comparison of before and after 2005. On the other hand, residing in a different country becomes more important after 2005. The effects of ties are remarkably stable with 25 out of 28 ties coefficients in columns (5) of Tables D.5 and D.6 being less than a standard error from the values in Table 1. The magnitudes of some ties hardly change: advisor cited has a coefficient of 1.396 before 2005 and 1.349 afterwards. The Table 2 finding that ties matter more for recent and obscure papers holds up in both periods but the different-field interaction is only statistically significant before 2005 (Tables D.7 and D.8).

Table D.9 reports the results of four additional specifications designed to explore the robustness of our main results. The first specification is closely related to Figure **??**. As in the figure, we interact a "bothUS" dummy with the geography and ties variables. The big difference is that the figure uses moving windows, whereas this regression uses the whole data. Moreover, the table reports all the geography interactions rather than just the distance effects. The main novel finding is that when both citing and potentially cited author teams are based in the US, the odds of a realized citation rise by 45%. As seen in the figure, the effect of distance is near

Table D.5: Baseline Results before 2005

	(1)	(2)	(3)	(4)	(5)	(6)
Specification:	Triad-fixed-effects logit (TFE-Λ)					
Sample	WOS	WOS	MGP	MGP	MGP	MGP
Geography:						
Distance > 0	-0.974*	-0.907*	-1.087*		-0.439*	
	(0.040)	(0.042)	(0.089)		(0.100)	
ln Distance	-0.071*	-0.055*	-0.089*		-0.060*	
	(0.005)	(0.005)	(0.010)		(0.011)	
Different country	-0.188*	-0.133*	-0.166*	-0.204*	-0.030	-0.038
	(0.019)	(0.019)	(0.042)	(0.044)	(0.043)	(0.045)
Different language	-0.069*	-0.036†	-0.057	-0.052	-0.002	-0.006
	(0.016)	(0.016)	(0.036)	(0.036)	(0.036)	(0.037)
Ties:						
Co-authors		1.638*			1.499*	1.510*
		(0.030)			(0.069)	(0.069)
Coincided past		0.678*			0.321*	0.318*
		(0.025)			(0.058)	(0.058)
Worked same place		0.519*			0.349*	0.347*
		(0.028)			(0.059)	(0.059)
Share Ph.D. (5 years)					0.302*	0.297*
					(0.095)	(0.095)
PhD siblings					0.685*	0.697*
					(0.141)	(0.141)
PhD cousins					0.349*	0.341*
					(0.113)	(0.113)
Advisor citing					0.938*	0.929*
					(0.222)	(0.222)
Advisor cited					1.394*	1.396*
					(0.140)	(0.140)
Academic grandparent citing					-0.376	-0.362
					(0.595)	(0.596)
Academic grandparent cited					1.058*	1.057*
					(0.223)	(0.222)
Academic uncle citing					0.358†	0.368†
					(0.152)	(0.153)
Academic uncle cited					0.651*	0.654*
					(0.106)	(0.106)
Alma Mater citing					0.303*	0.289*
					(0.072)	(0.073)
Alma Mater cited					0.087	0.082
					(0.076)	(0.076)
Observations	267322	267322	177000	177000	177000	177000
pseudo-R^2	0.041	0.077	0.033	0.034	0.091	0.091

Notes: Robust standard errors clustered by cited article in parentheses. Significance: *: 1%, †: 5%, $^{\sim}$: 10%.

