Geography, Ties, and Knowledge Flows: Evidence from Citations in Mathematics*

Keith Head† Yao Amber Li‡ Asier Minondo§

March 27, 2018

Abstract

Using data on academic citations, career and educational histories of mathematicians, and disaggregated distance data for the world's top 1000 math departments, we study how geography and ties affect knowledge flows among scholars. The ties we consider are coauthorship, past colocation, advisor-mediated relationships, and *alma mater* relationships (holding a Ph.D. from the institution where another scholar is affiliated). Logit regressions using fixed effects that control for subject similarity, article quality, and temporal lags, show linkages are strongly associated with citation. Controlling for ties generally *halves* the negative impact of geographic barriers on citations. Ties matter more for less prominent and more recent papers and show no decline in importance in recent years. The impact of distance—controlling for ties—has fallen and is statistically insignificant after 2004.

Keywords: knowledge diffusion, distance, borders, networks, academic genealogy
JEL: F1, O3, R1

*The authors thank the Mathematics Genealogy Project (MGP) for providing data from its database for use in this research and Mitch Keller's assistance in obtaining genealogy data from MGP. The authors also thank Nicolas Roy, from zentralblatt-math.org for providing a correspondence between MGP author identification and zbMATH author identification. Yao Amber Li gratefully acknowledges financial support from the Research Grants Council of Hong Kong, China (General Research Funds Project no. 643311), and Asier Minondo from the Spanish Ministry of Economy and Competitiveness (MINECO ECO2015-68057-R and ECO2016-79650-P, co-financed with FEDER) and the Basque Government Department of Education, Language policy and Culture (IT885-16). Seminar participants at Dartmouth, LSE, Oxford, UBC, Glasgow, Birmingham, and Nottingham made helpful suggestions. We also thank Andrew Bernard, Teresa Fort, Joshua Gottlieb, Bob Staiger, Bronwyn Hall, Wolfgang Keller, Anthony J. Venables, Quoc-Anh Do, Edwin Lai, Jim MacGee, Andrés Rodríguez-Clare, Daniel Sturm, Dave Donaldson, Ben Faber, and Pablo Fajgelbaum for valuable discussions. We thank Michal Fabinger for alerting us to the importance of arXiv.org for knowledge dissemination over the internet. We are particularly grateful to Andrei Levchenko for raising issues that led to the results in section 4.3. Finally, we thank Ho Yin Tsoi, Bo Jiang, Yiye Cui, and Song Liu for excellent research assistance during this project.

†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

1 Introduction

Mounting evidence points to the importance of geographic barriers to knowledge flows. Arguing that citations provide the "paper trail" for knowledge flows, Jaffe et al. (1993) establish that cites to patents are geographically localized.[1] Keller (2002) shows that research spillovers on productivity decay with distance and Comin et al. (2012) find the likelihood of adopting new technologies declines with distance to the origin of the invention. Ellison et al. (2010) show that industries that share ideas (proxied by R&D and patent citation flows) have a stronger tendency to coagglomerate in space. While much of the literature focuses on technology diffusion, spatial separation impedes the spread of many other types of information. For example, information frictions account for half of the distance effect in the Allen's (2014) study of differences in rice prices between Philippine islands. Urbanization continues to increase despite rising land prices and congestion, a fact Glaeser (2011) attributes to the spread of innovations "from person to person across crowded city streets."

All the above evidence notwithstanding, the notion that borders or distance could prove to be practical obstacles to flows of knowledge seems hard to square with the fact that information can move anywhere without incurring either tariffs or freight costs. As Keller and Yeaple (2013) put it "Knowledge, as an intangible, seems ideally suited to overcoming spatial frictions ..." Especially in the age of Google, whose self-described mission is to "Organize the world's information and make it universally accessible and useful," the microfoundations for geographic knowledge frictions are far from obvious. To the extent there is a standard explanation, it is often mentioned that tacit knowledge is easier to communicate face to face. However, one study shows that even the transmission of highly codified information benefits from proximity. Lissoni (2001) examined a cluster of mechanical firms in Brescia, Italy and found they engaged primarily in the transfer of CAD encoded designs.

In this paper we hypothesize that distance's impact on knowledge arises in large part due to spatially concentrated personal ties. Proximity facilitates tie formation and those ties foster knowledge flows. The general mechanism we envision is that an agent trying to solve a problem becomes aware of potential solutions by tapping the knowledge residing in their network of personal relationships. This hypothesis can only be tested in a specific context where interpersonal ties, geography, and knowledge flows can all be tracked in a systematic way. We argue that the rich data available on mathematicians makes them, despite their idiosyncrasies, an insightful group to study for this purpose. Our first key finding is that adding controls for a comprehensive set of career and educational linkages between authors of mathematics papers, leads to a halving of estimated geography effects.

[1] Successive studies including Peri (2005), Belenzon and Schankerman (2013), Singh and Marx (2013), and Li (2014) estimate robust negative distance effects on patent citation propensities.

The role of ties in attenuating the negative effect of distance on citations echoes Keller's (2001) finding that including trade flows and FDI in the equation for technological knowledge spillovers shrinks the estimated negative effect of distance. The paper proceeds to combine additional results to establish the microfoundations for why ties matter so much.

Prior work on patent citation has already pointed towards ties as an important determinant of knowledge flows. Invoking the idea of "social proximity" Agrawal et al. (2008) and Kerr (2008) show that inventors have a higher propensity to cite patents by those who share their ethnic origins (as revealed by their surnames). While social connections are known to be richer within ethnic groups, sharing surnames with the same ethnic origin does not imply a personal connection between citing and cited inventors. Co-ethnicity can reflect cultural similarities between inventors who do not know each other personally. In order to capture the effect of person-to-person ties on knowledge flows, we need data sources from which we can extract the histories of personal relationships. Patent applications provide enough information to determine past collaboration; Singh (2005) and Breschi and Lissoni (2009) find this type of tie increases citation. Agrawal et al. (2006) investigate a second tie, past colocation. They find that inventors who move institutions are still disproportionately cited in patent applications by their former colleagues.

To capture a richer set of social ties between individuals who potentially transmit knowledge to each other, we believe it useful to consider academics, for whom it is possible to identify ties based on educational histories. We take advantage of the fact that in mathematics, Ph.D. institutions and advisors have been tracked globally for a long time by the Mathematics Genealogy Project (MGP).[2] There is strong evidence from Waldinger (2010) that the quality of mathematics faculty causally increases subsequent academic success of their doctoral students. The MGP allows us trace the patterns of citation between advisors and advisees, classmates, and the academic "extended family."

No one would claim that the process through which mathematicians (or anyone else) form ties is entirely random. A concern for the estimation of the effect of ties is that the same unobservables that promote scholars to form ties with each other also affect the likelihood of them citing each other. The educational ties we focus on have the advantage of being predetermined with respect to the citation process, since it is rare for an academic to cite or be cited prior to obtaining doctoral education. Unlike colocation and collaboration, educational linkages do not change over time in response to shocks to the interests of citing authors. While there is substantial randomness involved in determining classmates, the matching between advisors and advisees is likely to be shaped by common interests. The worry is that author A may be more likely to cite a paper by a tied author B than author C who has no tie with B because A and B write on the

[2]Borjas and Doran (2012) use the MGP to identify immigrant mathematicians who received Ph.D.s from Soviet institutions.

same topics. The way we respond to this concern is to compare citation probabilities only between authors A and C who have written papers in the same 3-digit field of mathematics. We show that this control for article subject is essential. Without it, estimates of ties are substantially inflated. Controlling for 3 or 5 digit fields or even keywords, ties have reduced—but still large—estimated effects on citation. With the controls that give the lowest magnitude, 5-digit subject and a cocitation indicator, a single tie on average boosts the odds of citation by 46%.

In addition to the strength of its academic genealogy data, mathematics offers two additional advantages relative to other academic fields. First, mathematics employs a common language of communication. This suggests transmission of mathematics knowledge would be less influenced by linguistic and cultural factors. In many social sciences and humanities fields, there are journals that focus on certain regions or countries. For example, in the fields of history and literature, there are obvious reasons to expect national borders and language to influence citation patterns. A second advantage of studying mathematics comes from the citation norms of the discipline. New theorems build upon previous theorems, which must be cited. There also appears to be a norm against gratuitous citation, as evidenced by the relatively low number of references in each paper. Althouse et al. (2009) report that math papers cite 18 papers on average, compared to 30 in economics and 45–51 in sociology, psychology, business and marketing.

Our first set of results establish that ties are an important mechanism underlying estimated geography effects on citations. But what is the mechanism underlying the importance of ties? We present two lines of evidence to argue that ties matter because they transmit information. The first follows from the idea of Arrow (1969) that knowledge flows can be generally thought of as interactions between a teacher (sender) and a student (receiver). We find evidence that citations are stronger to the authors who are more likely to be senders of information. The odds of citation are seven times higher if a paper is written by the advisor of the citing author. The impact of the author being a former advisee is weaker, albeit still very large. Moving one step further apart in the advisor network, we find advisors of advisors have three times the normal odds of being cited, but there are no significant differences in their propensity to cite their advisees' advisees. The second line of evidence is that ties matter more for the types of papers where information is harder to acquire. Our estimates show that ties (and geographic separation) have stronger impacts for papers that were only recently published, or not heavily cited, or just in a different field.

The role of distance—after controlling for ties—even becomes statistically insignificant in recent years. This finding of declining geographic barriers extends the results of two earlier studies using very different methodologies. Keller (2002) estimates the rate of distance decay in the benefits that one country receives from R&D conducted in another

country. He finds that the distance decay rate fell by two thirds from the period 1970–1982 to 1983–1995. Griffith et al. (2011) analyze the number of days until the first citation of a newly granted patent. They find home inventors take fewer days on average to be the first to cite home-invented patents than foreign-based inventors. This home-bias declined substantially between 1975–1989 and 1990–1999. Our study shows that distance effects have fallen by two thirds from the early 1990s to the late 2000s. This extends the evidence from the previous literature to the decade in which internet usage becomes pervasive. Our investigation of time-varying coefficients also reveals that, despite the advances in scholar's ability to search for information over the internet, the impact of personal ties remains as strong as ever.

While we do not wish to draw conclusions that stray too far from the context of our estimation, the whole rationale for studying citations in mathematics is to obtain insights with broader applicability to knowledge flows. The extent that ties facilitate transfer of valuable knowledge in one context (math) provides a *prima facie* case for their potential importance in all cumulative, collaborative discovery processes. Collaboration in mathematics often takes the form of tied researchers making suggestions of previously proven theorems that could help prove new theorems. In other research contexts, from drug invention to financial engineering, there would be analogous ways that lessons learned by one person could help a tied person to solve a new problem.

Going beyond research, there is a wealth of suggestive evidence that entrepreneurs learn about potential business opportunities from their web of connections. For example, Kerr and Mandorff (2015) explain the remarkable concentration of ethnic groups in certain occupations (Gujarati-speaking Indians are over-represented in the motel industry by a factor of 108) by invoking knowledge acquired through social interactions. Learning from ties might also explain the robust empirical association between bilateral immigration stocks and trade flows.[3] Such work generally lacks individual-level evidence on the relevant social ties. Using our person-to-person measures of ties provides insight into the processes underlying the patterns seen in aggregated data.

