HKUST IEMS
INSTITUTE FOR EMERGING MARKET STUDIES

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

Keith HEAD, Yao Amber LI, Asier MINONDO

HKUST IEMS Working Paper No. 2015-30

September 2015

HKUST Institute for Emering Market Studies
The Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong
T: +852 3469 2215 E: iems@ust.hk W: iems.ust.hk

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

Keith HEAD, Yao Amber LI, Asier MINONDO

HKUST IEMS Working Paper No. 2015-30
September 2015

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 a 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.

Authors' contact information

Keith Head
Sander School of Business
University of British Columbia
E: keith.head@sauder.ubc.ca
Click here to enter text.

Yao Amber Li
Department of Economics
Hong Kong University of Science and Technology
E: yaoli@ust.hk

Asier Minondo
Deusto Business School
University of Deusto
E: aminondo@deusto.es

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

Keith Head† Yao Amber Li‡ Asier Minondo§

September 11, 2015

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: network, distance, border, geography, knowledge flows, academic citations, genealogy, matching

JEL: O3, F1, 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 zb-math 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 ECO2013-46980-P, co-financed with FEDER) and the Basque Government Department of Education, Language policy and Culture (IT629-13). 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,Tom Ross, and Daniel Sturm for valuable discussions. We thank Andrei Levchenko in particular for the questions that lead 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, Instituto Complutense de Estudios Internacionales. aminondo@deusto.es

1 Introduction

An improved understanding of the role of geography as an impediment to knowledge diffusion would point towards a unified answer to three major economic questions. First, why do countries differ such much in their productivity given that knowledge is generally considered non-rival and non-excludable? Second, why do borders and distance still have such large negative impacts on bilateral trade, given that conventional barriers such as tariffs and freight are now such small shares of the value of goods? Third, why is the share of people living in cities still rising despite major congestion costs and revolutions in long distance communication?[1]

Following the seminal work of Jaffe et al. (1993) showing localization of patent citations, there has been mounting evidence that proximity facilitates knowledge flows. Recent studies estimating negative distance effects on citation propensities include Peri (2005), Belenzon and Schankerman (2013), and Li (2014). No consensus has emerged on why distance and borders matter, a puzzle given that information is weightless and tariff-free. Keller (2004) points to the problem of tacit knowledge but it is hard to define or measure this concept without being circular; i.e. tacit knowledge is the knowledge that can only be transferred through face-to-face interactions. Furthermore, one case study found that much of the information flows within a cluster involved highly codified messages.[2]

In this paper we examine a different mechanism underlying geography's impact on knowledge. We hypothesize that proximity facilitates formation of interpersonal ties and these ties foster knowledge flows. This might be due to trust or reciprocity or some other mechanism. Our key finding is that adding controls for a rich set of career and educational linkages between authors of mathematics papers, leads to a halving of estimated geography effects. The role of distance—after controlling for ties—even becomes statistically insignificant in recent years.

Since citations play the role of "paper trails" measuring information flows, they have been the focus of a large body of research attempting to identify factors that limit transfers of knowledge between inventors. Most of this literature has studied *patent* citations, both because of data availability and the obvious interest in diffusion of innovations. Unfortunately, patent applications provide little information on the identity of inventors and their interrelationships. Past collaboration can be determined and Singh (2005) and Breschi and Lissoni (2009) find this type of tie increases citation. Agrawal et al. (2006)

[1] Keller (2002) shows that research spillovers on productivity are decreasing in distance, Head and Mayer (2013) show tariffs and transport costs can explain less than half of estimated border and distance effects in trade, and Gaspar and Glaeser (1998) find that modern telecommunication technologies are not strong substitutes for the face-to-face interactions available in cities.

[2] Lissoni (2001) examined a cluster of mechanical firms in Brescia, Italy and found they engaged in transfer of CAD encoded designs.

investigate a second tie, past colocation. They find that inventors who move institutions are still disproportionately cited by their former colleagues.

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). Sharing surnames with the same ethnic origin need not reflect a direct or even indirect personal connection between the citing and cited inventors.[3] Co-ethnicity also likely captures cultural similarities.

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. Such ties 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. Thus, unlike colocation and collaboration, educational linkages do not change over time in response to shocks to the interests of citing authors. This paper focuses on citations between mathematicians, whose Ph.D. institutions and advisors have been carefully compiled by the Mathematics Genealogy Project (MGP).[4] A second advantage of academic citations in studying knowledge flows is that there appears to be less reluctance to cite influential work than is the case with patent citations. Alcácer and Gittelman (2006) report that "approximately 40% of citing patents have all citations imposed by examiners." It would be unthinkable for an academic paper to contain no references except those imposed by referees and editors.

In addition to the strength of its academic genealogy data, mathematics offers two additional advantages relative to other academic fields. First, almost by definition, 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.

One of the novel aspects of our paper is the inclusion of asymmetric linkages. As knowledge flows are directional, it is useful to determine whether citations are stronger to the presumed sources of information. We find evidence that this is the case. The odds

[3]There are 92 million Wangs, the vast majority of which do not know each other.

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

of citation are seven times higher if a paper is written by the advisor of the citing paper. 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 2.7 times the normal odds of being cited, but there are no significant differences in their propensity to cite their advisees' advisees.

A major challenge faced by the literature investigating effects of geography, and especially ties, on citation is the difficulty of controlling adequately for the relevance of potentially cited papers. We contribute two methodological advances in tackling this difficult issue. First, we employ a novel set of fixed effects capturing the *triad* of cited paper, citing year and citing subject. This controls in a very general way for proximity in subject matter between citing and cited papers, as well as proximity in publication dates. Second, we show how different controls for subject similarity affect the estimates on the variables of interest. In particular we find that inadequate subject controls lead to large overestimates of linkage effects, but also that ties survive as significant forces in determining cites no matter how detailed we go in measuring similarity.

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 addition to the findings we have already mentioned, three empirical relationships are worth highlighting. First, a parsimonious power law governing distance decay fits the data almost as well as a 12-step non-parametric specification. Second, ties matter more for obscure, recent, and different-field papers—just as most information diffusion mechanisms would predict. Third, the average effect of a tie on the tendency to cite has not declined over the last three decades, even as distance effects have shrunk to insignificance. The final section returns to the questions that opened the paper, drawing out the implications of our results.

