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

Keith Head† UBC & CEPR

Yao Amber Li‡ HKUST & UWO

Asier Minondo§ University of Deusto

December 9, 2014
Preliminary draft
(but feel free to cite)

Abstract

Using data on academic citations, career histories of mathematicians, and disaggregated distance data for the world's top 1000 math departments, we study how geography and personal ties affect knowledge flows among scholars. The ties we consider are co-authorship, past co-location, advisor-advisee relationships, and *alma mater* relationships (holding a Ph.D. from the institution where another scholar is affiliated). Using matched choice-based sampling logit regressions, we find that linkages significantly facilitate knowledge flows. Controlling for ties generally halves the negative impact of geographic barriers on citations. Ties matter more for less prominent and more recent papers. Proximity continues to influence knowledge flows even in the era of Google searches.

Keywords: network, distance, border, geography, knowledge spillovers, paper citations, genealogy, matching

*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. We also thank Andrew Bernard, Teresa Fort, Joshua Gottlieb, Bob Staiger, Bronwyn Hall, Wolfgang Keller, Anthony J. Venables, Quoc-Anh Do, Jim MacGee, Tom Ross, Beata Javorcik, Peter Neary, Edwin Lai, and Daniel Sturm for helpful discussions. Finally, we thank Ho Yin Tsoi, Bo Jiang, Yiye Cui, and Song Liu for excellent research assistance during this project.

†Head: Corresponding author. Sauder School of Business, University of British Columbia, 2053 Main Mall, Vancouver, B.C., V6T 1Z2, Canada. Email: keith.head@sauder.ubc.ca. Tel: (604) 822-8492; Fax: (604) 822-8477.

‡Li: Department of Economics and Faculty Associate of the Institute for Emerging Market Studies (IEMS), Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong SAR-PRC. Email: yaoli@ust.hk. Tel: (852)2358 7605; Fax: (852)2358 2084. Research Affiliate of the China Research and Policy Group at University of Western Ontario.

§Minondo: Deusto University, Deusto Business School, Mundaiz 50 20012 Donostia, San Sebastián, Spain. Email: aminondo@deusto.es.

1 Introduction

Geographic barriers to knowledge diffusion offer a unified explanation for at least three major economic phenomena. First, they explain in part why some countries appear to be many times more productive than others.[1] Second, information exchange is argued to be the strongest force drawing a large and growing share of the world's population into cities, despite their higher rents and congestion.[2] Finally, informational barriers may explain why distance and borders have much larger effects on trade and investment flows than can be accounted for by freight costs and tariffs.[3]

The challenge of corroborating the information friction hypotheses is that, as Krugman (1991) put it, "knowledge flows ... are invisible; they leave no paper trail by which they can be measured and traced." Patent and academic citations provide a counter-example where flows of knowledge become visible via the obligation to cite relevant precedents. As the only systematically tracked measure of knowledge flows, citations have been much studied. Following the seminal paper of Jaffe et al., 1993, several papers have found that geographic separation has a significant negative effect on patent citations.[4] What the literature has yet to establish is *why* knowledge flows should be impeded by geographic barriers such as distance and borders.

In this paper we argue that the underlying reason why geographic separation reduces knowledge flows is that the professional networks of information generators and users are both very important for facilitating knowledge flows and strongly geographically biased. In estimations that do not control for network connections, this gives the appearance that geography has a direct impact. After conditioning on the presence or absence of ties between citing and cited authors, we that impacts of distance, borders, and language differences fall by half. Thus, our work suggests that a major mechanism through which separation impedes knowledge transmission is that geography shapes academic ties and these linkages in turn drive much dissemination of knowledge.

Localized information is one of the three Marshallian spillovers that motivate industries to concentrate geographically. If new ideas are transmitted through conversations, informal meetings or other types of personal interactions, physical proximity will play a key role in the transmission of knowledge. Glaeser (2011) remarks that pundits have long predicted that technological improvements would doom cities, yet "To defeat the human need for face-to-face contact, our technological marvels would need to defeat millions of

[1] Keller (2002) shows the effect of distance on productivity differences.

[2] What Glaeser (2011) refers to as the "urban ability to create collaborative brilliance" could not persist if all the ideas generated in cities costlessly diffused to places where it was cheaper to live.

[3] Allen (2014) finds that information frictions explain half the observed regional price variation in agricultural prices in the Philippines. Even "weightless" transactions like portfolio and direct investment have have distance decay elasticities that resemble those of goods.

[4] See Peri, 2005, Belenzon and Schankerman, 2013.

years of human evolution that has made us into machines for learning from the people next to us." The results we present suggest that mechanism underlying the effect of distance on knowledge flows is not a simple reduction in face-to-face interactions. Rather, we find that a paucity of network connections at large distances drives the distance effect. This suggests face-to-face interactions are more important for constructing networks than for facilitating contemporaneous knowledge flows.

The international trade literature has made repeated attempts to understand the large effect of geographic distance on bilateral trade. Grossman, 1998 pointed out that the estimated effects of distance are too large to be consistent with what we know about the magnitude (small) and form (price by volume or weight but not by value) of freight costs. Grossman suggested three mechanisms through which distance might matter: familiarity (information decaying with distance), localized tastes, and distribution networks. Head and Mayer (2013) synthesize evidence from a number of papers to argue that there must be some form of "dark matter" to account for large distance. They combine estimates of freight-distance elasticities and trade-price elasticities from the literature to apportion the trade-distance elasticity between freight-related costs and unobserved factors. Freight alone can account for just 4% to 28% of the distance cost. Even after considering time costs of transport, there is a substantial unexplained part of the distance effect that could be attributed to information decay. Head and Mayer (2013) evaluate estimated border effects using a similar approach. Customs duties at currently prevalent levels can only account for a small fraction of the trade reduction associated with moving goods across national borders.

In the context of citations, an academic or a patent inventor should cite all the relevant prior work. In practice, she can only cite the work she is familiar with (leaving the patent examiner or referees to supply the remainder). If the information decay story is correct, we would expect that knowledge flows are more likely between more closely connected, familiar persons. If we are able to find appropriate measures to control for "connections," we would expect a much more limited role for geographic separation as an impediment to the spread of knowledge.

The primary aim of this paper is to test this hypothesis using data from academic citations in mathematics. Previous studies of geographic decay of knowledge spread have used patent citation information. Papers have many properties that are shared by patents. On the one hand, papers, as patents, are published (granted) in prestigious journals as long as they advance the knowledge frontier. Second, papers, as patents, build upon and cite prior work. The paper citation records indicate who benefits from the efforts of whom. Therefore, cross-paper citations trace out the direction and the intensity of knowledge flows among scholars. Patents are closer to commercial application than academic citations and hence are arguably better measures of economically relevant

information. However, academic citations have a decisive advantage for the purposes of this study. Unlike inventors, academics and particularly mathematicians, have a rich set of well-documented linkages via past educational and career histories.

The field of mathematics is well-suited to our research question for a number of reasons. First, as a basic science, transmission of mathematics knowledge should be less affected by linguistic, cultural, and social factors. Almost by definition, mathematics employs a common language of communication. In many social sciences and humanities fields, there are journals that focus on issues specific to 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.

Because new theorems build upon previous theorems, the citation of intellectual predecessors is particularly important in mathematics. A second advantage of studying math citations is reduced frequency of "cheap" citation, as compared to the social sciences and applied science disciplines. That is, math cites are more likely to reflect knowledge flows rather than favors to friends or gestures to curry favor from potential referees. While there is no direct evidence to support this conjecture, it is a fact that the references sections of math papers are typically shorter, containing 18 papers on average, compared to 30 in economics and 45–51 in sociology, psychology, business and marketing.[5]

The biggest practical advantage of mathematicians for our purposes is the availability of data on graduate school attended and the Ph.D. advisor. This data forms the basis for most of our measures of ties between citing and cited authors. Moreover, there is data classifying mathematicians and their individual papers into detailed sub-fields.