The remainder of the paper is organized as follows. Section 2 posits a simple citation model to serve as the estimating framework for relating a paper-to-paper citation indicator to the ties and geography variables measured at the author level. Section 3 describes our data on citations, geography and ties and explains how we construct the estimating sample. Section 4 presents the results of our regressions. In the final section we reinterpret other research findings in light of our results. We also suggest the potential policy implications.

[3]Gould (1994) is the seminal paper. Rauch and Trindade (2002) show that countries with larger ethnic Chinese diaspora populations trade more with each other. Combes et al. (2005) use migration and investment data to infer that social and business networks create trade within France.

2 Specification of citation probability equation

To guide estimation and interpretation, we provide a simple model of the citation process, leading to a reduced-form estimating equation for the probability of one article citing another. We then specify the observed determinants of citation and a method for controlling for key unobservables.

At the article level, citation is a binary choice and we therefore follow the standard approach of defining a latent variable C^*_{id} which leads to a realized citation, $C_{id} = 1$ of paper d by paper i when a threshold κ is exceeded. Thus the probability of citation is $\mathbb{P}(C^*_{id} > \kappa)$.

Articles should cite the relevant preceding work. However, author teams can only cite papers if they are aware of them. These truisms suggest that citation probabilities should be increasing in the product of relevance and awareness. We therefore model $C^*_{id} = A_{id}R_{id}$ where A_{id} denotes the level of awareness of citing team i of paper d and R_{id} scores the relevance of the content of paper d for paper i. The marginal effect of awareness is zero for irrelevant ($R_{id} = 0$) papers and the marginal effect of relevance is zero under the condition of ignorance ($A_{id} = 0$).

We model awareness as an exponential function of a vector of indicators of geographic separation, $\mathbf{G}_{id}$, and of the educational and career linkages, $\mathbf{L}_{id}$, between members of the two author teams. Geographic proximity matters because it increases the frequency of face-to-face interactions (from "water-cooler" conversations to conference meetings). Information flows can overcome geographic barriers if authors of papers i and d are connected via overlapping career and/or educational histories. Past colocation or just indirect linkages such as having the same advisor at different times create a kind of connective tissue that facilitates knowledge flows. In summary we hypothesize that $\partial A/\partial \mathbf{G}^{(k)} < 0$ for all k elements of geographic separation and $\partial A/\partial \mathbf{L}^{(k)} > 0$ for all k indicators of ties between author teams.

We model relevance to depend on an article-d specific function of the subject area of the citing article, $s(i)$, the year the citing article is published, $t(i)$, and a random term, ε_{id}, representing idiosyncratic factors operating between the article pair. Thus, we have

$$R_{id} = \exp(\alpha_{s(i)t(i)d} + \varepsilon_{id}).$$

The d component of $\alpha_{s(i)t(i)d}$ embodies the *general* importance of article d to all mathematics articles. The "intellectual distance" between the subject of article i and article d enters via the $s(i)d$ component of α. The $t(i)d$ component captures the idea that relevance of article d to all subjects may decrease over time due to obsolescence of older ideas. The particular usefulness of the combined fixed effect is that it allows article d

to have time-varying patterns of relevance that differ across subject areas. Consider an example familiar to trade economists. Hopenhayn (1992) became more important for the subject of international trade after the publication of Melitz (2003). This subject-specific rise in relevance of an article would not be captured via time or subject or article fixed effects introduced separately. However, the triad fixed effect $\alpha_{s(i)t(i)d}$ is able to absorb it.

We can take monotonic transformations of C^* and the threshold without affecting probabilities so we take logs, leading to

$$\ln C^*_{id} = \mathbf{G}'_{id}\boldsymbol{\gamma} + \mathbf{L}'_{id}\boldsymbol{\lambda} + \alpha_{s(i)t(i)d} + \varepsilon_{id}, \tag{1}$$

The probability of citation is the probability $C^*_{id} > \kappa$ and is given by

$$\mathbb{P}(C_{id} = 1) = \mathbb{P}(-\varepsilon_{id} < \mathbf{G}'_{id}\boldsymbol{\gamma} + \mathbf{L}'_{id}\boldsymbol{\lambda} + \alpha_{s(i)t(i)d} - \ln\kappa) \tag{2}$$

For ε distributed logistically with parameters μ and σ the probability of citation takes the familiar logit form:

$$\mathbb{P}(C_{id} = 1) = \Lambda[(\mathbf{G}'_{id}\boldsymbol{\gamma} + \mathbf{L}'_{id}\boldsymbol{\lambda} + \alpha_{s(i)t(i)d} - \ln\kappa - \mu)/\sigma], \tag{3}$$

where $\Lambda(x) = (1 + \exp(-x))^{-1}$.

We use logit as the primary estimator (and discuss linear probability model results in the robustness section 4.5) since it constrains predicted citation probabilities to be non-negative. Logit coefficients provide the marginal effect on changes in the log odds.[4]

The $\alpha_{s(i)t(i)d}$ fixed effects are a critical part of our estimation strategy since there is no reason to expect the geography and ties variables to be orthogonal to the triadic relevance term. Indeed, it is likely that authors of more important articles would be better connected. Moreover, authors who tend to work on similar subjects are more likely to be connected. That is, intellectual separation between $s(i)$ and article d may be negatively related to $\mathbf{L}_{id}$. We therefore estimate our model controlling for $\alpha_{s(i)t(i)d}$, the triad of subject of i, year of i, and article d.

While we have modeled awareness as a function of geography and ties only, we could easily introduce $s(i)t(i)d$ effects and random article-pair effects. They would simply be incorporated into α and ϵ. This means, for example, that we allow for a completely general pattern of diffusion of awareness of article d on different subjects s.

Estimating $\alpha_{s(i)t(i)d}$ with a large number of articles is computationally difficult and

[4]In the context of rare events such as citations, marginal effects on probabilities can be tiny. Singh (2005) multiplies his marginal effects by one million for reporting purposes. We find odds ratios are more intuitive, but as with rare diseases, one must keep in mind that a large odds ratio does not imply a large change in the probability of a positive outcome.

raises concerns over the incidental parameters problem. Instead we take advantage of the logit feature that the total number of cites received by each triad is a sufficient statistic for $\alpha_{s(i)t(i)d}$. This permits estimation in terms of a conditional density to obtain consistent estimators of the $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$ parameters. Prior work has included fixed effects for time lags (Singh, 2005), cited patents (Thompson, 2006), and cited institutions (Belenzon and Schankerman, 2013). This is the first study to control for the triad of citing article subject, citing article publication year, and cited article.

The unit of observation for citations is the *article* pair. However, the geography and ties variables underlying $\mathbf{G}_{id}$ and $\mathbf{L}_{id}$ are measured at the *author*-pair level. Mathematics has traditionally been characterized by more sole authorship than other fields. The average number of authors in mathematics has risen over time[5] but remains just 1.88 in 2009.[6]

For multiple-author article pairs, we must decide how to aggregate geography and ties of coauthors. 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 $\mathbf{G}$ and $\mathbf{L}$ 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

[5]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.

[6]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).

ties and geography variables are easier to interpret. We show in a robustness table that the averaging approach yields results that are similar for geography variables but stronger for ties.

3 Data

In this section we describe the four sources of data we have used in this study and how we obtained the geography and ties indicators. We then show how we combined the different sources to construct the estimating sample using a matching methodology. We also detail important features of the estimation method that arise because citation is a rare event.

3.1 Sources of data

Our data set combines four main sources:

1. Web of Science (WOS):[7] citations, author affiliations, keywords.
2. Mathematics Genealogy Project (MGP): place and time of Ph.D., names of the dissertation supervisor(s).
3. Zentralblatt MATH (zbMATH): 5-digit mathematical subject classifications (MSC) for citing and cited articles.
4. Google Maps: longitudes and latitudes for 1000 mathematics institutions used to calculate geodesic distance data between citing and cited author teams.

Web of Science

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

[7] This database was previously called Thomson Reuters' ISI Web of Knowledge.

than 1975. With 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 Appendix A.1 we raise the fraction of affiliation identifications to 84%.

The WOS contributes three indicators of ties based on past coauthors and past affiliations. Each tie variable is based on actions taken prior to the publication year of the relevant citing article.

- "Coauthors" indicates whether author pairs have collaborated on a paper published in one of the 255 math journals included in WOS since 1975.
- Location history: "Coincided past" requires colocation at the same institution in the *same year* but the authors no longer work at the same place. "Worked same place" indicates that two authors worked at the same institution in *different years* in the past.

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.

The MGP data allow us to construct eleven additional measures of ties based on three types of relationships.

- Classmate relationships: "Share Ph.D." denotes author pairs who graduated from the same Ph.D. program within a 5-year period and who are therefore assumed to have overlapped.
- Academic "family" relationships: "Advisor citing" takes the value of 1 if the author of the citing article was the PhD advisor of the author of the cited article. For "Advisor cited" the citing author was the advisee. Academic siblings were both supervised by the same professor. Academic grandparents are the advisors of the citing or cited authors' advisors. Academic cousins are authors that share a grandparent. Academic uncles are the advisees of one's academic grandparent.
- Alma Mater relationships: These variables indicate when the citing or cited author is affiliated to the institution where the other author received her PhD. For example

"Alma Mater cited" takes a value of 1 when an Princeton alumnus cites a professor currently affiliated with Princeton.

All ties are computed as dummy variables, taking the value of one if the tie exists. These dummy variables are additive: if author i has co-authored with author d who is also i's Ph.D. advisor, there would be ones for both co-author and advisor cited.

Table 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 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.[8] 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,

[8]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/`

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 the results section).

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.

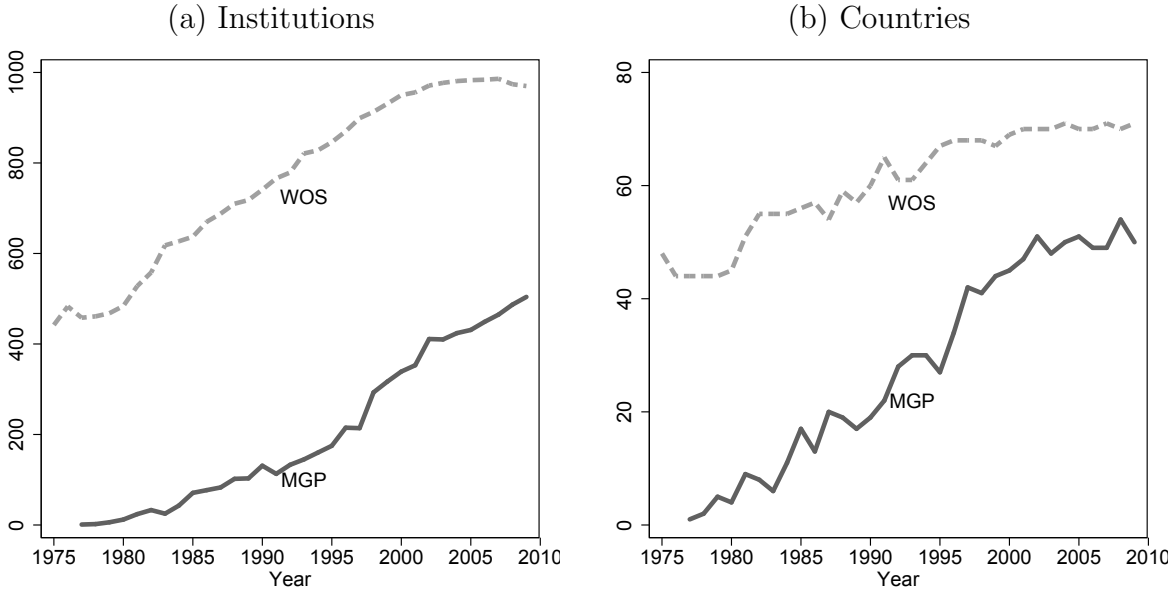
Using publications to track author locations over time, we calculate distances (and other measures of geographic separation) at the time the *citing* article is written. Past work using patents calculated distances between inventors using the cited inventors' addresses in the year the *cited* patent 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 Appendix A.1, we fill these gaps through interpolation and extrapolation, assuming that moves occur in the midpoint 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 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).