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

To ease computation, some earlier papers such as Belenzon and Schankerman (2013) have estimated linear probability models (LPM). While we will report one LPM to show that it yields similar signs, relative magnitudes and significance levels, we see logit as the superior estimator since it constrains predicted citation probabilities to be non-negative. Logit coefficients provide the marginal effect on changes in the log odds. Some audiences may find the marginal effects on the probabilities themselves to be more intuitive. The problem we see is that in the context of rare events, marginal probabilities can be tiny. Singh (2005) multiplies his marginal effects by one million for reporting purposes. We find odds ratios are more useful, 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. As we will discuss later, logit also has better properties than LPM under choice-based sampling.

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 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. 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 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 used them to obtain 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 combines four main sources:

1. Thomson Reuters' ISI Web of Science (WOS): 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 than 1975. With these limitations, the database contains information about 339,613 articles.[5] 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 Annex 2 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.

[5] Annex 1 lists the journals included in the database, the number of articles per journal and the earliest article of the journal included in the database.

- 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 the university and year in which each degree recipient completed their Ph.D., and 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 nine 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.

- 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 Oxford alumnus cites a professor currently affiliated with Oxford.

Mathematics subject classification data

We used Zentralblatt MATH (zbMATH) to obtain the Mathematics Subject Classification (MSC) for the articles in our sample.[6] 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,

[6]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.

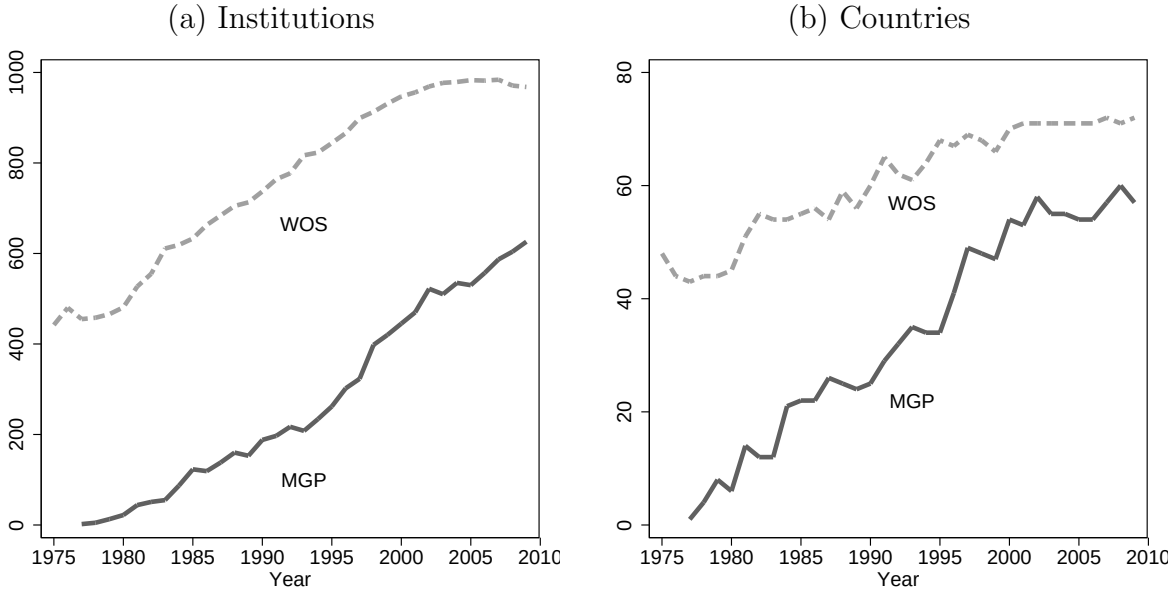
Using publications to track author locations over time, we calculate distances (and other measures of geographic separation) at the time the *citing* article is written. Past work using patents calculated distances between inventors using the cited inventors' addresses in the year the *cited* patent 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 Annex 2, 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

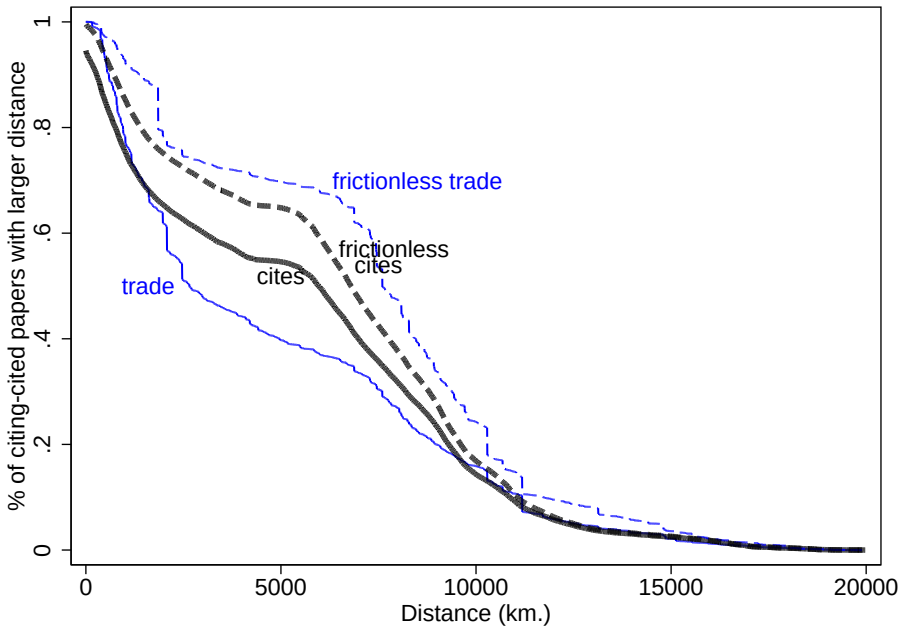


Meanwhile, the numbers of institutions and countries represented in the WOS citation data increase over time. Figure 1 shows that during the period 1975–2009 the set of institutions with citing or cited author affiliations rises to nearly 1000 (some institutions disappear) and the corresponding number of countries rises to 72. The sample containing MGP information on all authors starts very small but eventually represents 626 institutions located in 57 countries. This number of countries in our analysis is unprecedented in the citations literature, which has mainly focused on cross-metropolitan area citations within the United States.[7]

Before estimating any regressions, it is useful to see whether geographic impediments to knowledge flows can be seen in a fully non-parametric setting. Figure 2 provides a sketch of the geography of citation patterns in mathematics. For each distance, D, it depicts the fraction of citations that occur over distances greater than D. That is, it depicts the complementary cumulative distribution function (also known as the survival function) of citation distance. In order to be meaningful, the distribution should be contrasted with a benchmark. We construct such a benchmark by considering the

[7] Peri (2005) is one of the few studies of international citation data but the challenge of determining locations for individual patentors resulted in limiting the sample to 18 countries.