We track all articles published in all journals that are listed in the Mathematics category of the ISI Web of Science (WOS) to extract citations data.[6] We capture "connections" by constructing network variables based on data from Web of Science and the Mathematics Genealogy Project (MGP) . Those scholarly networks include co-authorship, advisor-advisee relationship, academic siblings (sharing the same Ph.D. advisor or obtaining Ph.D. from the same institution), employment connections (working in the same institution in different years or in the same year), and Alma Mater relationships (obtaining a Ph.D. from the institution where another scholar is affiliated). To our knowledge, this is the first time "genealogy" data of this type has been used in conjunction with geographic information to study knowledge flows.[7]

Finally, we use highly disaggregated geographic location data extracted from Google Maps to precisely measure distance between any two institutions for the world top 1000

[5] Althouse et al. (2009) Table 3.

[6] To focus on basic science and keep the data set manageable, we exclude applied mathematics journals.

[7] Borjas and Doran (2012) use the MGP data to show that mathematicians whose Ph.D. advisor was a Soviet émigré to the US tend to have higher productivity and receive more citations than other researchers.

institutions in mathematics. Our distances can be as few as 200 meters for adjacent institutions, allowing a much finer scale than the "same metro area" indicators used by Jaffe et al. (1993) and Singh (2005). Because we include institutions from around the world, we have a much greater range of distances than Belenzon and Schankerman (2013). Moreover, because we use data including authors from 57 different countries, we are able to investigate national border and language effects.

To evaluate the effect of geography and networks in paper citations, we follow the matching methodology of Jaffe et al. (1993) and Belenzon and Schankerman (2013) and compare the characteristics of our sample with the characteristics of a control group. Our main control group is constructed by the following criteria: for each citation received by a paper we randomly select another article that does not cite the paper but is published in the same year as the original citing paper and classified with the same 3-digit field classification. The union of the original sample and the control group constitutes the sample that is used in the econometric analysis.

There are three key findings. First, network factors significantly facilitate knowledge spillovers, given geographic barriers such as distance and borders. The most important network variables include co-authorship, past employment relationships (colleagues), and advisor-advisee relationship. Second, controlling for ties, reduces the effect of geographic separation by about one half. Third, the impact of geographic separation does not appear to have changed significantly during the internet era. Our main findings are robust to various econometric specifications and sampling strategies including different control groups and subsamples.

This paper makes contributions to three important literatures. First, it advances the study of knowledge diffusion by considering the effects of geography and networks simultaneously. The patent citations literature has not been able to partial out the roles of physical separation from network separation because of the limited availability of data on the personal ties relevant for patent citations. Second, this paper contributes to the networks literature by providing compelling empirical evidence that networks mediate knowledge transmission. Prior work has considered single measures of social connections measured via past co-location by Agrawal et al. (2006a) and co-ethnicity by Agrawal et al. (2008). Singh (2005) use past collaboration on patent-inventing teams. We also consider past co-location and past collaboration but extend the set of ties to include four types of ties derived from educational histories.

Last, and perhaps most important, this paper contributes to resolving the longstanding puzzle of why distance and border effects on international transactions are so large. The result that controlling for networks substantially attenuates both effects strongly implicates information decay as a key mechanism underlying the dark matter of distance

and borders.[8]

The remainder of the paper is organized as follows. Section 2 describes the data. Section 3 details how we measure the network linkages and presents our econometric specifications and . Section 4 reports main findings and a series of robustness checks. The last section concludes.

2 Data Description

2.1 Data Sources

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 citing and cited author teams.

Citation Data

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 authors, the affiliation of the authors, and the cited articles. The cited article is identified by the first author, the journal in which it was published, the year of publication, volume and first page.

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.[9] A shortcoming of WOS is that it does not provide the affiliation for a substantial

[8] Head and Mayer (2013) summarize other recent research finding support for the information decay hypothesis using very different methodologies.

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

number of authors. In particular, for 536,454 author-article combinations included in our database, we have the affiliation for 69% of combinations.

We augment affiliation information applying the procedures developed by Tang and Walsh (2010), as implemented in Agrawal et al. (2013).[10] This increases the author-article combinations with affiliation information for some authors from 69% to 80%. Of those, 84% have affiliations for all authors.

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.	137,877	92,992	492,812
With 5-digit MSC field	77,941	77,013	307,867
MGP data all authors	8,825	8,751	36,502

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

Table 1 shows how our sample declines from a very large set (almost five million) of realized cites to the much smaller sets (the last two rows) that we use in regressions.

We identify the affiliation of mathematicians every time they publish an article. As authors might not publish an article every year, there were many gaps in affiliation histories. To fill these gaps, we interpolate and extrapolate from the information we do have. 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).

[10]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.

The full information set of articles cites 1,247,171 different articles (including self-citations). In this set 987,056 articles were published in the period analyzed in this paper, 1975–2009. The ISI Web of Knowledge only identifies the first author of the cited articles. To identify the affiliation of the first author, and the identity and affiliation of the rest of co-authors (if there are), we matched the cited articles with our original database. As our database only includes the journals included in Mathematics category, we can only identify the authors and co-authors of the cited articles belonging to this set. In particular, only 321,447 cited articles of the 987,056 sample, were published in the 255 journals included in our database. We could get complete affiliation for the citing authors and the cited authors for 187,114 citing articles and 133,465 cited articles. After removing self-citations the number of citing articles declines to 168,112 citing articles (50% of the initial number) and 108,238 cited articles (9% of the initial number).

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.[11] The top 1000 affiliations account for 80% of the observations after all previous cleaning steps. This final sample contains 137,877 citing articles and 92,992 cited articles.

The WOS database does not provide information about the mathematical field that is covered in the article. We obtain this information from Zentralblatt MATH (see below). However, we could not find the field information for all the articles. This further reduces the sample to 77,941 citing articles and 77,013 cited articles.

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: "Worked same place" indicates that two authors shared the same affiliation in the past (but no longer do). "Coincided past" requires co-location at the same institution in the *same year*.

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 names of the Ph.D. program, the degree recipient, and his or her advisors. The MGP also shows the year of Ph.D. completion. We merged these

[11] When there is more than one author, the article is divided by the number of authors.

data with the citing authors and cited authors in our database. We were able to match the records by author for around 44% of records.

The MGP data allow us to construct six 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 parent/sibling 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. "Same Advisor" refers to academic "siblings" were both supervised by the same professor—regardless of whether they overlapped.

- 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.[12] 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 MSC year 2000 revision (MSC2000). In light of the criticism of 3-digit patent technology fields by Thompson and Fox-Kean (2005), we also use 5-digit codes, but the extra detail (2175 fields) comes at the cost of not always being able to find a control observation.

Geographic distance data

We extracted the latitude and longitude information for all top 1000 institutions from Google Maps (http://maps.google.com) enabling construction of highly disaggregated distance data between each institution pair. Much of the prior work uses coarse measures of location such as residing in the same metropolitan area. That implies that within the same city, all institutions share either zero distance or a common internal distance calculated by the area of the city. For example, within the Boston metro area, the distance between Harvard and MIT is only 3km but the distance of MIT to Brandeis University

[12] 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/

is 14km. We are able estimate non-parametrically the profile of information decay over fine and broad scales.

Because we use publications to track author locations over time, we are able to calculate distance (and other measures of geographic separation) at the time the *citing* article is written. This differs from past work using patents where distance is measured between the location of the citing team in the year of that patent and the locations of the cited team 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 that paper had by 2005 moved to a proximate institution or even become colleagues at the same institution.