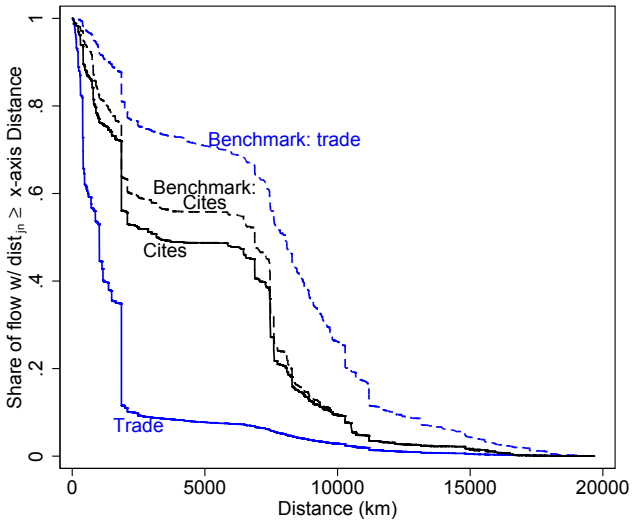
Meanwhile, the numbers of institutions and countries represented in the WOS citation data increase over time. Figure 1 shows that during the period 1975–2009 the set of institutions with citing or cited author affiliations rises to nearly 1000 (some institutions disappear) and the corresponding number 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.[9]

Before estimating any regressions, it is useful to see whether geographic impediments to knowledge flows can be seen in a fully non-parametric context. Figure 2 displays the geography of citation patterns in mathematics. It graphs survival functions for citation

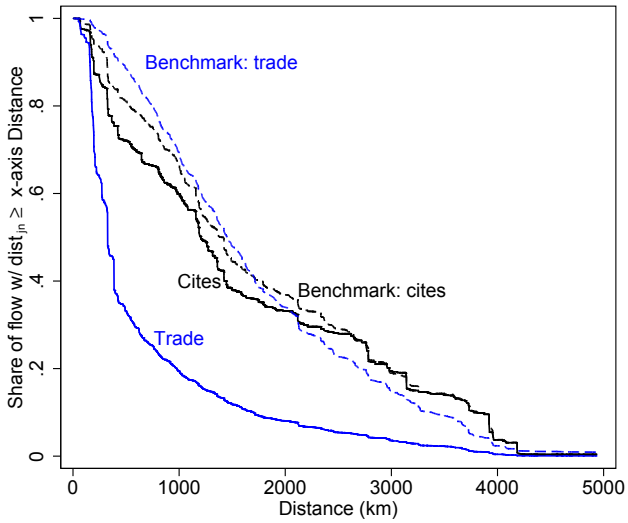
[9]Peri (2005) and Singh and Marx (2013) include international citations but the challenge of determining locations for individual patentees limited Peri's sample to 18 countries, whereas Singh and Marx (2013) limit their sample to cited patents with US-resident inventors.

Figure 2: Distribution of distances for math citations and trade in goods

(a) International



(b) Inter-state (US)



Note: Solid lines represent actual flows of citations or goods between origins and destinations. International flows of value-added in manufacturing (blue lines in left figure) come from the OECD/WTO TiVA database (2009) whereas state-to-state flows come from the FAF database (2007). Dashed lines are "frictionless" benchmarks for citations or trade (calculation detailed in the text).

flows as a function of distance between the authors of the citing and cited papers. $S^c(D)$ is the share of all cites that occur with distance $\geq D$. Citations from one nation to another are calculated by summing the citations from papers written by authors affiliated with institutions in country j to papers written by authors in country n.[10]

The benchmark for cites is a dartboard model that takes as given each country's outward citations, $C_j \equiv \sum_n C_{jn}$ and inward citations, $C_n \equiv \sum_j C_{jn}$. The international allocation of these citations is completely random; that is, each paper is equally likely to cite any other paper regardless of distance. Randomness implies that outgoing cites from j go to country n with probability given by n's share of all received cites. Thus, the aggregate flow of benchmark cites from j to n is given by $F^c_{jn} \equiv C_j(C_n/C_w)$, where C_w sums all cites in the world. F^c_{jn} can be thought of as the "frictionless" flow of citations from j to n. The survival curve for the benchmark is $\bar{S}^c(D) = \left(\sum_{\text{dist}_{jn} \geq D} F^c_{jn}\right)/C_w$. Figure 2 displays $S^c(D)$ and $\bar{S}^c(D)$ using solid and dashed black lines. The vertical gap between $\bar{S}^c(D)$ and $S^c(D)$ measures the frictions that divert citations away from the dartboard benchmark.

The blue lines in Figure 2 permit comparison with actual and benchmark flows of

[10]For papers with multiple authors from different countries, citations are allocated fractionally. Thus, a paper co-authored by two scholars from countries A and B to a paper written by two other authors from countries C and D would generate four international citation flows of 0.25 each. This fractional accounting of citations ensures that the sum of all citations in the world, C_w, is the same regardless of whether one sums across paper dyads or country dyads; that is $C_w = \sum_{jn} C_{jn} = \sum_{id} C_{id}$.

trade in goods. Research using gravity equations has established that distance is a major friction impeding trade in goods.[11] To facilitate comparisons with the citation data, we employ trade data sets that measure each origin's aggregate flows including those that remain within that origin. Thus, trade flows to self are value-added minus exports of value-added to the rest of the world.

Panel (a) displays flows of manufacturing value-added between and within 63 countries derived from the Trade in Value Added (TiVA) dataset made available through a joint effort of the World Trade Organization (WTO) and the Organization for Economic Cooperation and Development (OECD). One prominent aspect of Panel (a) is that a very large share of trade takes place within countries. The precipitous drops in the survival functions for both cites and goods seen at 1854km correspond to the CEPII internal distance of the United States (the average distance between 20 major cities).

To see what is happening within this important set of intra-national flows, we display the survival functions for state to state citations and trade in Panel (b). Citations are aggregated up to the state level just as they were for countries in Panel (a). The value of goods transported in 2007 between and within the 50 states and Washington, DC come from the Freight Analysis Framework (FAF) database. We display distances up to 5000km (excluding some Hawaii and Alaska dyads) because by that distance both benchmarks and actual flows are indistinguishable visually from zero.

What we learn from Figure 2 is that distance attenuates knowledge flows in mathematics leading them to occur over shorter ranges than one would expect in a frictionless world. This is true at international scale and also true within the United States. The gap between actual and benchmark citation flows is much smaller than what we observe for goods flows. This is consistent with the hypothesis that trade flows are attenuated by *both* transport costs and information decay. Moreover, distance decay effects in commercial activities may be larger than those that apply to researchers.

3.2 Construction of estimating sample

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 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

[11] See Head and Mayer (2014) for explanation of the gravity methodology and results and Head and Mayer (2013) for a version of the distance distribution figure that considers only gross trade flows between countries.

Table 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.

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.[12] 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.[13] 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 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%.[14]

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 2 raises concerns that the new sample might not be representative. We shall show in Table 3 that the MGP sample is

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

[13]Appendix A.2 describes how we identified and removed self-citations.

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

remarkably *similar* to the larger sample with respect to the means of the variables we can measure for both sets.

A standard "exogenous sampling" approach would entail picking a set of citing articles and constructing the universe of papers they might cite and predicting which potential cites are actually realized. Applying such an approach in the case of citations creates both conceptual and practical problems. First, it is hard to determine the appropriate "universe." Should we consider the applied math papers that might have cited a given paper, the physics papers, the economics papers? The data gathering challenge for a true universe of potential citing papers would be formidable. There would also be computational difficulties with incorporating so many non-citation observations. Citations are an example of a rare event problem. In the Web of Science sample (before imposing the requirement of MGP data on all authors), there are approximately 3 billion potential cites and about 269,000 realized cites. Thus, the rate of citation is only 9 per 100,000. In response to this problem, the patent citation literature has generally adopted a choice-based sampling approach following the matching methodology of Jaffe et al. (1993). For each realized citation (case), a single non-realized citation (control) is selected at random from a larger set of matched potential controls.[15]

We adopt the one case per control approach when using the whole WOS sample. However, the sample featuring our full set of ties has a small enough number of realized citations that we can incorporate *all* potentially cited papers that meet certain criteria. Our baseline matching criteria is that controls be published in the same year and the same 3-digit field as the original citing paper (case). The union of the realized citations and the control group constitutes the sample that is used in the econometric analysis.[16] The presence of triadic fixed effects means that we have effectively the full set of control observations. To see this imagine another field A in which none of the papers cite a given paper d. Then the A-d part of the triadic fixed effect would be a perfect predictor for non-citation so all such observations would be automatically dropped from the fixed effects logit estimation.

Table 3 displays the differences between the characteristics of realized citations and the 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.[17] In terms of ties, citing authors are many times more likely to coauthor with the (realized) cited authors. They are also more than

[15]Singh (2005) uses five controls per realized citation in his weighted exogenous sampling maximum-likelihood (WESML) estimator.

[16]Kerr and Kominers (2015) use an alternative method that randomly samples patent distances to calculate *expected* citations within a fixed ring.

[17]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.

Table 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.

twice as likely to have worked at the same university either at the same or different times. Since all these variables are correlated we will need to estimate regressions to determine the partial relationships.

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

4 Regression results

This section presents the main results regarding the effect of geography and ties on knowledge flows. All regressions are logits with fixed effects for each group defined by citing field (3-digit subject codes), citing year, and cited article. The assumption is that conditional on these fixed effects, variation in geography and ties can be viewed as random, allowing for a causal interpretation of the estimates. We recognize this is a strong assumption but provide evidence that our subject controls are effective at reducing bias due to endogeneity. The reported coefficients have the interpretation of marginal effects on the log odds. Standard errors are clustered at the cited article level to allow

for correlations in the errors across potentially citing articles for the same cited article.

There are four key findings. First, the effects of distance, borders, and language differences are about half as strong once educational and career links are taken into account. Second, 13 of the 14 measures of ties have positive effects that are significant at the 5% level in our final specification. On average the effect of adding a tie raise the odds of citation by 80%, with some ties having much bigger effects. While this magnitude depends on the specific way we control for subject of the citing paper, a large, highly significant association between ties and citations holds up with even the most stringent measure of subject (using the same keywords). Third, ties and geography affect different types of papers differently. In particular, less prominent and more recently published papers exhibit stronger effects. Finally, while the importance of distance has declined to the point of statistical insignificance in recent years, ties remain as valuable as ever.

