Centre for Economic Performance

Discussion Paper

ISSN 2042-2695

No.1731
November 2020

No inventor is an island: social connectedness and the geography of knowledge flows in the US

Andreas Diemer
Tanner Regan

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

30 YEARS CEP

Economic and Social Research Council

Abstract
Do informal social ties connecting inventors across distant places promote knowledge flows between them? To measure informal ties, we use a new and direct index of social connectedness of regions based on aggregate Facebook friendships. We use a well-established identification strategy that relies on matching inventor citations with citations from examiners. Moreover, we isolate the specific effect of informal connections, above and beyond formal professional ties (co-inventor networks) and geographic proximity. We identify a significant and robust effect of informal ties on patent citation. Further, we find that the effect of geographic proximity on knowledge flows is entirely explained by informal social ties and professional networks. We also show that the effect of informal social ties on knowledge flows: has become increasingly important over the last two decades, is higher for older or `forgotten' patents, is more important for new entrepreneurs or `garage inventors', and is somewhat stronger across distant technology fields.

Key words: knowledge flows, diffusion, social connectedness, informal networks
JEL codes: O33; R12; Z13

This paper was produced as part of the Centre's Urban Programme. The Centre for Economic Performance is financed by the Economic and Social Research Council.

Acknowledgements
We thank Adam Jaffe, Francesco Lissoni, Steve Gibbons, Marco Percoco, Olmo Silva, Felipe Carozzi, Riccardo Crescenzi, Enrico Berkes, and Jorge Perez Perez for fruitful comments and discussions. We also thank seminar participants at the 5th Geography of Innovation Conference in Stavanger, RSA North America Conference in Montreal, 2nd Regional and Urban Economics Workshop in Bogota, 9th European Meeting of the Urban Economics Association in Amsterdam, 6th Workshop in Economics of Innovation, Complexity and Knowledge in Turin, and the LSE Economic Geography Work in Progress Series. We are grateful to Michael Bailey at Facebook for early access to the Social Connectedness data. Andreas Diemer gratefully acknowledges financial support from the UK Economic and Social Research Council (ESRC).

Andreas Diemer, SOFI, Stockholm University. Tanner Regan, Centre for Economic Performance, London School of Economics and London Business School.

Published by
Centre for Economic Performance
London School of Economics and Political Science
Houghton Street
London WC2A 2AE

1 Introduction

Do inventors learn from the informal context that surrounds them? This paper empirically examines the role of social connectedness in the diffusion of knowledge among agents located across distant geographies. Social connectedness is conceptualised as the overall informal social environment of an agent, measured by the aggregate ties connecting the agent's neighbourhood to other neighbourhoods, net of her formal, professional, networks. The research question we address, therefore, is whether stronger informal social ties to other places can foster knowledge exchange with these places, above and beyond what would be explained by professional channels or by simple geographic proximity. While the paper is conceptually interested in the general case of knowledge flows, the empirical analysis focuses on patent citations. Citations provide a powerful measure of economically relevant knowledge exchange, otherwise difficult to observe in different settings. Moreover, they speak to the process of innovation and technological change, which is a key determinant of long run economic growth (Romer, 1986, 1990; Lucas, 1988; Aghion and Howitt, 1992).

This research relates to an old question in economics that considers the role of localised knowledge spillovers in promoting the agglomeration of people and industries in space (Marshall, 1890). As individuals come together and interact, they learn from each other and become more productive (Glaeser, 1999). Local knowledge exchange, or learning, is in fact one of the key drivers of urban agglomeration externalities (Duranton and Puga, 2004). With respect to innovation, the sharing and recombination of existing ideas in dense urban environments supports the creation of more ideas (Carlino et al., 2007). A large body of empirical research has attempted to validate the notion of knowledge spillovers, frequently using patent data and patent citations to measure innovation and knowledge transfers. In keeping with the notion of agglomeration, these studies typically focus on the geographical dimension of spillovers (Jaffe et al., 1993; Audretsch and Feldman, 1996; Thompson and Fox-Kean, 2005; Murata et al., 2013).[1] Yet numerous papers emphasise that the mechanisms underlying the spread of knowledge, whether intentional or unintentional, rely on interaction of people over networks (Saxenian, 1996; Bala and Goyal, 2000; Feldman, 2002; Powell and Grodal, 2005; Henderson, 2007). Throughout the paper, we refer to this process as *social learning*. Social learning, then, is geographically local only to the extent that physical proximity shapes the quantity and quality of social connections (Breschi and Lissoni, 2001; Storper and Venables, 2004). Breschi and Lissoni (2009), in particular, show that the localisation of patent citations is largely determined by the limited geographical reach of inventors' inter-firm mobility.