2.2 Descriptive Statistics

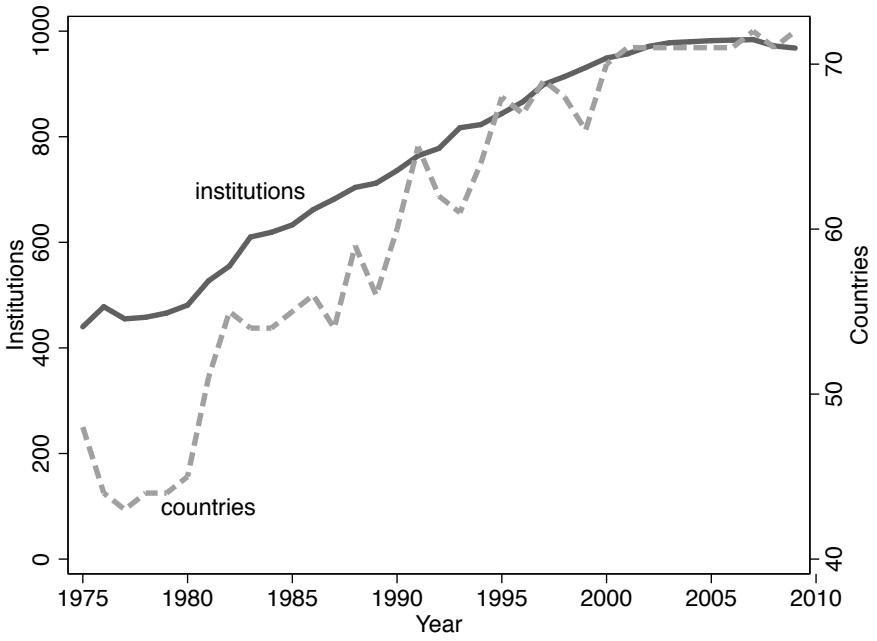
Here we provide some descriptive statistics of our database. First, we analyze the evolution of the number of articles, authors, institutions to which authors are affiliated and countries where those institutions are located in 1975–2009. There has been a notable increase in the number of articles and 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 2008. 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.

Meanwhile the number of institutions authors are affiliated to and the number of countries where these institutions are located grew markedly. Figure 1 shows that during the period 1975–2009 the set of institutions producing pure math rose by 2% per year from 448 to the final set of the top 1000 we use. The number of countries participating in the production of pure math rose from 48 to 72.

The institution with the highest number of articles (the largest node) is Moscow State University, with 1,615 articles, followed by Berkeley (1,494), Paris 11 (1,247), Paris 6 (1,247) and Kyoto University (1,167 articles). If we exclude citations to the same institution, the most important edges (bilateral citations), are those linking UCLA and Berkeley (498 citations), Princeton and Berkeley (484 citations), Princeton and MIT (478) citations, MIT and Berkeley (468 citations), and Princeton and Paris 11 (465 citations).

There is a steady rise in the number of authors per article. In 1975 the average number of authors per article was 1.24, whereas in 2009 it raised to 1.88. This trend is similar in other scientific areas, such as evolutionary biology (Agrawal et al., 2013) or economics (Hamermesh, 2013). Nevertheless, the average number of authors in mathematics remains

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



much lower than in most other sciences. [13]

3 Specification

We start by sketching a reduced form of the probability of one article citing another. We specify the observed determinants of citation and how to control as best as possible for unobservables. Then we important features of the estimation method that arise from the nature of the citation data.

3.1 Empirical model of the citation likelihood

The main empirical specification is based on a latent variable C^*_{id} that is linearly related to underlying determinants of citations:

$$C^*_{id} = \mathbf{G}'_{id}\boldsymbol{\gamma} + \mathbf{L}'_{id}\boldsymbol{\lambda} + \tau_{t(i)} + \alpha_d + \varepsilon_{id}. \tag{1}$$

G is a vector of geography-related variables pertaining to relationship between the institutions of the authors of the citing and cited articles. **L** is a vector of network-related

[13] For example, the average number of authors in evolutionary biology articles was four in 2005 (Agrawal et al., 2013), 3.75 authors per article in biomedical research during the period 1961–2000, 2.5 authors per article in physics in the period 1991-2000, 2.22 authors per article in computer science in the period 1991–2000 (Newmen, 2004), and 2.19 authors per article in economics in 2011 (Hamermesh, 2013).

indicators of the connections between authors of article i and d. The $\tau_{t(i)}$ are year effects corresponding to the year t (1975–2009) in which citing article i was published.

Without putting a forward a full-fledged model of the citation process, we think of this expression as combining the two key factors of citation: relevance and awareness.[14] The unobserved relevance of paper d for paper i comprises a general relevance term α_d and a dyadic idiosyncratic term ε_{id}. Meanwhile, awareness depends on geographic separation because it increases frequency of face-to-face interactions (from "water-cooler" conversations to conference meetings). The novel element of our model is that information can overcome geographic barriers if authors A and B are connected via overlapping career and/or educational histories. Past co-location or just indirect linkages such has having the same advisor at different times create a kind of connective tissue that facilitates knowledge flows. In summary we hypothesize that $\boldsymbol{\gamma} < 0$ and $\boldsymbol{\lambda} > 0$.

We use C_{id} is a dummy variable set equal to 1 if article i cites article d and zero otherwise. We assume that paper i cites paper d when C^*_{id} exceeds a threshold (normalized to zero). The probability of citation is therefore

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

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

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

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

Estimating α_d is not feasible (IPP) so instead we condition on the total number of cites received by each article: $C_d = \sum_i C_{id}$. The resulting estimator is referred to as the fixed effects (or conditional) logit Exponentiated coefficient express geography and tie effects in terms of the change in citation odds ratios.

Alternatively one can assume ε_{id} is uniformly distributed in which case equation 2 becomes a linear equation. This linear probability model (LPM) can also be thought of as a first order approximation of equation 3 that delivers immediate estimates of the average marginal effects. The use of LPM also permits us to compute standard errors clustered at the level of the cited article, which allows for correlations in the errors across citing articles for the same 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. Thus, it is

[14] Citations made for non-academic reasons are assumed away. We also do not model the possible interdependence between relevance and awareness, namely that because A is ignorant of B's work, she writes a paper on a different topic than she would have written with fuller information.

necessary to aggregate author pairs for a given article pair. For example suppose paper i has authors A and B, whereas the authors of paper d are C and D. Then there are four combinations (A-C, A-D, B-C, B-D) of primitive **G** and **L** variables (e.g. distance between A's and C's respective institutions or whether A and C have collaborated on a paper in the past). 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 co-authors. 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 ties, 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 averaging is an alternative which implicitly assumes that more linkages increase information flow. Thus 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.