4.1 Baseline

Table 4 reports the result of baseline logit regressions.[18] The first specification includes only the four geographic explanatory variables: an indicator for distance greater than zero (not being at the same institution), log distance (interacted with the positive distance indicator), and indicators for residing in different countries and from countries that have different official languages. The two-part distance function is necessary because there is no good way to directly measure the distance between two scholars at the same institution. The first of the two parts implicitly estimates this distance. The indicator for distance greater than zero is equivalent to a "different university" dummy. The two-part formulation has a jump from zero to positive distances, but thereafter the elasticity of citations odds with respect to distance is constant. While a constant elasticity of distance in trade equations is the standard assumption underlying gravity equations, there is little *a priori* reason to expect this relationship to carry over to citations. Therefore we re-estimate specifications (3) and (5) with distance-interval step functions in columns (4) and (6).

The second specification adds ties constructed from the WOS database. The third to sixth specifications restrict the sample to the articles with full information from the MGP database. The overall estimating sample does not decline much because the MGP sample uses all available controls (non-citations in the same subject-year), whereas the WOS sample has just one control per case. As in the first two columns, we first show the effects of geography without ties (columns 3 and 4) and then with the full set of ties available in the MGP data (columns 5 and 6).

[18]The entire table is re-estimated using the linear probability model in appendix table C.5, with the results compared in section 4.5.

Table 4: Baseline: matching by MSC-3d, full author information

	(1)	(2)	(3)	(4)	(5)	(6)
Specification:	Triad-fixed-effects logit (TFE-Λ)					
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 | Dist > 0	-0.073⋆	-0.052⋆	-0.068⋆	Figure 3	-0.037⋆	Figure 3
	(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~	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%, ~: 10%.

Specification (1) presents significantly negative coefficients on distance and borders, suggesting that physical distance and borders indeed impede knowledge flows. We estimate smaller border and distance effects than those obtained by Singh and Marx (2013) using citations of US patents. Whereas we find that crossing a national border reduces citation odds by $\exp(-0.198) - 1 = -18\%$ they find a 41% reduction (specification (6) of Table 5).[19] Our distance elasticity is -0.073 whereas theirs is -0.137. While it is tempting to attribute this halving of geography effects to differences between academic and commercial diffusion of ideas, other evidence on patent effects obtains similar magnitudes to our column (1). As the coefficients show the marginal effects on the log odds of citation and citation is rare, the dependent variable approximates the log probability and should therefore be proportionate to the log citation flow in aggregated data. This means we can compare our estimates directly to the results from the gravity-type regressions on patent citations estimated by Peri (2005) and Li (2014). The different country (border) effect we estimate is -0.198, whereas the baseline estimate of Peri (2005) is -0.19. Li (2014), also estimating a patent citation gravity equation, reports distance elasticities (after controlling for subnational borders) from -0.03 to -0.067, which are slightly weaker than those reported in our column (1). All these results support the conclusion that border and distance decay of citations are considerably smaller than the effects typically estimated for trade in goods. Nevertheless, it may be surprising to many that geography has a significant impact on academic citations at all. We now show that the estimated effects are substantially reduced by controlling for ties.

The second specification shows that the three measures of career ties (past coauthorship, past colocation, and past work at the same institution) all have strong positive associations with citation. As exponentiating the coefficients in a logit expresses the effects in terms of the change in citation odds ratios the 0.712 coefficient on past colocation implies that even after colleagues have moved to separate institutions, they have 104% higher odds of citing each other ($\exp(0.712) - 1 = 104\%$). Prior coauthors are even more likely to cite each other. We also see that the inclusion of career ties lowers geography effects somewhat.

Comparing columns (1) and (3) we see that estimating the same specification on the MGP-restricted sample does not change the geography coefficients by more than one would expect given the standard errors.[20] Comparing columns (3) and (5) we see one of the headline results of this paper: Controlling for ties shrinks the negative effects of geographic separation by about 50%. The ratios of the four geography coefficients in

[19] The gap between our results would be narrowed by including an additional -10% citation odds reduction from not sharing a common language, which would be the case on the majority of cross-border country pairs.

[20] Additional investigation of the possibility of MGP sample selection bias is reported in the robustness subsection 4.5.

column (5) to the corresponding coefficients of column (3) are 0.46, 0.54, 0.39, and 0.30. The omitted variable bias formula tells us that this means that ties and geography are correlated and that the pure partial effect of being far away or in a foreign country is overestimated in regressions that omit controls for ties.

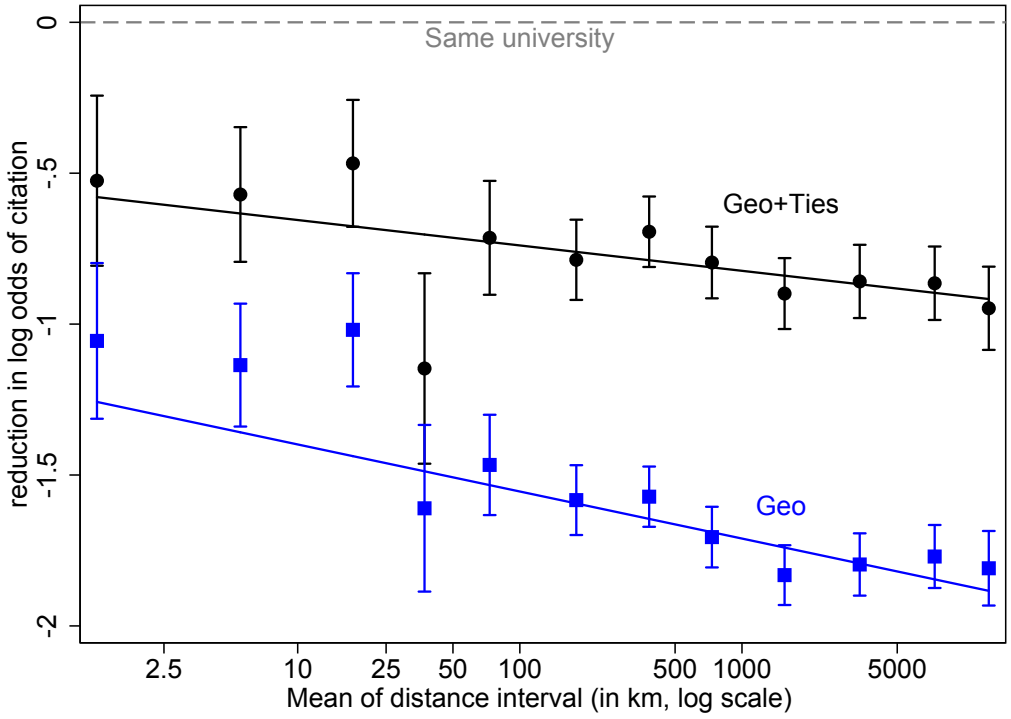
Table 4, columns (5) and (6) show that, with just one exception, ties have systematically positive effects on citation probability. All of the estimates are statistically significant at the 5% level except "grandparent citing" which has an imprecisely measured negative effect and "uncle citing" which has a borderline significant result in column (5). The average over all fourteen ties coefficients is 0.59, implying that the average tie raises the odds of citation by 80%. The addition of the full set of ties in column (5) dramatically increases the fit of the logit to the data: the pseudo R^2 nearly triples from 0.033 to 0.091.[21]

There are three tie relationships where one can identify the more "senior" of the two authors: advisors, uncles, and grandparents. In each case we observe the author in the teaching role is more likely to be cited than to cite. Advisees massively over-cite their advisor's papers by a factor of four (the second largest impact of the 14 types of ties). In the reverse direction, we find advisors over-cite their advisees' articles by a factor of three. The academic "nephew" overcites his "uncle" by 86% but the reverse direction features a bias of just 26%. The most pronounced asymmetry emerges when we skip a generation. Authors over-cite their advisor's advisors (academic grandparents) by a factor of three. Yet this intergenerational flow is not reciprocated; the grandparents' propensity to cite advisees of their advisees is not significantly different from zero. These vertical patterns support the hypothesis that citations transmit knowledge.

Figure 3 illustrates the coefficients on each of the 12 steps in the non-parametric estimation of distance effects conducted in specification (6) of Table 4, represented with black circles. The vertical axis depicts the reduction in the log odds of citation associated with each step, relative to working at the same institution. We also show with blue squares the corresponding estimates for the 12-step specification omitting ties. For each set of steps, we overlay the implied reduction in the log odds of citation based on the 2-part coefficients from specifications (3) and (4). The key finding illustrated in the figure is that after the dramatic fall associated with positive distance, the subsequent declines are consistent with a constant elasticity decay rate. Controlling for ties moves the decay function up (lower effect of being at different institutions) and flattens it. After controlling for ties, the two-part prediction lies within two standard errors for 11

[21] Pseudo R^2 is measured as $1 - \mathcal{L}_1/\mathcal{L}_0$ where $\mathcal{L}_0$ is the likelihood of the constant-only model. Hence it rises with the number of estimated parameters. It is therefore worth noting that the inclusion of ties reduces the Akaike Information Criterion (AIC) by 7995 points compared to column (3), indicating that the rise in the likelihood from adding ties is large enough to offset the penalty AIC imposes for adding 14 parameters.

Figure 3: Non-parametric estimated geography effects



Note: Black circles and whiskers correspond to estimates and confidence intervals for 12 distance bins for specification (6) in Table 4. Blue square estimates are from a specification that omits ties.

out of 12 steps.[22] Clearly there is a big discontinuity between zero and positive distances corresponding to a same-university effect. Conditional on positive distance, the figure shows that it is hard to distinguish empirically between a decay function that is flat after 1000 kilometers and one that exhibits regular decay with a constant elasticity of -0.037. Since the 2-part approach adequately captures distance effects, we use it for all the subsequent estimations.

The negative effect of geographic barriers on citation probabilities is presumed to arise because these barriers reduce the frequency of face-to-face interactions. In academics (as well as other areas) co-attendance at conferences provides one of the most important opportunities to meet in person with scholars doing related work. We collected data on papers presented between 1990 and 2009 at one of the most important conferences, the Joint Mathematics Meetings (JMM). Held annually in the United States, an average of 1459 participants present 1037 papers.

The first exercise we conduct, reported in Table B.1, is to show that there is a strong

[22]The exceptional case is the 25–50km bin, which is driven by the dyad Rutgers-CUNY (45km apart). Both of these math departments are very active in the Set Theory 3-digit code but they do not cite each other's papers. The apparent cause is that while Rutgers papers span the field, CUNY authors specialize in two sub-fields, Consistency and Independence Results and Large Cardinals, which comprise 52 out of CUNY's 58 papers.

and precisely estimated negative effect of distance on the probability of attending a conference. Because the conference venue moves each year, the data exhibit substantial variation in distance for a given scholar. This permits estimation of the logit with author-specific fixed effects. The distance elasticity in this specification is -0.136 with a 0.016 standard error (clustered by author).[23]

The second exercise, reported in Table B.3, uses the conference data to show the impact of attendance on citation. Using the Table 4 column 5 specification, we add indicators of coinciding at the same conference (as presenters or session organizers). While just coinciding has a negligible effect on citation, coinciding when the (potentially) cited paper is presented increases the odds of citation by a factor of 8.3. Table B.3 also shows a positive effect of presenting at the same session, regardless of whether it was the citing or cited paper. Contrary to our own observation of presenters being encouraged to cite the work of co-attendees at a session, we find no significant evidence of a citing paper effect in this data. While these two exercises are confined to the one conference for which we could obtain long-term conference participation data, they illustrate a broader mechanism that we view as underlying distance effects on citation. Proximate authors are more likely to present at the same conferences and, when they do so, this makes the citing authors aware of new relevant research which they build upon in their own work.[24]

Why does controlling for academic linkages lead to the large reduction in distance effects shown in Table 4? It must be that ties are negatively correlated with geographic barriers. We illustrate this in Figure 4(a), which shows that linked authors tend to be closer to each other than authors who have no ties. For example, about 33% of tied authors are more than 5000 kilometers apart, compared to almost 60% of non-tied authors. Similarly, tied authors are much more likely than non-tied authors to reside in the same country (51% vs 16%) or countries that share a common language (65% vs 32%).