Figure 2: Distribution of citation (1975–2009) and trade (2000) distances



aggregate inward cites received by an institution and then assigning them to citing institutions based on their shares of aggregate outward cites. This is similar to the idea of frictionless gravity employed in Head and Mayer (2013) and for comparison we include the true and frictionless benchmarks for goods trade in the same figure. What we learn is that distance attenuates knowledge flows in mathematics leading them to occur over shorter ranges than one would expect in a frictionless world. The gap is 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.

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 1 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 ISI Web of Knowledge only identifies the first author of the cited articles. To identify the institutional affiliation of the first author, and the identity and affiliations of any coauthors, we matched the cited articles with our original database providing more complete information on the citing authors. As our database is restricted to the 255 journals included in Mathematics category, we can only identify the authors and coauthors of the cited articles belonging to this set. Only one third of the cited papers

Table 1: 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	221,942	162,483	1,044,952
Full affiliation data	187,114	133,465	749,823
Excluding self-citations	168,112	108,238	562,433
Authors at top 1000 inst.	137,887	92,993	492,813
With 5-digit MSC field	79,865	77,725	314,196
MGP data all authors	14,642	14,187	33,327

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

(containing about half the citations) were published in the pure math journals included in our database.[8] 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 since it is hard to interpret them as flows of information, especially for single-authored papers. This subtracts a surprisingly high one quarter of the realized citations.

There are 11,764 different affiliations for the citing authors and 7,750 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 88% 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 39%.

Applying the filters described above leaves us with 314 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 1 raises concerns that the new sample might not be representative. We shall show in Table 2 that the MGP sample is 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

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

"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 paper 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 4 billion potential cites and about 315,000 realized cites. Thus, the rate of citation is only 8 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.[9]

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. 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 2 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 (8% chance vs 2% chance in the first row of columns (1) and (3)), 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.[10] In terms of ties, citing authors are many times more likely to coauthor with the (realized) cited authors. They are also more than twice as likely to have worked at the same university either at the same or different times. Since all these variables are correlated we will need to estimate regressions to determine the partial relationships.

A comparison of columns (1) and (2) of Table 2 shows that imposing the criteria that all citing and cited authors have MGP data leaves a much smaller sample of realized citations but the average characteristics of the MGP and non-MGP samples are very

[9] Singh (2005) is an exception in that he uses five controls per realized citation. Also he uses the weighted exogenous sampling maximum-likelihood (WESML) estimator suggested by Manski and Lerman (1977).

[10] The calculation is $\exp(6.998 - 7.754) = 0.47$ for MGP authors and $\exp(7.111 - 7.815) = 0.49$ for non-MGP.

Table 2: Non-MGP vs. MGP Sample

	Only realized citations		only control citations	
mean of variables	non-MGP (1)	MGP (2)	non-MGP (3)	MGP (4)
Different institution (Distance > 0)	0.923	0.915	0.984	0.987
ln Distance | Distance > 0	7.111	6.998	7.815	7.754
Different country	0.645	0.637	0.759	0.764
Different language	0.513	0.484	0.615	0.590
Co-authors	0.101	0.089	0.018	0.014
Coincided past	0.078	0.082	0.027	0.025
Worked same place	0.059	0.056	0.031	0.032
Observations	314196	33327	314196	487791

similar. There is no clear evidence that MGP authors are more connected than non-MGP authors for those measures of ties we have data on for both samples. Comparing columns (3) and (4)—the control observations—the similarity between MGP and non-MGP authors persists. The number of observations in column (4) is considerably higher than column (3) because the latter only contains one control per case in column (1) whereas there are on average 15 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. Except where noted, all regressions are logits with triad fixed effects.[11] We report coefficients (not marginal effects or incidence ratios). 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, 10 of the 12 measures of ties have positive and significant impacts. On average the effect of adding a network tie more than doubles the odds of citation. 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

[11] They are estimated with Stata's clogit command using citing year, citing subject and cited article triads as the "group" variable.

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 3 reports the result of baseline regressions. 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. As a proxy for (not being) current colleagues, it could also have an interpretation as a measure of professional ties. Nevertheless, we will group it with the geography variables.

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. For this reduced set of realized citations, we add nine additional ties based on educational histories of the citing and cited authors. 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 (column 3) and then with ties (column 4).

Specification (1) presents significantly negative coefficients on distance and borders, suggesting that physical distance and borders indeed impede knowledge flows. 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 -0.205, whereas the baseline estimate of Peri (2005) is -0.19. Thus, for both math and patent citations, national borders lower citations by about 20%. Our column (1) estimate of the distance elasticity on math citation is -0.07, just outside the range of -0.03 to -0.067 that Li (2014) reports in regressions with a full set of fixed effects.

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

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

	(1)	(2)	(3)	(4)	(5)	(6)
Specification:	Triad-fixed-effects logit (TFE-Λ)				TFE-LPM	TFE-Λ
Sample	WOS	WOS	MGP	MGP	MGP	MGP
Geography:						
Distance > 0	-1.045⋆	-1.011⋆	-1.292⋆	-0.696⋆	-0.105⋆	
	(0.028)	(0.029)	(0.061)	(0.067)	(0.007)	
ln Dist | Dist > 0	-0.071⋆	-0.049⋆	-0.065⋆	-0.036⋆	-0.002⋆	Figure 3
	(0.003)	(0.003)	(0.007)	(0.007)	(0.000)	
Different country	-0.205⋆	-0.149⋆	-0.238⋆	-0.105⋆	-0.005⋆	-0.124⋆
	(0.013)	(0.013)	(0.030)	(0.030)	(0.002)	(0.031)
Different language	-0.104⋆	-0.060⋆	-0.087⋆	-0.028	-0.002	-0.030
	(0.011)	(0.011)	(0.025)	(0.025)	(0.001)	(0.026)
Ties:						
Co-authors		1.668⋆		1.522⋆	0.170⋆	1.527⋆
		(0.020)		(0.048)	(0.007)	(0.048)
Coincided past		0.682⋆		0.423⋆	0.028⋆	0.422⋆
		(0.018)		(0.040)	(0.004)	(0.040)
Worked same place		0.420⋆		0.355⋆	0.019⋆	0.350⋆
		(0.018)		(0.040)	(0.003)	(0.040)
Share Ph.D. (5 years)				0.517⋆	0.069⋆	0.512⋆
				(0.063)	(0.008)	(0.063)
PhD siblings				0.707⋆	0.112⋆	0.710⋆
				(0.095)	(0.010)	(0.095)
PhD cousins				0.398⋆	0.025⋆	0.395⋆
				(0.078)	(0.006)	(0.078)
Advisor citing				1.419⋆	0.234⋆	1.418⋆
				(0.120)	(0.020)	(0.120)
Advisor cited				1.946⋆	0.315⋆	1.946⋆
				(0.069)	(0.011)	(0.069)
Academic grandparent citing				-0.201	-0.038	-0.179
				(0.384)	(0.049)	(0.383)
Academic grandparent cited				1.004⋆	0.109⋆	0.997⋆
				(0.149)	(0.019)	(0.149)
Alma Mater citing				-0.060	-0.005	-0.067
				(0.054)	(0.005)	(0.054)
Alma Mater cited				0.155⋆	0.010†	0.154⋆
				(0.052)	(0.005)	(0.053)
Observations	628392	628392	521118	521118	521118	521118
AIC	454535	435243	153801	145058	n/a	145023