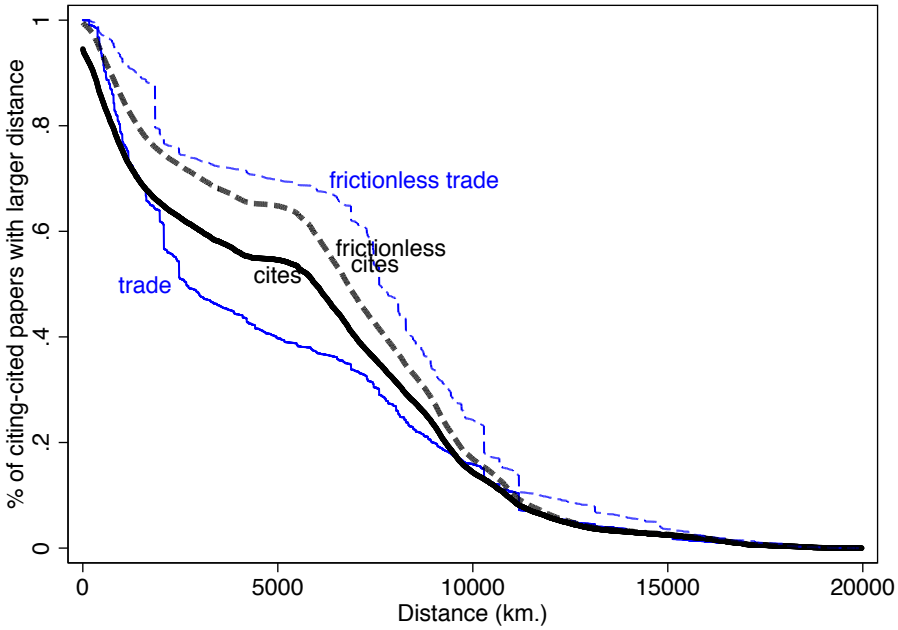
We consider three geography variables, distance, borders, and language difference. Each variable is expressed such that a large value indicates greater separation. The national border dummy takes the value of 1 at the author-pair level if the citing author is affiliated to an institution located in different country from the cited author. The language dummy is based on the official language of the country hosting each authors' institution (this need not be the native language of the author in question).

Figure 2 provides a first 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.

3.2 Estimation method: matched choice-based sampling

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. The problem is that citations are an example of a rare event problem that makes the standard approach computationally challenging. There are approximately 6 billion potential cites and about 300,000 realized cites. Thus, the rate of citation is only 5 per 100,000. In response to this problem, the patent citation literature has generally adopted a choice-based sampling. For each realized citation (case), pick

Figure 2: Distribution of citation distances



one non-realized citation (control) at random.

Selecting the control to match the case on important dimensions should help. We follow the matching methodology of Jaffe et al. (1993) by comparing the characteristics of our sample of realized citations with the characteristics of a control group. The control group is constructed as follows: for each citation received by a paper (the "case") we randomly select another observation that does not cite the paper (the "control"). The control is required to be published in the same year and the same 3-digit field as the original citing paper (case). The union of the original sample and the control group constitutes the sample that is used in the econometric analysis. We investigate in depth the consequences of alternative methods of choosing the controls involving less and more stringent criteria for relevance.

While earlier work such as Jaffe et al. (1993) and Agrawal et al. (2006b) simply compared the citation propensities of cases and controls, we use the regression approach deployed by Singh (2005) and Belenzon and Schankerman (2013), with the main difference that we condition on citing article quality. Controlling for quality is important since there is no reason to expect the geography and ties variables to be orthogonal to paper quality.

In the appendix we report the results of a Monte Carlo investigation we conducted into the properties of choice-based sampling in the presence of non-logit errors and article-specific components to the error term that are correlated with the variable of interest. We find that choice-based conditional logit works reasonably well at replicating the population estimates.

4 Results

This section presents the main results regarding the effect of geography and network as well as their interaction effects in knowledge spillovers. There are four key findings. First, the effects of distance, borders, and language differences are half as strong once network indicators are taken into account. Second, all nine of the measures of ties we included had positive and significant impacts. On average the effect of adding a network tie more than doubles the odds of citation. Third, networks and geography affect different types of papers differently. In particular, less prominent and more recently published papers exhibit stronger effects. Finally, there is little evidence of trends in the geography effects and no evidence that the internet is making distance irrelevant.

4.1 Baseline results

Table 2 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 or from countries that have different official languages. The second specification adds the first set of ties, namely, "coauthors," "coincided," and "work in the same institution," constructed directly from WOS database. The third to sixth specifications restrict the sample to the articles with full information from the MGP database. This permits us to add six additional network variables, based on educational history. As with the first two columns, we first show the effects of geography without network controls (column 3) and then with them (column 4). There is a big drop in the number of observations when we introduce the MGP network variables due to the matching of mathematicians' names between two data sets.

Specification (1) presents significantly negative coefficients on distance and borders, suggesting that physical distance and borders indeed impede the knowledge spillovers, which is consistent with the findings in the existing literature. The most interesting and novel finding is that controlling for ties substantially attenuates the negative effects of geographic separation. Incorporating network variables substantially halves the effect of distance and borders.

Table 2 shows that all network variables significantly positively affect citation probability. It is also interesting to compare the magnitudes of the effects of network ties on knowledge flows.[15] The co-authorship relation increases the odds of citation by $\exp(1.709) - 1 = 452\%$. Advisees are more likely to cite their advisor's papers: if any of the authors of article d is advisor of any of the authors of article i, that increases the

[15]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.

Table 2: Baseline: matching by MSC-3d, full author information

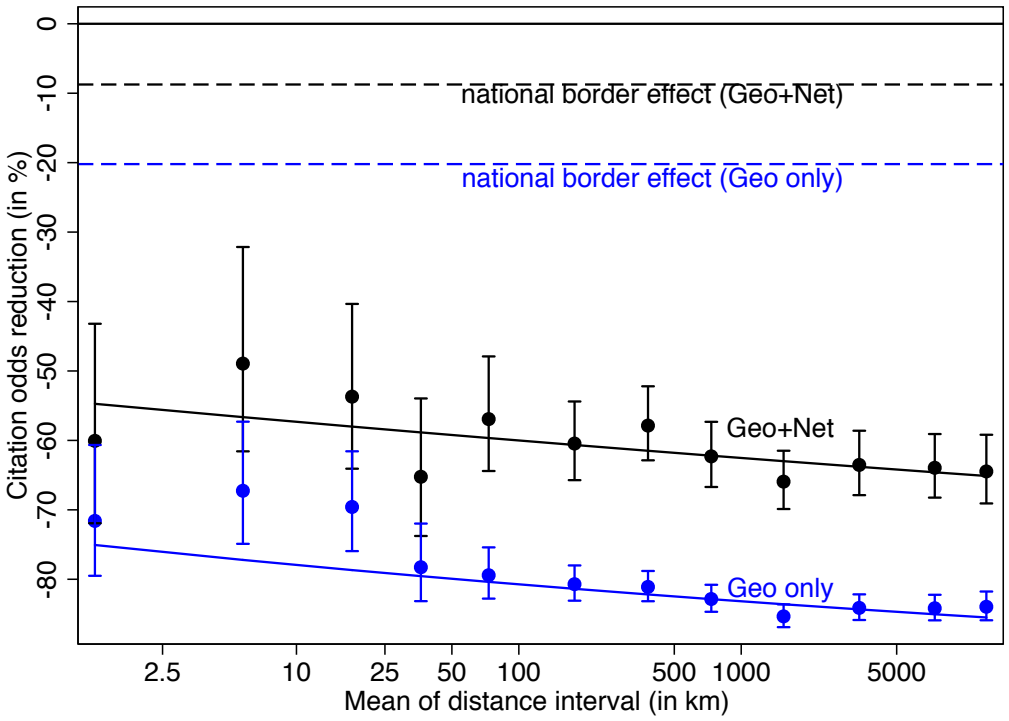
	(1)	(2)	(3)	(4)	(5)	(6)
Specification:	Article-fixed-effects logit (AFE-Λ)				AFE-LPM	AFE-Λ
Sample	WOS	WOS	MGP	MGP	MGP	MGP
Geography:						
Distance > 0	-1.065*	-1.031*	-1.376*	-0.786*	-0.159*	
	(0.023)	(0.024)	(0.069)	(0.075)	(0.019)	
ln dist $\mid$ dist > 0	-0.070*	-0.048*	-0.059*	-0.028*	-0.008*	
	(0.003)	(0.003)	(0.007)	(0.008)	(0.002)	
Different country	-0.201*	-0.146*	-0.193*	-0.078*	-0.024*	-0.092*
	(0.011)	(0.011)	(0.030)	(0.031)	(0.010)	(0.032)
Different language	-0.103*	-0.062*	-0.116*	-0.062*	-0.020*	-0.056*
	(0.009)	(0.009)	(0.024)	(0.025)	(0.008)	(0.025)
Co-authors		1.651*		1.709*	0.355*	1.713*
		(0.016)		(0.056)	(0.011)	(0.056)
Coincided past		0.669*		0.521*	0.127*	0.519*
		(0.015)		(0.045)	(0.011)	(0.045)
Worked same place		0.397*		0.357*	0.099*	0.353*
		(0.014)		(0.044)	(0.013)	(0.044)
Share Ph.D. (5 years)				0.279*	0.080*	0.279*
				(0.084)	(0.021)	(0.084)
Same advisor				1.122*	0.258*	1.116*
				(0.074)	(0.019)	(0.074)
Advisor citing				1.606*	0.317*	1.608*
				(0.201)	(0.031)	(0.201)
Advisor cited				1.819*	0.323*	1.820*
				(0.079)	(0.014)	(0.079)
Alma Mater citing				0.222*	0.043*	0.214*
				(0.061)	(0.015)	(0.061)
Alma Mater cited				0.205*	0.054*	0.200*
				(0.058)	(0.014)	(0.058)
Observations	615734	615734	73004	73004	73004	73004
AIC	667664	648263	70839	66789	98275	66776