Figure 4(b) reveals the phenomenon that helps to understand our baseline results: Mathematicians tend to remain close to the university where they obtained their doctorates. Thirty percent either do not leave or have returned and only 18% move more than 5,000km away.[25] Proximity to the *alma mater* is likely to beget proximity to one's advisor (and his advisor), former classmates, etc. The story underlying our baseline results is a

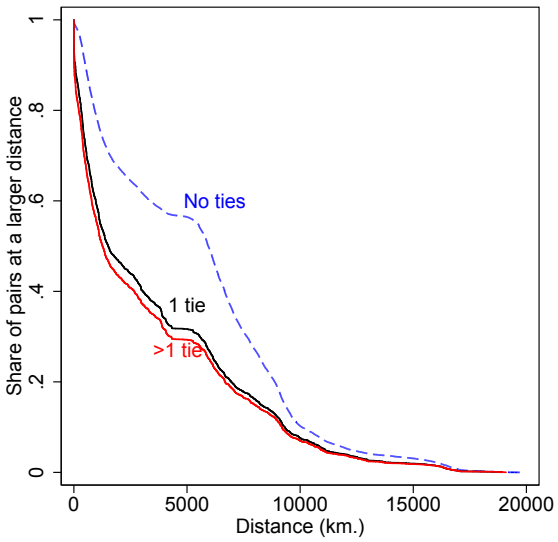
[23]This estimation includes only those authors who attended at least one meeting but not every one (no perfect predictors). Appendix B also presents an estimation without author fixed effects that includes all potential attendees. The distance effect in this estimation is not as strong (-0.05) but negative and significant border and language effects show up in this specification.

[24]Our results align with the finding of Iaria et al. (2018) that the ban on Central scientists from participating at international conferences during and after World War I was associated with a drop in citations between Allied and Central scientists.

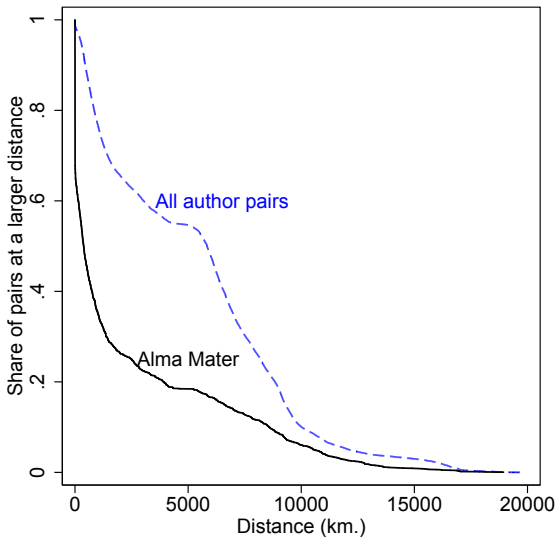
[25]There is substantial heterogeneity in the tendency to work at the Ph.D. granting institution, with the just 20% of US-educated authors staying/returning compared to 46% in Spain. The sample comprises 2213 MGP authors who published in pure mathematics journals in 2009.

Figure 4: The spatial concentration of tied authors

(a) Geography of ties



(b) Distances from Doctoral Institution



Note: Panel (a) constructs the complementary CDF for distances between authors of 2009 papers, distinguishing author dyads by the presence of ties. The blue line in panel (b) aggregates all the author pairs from panel (b). The black line plots the distribution of distances to the author's Ph.D. granting institution (*alma mater*).

simple but important illustration of omitted variable bias. Ties are very important for citation but ties are negatively correlated with distance. Thus a failure to control for ties leads to the inference that distance has a greater *direct* impact on knowledge flows than is truly the case. Authors are unlikely to cite papers written by faraway authors partly because they are less likely to have interacted at conferences, but an equally important factor is that they are less likely to have an academic or career tie with each other.

4.2 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 5 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 the previous table.

Table 5: 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 5 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 5 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 5 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.[26]

Appendix Table C.2 removes the ties indicators, but is otherwise identical to Table 5. Failure to control for ties dramatically magnifies the estimated impact of the geography variables. Generally 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.

4.3 Evidence for information mechanisms

The results we have obtained so far point to an important role for educational and career ties in fostering citations. The underlying mechanism we imagine is one of communication along the network of ties that causes one set of authors to become aware of useful theorems and conjectures provided by other authors. This information transfer mechanism predicts that the presence of ties should matter more for certain types of papers than others. Specifically, we conjecture that authors rely more on their ties to find out about work

[26]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.

that is less widely known, more recently written, and further from the author's field of expertise. To the extent that face-to-face interactions matter more for such papers, geographic barriers should be stronger as well.

Table 6: Obscure, Recent, and Different-field papers are more impacted by ties and geography

Specification:	(1)	(2)	(3)	(4)	(5)	(6)	(7)
		Obscure		Recent		Different field	
		base	interact	base	interact	base	interact
Geography:							
Distance > 0	-0.571*	-0.541*	-0.011	-0.321*	-0.339†	-0.804*	0.147
	(0.073)	(0.083)	(0.166)	(0.107)	(0.137)	(0.136)	(0.198)
ln Dist $\mid$ Dist > 0	-0.037*	-0.031*	-0.037$^{\sim}$	-0.021$^{\sim}$	-0.028$^{\sim}$	-0.026$^{\sim}$	-0.004
	(0.008)	(0.009)	(0.019)	(0.011)	(0.015)	(0.014)	(0.021)
Different country	-0.090*	-0.082†	-0.061	-0.095†	0.002	-0.072	0.022
	(0.031)	(0.034)	(0.079)	(0.044)	(0.060)	(0.056)	(0.088)
Different language	-0.025	-0.028	0.014	-0.017	-0.015	-0.037	-0.075
	(0.026)	(0.029)	(0.064)	(0.037)	(0.048)	(0.047)	(0.072)
Ties:							
Average effect of ties	0.652*	0.619*	0.135*	0.543*	0.176*	0.572*	0.223*
	(0.018)	(0.020)	(0.059)	(0.027)	(0.036)	(0.036)	(0.057)
Observations	441792	441792		441792		225768	
pseudo-R^2	0.091	0.092		0.093		0.100	

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.

We develop three proxies for papers that researchers are less likely to know about. First, we categorize papers as "obscure" if they receive less than or equal to the median number of cites (three). Our second proxy for low awareness is the gap in time between when the citing and potentially cited papers were published. A paper is "recent" if the gap is less than or equal to the median gap in our data (nine years). The third awareness measure follows from the observation that authors are more familiar with work in their own fields than in other subject areas. We classify papers as different field if their 2-digit mathematical subject classifications (MSC) differ (for example 11 Number Theory vs 14 Algebraic Geometry). As we show in a subsequent table, these specific rules for categorizing obscure, recent, and different field are not critical for the results.

Table C.3 in the appendix provides summary statistics on these variables. Not surprisingly, there are lower average number of cites for obscure papers and recent papers. We see approximate balance between the average number of cites to the same and to

different fields. There are more observations in total featuring cites within the same field so this suggests that cross-field citations go mainly to more prominent papers. In terms of ties, on average the differences between obscure and recent papers are small. The fact that ties are higher for same-field papers probably reflects greater ties within the same field. This is an important reason why our fixed effects control for the citing paper's 3-digit subject code.

Table 6 reports the detailed results for the three awareness proxies. To reduce the number of parameters to be displayed and discussed, we report the average of 13 ties indicators, followed in the next column by the averages over 13 interaction terms.[27] Column (1) reports the corresponding regression *without* interactions for comparison purposes. Compared to column (3) of Table 5, the average effect of ties rises because we excluded the grandparent citing indicator in this table (it has a negative and insignificant effect in Table 4).

The first set of interactions in Table 6 shows the results of interacting geography and ties with an indicator for obscure papers. Column (2) shows the base effects corresponding to non-obscure papers and column (3) shows the coefficient on each corresponding interaction. We find that the more prominent papers (>3 cites) have a 18% ($=0.135/(0.619+0.135)$) smaller coefficient on the average of ties than the lesser known papers. This is consistent with the interpretation that ties facilitate awareness. Papers that are big successes require less help from networks to promote transmission. The coefficient on log distance is about 1.2 times as large for obscure papers.

When the interaction is changed to distinguish recent versus older papers, the results are similar as shown in columns (4) and (5). Recent papers have a 24% higher coefficient on the average effects of ties. Distance decays are estimated at $-0.021-0.028=-0.049$ for papers in their first nine years after publication (the median age of papers in our sample) and -0.021 thereafter. These numbers are remarkably similar to those reported by Li (2014) in a gravity-style study of inter-city patent citation flows. She finds that the distance elasticity declines monotonically with age from a -0.028 in the first five years to -0.014 for patents granted 20 or more years before. These combined findings of significantly higher geographic concentration of "new knowledge" are intuitively appealing and provide some guidance for models of knowledge diffusion.[28]

Ties also have larger impacts for papers in different fields, with a coefficient, reported in column (7) that is 28% larger than for same-field papers. None of the different-field

[27] We drop "grandparent citing" in this table because of a logit perfect predictor problem. In the different-field specification, there were only 9 grandparent citing instances and all of them were for control observations, rather than realized cites. We reinstate grandparent-citing back in a robustness check where it is a component in a sum of ties variable.

[28] A recent paper studying patents finds corroborating results. Packalen and Bhattacharya (2015) show that denser cities are responsible for patents that make use of newer knowledge, as measured by textual analysis of the patent applications.

geographic interactions are statistically significant, suggesting that face-to-face communication matters more for obscure and recent papers than for different fields.

All three sets of interactions therefore support the premise that *scholars draw more heavily on their connections when obtaining less familiar information.* The positive interactions between ties and information proxies have similar magnitudes and strong statistical significance in five alternative specifications described in subsection 4.5. These robustness regressions also find statistically significant (10% or better) negative effects for the geography-information interactions in 11 out of 30 estimates. The remaining estimates are mainly negative but not statistically different from zero.

Tables 6 and C.6 show strong and robust evidence that ties matter more for three types of papers where awareness poses a more serious challenge. We also find that geographic barriers pose a greater impediment to citation for recent papers in nearly every specification. This evidence supports the interpretation of ties as facilitating information transfer rather than an alternative mechanism involving "citation cliques." Under this alternative, scholars have perfect awareness of the relevant research in their field but choose to cite specific prior work because it was written by the scholars for whom they have some kind of social affiliation. If ties are just proxies for intra-group loyalties, it is not obvious why such forces should be relatively more important specifically for the types of papers where the awareness gap is predictably larger.[29]

There is a third mechanism, combining elements of information and affiliation, which is also consistent with our results. In this story, mathematicians are aware of relevant work but uncertain of whether the proofs those papers contain are all correct. Since the validity of one's own results hinges on the correctness of the proofs of the cited theorems, the mathematicians we have spoken to claim to check all proofs, regardless of the author. In practice, this may not always occur. There could be cases where, for example, an author would cite her advisors papers because she knows his proofs have always stood up to scrutiny. This trust mechanism would likely be stronger for lesser known and more recent papers because they are less likely to have been thoroughly checked by others. Trust could also matter more for papers outside one's field because those involve unfamiliar techniques that make it difficult for an outsider to verify the proof.