Table D.6: Baseline Results after 2005

	(1)	(2)	(3)	(4)	(5)	(6)
Specification:	Triad-fixed-effects logit (TFE-Λ)					
Sample	WOS	WOS	MGP	MGP	MGP	MGP
Geography:						
Distance > 0	-1.043*	-0.966*	-1.395*		-0.705*	
	(0.040)	(0.042)	(0.082)		(0.092)	
ln Distance	-0.075*	-0.049*	-0.046*		-0.011	
	(0.004)	(0.005)	(0.010)		(0.010)	
Different country	-0.213*	-0.150*	-0.299*	-0.337*	-0.152*	-0.168*
	(0.018)	(0.019)	(0.041)	(0.043)	(0.042)	(0.044)
Different language	-0.137*	-0.094*	-0.108*	-0.105*	-0.048	-0.042
	(0.015)	(0.015)	(0.034)	(0.034)	(0.034)	(0.035)
Ties:						
Co-authors		1.699*			1.630*	1.637*
		(0.028)			(0.063)	(0.063)
Coincided past		0.742*			0.434*	0.436*
		(0.025)			(0.057)	(0.057)
Worked same place		0.440*			0.333*	0.332*
		(0.026)			(0.056)	(0.056)
Share Ph.D. (5 years)					0.636*	0.631*
					(0.082)	(0.083)
PhD siblings					0.634*	0.628*
					(0.124)	(0.124)
PhD cousins					0.390*	0.393*
					(0.105)	(0.105)
Advisor citing					1.255*	1.250*
					(0.231)	(0.231)
Advisor cited					1.355*	1.349*
					(0.131)	(0.131)
Academic grandparent citing					-0.184	-0.182
					(0.517)	(0.511)
Academic grandparent cited					1.004*	1.001*
					(0.180)	(0.180)
Academic uncle citing					0.082	0.087
					(0.167)	(0.166)
Academic uncle cited					0.580*	0.582*
					(0.096)	(0.096)
Alma Mater citing					0.173†	0.172†
					(0.074)	(0.074)
Alma Mater cited					0.157†	0.161†
					(0.075)	(0.075)
Observations	269732	269732	264792	264792	264792	264792
pseudo-R^2	0.048	0.092	0.033	0.034	0.092	0.092

Notes: Robust standard errors clustered by cited article in parentheses. Significance: *: 1%, †: 5%, $^{\sim}$: 10%.

Table D.7: Obscure, Recent, and Different-field papers are more impacted by ties and geography (before 2005)

Specification:	(1)	(2)	(3)	(4)	(5)	(6)	(7)
		Obscure		Recent		Different field	
		base	interact	base	interact	base	interact
Geography:							
Distance > 0	-0.438*	-0.383*	-0.400	-0.311*	-0.143	-0.610*	0.072
	(0.100)	(0.109)	(0.267)	(0.164)	(0.199)	(0.203)	(0.298)
ln Dist $\mid$ Dist > 0	-0.060*	-0.062*	0.017	-0.035†	-0.043†	-0.062*	-0.008
	(0.011)	(0.011)	(0.031)	(0.016)	(0.020)	(0.021)	(0.032)
Different country	-0.030	-0.031	0.022	-0.052	0.033	0.036	0.053
	(0.043)	(0.046)	(0.126)	(0.063)	(0.083)	(0.080)	(0.128)
Different language	-0.002	0.008	-0.096	0.004	-0.016	0.011	-0.149
	(0.036)	(0.039)	(0.100)	(0.052)	(0.069)	(0.069)	(0.100)
Ties:							
Average effect of ties	0.638*	0.619*	0.208$^{\sim}$	0.547*	0.156*	0.499*	0.423*
	(0.027)	(0.028)	(0.040)	(0.027)	(0.050)	(0.055)	(0.091)
Observations	177000	177000		177000		76152	
pseudo-R^2	0.091	0.092		0.093		0.098	

Notes: 1. Robust standard errors clustered by cited article in parentheses. Significance: *: 1%, †: 5%, $^{\sim}$: 10%. 2. Average effect of ties is the mean of the base and interaction coefficients of 13 ties (3 WOS and 10 MGP). "Obscure" indicates that total citations received for this article are less than or equal to the median number of citations received among all articles, "recent" corresponds to citation lags less than or equal to the median, and "different field" equals 1 if citing article and cited article belong to different 2-digit MSCs.

zero ($-0.044 - 0.040 = -0.004$) within the US. A surprising effect shown in this column is that being at the same university matters more for both-US pairs, but this interaction is only significant conditional on ties, which matter less in the US.