The increasing availability of data on networks and interaction has thus spurred a growing

[1] See Audretsch and Feldman (2004) and Carlino and Kerr (2015) for comprehensive reviews.

empirical literature that evaluates the role of social ties in the exchange of knowledge. Some papers consider inventor networks constructed from data on co-patenting (Agrawal et al., 2006; Breschi and Lissoni, 2009), others examine social proximity measures inferred from belonging to similar ethnic groups (Agrawal et al., 2008; Kerr, 2008). Our efforts focus on the role of informal connections across places, above and beyond what would be explained by professional relationships (co-inventor networks). We exploit a new broad and direct measure of social connectedness across counties in the US based on aggregate counts of the universe of online friendships on Facebook, a popular social media platform (Bailey et al., 2018b). These new data allow us for the first time to study knowledge flows over informal networks on a large scale without relying on indirect proxies.

Our empirical analysis uncovers two new findings on the importance of informal networks in knowledge flows. First, We identify a significant and robust effect of the social proximity of places on their propensity to cite one another. This is independent of geographical distance or professional linkages between inventors and takes into account the endogenous location of relevant knowledge due to the pre-existing geography of production. According to our preferred estimate, two counties at the 75th percentile of social connectedness are on average 1.1 percentage points more likely to cite one another than a pair of counties at the 25th percentile.[2] Second, we show that, conditional on social connectedness, physical proximity (across regions, not within) has no economically or statistically significant effect on knowledge flows. This is not true for professional ties, which maintain a large and significant effect on knowledge flows even conditional on social and geographical proximity. All of these findings are robust to inclusion of a long list of bilateral controls, patent and citation fixed effects, and a series of conservative robustness checks.

Next we examine margins of heterogeneity in the effect of social connectedness on knowledge flows. Our analysis uncovers several new findings on the ways in which social connectedness matters for knowledge flows. First, we find that the relevance of informal social ties has been increasing over time since the early 2000s. Second, we find no evidence that geography and social proximity act as substitutes; social connectedness is similarly strong across counties within a commuting zone as it is across states. Third, we find that geographical distance is important for young patents (<3 years), while social connectedness is more important for older (>3 years) patents. Fourth, we find that social proximity is most important for entrepreneurs or ‘garage inventors’; social connectedness is strongest among young assignees (<2 years). Fifth, we find some evidence, albeit noisy, that social connectedness is increasingly important for patents that are more technologically distant.

[2] While this result may appear somewhat abstract, a look at the data reveals that two otherwise neighbouring counties may occasionally display such a difference in connectedness strength with the same third county. The relationship we document is thus economically meaningful as it can potentially be achieved with limited geographical mobility.