Robust standard errors clustered by cited article in parentheses. Significance: ⋆: 1%, ⋆: 5%, †: 10%

associations with citation. As exponentiating the coefficients in a logit expresses the effects in terms of the change in citation odds ratios the 0.682 coefficient on past colocation implies that even after colleagues have moved to separate institutions, they have 98% higher odds of citing each other ($\exp(0.682) - 1 = 98\%$). 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. This is despite losing almost 90% of the realized citations (see Table 1, bottom two rows). This finding offers further assurance that the MGP sample restriction does not induce selection bias.

Comparing columns (3) and (4) we see the headline result of this paper: Controlling for ties substantially attenuates the negative effects of geographic separation. On average, the four geography coefficients in column (4) are less than half the corresponding coefficients of column (3). 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 3, column (4) shows that, with just two exceptions, network connections have significant positive effects on citation probability. The average over all twelve ties coefficients is 0.682, implying that the average tie nearly doubles the odds of citation.

Systematic asymmetries between the impacts of different types of ties appear in Table 3.[12] Advisees massively over-cite their advisor's papers: if any of the authors of article d is advisor of any of the authors of article i, the odds that article i cites article d increase by a factor of seven (the largest impact of the 12 types of ties). In the reverse direction, we find advisors over-cite their advisees' articles by a factor of four. An even more pronounced asymmetry emerges when we skip a generation. Authors over-cite their advisor's advisors (academic grandparents) by a factor of 2.7. Yet this intergenerational flow is not reciprocal; the grandparents are neutral with respect to papers written by the advisees of their advisees. The same directional asymmetry appears for Alma Mater relations. The graduate of a school is significantly more likely to cite professors at that school but not vice-versa. All these asymmetries point in the same direction—young scholars disproportionately cite previous generations but the reverse is generally not true. This suggests information mainly flows from older to younger mathematicians.

Specification (5) re-estimates with the same set of covariates as (4) but using the linear probability model, i.e. an OLS regression of triad-demeaned citation on triad-demeaned explanatory variables. The LPM is usually a reliable estimator of the average

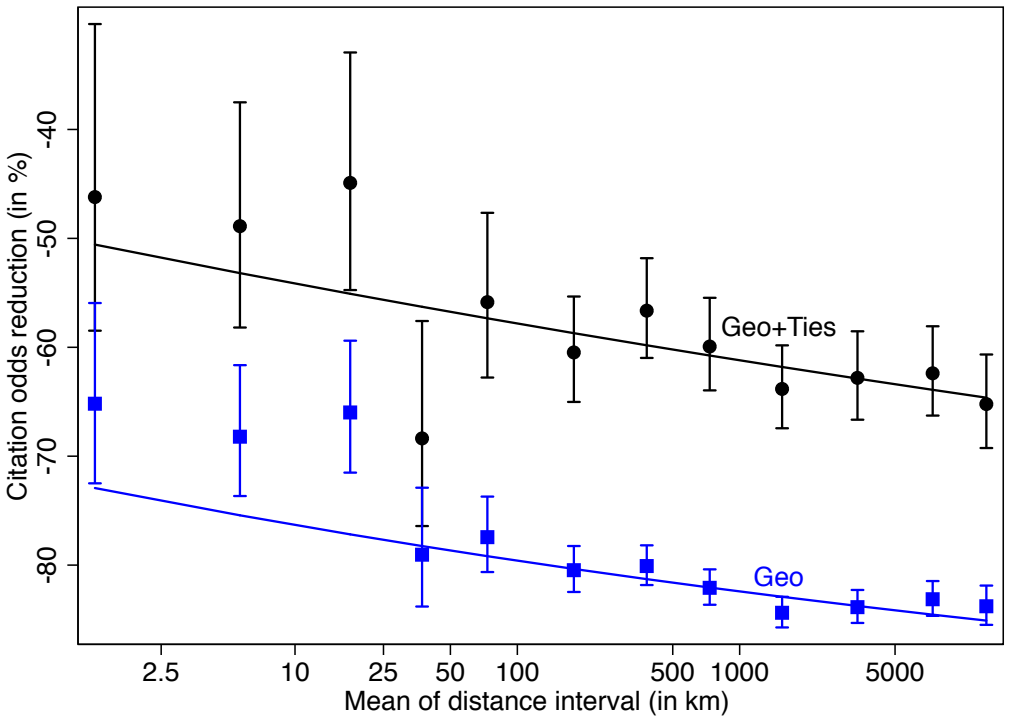
[12]For ease of expression, we will discuss these effects as if they were entirely causal. Later we will discuss concerns over omitted variable biases that might inflate these magnitudes.

marginal effects (AME) of x_i on the probability of a positive outcome, p_i: $b_{\text{ols}} \approx \text{AME} \equiv (1/N)\sum_i \partial p_i/\partial x_i$. With logit, $p_i = (1+\exp[-\beta x_i])^{-1}$, which implies

$$b_{\text{ols}} \approx (1/N)\sum_i p_i(1-p_i)\beta \approx \bar{p}(1-\bar{p})\beta$$

In the extended sample $\bar{p}(1-\bar{p}) = 0.06$. This explains why the LPM coefficients are so much smaller than the corresponding logit coefficients from column (4). The LPM to logit ratio of the log distance coefficients is 0.055, very close to what the formula predicts.[13]

Figure 3: Non-parametric estimated geography effects



The first five specifications employ a parsimonious parametrization of distance effects. This 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 estimate specification (6) replacing the two-part distance formulation with a 12-step approach. Comparing the coefficients on ties in specifications (4) and (6), we find only small, uninteresting differences, suggesting that the 2-part approach may be