Column (2) replaces the min/max approach to aggregating geographic and network variables across coauthors with averages over all the author pairs. The coefficient for average effect of ties is 28% larger (0.837 vs 0.652).[12] This suggests the *existence of more than one tie among the author-pairs is reinforcing*. On the other hand, the geography effects do not change much: the continuous effect of distance is -0.041 with averaging versus -0.037 under min/max. The overall fits of the two methods, as measured by the pseudo-R^2 are almost the same (0.092 vs 0.091). The similarity in results is partly due to the fact, discussed earlier, that there is relatively little coauthorship in mathematics. Column (3) measures the geographic variables at the time the cited article was published rather than when it was cited. Thus, it does not capture movement of the authors following the publication of the cited article. The contemporaneous

[12]The baseline coefficient for the average effect of ties comes from Table 2 column(1).

Table D.8: Obscure, Recent, and Different-field papers are more impacted by ties and geography (after 2005)

Specification:	(1)	(2)	(3)	(4)	(5)	(6)	(7)
		Obscure		Recent		Different field	
		base	interact	base	interact	base	interact
Geography:							
Distance > 0	-0.705*	-0.729*	0.325	-0.320†	-0.561*	-0.946*	0.164
	(0.092)	(0.108)	(0.207)	(0.133)	(0.180)	(0.156)	(0.240)
ln Dist $\mid$ Dist > 0	-0.011	0.005	-0.085*	-0.007	-0.007	0.002	0.001
	(0.010)	(0.012)	(0.024)	(0.015)	(0.020)	(0.018)	(0.027)
Different country	-0.152*	-0.139*	-0.082	-0.134†	-0.047	-0.154†	-0.012
	(0.042)	(0.047)	(0.101)	(0.058)	(0.084)	(0.075)	(0.114)
Different language	-0.048	-0.068$^{\sim}$	0.098	-0.037	-0.020	-0.072	-0.026
	(0.034)	(0.039)	(0.082)	(0.047)	(0.067)	(0.057)	(0.094)
Ties:							
Average effect of ties	0.666*	0.619*	0.133†	0.534*	0.197*	0.628*	0.085
	(0.022)	(0.026)	(0.058)	(0.032)	(0.051)	(0.038)	(0.066)
Observations	264792	264792		264792		149616	
pseudo-R^2	0.092	0.094		0.095		0.103	

Notes: 1. Robust standard errors clustered by cited article in parentheses. Significance: *: 1%, †: 5%, $^{\sim}$: 10%. 2. Average effect of ties is the mean of the base and interaction coefficients of 13 ties (3 WOS and 10 MGP). "Obscure" indicates that total citations received for this article are less than or equal to the median number of citations received among all articles, "recent" corresponds to citation lags less than or equal to the median, and "different field" equals 1 if citing article and cited article belong to different 2-digit MSCs.

geography used in the earlier specification leads to a similar fit (0.091 vs 0.090).The larger distance effect estimated for original geography is within a two-standard-error margin. Column (4) vastly increases the sample size by using observations that had previously been rejected because affiliation information or MGP data was missing for at least one of the co-authors. The distance greater than zero and the average effect of ties coefficients are significantly smaller than in the baseline specification. The remaining coefficients are within the two standard errors margin.