This work dialogues with three main strands of literature. First, it speaks to research on urban economics and agglomeration economies by outlining micro-level channels by which learning might occur, and specifically how social connectedness might provide non-agglomerative mechanisms for transmission of knowledge unrestricted to physically proximate agents. It also emphasises how this type of knowledge tends be technologically more distant, which offers a new perspective on debates about specialised and diversified industrial clusters, opening to new research questions about complementarity and substitution between internal and external sources of knowledge. Second, this analysis contributes to the innovation literature pioneered by Jaffe et al. (1993), hereafter JTH, which relies on patent citations to capture the geographic localisation of knowledge exchange and spillover of ideas. The use of data on social connectedness in this paper allows to study informal social interactions at an unprecedented spatial scale. In fact, imposing an a priori spatial boundary to social learning would seem excessively restrictive considering the tremendous progress in ICT and the fall in travel costs observed in the past few decades. This work thus examines the conditions under which social learning might occur independently of geographical constraints, particularly beyond the local level. Finally, this paper contributes to the growing scholarship on the role of social networks in the innovation process, by explicitly looking at informal social connections defined as the broader social environment to which inventors are exposed in their daily work. There is ample scholarship documenting the importance of such ties among scientists and inventors. The extant literature, however, emphasises the importance of professional ties over informal ones (Breschi and Lissoni, 2009). Yet limiting the analysis to professional networks draws a potentially incomplete picture due to the likely discipline-biased nature of such ties, which tend to convey specialised knowledge. By contrast, it is possible that informal types of connections lead to an entirely different type of knowledge exchange, due precisely to their more diverse composition. This distinction reminds conceptually of Granovetter (1973, 1983)'s 'strength of weak ties' hypothesis.[3] However, research linking informal networks to innovation dynamics is scant (Powell and Grodal, 2005). A notable exception is the work of Bailey et al. (2018b), who rely on the same data used in this paper to explore empirical correlations of social connectedness with a broad set of outcomes, including patent citations. Their analysis relies on the case-control matching strategy by JTH, finding that connectedness positively correlates with innovative activity and knowledge flows. This paper aims to complement their work in several ways. It focuses exclusively on the learning outcome, carefully conceptualising the underlying relationship notably with respect to informal ties and light conveyors. It also improves the estimation framework by attempting to identify the causal effect of connectedness on knowledge flows using examiner added citations as a control group, and by controlling

[3] According to this notion, it is more distant relationships (acquaintances, and friends of friends) that convey the most novel and valuable type of knowledge.

for inventor mobility, their professional networks, and other confounding channels. Finally, we document several important dimensions of heterogeneity in the effects of social connectedness that are consistent with the conceptual discussion.

The rest of the paper is organised as follows. Section 2 frames the problem conceptually. Section 3 discusses the data and the empirical methods. Section 4 presents the results of the analysis. Section 5 concludes highlighting limitations and future work.

2 Conceptual Framework

The conceptual discussion of knowledge flows over social networks that follows focuses on the kind of scientific and technical knowledge found in patents (ideas or inventions).[4] The reasons for this are twofold. First, patents embody knowledge that is economically relevant and that determines, at least by some approximation, the rate of innovation, productivity, and growth of an economy.[5] There is therefore an economic interest in studying this particular kind of knowledge. Second, patent citations 'leave a trail', allowing to track flows of knowledge which are otherwise notoriously difficult to observe. Patent citations are thus the empirical proxy for learning used in this analysis.

How might social interaction affect the flow of ideas and technological knowledge, as captured by patent citations? Three distinct mechanisms come to mind, which we generally refer to as mobility, meetings, and exposure. Among these, we distinguish between heavy and light knowledge conveyors in the process of social learning. Heavy knowledge conveyors are associated with interaction of inventors with colleagues in professional networks, or with geographic mobility of inventors themselves. Light conveyors, by contrast, are related to interaction in informal networks, and refer to less structured channels such as chance meetings, referrals, perceptions, and salience of market opportunities. Table 1 gives an overview.

Table 1: Overview of possible mechanisms for the effects of social learning

	Mobility	Meetings	Exposure
Heavy	Endogenous inventors' location	Endogenous inventors' networks	N/A
Light	N/A	Chance meetings and referrals	Salience of market opportunities

The distinction between heavy and light conveyors is important because, in our empirical

[4] A systematic survey and discussion of possible transfer mechanisms over social networks for different types of knowledge falls beyond the scope of this paper.