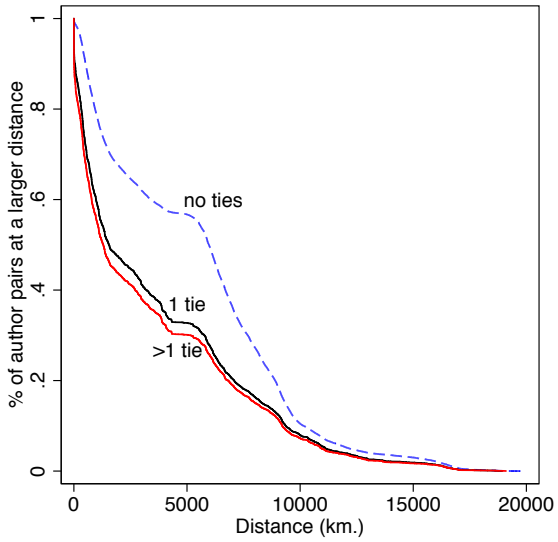
[13] A problem would arise if we applied the LPM to the case-based samples of columns (1) and (2). With just one control per case, $\bar{p} = 0.5$ and $\bar{p}(1-\bar{p}) = 0.25$. As mentioned before the true $\bar{p}$ is about eight per hundred thousand, implying $\bar{p}(1-\bar{p})$ is about the same, leading LPM to overstate the AME by a factor of 3000.

sufficient to capture distance effects.

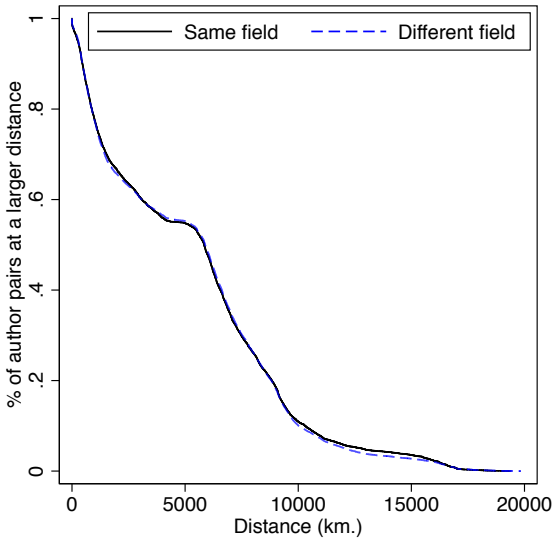
Figure 3 illustrates the coefficients on each of the 12 steps in the non-parametric estimation of distance effects conducted in specification (6) of Table 3, depicted with black circles. The vertical axis depicts the percentage reduction in the odds of citation (relative to a distance of zero) associated with each step.[14] 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 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. The two-part prediction lies within the confidence interval for 11 out of 12 steps, *after* controlling for ties. 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 -0.036. Failure to control for ties (as in the blue step function) leads to greater divergence between the two-part and 12-step specifications.

Figure 4: Ties, but not specializations, agglomerate

(a) Geography of ties



(b) Geography of specializations



Why does controlling for academic linkages lead to such striking changes in the size and shape of distance effects? The answer must be that ties are geographically biased.

[14] As with the odds effects reported above for ties, this is obtained by exponentiating the coefficients and subtracting one.

We illustrate this effect 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.

On the other hand, Figure 4 (b) depicts distance distributions for authors in the same field versus authors in different fields that are essentially the same. This suggests the absence of agglomeration by specialty. To some extent, dispersion across broad subjects is expected given that most math departments try to have reasonable coverage of the various areas of mathematics. But it is surprising that we do not see more concentration given that there are over 400 3-digit fields.

4.2 Alternative controls for relevance

Our maintained hypothesis is that networks spread ideas by facilitating information flows. An alternative hypothesis is that our network indicators are just proxies for author-pairs who have common research interests. In that case it would be lack of relevance rather than lack of awareness that impedes citation. X cites Y instead of C (control) not because of the network connection between X and Y, but because Y's results pertain more closely to X's work than those of C. In this story, X and Y establish linkages as the *consequence* of X and Y researching similar subjects.

Our triadic fixed effects methodology is intended to neutralize the relevance story to focus on the awareness channel. Table 4 shows how the results vary as we use more stringent criteria for the subject component of the fixed effect (and corresponding set of control observations). The purpose is to see whether the effects of geography and ties are stable or if one or the other deteriorates when cases are matched to more similar controls. To trim down the number of effects to be compared across specifications, we average the coefficients of all twelve tie indicators.[15] For each specification we also estimate a version that removes the ties indicators. This shows whether differences in the way we control subject can change our main result that controlling for linkages roughly halves the impacts of the geography variables. 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 52,744. 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 (4) from the previous table.

[15]Tests strongly reject the constraint that all networks enter with the same coefficient, which is why we do not estimate the model based on average or summed linkages.

Table 4: Sensitivity of results to alternative controls for article relevance

Control group:	(1) nil	(2) journal	(3) MSC-3d	(4) MSC-3d	(5) MSC-5d	(6) keyword
Panel A: including ties						
Distance > 0	-0.954*	-0.863*	-0.696*	-0.712*	-0.486*	-0.437*
	(0.058)	(0.055)	(0.067)	(0.068)	(0.086)	(0.154)
ln Dist $\mid$ Dist > 0	-0.038*	-0.031*	-0.036*	-0.033*	-0.029*	-0.057*
	(0.006)	(0.006)	(0.007)	(0.007)	(0.010)	(0.016)
Different country	-0.046†	-0.049†	-0.105*	-0.115*	-0.079*	-0.141*
	(0.026)	(0.026)	(0.030)	(0.030)	(0.039)	(0.064)
Different language	-0.031	0.016	-0.028	-0.020	-0.010	-0.089†
	(0.022)	(0.022)	(0.025)	(0.025)	(0.033)	(0.050)
Average effect of ties	1.989*	1.336*	0.682*	0.666*	0.450*	0.533*
	(0.056)	(0.041)	(0.037)	(0.035)	(0.040)	(0.075)
Cocitation				3.291*	2.139*	1.724*
				(0.054)	(0.071)	(0.186)
Observations	1346588	1345177	521118	521118	86936	26606
AIC	294915	260768	145058	139201	48811	15637
Panel B: excluding ties						
Distance > 0	-1.915*	-1.704*	-1.292*	-1.306*	-0.994*	-1.086*
	(0.048)	(0.048)	(0.061)	(0.061)	(0.080)	(0.135)
ln Dist $\mid$ Dist > 0	-0.075*	-0.063*	-0.065*	-0.062*	-0.052*	-0.077*
	(0.006)	(0.006)	(0.007)	(0.007)	(0.009)	(0.016)
Different country	-0.238*	-0.215*	-0.238*	-0.243*	-0.180*	-0.314*
	(0.025)	(0.025)	(0.030)	(0.030)	(0.038)	(0.061)
Different language	-0.126*	-0.060*	-0.087*	-0.077*	-0.052	-0.117*
	(0.021)	(0.022)	(0.025)	(0.025)	(0.033)	(0.049)
Cocitation				3.350*	2.183*	1.739*
				(0.052)	(0.069)	(0.185)
Observations	1346588	1345177	521118	521118	86936	26606
AIC	341032	291013	153801	147370	50920	16538

Notes: Average effect of ties refer to the mean effect of 12 (3 WOS and 9 MGP) ties.