Standard errors in parentheses. Significance: *: 1%, *: 5%, †: 10%

odds that article i cites article d by 516%. Advisors over-cite their advisees' articles by a somewhat small amount. The average over all nine coefficients is 0.836, implying an odds increase of 130%.

Specification (5) re-estimates specification (4) using the linear probability model (linear regression of the citation indicator on the controls after demeaning based on article averages). In the context of choice based sampling (CBS) this method no longer gives a direct estimate of marginal effects. However, it should provide a reasonably accurate estimate of relative magnitudes and t-statistics. Also the linear regression allows for clustering at the article level so the significance levels reported here should be more conservative than in the fixed effects logit which does not have clustered standard errors. In general the coefficients are proportional and the factor appears to be not far off the expected 4:1 ratio. Significance levels remain similar suggesting that the bias in the standard errors from correlation of errors across cites of the same article is relatively small.

Figure 3: Estimated geography effects with and without tie indicators

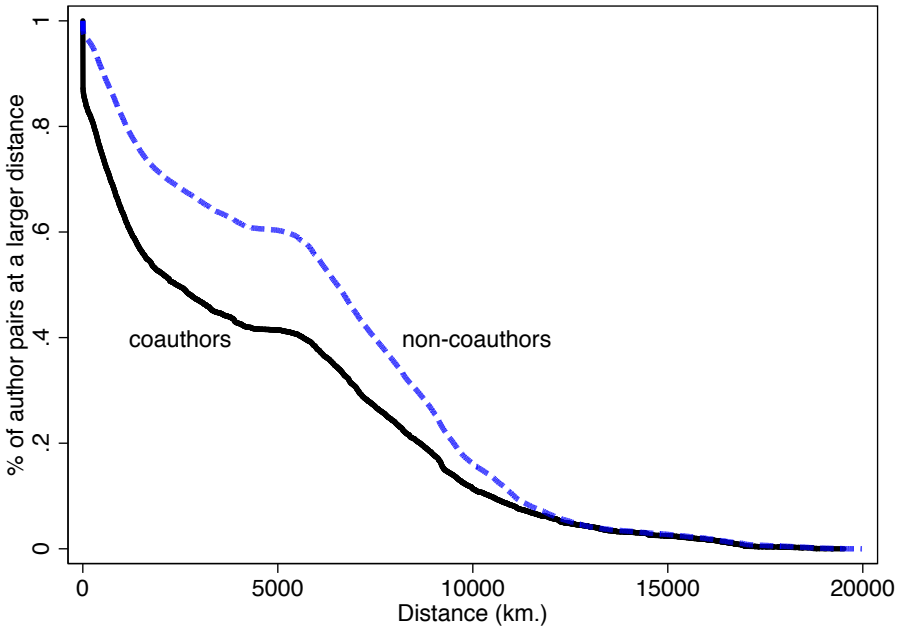


The first five specifications employ a parsimonious parametrization of distance effects. This two-part formulation has a jump from zero to positive distances, but thereafter the elasticity of citations odds with respect to distance is constant. While a constant elasticity of distance in trade equations 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 estimates in specifications (4) and (6), we find only small, uninteresting differences, suggesting that the 2-part approach may be 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 2, depicted in black. The vertical axis depicts the percentage reduction in the odds of citation (relative to a distance of zero) associated with each step.[16] We also show in blue the corresponding estimates for specification (3) with the two-part replaced with the 12-step. For each set of steps, we overlay the implied reduction in the odds of citation associated with the two-part parametrizations. The key finding illustrated in the figure are that after the dramatic fall associated with positive distance, the subsequent declines are in line with the prediction of the two-part model. Controlling for networks moves the decay function up (lower effect of being at different institutions) and flattens it. The figure shows that it is hard to distinguish empirically between a decay function that is flat after 1000 kilometers and one that continues regular decay at the small elasticity of -0.03. Failure to control for networks (as in the blue step function) suggests more curvature with distance mattering more at first.

Figure 4: Distribution of distances between coauthors



Why does the inclusion of the nine network linkage indicators lead to such striking

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

changes in the size and shape of distance effects? The answer must be that ties are geographically biased. There is already literature (to be added) showing that co-authorship is impeded by separation. We illustrate this effect in Figure 4, which shows that coauthors tend to be closer to each other than authors who have not collaborated.