We see the awareness and trust mechanisms as both emphasizing ties as conduits of information. In the first case, the information is about the *existence* of a useful theorem; in the second case the information is about the *reliability* of the theorem. This echoes the situation in international trade where Rauch (2001) summarizes a number of studies showing that "transnational business and social networks promote international trade by alleviating problems of contract enforcement and providing information about trading

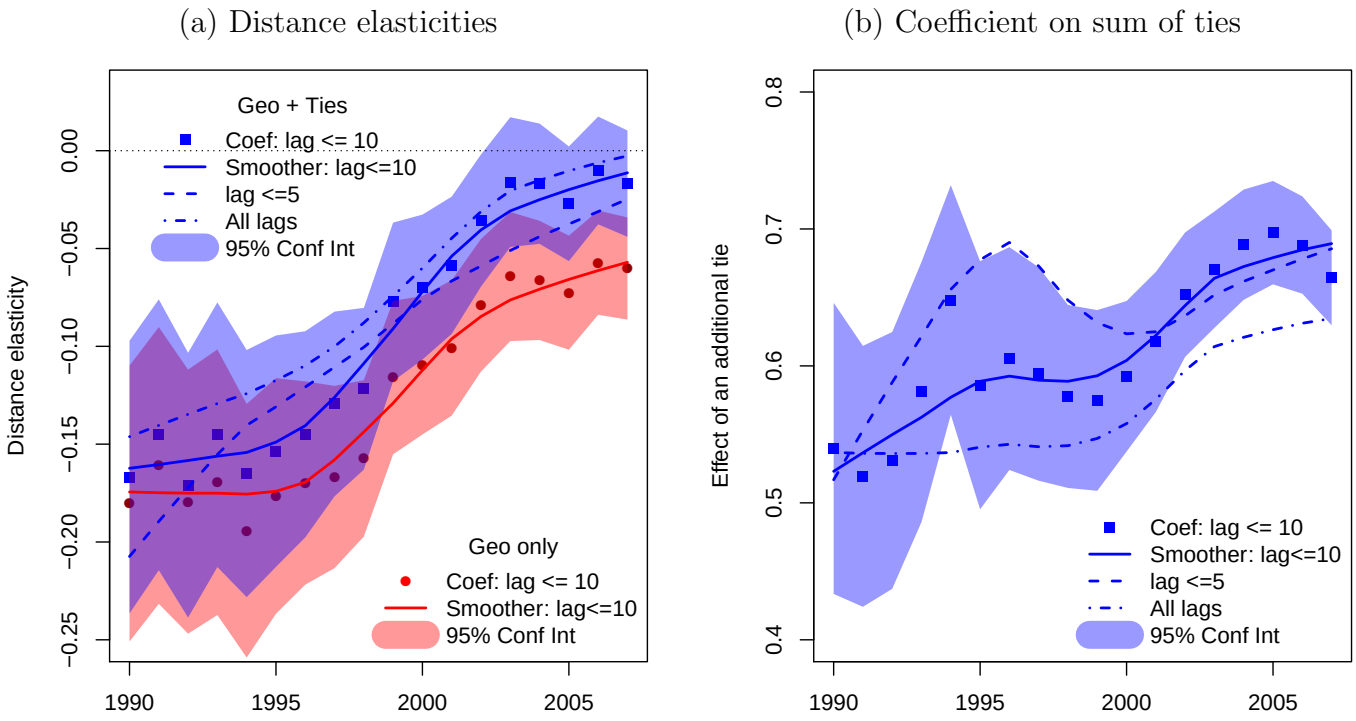
[29]The fixed effects control for the overall tendency to cite each article d so the interactions measure how ties boost the *relative* tendency to cite specific types of papers.

opportunities." Thus, networks help exporters by making them aware of the specific needs of foreign buyers, while also promoting trust that buyer and seller will comply with the terms of their contract with each other.

4.4 Time-varying effects of distance and ties

The estimates presented so far pool citations made from 1980 to 2009. This section investigates whether the effects of distance and ties on more recent citations differ from the past. The results of Keller (2002) and Griffith et al. (2011) show a decline in the importance of geographic separation between the 1980s and late 1990s. We extend the investigation these authors initiated by including more recent data and also estimating the time-varying effect of ties. We examine changes since 1990 because our 1980s citation data are too sparse. Estimation of time-varying coefficients from 1990 to 2009 is of great interest given the many relevant advances observed over this period.

Figure 5: Distance effects shrinking while ties matter more over time



Note: Plotted coefficients are marginal effects on the log odds of citation. Confidence intervals based on standard errors clustered at cited-article level. Red-shaded interval corresponds to estimates that do not control for ties. Blue-shaded estimates control for the *sum* of 14 ties. Estimation window moves by one year for each point, with citing papers published in years $t-2$ to $t+2$ and citation lags (time between publication of citing and cited papers) less than or equal to 5 or 10 years.

To investigate whether the impact of distance and ties have been changing, we estimate regressions based on a moving sample window. We construct the estimation windows by first restricting the citing papers to be published within a 5-year period centered around year t. This implies citing years, t_c, in the interval $t+2 \geq t_c \geq t-2$. To make the sample

size in later years comparable to that of earlier years, we impose a fixed maximum citation lag L set equal to 5 or 10 years. This implies cited years, t_d in the interval $t_c \geq t_d \geq t_c - L$. The first mid-year t we use is 1990 and the last is 2007 (since our data set runs to 2009).

Figure 5 shows the effects of distance in panel (a) and ties in panel (b). In both panels we use blue squares to depict the point estimates for 10 year maximum citation lags. A solid LOWESS smoother passes through the point estimates. The dashed smoother line depicts the results for a 5-year citation lag and the dot-dash line corresponds to an estimation with no restriction on citation lag (all years). The points in panel (a) are estimated distance elasticities, that is, the marginal effect on the log odds of citation of increasing log distance between citing and cited authors. The time-pattern of distance effects depends on whether the regressions controls for ties or not. We depict these differing results using blue for estimates that control for the sum of 14 ties and red for those that do not. 95% confidence intervals (as before standard errors are clustered at the cited article level) are shaded blue and red for estimates that do and do not (respectively) control for the sum of ties.[30]

All the specifications plotted in Figure 5(a) show absolute distance elasticities becoming much smaller over time since the early 1990s. In the geography-only specification shown in red, distance remains a statistically significant impediment to citation up to and including the final interval, 2005–2009, when its elasticity is -0.06 (standard error: 0.013). However, the magnitude falls by two thirds from its 1990 value of -0.18. The confidence intervals also shrink over time, since increasing numbers of digitalized articles raise $\sqrt{N}$ in the standard error calculation. Controlling for the sum of ties, we see the absolute elasticities are uniformly smaller in all periods, with the largest gap between the smoother lines appearing in the last estimation windows. Starting around 2005, the confidence intervals mainly include zero. The final estimated distance elasticity controlling for ties is -0.017 (standard error: 0.012).

Panel (b) of Figure 5 shows the evolution of the coefficient on the sum of ties. The impact of ties on citation has been mainly rising over the 1990s and 2000s. The increase in citation odds from adding a tie rises from 72% in 1990 to 94% in 2007.[31]

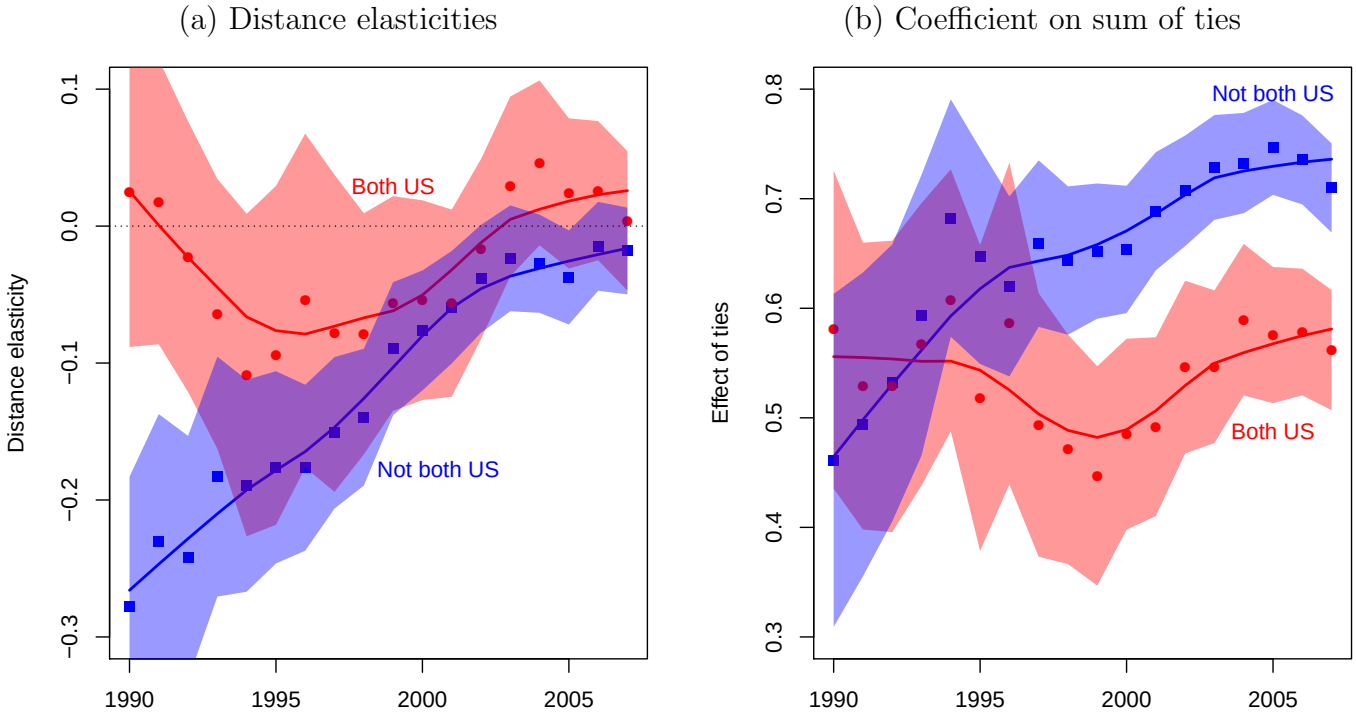
In all the results presented to this point we have used a world-wide sample. This contrasts with much of the work we cited in the introduction on the geography of knowledge flows that uses citations *within the United States*. It is therefore worth investigating whether the patterns shown in Figure 5 reflect global phenomena or whether the US is special.

Figure 6 graphs the results for a moving-window specification similar to that depicted in Figure 5 except that it estimates separate distance> 0, and (sum of) ties coefficients

[30] The purple area corresponds to the intersection of the two intervals.