Robust standard errors clustered by cited article in parentheses.

Significance: † $p < 0.1$, * $p < 0.05$, * $p < 0.01$

The results shown in specification (1) of Table 4 make it clear that the use of subject fixed effects and corresponding control observations is a very important element of the method. With random controls, the average coefficient on ties rises from 0.682 to 1.989. This means that the presence of a linkage goes from multiplying the odds of citation by 1.98 up to 7.3. 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 3.8).

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 27) 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 4 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 5-digit citing subject. The cost of tighter matching is that we now find far fewer control observations—the sample falls by 83% to 86,936 observations. The coefficient on ties declines by almost a third but the effects are still large and precisely estimated. A comparison of the column (5) coefficients in the lower panel shows that distance, different university, and national borders decline by about 50% when adding the controls for ties, whereas the language coefficient is 80% lower. Thus, the main messages of the paper hold up when using 5-digit field controls.

The final estimation of Table 4 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 71% 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 double. The 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.[16]

4.3 Interactions as evidence for information mechanisms

The results we have obtained so far point to an important role for educational and career ties in fostering citations. A key question is whether we can actually distinguish information transfer mechanisms from an alternative mechanism that we call "citation cliques." Suppose that 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 "tribal" affiliation. In this story ties are proxies for intra-group loyalties, rather than information transfer.

Table 5: 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	-0.776*	-0.771*	0.056	-0.502*	-0.417*	-0.976*	0.181
	(0.065)	(0.071)	(0.165)	(0.095)	(0.120)	(0.123)	(0.180)
ln Dist $\mid$ Dist > 0	-0.033*	-0.027*	-0.056*	-0.018†	-0.026†	-0.030*	-0.004
	(0.007)	(0.008)	(0.020)	(0.011)	(0.014)	(0.014)	(0.020)
Different country	-0.077*	-0.083*	0.051	-0.078†	-0.004	-0.066	0.046
	(0.030)	(0.032)	(0.085)	(0.042)	(0.056)	(0.055)	(0.082)
Different language	-0.024	-0.030	0.056	-0.024	-0.002	-0.019	-0.086
	(0.025)	(0.027)	(0.067)	(0.035)	(0.046)	(0.045)	(0.068)
Ties:							
Combined ties	0.626*	0.606*	0.159*	0.556*	0.116*	0.577*	0.144*
	(0.012)	(0.013)	(0.030)	(0.018)	(0.022)	(0.020)	(0.033)
Observations	521118	521118		521118		265662	
AIC	147406	147327		147193		72650	

Notes: Combined ties refer to the sum effect of 12 (3 WOS and 9 MGP) ties. "Obscure" is defined to be equal to 1 if total citations received for this article is less than or equal to 5 (the median number of citations received among all articles) and 0 otherwise; "recent" is defined to be equal to 1 if the age of this citation flow (i.e., the citation lag) is less than or equal to 9 (the median age of all citations) and 0 otherwise; "different field" is defined to be equal to 1 if citing article and cited article belong to different 2-digit MSC and 0 otherwise.

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

To address this issue, we consider in Table 5 three types of interactions designed to uncover informational mechanisms. First, we look at whether geography and ties matter

[16]The trade-off between fineness of comparisons and sample attrition we have encountered here recalls the exchange between Thompson and Fox-Kean (2005) and Henderson et al. (2005) over the appropriate level of detail in the technology field classification used for constructing the control patent sample.

more for papers that are less well-known in columns (2) and (3). In columns (4) and (5), we examine citing papers that were published more recently after the cited paper. Finally in columns (6) and (7), we consider citing papers that are in different subjects from the cited paper. The idea here is that authors are more aware of work in their own fields but rely more on their networks to learn about work in other subject areas. To reduce the number of parameters to consider, we replace the individual ties indicators with a single count of the total number of ties between author teams. Column (1) reports the corresponding regression *without* interactions for comparison purposes.

The first set of interactions in Table 5 show the results of interacting geography and ties with an indicator for papers that are "obscure," measured here by having fewer than or equal to five cites, which is the median in our data. 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 (> 5 cites) have a 21% ($= 0.159/(0.606+0.159)$) smaller coefficient on the sum of ties than the lesser known papers. This is consistent with our view that ties facilitate awareness. Papers that are big successes require less help from networks to facilitate transmission. The coefficient on log distance is about 3 times as large for obscure papers.

We find similar results in columns (4) and (5) when the interaction is changed to distinguish recent versus older papers. Recent papers have a 21% higher coefficient on the sum of ties. Distance decays are estimated at $-0.018 - 0.026 = -0.044$ for papers in their first nine years after publication (the median age of papers in our sample) and -0.018 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.[17]

The last interactions we investigated obtain no statistically significant geographic interactions with the different-field indicator. However, as expected, ties do have larger impacts for different-field papers, with a coefficient that is 25% larger than for same-field papers. All three sets of interactions therefore support the premise that *scholars draw more heavily on their connections when obtaining less familiar information.* We see no reason why intellectual cliques would operate in this way. Our results do not rule out a role for non-informational functions of ties but they do suggest that information transfer is a quantitatively important way that ties influence citation patterns.

[17] 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.