Figure 5: Geographic bias of ties

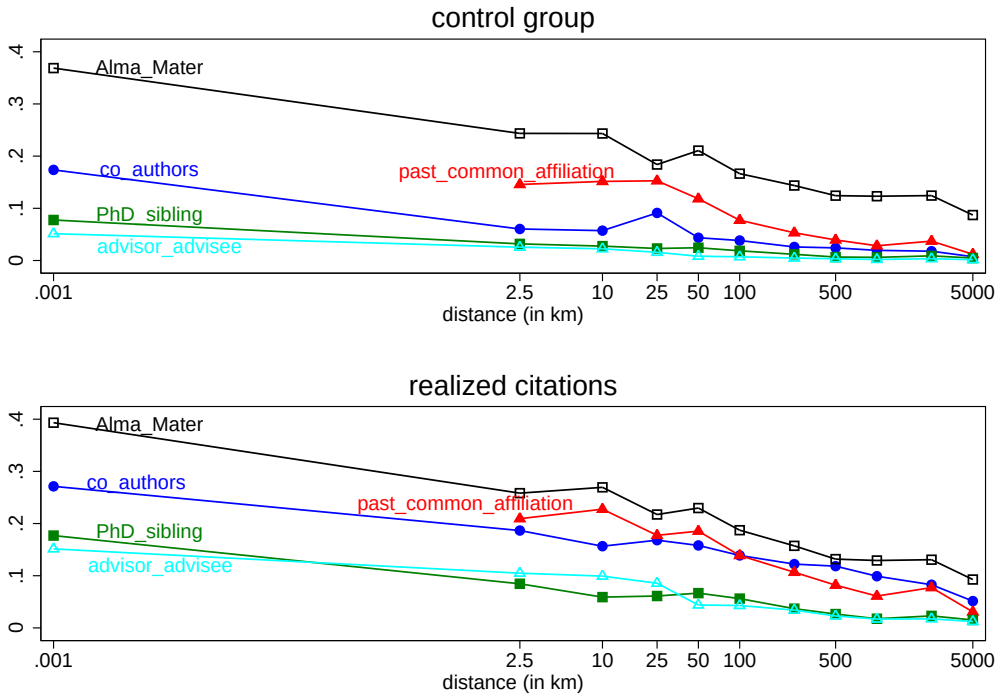
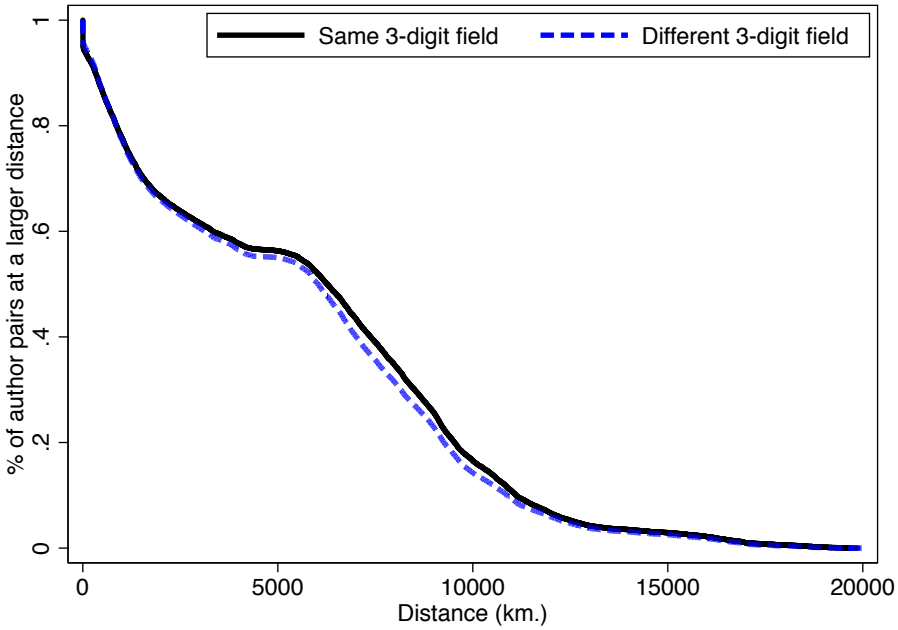


Figure 5 shows that each of our measures of linkages between mathematicians decrease on average with distance. We show in separate panels the distance relationships for our realized citations ($C_{it} = 1$) and our controls ($C_{it} = 0$). This upper panel can be thought of as a more random sample but, in any case, both panels tell the same story. They provide a graphical representation of the phenomenon already uncovered by the regressions: Ties and geography are strongly related. Excluding ties from the citation regression leads to large over-estimates of the partial correlation of knowledge flows and geographic separation.

On the other hand, Figure 6 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 since there are over 400 3-digit fields.

Figure 6: Distribution of distances between authors in same field (MSC-3d)



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 lack of "awareness" is not the impediment to citation; rather, it is lack of relevance. X cites Y rather than C (control) not because of the network connection between X and Y, but because Y's results are more *relevant* for X than those of C. In this story, the network connection was established as the *consequence* of the relatedness of X and Y's research. A goal of our empirical methodology will be to neutralize the relevance story to focus on the awareness channel.

Table 3 shows how the results vary as we use more stringent criteria for the control observation. 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 nine tie indicators.[17] For each specification we also estimate a version that removes the ties indicators altogether so as to see whether the match procedure affects 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. Column (3) reproduces column (4) from the

[17]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 3: Sensitivity of results to alternative controls for article relevance

	(1)	(2)	(3)	(4)	(5)	(6)
Control group:	nil	journal	MSC-3d		MSC-5d	keyword
Panel A: including ties						
Distance > 0	-1.060*	-0.919*	-0.786*	-0.792*	-0.559*	-0.263†
	(0.074)	(0.064)	(0.075)	(0.078)	(0.087)	(0.156)
ln dist $\mid$ dist > 0	-0.041*	-0.029*	-0.028*	-0.024*	-0.027*	-0.066*
	(0.006)	(0.006)	(0.008)	(0.008)	(0.009)	(0.017)
Different country	-0.076*	-0.032	-0.078*	-0.075*	-0.055	-0.156*
	(0.025)	(0.024)	(0.031)	(0.033)	(0.039)	(0.068)
Different language	-0.023	0.002	-0.062*	-0.057*	-0.026	-0.087
	(0.020)	(0.020)	(0.025)	(0.026)	(0.031)	(0.053)
Average effect of ties	2.336*	1.460*	0.871*	0.836*	0.569*	0.668*
	(0.087)	(0.032)	(0.026)	(0.027)	(0.024)	(0.048)
Cocitation				3.074*	2.239*	2.106*
				(0.056)	(0.055)	(0.142)
Observations	119628	118756	73004	73004	48940	18138
AIC	107342	110572	66789	60773	42057	13656
Panel B: excluding ties						
Distance > 0	-2.055*	-1.752*	-1.376*	-1.380*	-1.030*	-0.861*
	(0.066)	(0.058)	(0.069)	(0.071)	(0.081)	(0.144)
ln dist $\mid$ dist > 0	-0.078*	-0.060*	-0.059*	-0.053*	-0.049*	-0.088*
	(0.006)	(0.006)	(0.007)	(0.008)	(0.009)	(0.016)
Different country	-0.230*	-0.173*	-0.193*	-0.187*	-0.163*	-0.277*
	(0.023)	(0.023)	(0.030)	(0.031)	(0.038)	(0.065)
Different language	-0.102*	-0.060*	-0.116*	-0.113*	-0.061*	-0.114*
	(0.019)	(0.019)	(0.024)	(0.025)	(0.031)	(0.052)
Cocitation				3.116*	2.271*	2.146*
				(0.055)	(0.054)	(0.141)
Observations	119628	118756	73004	73004	48940	18138
AIC	120908	121420	70839	64337	43498	14257

Average effect of ties refer to the mean effect of 9 (3 WOS and 6 MGP) ties.

Standard errors in parentheses. Significance: † $p < 0.1$, * $p < 0.05$, * $p < 0.01$

previous table.

The results shown in specification (1) of table 3 make it clear that the matching of controls to cases is a very important element of the method. With random controls, the average coefficient on ties rises from 0.836 to 2.336. This means that the presence of a linkage goes from multiplying the odds of citation by 2.3 up to 10.3. This is 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 4.3).

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 22) and inclusion of the cocitation dummy lowers the estimated network effects. However, the reduction is minor (4%) and the network effects remain strong and statistically significant. In column (5) we change the data set by imposing that the control observation must be a paper in the same 5-digit field as the case. An example of 3 digit code is 15A, "basic linear algebra." Within that "inequalities involving eigenvalues and eigenvectors" is a 5-digit code. The cost of this tighter matching fit is that we now frequently fail to find a control observation—the sample falls to 48,940 observations. The effects of ties declines by almost a third but the effects are still estimated precisely. A comparison of the column (3) coefficients in the lower panel shows that distance effects still decline by about 50% when adding the controls for ties. Border and language effects shrink by even more and become insignificant. Hence all the main messages of the paper hold up when using 5-digit field controls. The final estimation of this table selects 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 is just a quarter of the same 3-digit sample. This attrition seems unacceptably high. It leads to a doubling of the average standard error for network effects and distance effects. It does not further decrease the network coefficient, suggesting to us that finer controls would not necessarily wipe out the estimated effects of ties. Indeed an unavoidable trade-off emerges between tighter matching restrictions and sample size. We view the 3-digit controls as hitting the "sweet spot."