Table D.9: Additional robustness tests

Sample:	(1) bothUS	(2) average	(3) original geography	(4) available author
Panel A: including ties				
Distance > 0	-0.397*	-0.523*	-0.394*	-0.459*
	(0.081)	(0.087)	(0.073)	(0.042)
$\times$ *bothUS*	-0.575*			
	(0.158)			
ln Dist $\mid$ Dist > 0	-0.044*	-0.041*	-0.049*	-0.031*
	(0.009)	(0.009)	(0.007)	(0.005)
$\times$ *bothUS*	0.040†			
	(0.017)			
Different country	-0.029	-0.144*	-0.136*	-0.110*
	(0.039)	(0.035)	(0.041)	(0.019)
Different language	-0.007	-0.027	-0.031	-0.017
	(0.027)	(0.029)	(0.023)	(0.015)
bothUS	0.372*			
	(0.112)			
Average effect of ties	0.639*	0.837*	0.571*	0.548*
	(0.014)	(0.044)	(0.034)	(0.018)
$\times$ *bothUS*	-0.126*			
	(0.022)			
Observations	441792	441792	441792	1449153
pseudo-R^2	0.081	0.092	0.090	0.069
Panel B: excluding ties				
Distance > 0	-1.243*	-1.332*	-1.043*	-1.121*
	(0.075)	(0.076)	(0.063)	(0.038)
$\times$ *bothUS*	-0.086			
	(0.155)			
ln Dist $\mid$ Dist > 0	-0.081*	-0.072*	-0.074*	-0.059*
	(0.009)	(0.009)	(0.007)	(0.004)
$\times$ *bothUS*	0.054*			
	(0.017)			
Different country	-0.264*	-0.324*	-0.444*	-0.231*
	(0.039)	(0.034)	(0.039)	(0.018)
Different language	-0.074*	-0.092*	-0.110*	-0.070*
	(0.026)	(0.028)	(0.023)	(0.015)
bothUS	-0.380*			
	(0.093)			
Observations	441792	441792	441792	1449153
pseudo-R^2	0.033	0.028	0.031	0.023

Notes: Average effect of ties refer to the mean effect of 14 (3 WOS and 11 MGP) ties, except that in the first column, we use the sum of the 14 ties variables instead of average effect of ties, for the simplicity of the interaction term with bothUS dummy. Significance: *: 1%, †: 5%, $^{\sim}$: 10%. Robust standard errors clustered by cited article in parentheses.

E Supplementary Tables

Table E.1: Baseline estimation (including standard errors)

	(1)	(2)	(3)	(4)	(5)	(6)
Sample	WOS	WOS	MGP	MGP	MGP	MGP
Geography:						
Distance > 0	-1.008*	-0.936*	-1.243*		-0.571*	
	(0.029)	(0.031)	(0.065)		(0.073)	
ln Dist $\mid$ Dist > 0	-0.073*	-0.052*	-0.068*	Figure 2	-0.037*	Figure 2
	(0.003)	(0.003)	(0.008)		(0.008)	
Different country	-0.198*	-0.140*	-0.232*	-0.270*	-0.090*	-0.103*
	(0.014)	(0.014)	(0.031)	(0.032)	(0.031)	(0.033)
Different language	-0.104*	-0.066*	-0.082*	-0.079*	-0.025	-0.025
	(0.011)	(0.012)	(0.026)	(0.026)	(0.026)	(0.027)
Ties:						
Co-authors		1.672*			1.572*	1.581*
		(0.021)			(0.050)	(0.050)
Coincided past		0.712*			0.378*	0.378*
		(0.019)			(0.043)	(0.043)
Worked same place		0.478*			0.342*	0.339*
		(0.020)			(0.043)	(0.043)
Share Ph.D. (5 years)					0.463*	0.457*
					(0.067)	(0.067)
PhD siblings					0.663*	0.666*
					(0.100)	(0.100)
PhD cousins					0.365*	0.362*
					(0.082)	(0.082)
Advisor citing					1.090*	1.079*
					(0.164)	(0.164)
Advisor cited					1.377*	1.375*
					(0.102)	(0.103)
Academic grandparent citing					-0.284	-0.254
					(0.392)	(0.390)
Academic grandparent cited					1.028*	1.023*
					(0.155)	(0.155)
Academic uncle citing					0.227$^\sim$	0.236†
					(0.118)	(0.118)
Academic uncle cited					0.616*	0.619*
					(0.076)	(0.076)
Alma Mater citing					0.239*	0.233*
					(0.055)	(0.055)
Alma Mater cited					0.120†	0.119†
					(0.056)	(0.057)
Observations	537054	537054	441792	441792	441792	441792
pseudo-R^2	0.044	0.085	0.033	0.034	0.091	0.091

Robust standard errors clustered by cited article in parentheses. Significance: *: 1%, †: 5%, $^\sim$: 10%.