4.4 Subsamples and other robustness checks

Table 6: Robustness

Sample:	(1) US only	(2) non-US	(3) average	(4) original geography	(5) available author
Panel A: including ties					
Distance > 0	-0.918*	-0.576*	-0.613*	-0.558*	-0.523*
	(0.140)	(0.109)	(0.080)	(0.068)	(0.040)
ln Dist $\mid$ Dist > 0	-0.017	-0.049*	-0.039*	-0.042*	-0.032*
	(0.016)	(0.014)	(0.008)	(0.007)	(0.004)
Different country		-0.190*	-0.165*	-0.131*	-0.123*
		(0.066)	(0.033)	(0.030)	(0.018)
Different language		-0.037	-0.035	-0.004	-0.024†
		(0.051)	(0.028)	(0.025)	(0.015)
Average effect of ties	0.507*	0.664*	0.993*	0.668*	0.645*
	(0.104)	(0.075)	(0.051)	(0.039)	(0.021)
Observations	40545	97931	521118	521118	1759828
AIC	18152	32775	145036	145184	487843
Panel B: excluding ties					
Distance > 0	-1.146*	-1.269*	-1.384*	-1.174*	-1.147*
	(0.140)	(0.101)	(0.072)	(0.060)	(0.037)
ln Dist $\mid$ Dist > 0	-0.058*	-0.069*	-0.067*	-0.068*	-0.057*
	(0.017)	(0.013)	(0.008)	(0.007)	(0.004)
Different country		-0.498*	-0.334*	-0.264*	-0.238*
		(0.066)	(0.033)	(0.030)	(0.018)
Different language		-0.072	-0.103*	-0.062*	-0.076*
		(0.050)	(0.027)	(0.025)	(0.015)
Observations	40545	97931	521118	521118	1759828
AIC	19514	35413	154648	154328	510487

Notes: Average effect of ties refer to the mean effect of 12 (3 WOS and 9 MGP) ties.
Robust standard errors clustered by cited article in parentheses.
Significance: † $p < 0.1$, * $p < 0.05$, * $p < 0.01$

Table 6 contains a range of checks on the impact of various features of the estimations thus far. First, we have used a world-wide sample instead of the US-only sample that many studies of citations have used. How much are our results driven by the observations where both citing and cited papers have US-affiliated authors? The answer is not much. The key result that inclusion of ties shrinks geography coefficients by about 50% (on average) holds up well in the non-US sample shown in column (2). As one might expect, continuous distance effects in the US-only sample are smaller than in the rest of world

sample. Indeed, after controlling for ties, distance has an insignificant impact on citations. However, the US sample exhibits especially large reductions in citation odds when authors are from different institutions. It is not obvious why intra-institutional citation bias would be stronger in the US.

Column (3) replaces the min/max approach to aggregating geographic and network variables across coauthors with averages over all the author pairs. The average coefficient for ties is almost 50% larger (0.993 vs 0.682). 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 -0.039 with averaging versus -0.036 under min/max. The overall fits of the two methods, as measured by the Akaike Information Criterion (AIC), are almost the same (145036 vs 145058). The similarity in results is partly due to the fact that there is relatively little coauthorship in mathematics. The average number of authors in mathematics, 1.88 in 2009, is lower than in most other disciplines.[18]

Column (4) 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 d. The contemporaneous geography used in the earlier specification leads to a better fit as measured by the AIC. The slightly larger distance effect estimated for original geography does not represent a statistically significant difference.

Column (5) 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. Using any available author triples the sample but does not change the coefficients much.

4.5 Has the Internet facilitated knowledge flows?

During the 1990s and 2000s a series of improvements in long distance communication were introduced which have the potential to diminish the role of distance in impeding knowledge flows. Advances affecting information flows generally include the spread of email in the late 1980s, the rise of web browsers in the mid 1990s, and the introduction of the Google search engine in 1998. Of particular importance to scientists was the creation of arXiv.org, a repository of pre-prints, which has included mathematics since 1992.[19] These technologies would be expected to have reduced the importance of face-to-face interactions, implying declining geographic separation effects in the 1990s and 2000s.

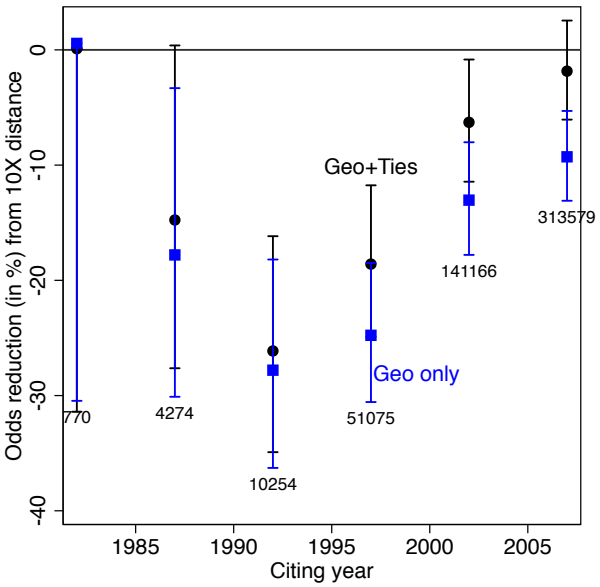
[18]For example, the average number of authors in evolutionary biology articles was 4 in 2005 (Agrawal et al., 2013), 3.75 in biomedical research (1961–2000), and 2.5 in physics (1991–2000,) 2.22 in computer science (1991–2000) (Newman, 2004), and 2.19 in economics (2011) (Hamermesh, 2013).

[19]We thank Michal Fabinger for alerting us to the importance of arXiv.org for knowledge dissemination over the internet.

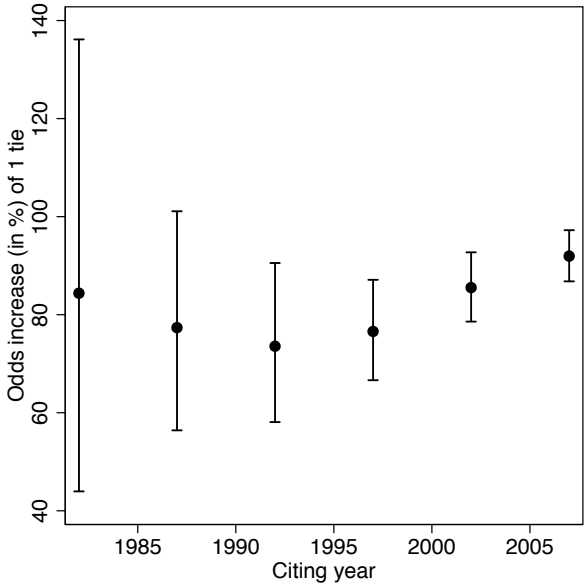
The estimates presented so far pool across a three-decade period (1980–2009). However, the majority of the sample comes from the last six years. To look for evidence of technology effects on knowledge diffusion, we break the sample into six sub-periods.

Figure 5: Time-varying distance and network effects (with confidence intervals)

(a) Multiply distance by ten



(b) Adding one tie



Note: Brackets are 95% confidence intervals. Sample size shown below brackets.