Table 4: Influential papers are less impacted by ties and geography

	(1)	(2)	(3)	(4)
	Cites ≥ 40	(Top 5%)	Cites < 40	(Bottom 95%)
Distance > 0	-1.118*	-0.726*	-1.486*	-0.781*
	(0.110)	(0.116)	(0.089)	(0.099)
ln Dist $\mid$ Dist > 0	-0.050*	-0.024*	-0.065*	-0.032*
	(0.011)	(0.012)	(0.010)	(0.010)
Different country	-0.107*	-0.010	-0.259*	-0.136*
	(0.046)	(0.047)	(0.039)	(0.041)
Different language	-0.111*	-0.080*	-0.119*	-0.050
	(0.039)	(0.039)	(0.031)	(0.033)
Average effect of ties		0.628*		0.978*
		(0.047)		(0.033)
Observations	28176	28176	44788	44788
AIC	32834	31612	37969	35117

Average effect of ties refer to the mean effect of 9 (3 WOS and 6 MGP) ties.
Standard errors in parentheses. Significance: † $p < 0.1$, * $p < 0.05$, * $p < 0.01$

4.3 Subsamples and other robustness checks

Table 4 divides the sample into two categories based on the total amount of cites each paper received, something we condition on in all regressions. This specification investigates whether geographic separation and ties have different impacts on citations for more prominent papers. To make the comparison stark we define influential papers as those in the top 5% of the citation distribution. In our data, 40 cites is sufficient to meet this threshold. We find about a third smaller network effects for the more prominent papers. This is consistent with our view that networks facilitate awareness. Papers that are big successes require less help from networks to facilitate transmission. Interestingly the results that controlling for ties halves the distance effects continues to hold here for both the prominent and lesser known papers.

Table 5 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 most 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. Every key result holds up well in the non-US sample, except that the different language effects do not appear to be much reduced by inclusion of the network variables. Surprisingly the US-only sample does not exhibit strong continuous distance effects. However it shows huge reductions in citation odds from being at different institutions. Why intra-institutional bias would be stronger in the US is a mystery.

Table 5: Robustness

Sample:	(1) US only	(2) non-US	(3) average	(4) original geography	(5) available author	(6) available 5y window
Panel A: including ties						
Distance > 0	-0.964*	-0.592*	-0.717*	-0.590*	-0.611*	-0.665*
	(0.147)	(0.124)	(0.095)	(0.076)	(0.043)	(0.074)
ln Dist $\mid$ Dist > 0	-0.023	-0.045*	-0.029*	-0.037*	-0.031*	-0.035*
	(0.017)	(0.014)	(0.009)	(0.008)	(0.004)	(0.008)
Different country		-0.158*	-0.127*	-0.113*	-0.074*	-0.152*
		(0.071)	(0.035)	(0.031)	(0.017)	(0.031)
Different language		-0.101†	-0.080*	-0.027	-0.048*	-0.122*
		(0.052)	(0.027)	(0.025)	(0.014)	(0.025)
Average effect of ties	0.704*	0.913*	1.289*	0.877*	0.795*	0.900*
	(0.046)	(0.042)	(0.042)	(0.026)	(0.015)	(0.023)
Observations	12754	21836	73004	73004	236046	78798
AIC	10909	17471	66867	66867	236148	63879
Panel B: excluding ties						
Distance > 0	-1.163*	-1.177*	-1.431*	-1.248*	-1.150*	-1.238*
	(0.139)	(0.109)	(0.086)	(0.070)	(0.040)	(0.067)
ln Dist $\mid$ Dist > 0	-0.052*	-0.075*	-0.056*	-0.066*	-0.057*	-0.071*
	(0.016)	(0.013)	(0.008)	(0.007)	(0.004)	(0.007)
Different country		-0.455*	-0.290*	-0.219*	-0.168*	-0.242*
		(0.067)	(0.033)	(0.030)	(0.017)	(0.030)
Different language		-0.114*	-0.140*	-0.089*	-0.102*	-0.189*
		(0.050)	(0.026)	(0.024)	(0.013)	(0.024)
Observations	12754	21836	73004	73004	236046	78798
AIC	11656	18848	71353	71067	246811	68329

Average effect of ties refer to the mean effect of 9 (3 WOS and 6 MGP) ties.

Standard errors in parentheses. Significance: † $p < 0.1$, * $p < 0.05$, * $p < 0.01$

Column (3) replaces the min/max approach to aggregating geographic and network variables across co-authors with averages over all the author pairs. The continuous effect of distance is almost the same but the effects of ties become stronger. 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. Somewhat surprisingly this leads to stronger distance effects, though the effect of distance being positive is smaller. The contemporaneous geography used in the earlier specification leads to a better fit as measured by the Akaike Information Criterion (AIC).

Column (5) vastly increases the sample size by using observations that had previously been rejected because affiliation information was missing for some of the authors. Using any available author triples the sample but does not change the coefficients much. Finally, we take a subset of these observations based on the restriction that article i was published no more than five years after the article d that it cites. The main effect of this change is to raise both distance and network effects, suggesting that both are more important during the first few years after publication.

4.4 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. We would highlight the spread of email in the late 1980s, the rise of web browsers in the mid 1990s, and the introductions of the Google search engine and Google Scholar in 1998 and 2004. These technologies should have reduced the importance of face-to-face interactions as therefore we expected the absolute magnitudes of the geography variables to contract in the 1990s and 2000s.

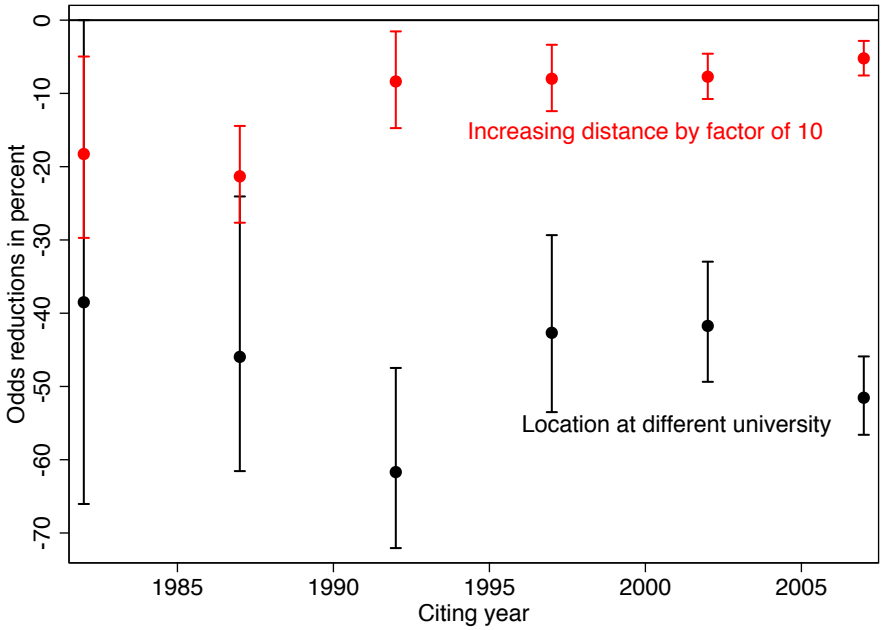
The estimates presented so far pool across a three-decade period (1980–2009). However the majority of the sample comes from the last six years. To look for evidence of technology effects on knowledge diffusion, we break the sample into six sub-periods.

Figure 7 shows the two-part distance effects generated from regressions done on five-year intervals based on the citing paper's year of publication. We use the same year/same 3-digit field criteria for the control sample. To boost sample size we include papers if we can obtain affiliation and MGP information for at least one citing and cited author.