Table E.2: Summary statistics for categories of papers included in the information mechanisms analysis.

	Obscure?		Recent?		Field	
	yes	no	yes	no	different	same
# of observations	76978	364814	231929	209863	82795	142973
Avg. dist. between cites	4711	4620	4567	4712	4667	4579
Avg. # of cites	1.29	6.88	4.01	7.99	5.22	4.86
Avg. # of ties						
Total	0.20	0.21	0.22	0.20	0.17	0.24
Co-authors	0.02	0.02	0.02	0.02	0.01	0.02
Coincided past	0.03	0.03	0.03	0.03	0.03	0.03
Worked same place	0.03	0.03	0.03	0.03	0.03	0.03
Share PhD (5 years)	0.01	0.01	0.01	0.01	0.01	0.01
PhD siblings	0.01	0.01	0.02	0.01	0.01	0.02
PhD cousins	0.02	0.02	0.03	0.02	0.02	0.03
Advisor citing	0.00	0.00	0.00	0.00	0.00	0.00
Advisor cited	0.00	0.01	0.01	0.01	0.01	0.01
Grandparent citing	0.00	0.00	0.00	0.00	0.00	0.00
Grandparent cited	0.00	0.00	0.00	0.00	0.00	0.00
Uncle citing	0.01	0.00	0.01	0.00	0.00	0.00
Uncle cited	0.01	0.02	0.02	0.02	0.02	0.02
Alma Mater citing	0.02	0.02	0.02	0.02	0.02	0.02
Alma Mater cited	0.02	0.02	0.02	0.02	0.02	0.02

Notes: Sample includes both realized and non-realized citations.

References

Agrawal, A., Goldfarb, A., and Teodoridis, F. (2016). Does knowledge accumulation increase the returns to collaboration? *American Economic Journal: Applied Economics*.

Agrawal, A., McHale, J., and Oettl, A. (2013). Collaboration, stars, and the changing organization of science: Evidence from evolutionary biology. NBER Working Papers 19653, National Bureau of Economic Research, Inc.

Belenzon, S. and Schankerman, M. (2013). Spreading the word: Geography, policy, and knowledge spillovers. *The Review of Economics and Statistics*, 95(3):884–903.

Hamermesh, D. S. (2013). Six Decades of Top Economics Publishing: Who and How? *Journal of Economic Literature*, 51(1):162–72.

Henderson, R., Jaffe, A., and Trajtenberg, M. (2005). Patent citations and the geography of knowledge spillovers: A reassessment: Comment. *American Economic Review*, 95(1):461–464.

Murata, Y., Nakajima, R., Okamoto, R., and Tamura, R. (2014). Localized Knowledge Spillovers and Patent Citations: A Distance-Based Approach. *The Review of Economics and Statistics*, 96(5):967–985.

Newman, M. E. (2004). Who is the best connected scientist?A study of scientific co-authorship networks. In Ben-Naim, E., Frauenfelder, H., and Toroczkai, Z., editors, *Complex Networks*, volume 650 of *Lecture Notes in Physics*, pages 337–370. Springer.

Peri, G. (2005). Determinants of knowledge flows and their effect on innovation. *The Review of Economics and Statistics*, 87(2):308–322.

Singh, J. (2005). Collaborative networks as determinants of knowledge diffusion patterns. *Management science*, 51(5):756–770.

Singh, J. and Marx, M. (2013). Geographic constraints on knowledge spillovers: Political borders vs. spatial proximity. *Management Science*, 59(9):2056–2078.

Tang, L. and Walsh, J. P. (2010). Bibliometric fingerprints: Name disambiguation based on approximate structure equivalence of cognitive maps. *Scientometrics*, 84(3):763–784.

Thompson, P. and Fox-Kean, M. (2005). Patent citations and the geography of knowledge spillovers: A reassessment. *American Economic Review*, 95(1):450–460.