Figure 5 shows the effects of distance (panel a) and ties (panel b) estimated in regressions on 5-year intervals (1980–1984,...,2005–2009) based on the citing paper's publication year. The distance effect here is the percent reduction in the odds of citation caused by increasing distance by a factor of ten. The ties effect is the percent increase in the odds of citation from adding the first tie. We use the same underlying sample, method, and control variables as in column (1) of Table 5. The brackets show 95% confidence intervals (as before standard errors are clustered at the article d level). Panel (a) shows the distance effect controlling for ties in black and without ties in blue.

Sample sizes appear to be inadequate in the 1980s to obtain precise estimates of distance effects. From the early 1990s, we see distance effects shrink dramatically. In the specification controlling for ties, the coefficient on log distance becomes insignificantly different from zero in 2005–2009. The timing coincides with internet-based improvements in communication.

The blue brackets reveal that controlling for ties is an essential part of the story. In the last period a 10-fold increase in distance decreases citation odds by about 9% if we use coefficients from regressions that exclude the sum of ties variable. This is about one third of the 28% decrease in the early 1990s, but still a significant effect. The omitted

variable bias from not controlling for ties appears to be increasing over time since the 1990–1994 period. In panel (b) we see that ties themselves have seen an increase in their importance over the same period. The story these results are suggesting is that scholars can now communicate costlessly over distance but they communicate preferentially with those whom they have established connections in the past. And those linkages tend to be relatively close.[20]

Our findings concord with two studies using very different methodologies. 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. 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 from the 1970s to the early 1990s. Our study is unique in examining how distance effects changed in the 2000s, a period of great interest because advances in communication and search such as Skype (2003) and Google Scholar (2004).

The regressions underlying Figure 5 also estimate the citation odds penalty for being at different institutions. We find that the coefficient on the dummy for distance>0 increases in absolute value by a log point (from 0.16 to -0.90 when we control for ties and from -0.41 to -1.47 when we do not). This contrasts with the contention of Kim et al. (2009) that there has been a "reduced importance of physical access to productive research colleagues which in turn seems due to innovations in communication technology." Note that Kim et al. (2009) measure productivity spillovers rather than citations. Moreover, they attribute the lower importance of being at an elite university to the widening networks of coauthors. Our results clarify that being colleagues at the same university continues to greatly influence the likelihood of being aware of a relevant research contribution.

5 Conclusion

This paper advances the literature on how networks facilitate knowledge flows by going beyond career ties (past collaboration and colocation) and co-ethnicity indicators used in past work. The educational ties we investigate were set at the beginning of the academic's career and are therefore more predetermined than career ties and more directly interpersonal than shared ethnicity. Given the absence of experimental variation in ties, we cannot interpret the coefficients as causal effects. Their robustness in the face of increasingly stringent fixed effects capturing relevance of a cited paper to potential citing

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

papers demonstrates that ties are not just proxies for pairs of researchers working on the same topic. Even our most conservative estimate (controlling for 5-digit subject and a cocitation indicator) shows large effects: a single tie on average boosts the odds of citation by 57%. Furthermore, we find interactions between ties and variables designed to capture information so we do not think the ties effect can arise solely from academic cliques.

The relationship between geography and knowledge flows supported by our results has important nuances. On the one hand, colleagues at the same institution cite each other much more than one would expect based on the topic they work on and the general importance of the work (which we capture in a composite fixed effect). Indeed, former colleagues continue to cross-cite even after they relocate to new institutions. On the other hand, we find that in the US-only sample the marginal effect of greater distance between institutions is insignificantly different from zero. This is also true in the most recent five years of the world-wide sample. Both of these "zero" distance effects show up in regressions that control for 12 linkages based on career and educational histories. Failure to control for such ties restores the significant negative effect of distance on citation odds. Evidently, in mathematics "what you know" depends a great deal on "who you know." It is increasingly unrelated to "where you work"—except insofar as "where you work" influences "who you know."

We now loop back to the the questions we posed at the beginning of the paper. With all the normal caveats about external validity, we offer some answers based our results. Cities may be valuable not just because of daily face-to-face interactions, but because they are good places to build networks. Such a view points to a different interpretation of the de la Roca and Puga (2012) finding that wages rise with experience in big cities but retain much of this growth even when individual goes back to a smaller city. While they attribute the wage premium to rising ability, our framework suggests it could have been an expanded set of professional ties. 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. Finally, our results suggest that it is not remoteness *per se* that prevents less developed countries from reaching the technological frontier, but sparse professional and educational ties to knowledge generators in the developed world.

References

Agrawal, A., Cockburn, I., 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., Kapur, D., McHale, J., September 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., Oettl, A., Nov. 2013. Collaboration, stars, and the changing organization of science: Evidence from evolutionary biology. NBER Working Papers 19653, National Bureau of Economic Research, Inc.

Alcácer, J., Gittelman, M., November 2006. Patent citations as a measure of knowledge flows: the influence of examiner citations. The Review of Economics and Statistics 88 (4), 774–779.

Althouse, B. M., West, J. D., Bergstrom, C. T., 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.

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

Borjas, G. J., Doran, K. B., 2012. The collapse of the Soviet Union and the productivity of American mathematicians. The Quarterly Journal of Economics 127 (3), 1143–1203.

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

de la Roca, J., Puga, D., 2012. Learning by working in big cities. Discussion Papers 9243, CEPR.

Feyrer, J., 2009. Distance, trade, and income–The 1967 to 1975 closing of the Suez Canal as a natural experiment. Tech. Rep. 15557, National Bureau of Economic Research.

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

Griffith, R., Lee, S., 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., March 2013. Six decades of top economics publishing: who and how? Journal of Economic Literature 51 (1), 162–72.

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

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

Hopenhayn, H. A., September 1992. Entry, exit, and firm dynamics in long run equilibrium. Econometrica 60 (5), 1127–50.

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

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

Keller, W., September 2004. International technology diffusion. Journal of Economic Literature 42 (3), 752–782.

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

Kim, E. H., Morse, A., Zingales, L., 2009. Are elite universities losing their competitive edge? Journal of Financial Economics 93 (3), 353–381.

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.

Manski, C. F., Lerman, S. R., 1977. The estimation of choice probabilities from choice based samples. Econometrica, 1977–1988.

Melitz, M. J., November 2003. The impact of trade on intra-industry reallocations and aggregate industry productivity. Econometrica 71 (6), 1695–1725.

Newman, M. E., 2004. Who is the best connected scientist? A study of scientific coauthorship networks. In: Ben-Naim, E., Frauenfelder, H., Toroczkai, Z. (Eds.), Complex Networks. Vol. 650 of Lecture Notes in Physics. Springer, pp. 337–370.

Packalen, M., Bhattacharya, J., 2015. Cities and ideas. Tech. rep., National Bureau of Economic Research.

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

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

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