[31] Exponentiate the 10-year lag coefficients shown in Figure 5 and subtract one to obtain these amounts.

Figure 6: Effects of distance and ties smaller within US



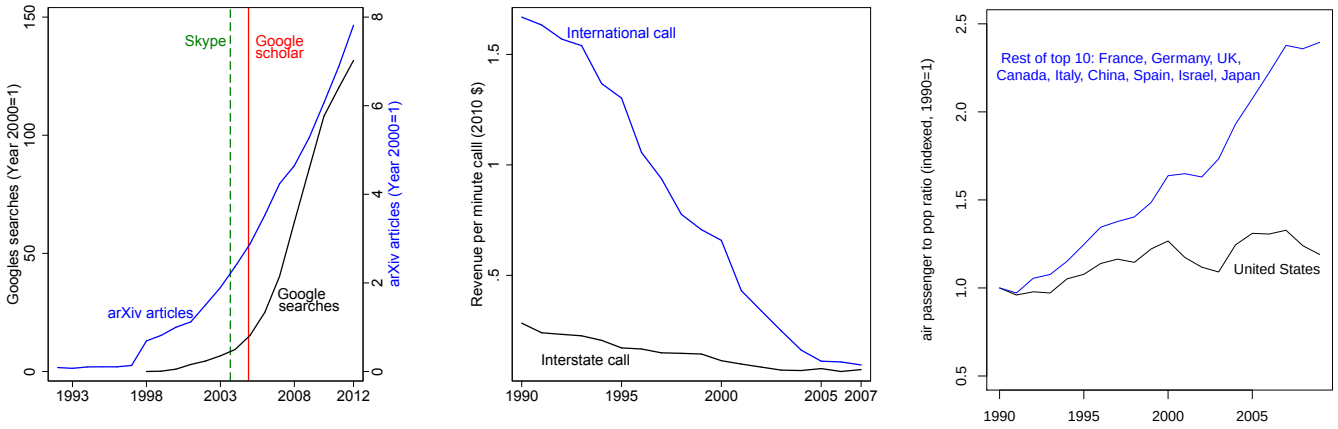
Note: Red-shaded interval corresponds to estimates where citing and cited authors reside in US. Blue-shaded estimates have at least one of citing or cited team from country other than US. Citation lags (time between publication of citing and cited papers) less than or equal to 10 years. Other details same as Figure 5.

for pairs where both i and d are US residents and others (1−both US_{id}). Panel (a) reveals that the shrinking distance effects depicted in Figure 5(a) derive from author pairs where at least one set is not US-based. Distance effects between US pairs have not been significantly different from zero throughout the period of study. Another difference between US pairs and others is that the latter exhibit rising effects of ties, becoming significantly larger than those between US pairs since the 2000s.

It is obviously tempting to try to explain the temporal patterns in the coefficients with reference to the technological advances we have observed since the 1990s. However, it is not possible to identify one cause or the other with so many trends at work during this period. Advances affecting information flows include, but are not limited to, the rise of web browsers in the mid 1990s, and the introduction of the Google search engine in 1998 and Google Scholar in 2004. Of particular importance to scientists was the creation of arXiv.org, a repository of pre-prints, which has included mathematics since 1992.

Figure 7(a) plots the growth of the number of arXiv papers in mathematics over time and compares (it on a second scale) with the spectacular increase in Google searches in the 2000s. Panel (a) also depicts the introductions of Skype and Google Scholar. The combination of all these technologies would be expected to have reduced the importance of face-to-face interactions, implying declining geographic separation effects since 1990.

Figure 7: Trends in internet use (arXiv, Google), communication costs, and air transport



(a) Internet use (b) Call rates (c) Air passengers

Note: (a) From `arXiv.org` we take annual counts of mathematics articles. Google searches from 2000 to 2012 come from `http://www.internetlivestats.com/google-search-statistics/`. The 1998 and 1999 numbers are extrapolated from reports from September of those years. Series indexed relative to 2000. (b) Average revenue per minute (FCC) (c) Ratio of air passengers carried over population, indexed relative to 1990 (WDI), Top 10 countries ranked by citing authors in 2009.

The smoother line for the distance elasticity in Figure 5(a) begins to trend up in the late 1990s, coinciding with the rise of arXiv shown in Figure 7(a). However, the stable importance of ties between US authors and the growing role of ties elsewhere shown in Figure 6(b) is not consistent with the view that arXiv and Google searches have been making all information universally accessible. Furthermore, the rise of internet article depositories and search engines cannot explain why distance effects between US author teams have been insignificantly different from zero during the whole period.

While internet advances capture the most attention, other contemporaneous changes could reasonably affect the importance of geography and ties in knowledge transmission. Figure 7(b) shows the dramatic decline in the costs of making international calls to and from the US.[32] In real terms international calls fell by 95% between 1990 and 2007, compared to a 75% decline over the same period for interstate calls. As we cannot find comparable series on domestic and international air fares, we use data on the volume of travel as a proxy. Figure 7(c) shows that air travel has been rising relative to the population size.[33] The rate of growth outside the US has been much larger, partly because the US started from a much higher base. In 1990 the US had 1.3 air passengers per capita compared to 0.11 for the nine other countries that comprise the top 10 countries in mathematics (measured by number of citing authors in 2009). Two decades later the US ratio had risen by 20% whereas the other countries rose by 140%.

The data shown in Figures 7(b) and (c) suggest an alternative interpretation of ad-

[32]Data from Federal Communications Commission (2010) Table 13.4, deflated by the CPI.
[33]Data from World Development Indicators series "Air transport, passengers carried."

vances since the 1990s. Perhaps cheaper phone calls and improved air travel makes it easier for scholars to stay in touch with their ties. Improved contact allows them to share the kind of complex knowledge that is hard to procure via Google searches. Thus communication cost reductions lower the need for face-to-face interactions but raise the opportunities for drawing upon one's ties.[34] Similarly lower costs for flying to conferences or visiting collaborators could also contribute to the explanation of why distance matters less, but ties matter more.

The greater drop in the effect of distance and the larger increase in the effect of ties for non-US based authors is in line with the big decline in call costs between the US and other countries shown in Figure 7(b) and the rise of air travellers per capita in the rest of countries relative to the US shown in Figure 7(c). While we find this story linking the coefficient patterns in Figures 5 and 6 to the trends in Figure 7 to be plausible, future work with different identification strategies would be needed to confirm it.

4.5 Subsamples and other robustness checks

This subsection reports the findings of additional robustness checks. The underlying tabular results are reported in Appendix Tables C.1–C.13. 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. Appendix Table C.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.4 in the Appendix adds an indicator for No Shared Association to the set of geographic barriers employed in Table 4. 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

[34]This story is consistent with the model of complementarity between proximity and communication technology in Gaspar and Glaeser (1998).

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 C.5 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.[35] All 51 coefficients in this table have the same sign as the corresponding coefficient in Table 4. Significance levels are the same for 47 coefficients. In general, an effect 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 C.6 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 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). 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

[35]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 4 is 18.5 times larger than the one in Table C.5.

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 C.6 are positive), the geography/distance interactions are negative when they differ significantly from zero.

Tables C.7 to C.12 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 5 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 C.7 and C.8 being less than a standard error from the values in Table 4. The magnitudes of some ties hardly change: advisor cited has a coefficient of 1.396 before 2005 and 1.349 afterwards.

The main finding of Table 5—that the estimated impact of ties falls with tighter controls for article subject until it rises slightly with keywords—is replicated in the before and after 2005 periods shown in Tables C.9 and C.10. The Table 6 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.

Table C.13 reports the results of four additional specifications designed to explore the robustness of our main results. The first specification is closely related to Figure 6. 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 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).[36] 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 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.

5 Conclusion

Our results add further evidence to the diverse strands of the literature finding geographic separation impedes knowledge flows. Geography matters in large part because of its role in shaping the personal ties between citing and cited scholars. In the full sample, including 14 linkages based on career and educational histories as controls cuts geography coefficients approximately in half. For the subsample where both citing and cited authors reside in the US, a region where communication and travel costs have long been relatively low, the marginal effect of greater distance between institutions is insignificantly different from zero. The distance effect also disappears in the most recent five years of the world-wide sample. These "zero" partial effects of distance are obtained only in those regressions that control for ties.

Despite the increase in global access to knowledge provided by the internet, the strength of the impact of ties on citation probabilities has not been declining. Because ties matter most for papers where awareness gaps are most acute (recent, obscure, and different-field articles), we infer that ties matter because connected scholars transmit

[36]The baseline coefficient for the average effect of ties comes from Table 6 column(1).

knowledge to each other. This view is further supported by the finding that scholars whose formal role is to impart knowledge (advisors and the academic parents and siblings of advisors) have larger impacts on subsequent academic generations than vice versa. In sum, the evidence suggests that "what you know" depends a great deal on "whom you know." It is increasingly unrelated to "where you work"—except insofar as where you work influences whom you know.

To the extent that we can generalize from the study of mathematicians, our study suggests novel interpretations of existing empirical findings. Cities may be valuable not just because of daily face-to-face interactions, but also because they are good places to build networks. Such a view points to a different takeaway from the De La Roca and Puga's (2017) finding that wages rise with experience in big cities but retain much of this growth even when the individual returns to a smaller city. While the authors attribute the wage premium to increased ability, our framework suggests it might also have arisen via an expanded set of professional ties. Since ties created at short distances can be maintained over longer distances, the ties explanation is also consistent with the continued prosperity of city leavers.

In trade, Feyrer (2009) estimates that changes in distance caused by the Suez canal closure have much lower impacts than cross-sectional differences in distance. Our interpretation would be that the lengthening of the shipping route has no impact on the ties between importer and exporter that predated the closure. The puzzle posed by Head and Mayer (2013) calculations that observable barriers such as tariffs and freight charges can only explain less than half the estimated magnitudes of border and distance effects has a simple resolution in light of our results. Traders depend on their networks and those networks are nationally and spatially biased.

It bears repeating that the broader lessons we draw from observing mathematicians are necessarily tentative; they beg for corroboration in other contexts. This is especially true when it comes to policy implications. However, a ties-centered view of knowledge flows does suggest certain types of government actions could be fruitful. To promote more geographically dispersed networks, universities could be strongly discouraged from hiring their own students straight out of graduate school. Another policy to broaden ties of researchers is for the government to fund and promote doctoral study abroad. Invitations to foreign faculty for short and long term visits often lead to the formation of new collaborative ties. Analogous versions of these policies can expand the networks for non-academics. For example, easing visa requirements to facilitate medium-run stays by employees of multinationals should thicken the set of connections between foreign and domestic knowledge creators.

References

Agrawal, A., Cockburn, I., and McHale, J. (2006). Gone but not forgotten: Knowledge flows, labor mobility, and enduring social relationships. *Journal of Economic Geography*, 6(5):571–591.

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

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

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.

Allen, T. (2014). Information frictions in trade. *Econometrica*, 82(6):2041–2083.

Althouse, B. M., West, J. D., Bergstrom, C. T., and Bergstrom, T. (2009). Differences in impact factor across fields and over time. *Journal of the American Society for Information Science and Technology*, 60(1):27–34.

Arrow, K. J. (1969). Classificatory Notes on the Production and Transmission of Technological Knowledge. *American Economic Review*, 59(2):29–35.

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

Borjas, G. J. and 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.

Breschi, S. and 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.