More recent papers have a larger pool to cite from and on average cite from older papers. Table 5 shows that geography and network effects are stronger for "young" papers. To avoid the bias this might produce, we restrict the sample to citations (and controls) where the citing and cited years are no more than five years apart.

The figure depicts in black the percent reduction in the odds of citation from distance

Figure 7: Time-varying distance effects (with 1 std. err. bands)



changing from zero (same university) to greater than zero as well as (in red) the corresponding reduction attributable to a 10-fold increase in distance between citing and cited authors. To provide an idea of the precision for these effects, we show using brackets the percent reductions for based on shifting the coefficients by *one* standard deviation in each direction. We opt against showing confidence intervals here because they were just too wide in the early periods when sample sizes were small. Even though we restrict the sample to the 5-year windows, the number of observations for the later sub-periods are much larger, resulting in tighter error bands depicted in the figure.

It is difficult to discern clear evidence of technology effects in Figure 7. The last two decades show very stable effects of distance on citations despite the diffusion of all the technologies described above. It is possible that spread of email explains why the first decade has bigger distance effects but the confidence intervals are too wide to be conclusive. The finding of stable distance effects corroborates evidence from Börner et al. (2006) who examined citations between all papers published in the *Proceedings of the National Academy of Sciences* from 1982 to 2001. Considering 500 US institutions, the authors estimate a univariate citation-distance power law whose decay parameter changes little over the twenty years of their study. Our estimates control for ties but this affects only the level of the distance effect, not its evolution over time. The stability we see from 1990–2009 is maintained in (unreported) regressions that do not control for networks.

The results for not being at the same institution (distance > 0) do not offer much support for 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." Controlling for networks, not being at the same university leads to about a 50% reduction in citations in 2005–2009. This is not any smaller than it was in the 1980s. It should be noted 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 co-authors. Our results clarify that holding network linkages constant, being colleagues at the same university continues to greatly influence the likelihood of being aware of a relevant research contribution.

5 Conclusion

Information is not heavy or perishable, nor is it subject to freight costs or tariffs. Despite these characteristics, it does not appear to costlessly diffuse across the world. In this paper we corroborate prior work finding that geography separation (national borders, distance) impedes knowledge flows. However, after controlling for the ties that academics had established in the past, current geography has a much smaller partial correlation with citations. Instead, the cost of distance appears to be mediated in large part via network linkages.

References

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

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

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. *Working paper*.

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.

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.

Börner, K., Penumarthy, S., Meiss, M., and Ke, W. (2006). Mapping the diffusion of scholarly knowledge among major us research institutions. *Scientometrics*, 68(3):415–426.

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

Grossman, G. (1998). *Comment on Determinants of Bilateral Trade: Does Gravity Work in a Neoclassical World?* Chicago: University of Chicago Press.

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.

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

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

Krugman, P. R. (1991). *Geography and trade*. MIT press.

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

Newmen, M. (2004). Who is the best connected scientist? a study of scientific coauthorship networks [j]. *Lecture Notes in Physics*, 650:337–370.

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

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

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

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

Xie, Y. and Manski, C. F. (1989). The logit model and response-based samples. *Sociological Methods & Research*, 17(3):283–302.

A Monte carlo study of choice-based sampling (CBS) methods

Choice-based sampling (CBS) is also referred to as the case-control method (especially in epidemiology). Manski and Lerman (1977) report that CBS is generally not consistent. Fortunately, the exception is logit where, under the maintained error assumption, every parameter besides the constant is estimated consistently, and even the constant can be adjusted to obtain an unbiased estimate. Manski and Lerman (1977) suggest a method of weighted maximum likelihood (WML) that is consistent. Xie and Manski (1989) show WML-logit works better than CBS-logit when errors are normal instead of logistic.

Prior Monte-Carlo investigations have not considered the treatment of fixed effects in CBS models and have not allowed for error distributions other than normal and logistic. Here we report results from Monte Carlo experiments that feature fixed effects correlated with the right hand side variable of interest and we show results for Gumbel error terms.

We investigate the following data generating process:

$$r_{id} = 1 + x_{id} + \alpha_d + \varepsilon_{id}$$

We assume citations occur if and only if $r_{id} > 0$. Therefore the probability of a citation is given by

$$\mathbb{P}(C_{id} = 1) = \mathbb{P}(-\varepsilon_{id} - \alpha_d < 1 + x_{id})$$

For α_d constant and ε distributed logistically or normally these probabilities correspond to $\mathbb{P}(C_{id}=1)=\Lambda((1+x_{it}-\mu_L)/\sigma_L)$ and $\mathbb{P}(C_{id}=1)=\Phi((1+x_{it}-\mu_N)/\sigma_N)$, where $\Lambda()$ and $\Phi()$ are the standardized logistic and normal CDFs.

The rows of table 6 correspond to different data generating processes (DGP). There are three different assumptions about the error term, first the familiar logistic and normal and then the Gumbel distribution, another commonly used bell-shaped distribution. Unlike the first two, the Gumbel is not symmetric. A second source of differences across rows is based on the presence or absence of article fixed effects in the DGP. This corresponds to whether the variance of α_d is assumed to be 0 or 0.5.

The columns of table 6 correspond to different estimation choices. At the highest level, the table is divided by whether we estimate the model on the full set of randomly generated data or on the choice-based sample. In the former case there are 100 potentially citing papers for each 100 potentially cited papers, yielding a total sample size of 10,000. In the choice based sample, we keep one randomly selected potential (but not actual) citing paper for each actual citing paper. Thus if there N_1 cases of $C_{id}=1$, the choice-based sample has size $2N_1$.

The columns also differ in terms of whether the econometric method conditions on article fixed effects ("xtlogit, fe" in Stata) and whether weights are used following Manski and Lerman (1977) and Xie and Manski (1989).

Table 6: Means of 1000 repetitions for logit-based estimators

Error Distbn.	Std. Dev. α	"Population"		Choice-based sample		
		(1)	(2)	(3)	(4)	(5)
Article FEs?		no	yes	no	no	yes
Choice weights?		—	—	no	yes	no
Logistic	0	1.001	1.001	1.008	1.017	1.008
(std. error)		(0.066)	(0.067)	(0.108)	(0.133)	(0.112)
	0.5	1.101	1.002	0.966	0.974	1.007
Normal	0	1.222	1.226	1.312	1.248	1.316
	0.5	1.268	1.164	1.192	1.127	1.213
Gumbel	0	1.036	1.035	1.055	1.031	1.030
	0.5	1.162	1.062	1.002	1.039	1.049

Key findings from the Monte Carlo exercise reported in Table 6:

1. Logit on the choice-based sample (CBS) estimates the coefficient on x correctly. Although the constant is dramatically upward biased, it can easily be corrected.

2. The weighted Logit estimator does better under normally distributed errors, as shown in asymptotic bias calculations and Monte Carlo exercise of Xie and Manski

(1989).

3. Surprisingly, it does not perform notably well with Gumbel errors.

4. Article fixed effects can be conditioned out using Stata's xtlogit, fe command. This leads to consistent estimates when the error is logistic.

5. The weighted logit does not appear to be applicable to the fixed effects logit because Stata requires that the weights be constant within groups, whereas weighted logit varies the weight according to whether an observation is case ($C_{id} = 1$) or a control ($C_{id} = 0$).

6. When there are two problems—article fixed effects (α_d) correlated with the explanatory variable (x_{id}) and normal instead of logistic errors—there are competing biases. Weighted logit does better but fixed effects logit is not that bad. When there are Gumbel errors, fixed effects logit actually outperforms weighted logit without fixed effects.