Combes, P.-P., Lafourcade, M., and Mayer, T. (2005). The trade-creating effects of business and social networks: Evidence from France. *Journal of International Economics*, 66(1):1–29.

Comin, D. A., Dmitriev, M., and Rossi-Hansberg, E. (2012). The spatial diffusion of technology. Working Paper 18534, National Bureau of Economic Research.

De La Roca, J. and Puga, D. (2017). Learning by working in big cities. *The Review of Economic Studies*, 84(1):106–142.

Ellison, G., Glaeser, E. L., and Kerr, W. R. (2010). What causes industry agglomeration? Evidence from coagglomeration patterns. *American Economic Review*, 100(3):1195–1213.

Federal Communications Commission (2010). Trends in telephone service. Statistical report, Wireline Competition Bureau, Washington, D.C.

Feyrer, J. (2009). Distance, trade, and income–the 1967 to 1975 closing of the Suez Canal as a natural experiment. Technical Report 15557, National Bureau of Economic Research.

Gaspar, J. and Glaeser, E. L. (1998). Information technology and the future of cities. *Journal of Urban Economics*, 43(1):136–156.

Glaeser, E. (2011). *Triumph of the city: How our greatest invention makes US richer, smarter, greener, healthier and happier*. Pan Macmillan.

Gould, D. M. (1994). Immigrant Links to the Home Country: Empirical Implications for U.S. Bilateral Trade Flows. *The Review of Economics and Statistics*, 76(2):302–316.

Griffith, R., Lee, S., and Van Reenen, J. (2011). Is distance dying at last? Falling home bias in fixed-effects models of patent citations. *Quantitative Economics*, 2(2):211–249.

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

Head, K. and Mayer, T. (2013). What separates us? Sources of resistance to globalization. *Canadian Journal of Economics*, 46(4):1196–1231.

Head, K. and Mayer, T. (2014). Gravity equations: Workhorse, toolkit, and cookbook. In Helpman, E., Gopinath, G., and Rogoff, K., editors, *Handbook of International Economics*, volume 4. Elsevier.

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.

Hopenhayn, H. A. (1992). Entry, Exit, and Firm Dynamics in Long Run Equilibrium. *Econometrica*, 60(5):1127–50.

Iaria, A., Schwarz, C., and Waldinger, F. (2018). Frontier knowledge and scientific production: Evidence from the collapse of international science. *The Quarterly Journal of Economics (forthcoming)*.

Jaffe, A. B., Trajtenberg, M., and Henderson, R. (1993). Geographic localization of knowledge spillovers as evidenced by patent citations. *The Quarterly Journal of Economics*, 108(3):577–98.

Keller, W. (2001). The Geography and Channels of Diffusion at the World's Technology Frontier. NBER Working Papers 8150, National Bureau of Economic Research, Inc.

Keller, W. (2002). Geographic localization of international technology diffusion. *American Economic Review*, 92(1):120–142.

Keller, W. and Yeaple, S. R. (2013). The Gravity of Knowledge. *American Economic Review*, 103(4):1414–44.

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

Kerr, W. R. and Kominers, S. D. (2015). Agglomerative Forces and Cluster Shapes. *The Review of Economics and Statistics*, 97(4):877–899.

Kerr, W. R. and Mandorff, M. (2015). Social Networks, Ethnicity, and Entrepreneurship. NBER Working Papers 21597, National Bureau of Economic Research, Inc.

Li, Y. A. (2014). Borders and distance in knowledge spillovers: Dying over time or dying with age? Evidence from patent citation. *European Economic Review*, 71:152–172.

Lissoni, F. (2001). Knowledge codification and the geography of innovation: The case of Brescia mechanical cluster. *Research Policy*, 30(9):1479–1500.

Melitz, M. J. (2003). The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity. *Econometrica*, 71(6):1695–1725.

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 coauthorship networks. In Ben-Naim, E., Frauenfelder, H., and Toroczkai, Z., editors, *Complex Networks*, volume 650 of *Lecture Notes in Physics*, pages 337–370. Springer.

Packalen, M. and Bhattacharya, J. (2015). Cities and ideas. Technical report, National Bureau of Economic Research.

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

Rauch, J. E. (2001). Business and Social Networks in International Trade. *Journal of Economic Literature*, 39(4):1177–1203.

Rauch, J. E. and Trindade, V. (2002). Ethnic Chinese Networks In International Trade. *The Review of Economics and Statistics*, 84(1):116–130.

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. (2006). Patent citations and the geography of knowledge spillovers: Evidence from inventor- and examiner-added citations. *The Review of Economics and Statistics*, 88(2):383–388.

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

Waldinger, F. (2010). Quality matters: The expulsion of professors and the consequences for PhD student outcomes in Nazi Germany. *Journal of Political Economy*, 118(4):787–831.

Online Appendix (Not for Publication)

A Data construction

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 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.2 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.

B 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.[37]

For each annual meeting, we extract the information contained in the full program web page.[38] 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.

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 B.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.

Second, we analyze whether coinciding at a conference raises the likelihood of citation.

[37] 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.

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

Table B.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.

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 B.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 B.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 B.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.

Table B.3: Baseline regression with conference variables

	(1) Logit	(2) LPM
Coincided conference	-0.032	-0.003
	(0.065)	(0.004)
Coincided conference+session	$0.423^{\star}$	$0.032^{\dagger}$
	(0.151)	(0.013)
Coincided conference cited paper presented	$2.122^{\dagger}$	$0.205^{\dagger}$
	(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.
$^{\sim}p < 0.1, ^{\dagger}p < 0.05, ^{\star}p < 0.01$

C Supplementary Tables

Table C.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%.

Table C.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.

Table C.3: 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.

Table C.4: 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 | 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~	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%, ~: 10%.

Table C.5: 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 C.6: 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.

Table C.7: 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%, ~: 10%.

Table C.8: 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%, ~: 10%.

Table C.9: Sensitivity of results to alternative controls for article relevance before 2005

	(1)	(2)	(3)	(4)	(5)	(6)
Control group:	nil	journal	MSC-3d		MSC-5d	keyword
Panel A: including ties						
Distance > 0	-0.715*	-0.645*	-0.438*	-0.472*	-0.275†	-1.062*
	(0.078)	(0.075)	(0.100)	(0.101)	(0.135)	(0.316)
ln Dist $\mid$ Dist > 0	-0.059*	-0.045*	-0.060*	-0.057*	-0.066*	-0.013
	(0.008)	(0.009)	(0.011)	(0.011)	(0.015)	(0.033)
Different country	-0.013	-0.004	-0.030	-0.041	0.029	-0.062
	(0.034)	(0.034)	(0.043)	(0.044)	(0.060)	(0.133)
Different language	0.026	0.053$^{\sim}$	-0.002	0.015	0.040	-0.214†
	(0.029)	(0.029)	(0.036)	(0.036)	(0.052)	(0.100)
Average effect of ties	1.646*	1.179*	0.638*	0.620*	0.457*	0.692*
	(0.028)	(0.025)	(0.027)	(0.027)	(0.035)	(0.107)
Cocitation				3.325*	2.180*	2.986*
				(0.066)	(0.093)	(0.502)
Observations	727976	613686	177000	177000	32129	5414
pseudo-R^2	0.178	0.142	0.091	0.151	0.112	0.145
Panel B: excluding ties						
Distance > 0	-1.757*	-1.567*	-1.087*	-1.117*	-0.699*	-1.298*
	(0.062)	(0.063)	(0.089)	(0.090)	(0.125)	(0.275)
ln Dist $\mid$ Dist > 0	-0.091*	-0.077*	-0.089*	-0.086*	-0.094*	-0.070†
	(0.008)	(0.008)	(0.010)	(0.010)	(0.014)	(0.031)
Different country	-0.206*	-0.165*	-0.166*	-0.172*	-0.057	-0.215$^{\sim}$
	(0.032)	(0.033)	(0.042)	(0.043)	(0.059)	(0.127)
Different language	-0.078*	-0.022	-0.057	-0.036	-0.008	-0.198†
	(0.028)	(0.029)	(0.036)	(0.036)	(0.051)	(0.097)
Cocitation				3.371*	2.207*	2.762*
				(0.063)	(0.091)	(0.459)
Observations	727976	613686	177000	177000	32129	5414
pseudo-R^2	0.043	0.036	0.033	0.099	0.070	0.075

Notes: Average effect of ties refer to the mean effect of 13 (3 WOS and 10 MGP) ties. Significance: *: 1%, †: 5%, $^{\sim}$: 10%. Robust standard errors clustered by cited article in parentheses.

Table C.10: Sensitivity of results to alternative controls for article relevance after 2005

Control group:	(1) nil	(2) journal	(3) MSC-3d	(4) MSC-3d	(5) MSC-5d	(6) keyword
Panel A: including ties						
Distance > 0	-1.031*	-0.978*	-0.705*	-0.700*	-0.433*	-0.348$^{\sim}$
	(0.088)	(0.085)	(0.092)	(0.092)	(0.117)	(0.189)
ln Dist $\mid$ Dist > 0	-0.024†	-0.009	-0.011	-0.010	-0.007	-0.059*
	(0.009)	(0.009)	(0.010)	(0.010)	(0.013)	(0.020)
Different country	-0.075$^{\sim}$	-0.098†	-0.152*	-0.155*	-0.183*	-0.106
	(0.039)	(0.039)	(0.042)	(0.042)	(0.054)	(0.079)
Different language	-0.073†	-0.015	-0.048	-0.053	-0.019	-0.094
	(0.032)	(0.032)	(0.034)	(0.034)	(0.044)	(0.062)
Average effect of ties	1.620*	1.151*	0.666*	0.658*	0.475*	0.573*
	(0.032)	(0.025)	(0.022)	(0.022)	(0.028)	(0.049)
Cocitation				3.144*	2.077*	1.177*
				(0.105)	(0.129)	(0.241)
Observations	487310	522139	264792	264792	43797	17266
pseudo-R^2	0.187	0.146	0.092	0.107	0.087	0.107
Panel B: excluding ties						
Distance > 0	-1.980*	-1.800*	-1.395*	-1.383*	-1.082*	-0.954*
	(0.072)	(0.071)	(0.082)	(0.082)	(0.106)	(0.162)
ln Dist $\mid$ Dist > 0	-0.068*	-0.049*	-0.046*	-0.046*	-0.029†	-0.084*
	(0.009)	(0.009)	(0.010)	(0.010)	(0.013)	(0.019)
Different country	-0.294*	-0.290*	-0.299*	-0.299*	-0.303*	-0.283*
	(0.036)	(0.037)	(0.041)	(0.041)	(0.053)	(0.075)
Different language	-0.167*	-0.093*	-0.108*	-0.111*	-0.064	-0.143†
	(0.031)	(0.031)	(0.034)	(0.034)	(0.043)	(0.060)
Cocitation				3.253*	2.186*	1.211*
				(0.101)	(0.125)	(0.238)
Observations	487310	522139	264792	264792	43797	17266
pseudo-R^2	0.049	0.040	0.033	0.051	0.043	0.051

Notes: Average effect of ties refer to the mean effect of 13 (3 WOS and 10 MGP) ties. Significance: *: 1%, †: 5%, $^{\sim}$: 10%. Robust standard errors clustered by cited article in parentheses.

Table C.11: 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.

Table C.12: 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.

Table C.13: Robustness

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.