[5] A growing literature in macroeconomics discusses endogenous growth models that are micro-founded onto the notion that social interaction spurs knowledge diffusion and innovation (Comin et al., 2012; Lucas and Moll, 2014; Akcigit et al., 2018; Buera and Oberfield, 2020).

framework, we are interested in isolating the effect of the latter, which we argue operates through informal networks. We interpret informal networks in a broad sense, as the social environment in which inventors work, net of their professional ties (see Section 3.1.1 below for further details). In what follows, we refer to this concept simply as 'social connectedness'. The broad measure of social connectedness we adopt, however, is potentially driving both heavy and light conveyors of knowledge. For instance, social networks are known to correlate with labour mobility (Buechel et al., 2019). In line with the findings of Breschi and Lissoni (2009), mobile inventors carry ideas with them as they move across firms and places.[6] Social connectedness might matter, then, to the extent that it favours inventor mobility and influences their choice of location. It is also possible that social connectedness determines professional networks (typically defined empirically as networks of co-inventors), and therefore technical collaboration networks can be endogenous to one's informal social network. Powell and Grodal (2005) refer to such ties as 'emergent networks', that is, unintentional networks that develop on the grounds of ongoing relationships of a different nature (friendship ties, common ethnicity, co-location or reoccurring meetings).[7] Professional networks are of paramount importance in innovation, as patents embody technical, often discipline-specific, ideas that require prior knowledge to be absorbed (Cohen and Levinthal, 1990).

The channels outlined so far point to relatively well-specified ways in which social connectedness can affect citation probability through heavy knowledge conveyors. However, interpreted this way, any observed impact of informal networks would effectively be nothing more but a reduced-form empirical correlation of limited interest if one can readily observe inventor collaboration networks or inventor mobility. In fact, the correlation should disappear once controlling for these variables (a task we take up in our empirical model). Is there, at least conceptually, a residual role for social connectedness to influence the flow of technical knowledge through lighter channels?

One light channel linking social connectedness to patent citations is through 'Meetings' (Table 1 column 2). Light meetings, unlike heavy meetings, refer to *chance* meetings and referrals. Research is increasingly considering the importance of serendipitous encounters in directing inventive activity and knowledge exchange. Catalini (2018) studies the exogenous reallocation of university researchers due to building renovation. Atkin et al. (2020) rely on cell-phone data to uncover the effect of unplanned meetings between workers from different firms on the propensity of these firms to cite each other's patents. Roche (2019) shows that chance interactions promoted by better connecting local road networks foster serendipitous knowledge exchanges within neighbourhoods, which can explain differences

[6] Lissoni (2018) provides a recent discussion with respect to international mobility and migration.

[7] Whilst not directly focusing on informal networks, Crescenzi et al. (2016) and Crescenzi et al. (2017) do show that social proximity in co-invention networks influences the probability of forming a collaboration in the future.

in their innovative performance. With respect to access to external sources of knowledge, intuitively, the probability of chance meetings occurring between individuals from different places increases in the number of ties connecting these places. In practice, this could happen through visits to distant friends, or even digitally via interaction on social media and online communication platforms.

Another light channel linking social connectedness to patent citations is through 'Exposure' (Table 1 column 3). Preliminary evidence emerges from a survey of inventors carried out by Jaffe et al. (2002), which aimed to shed light onto the black-box process of idea exchange in technical and scientific fields. The authors find that, asked about what factors had a significant influence on the development of their inventions, almost 60% of respondents cited the 'awareness of a commercial opportunity' while another 20% mentioned 'word of mouth or personal interaction'. Notably, 'joint work with others' was only mentioned by less than 15% of respondents. Moreover, 'word of mouth' and 'viewed a presentation or demonstration' also accounted for more than 30% of responses when asking about how citing inventors learned about the previous patent, compared to only about 5% of inventors who cited 'direct communication with the inventor'.[8] Taken together, these qualitative findings suggest that there might be something related to *salience* of ideas and identification of market opportunities in the process of scientific learning. This channel is not necessarily technical in nature nor is its scope confined to professional connections. More concretely, exposure-induced learning could be driven by preferences on the demand side (determining market opportunities for ideas both for consumers or downstream firms) or through awareness of supply side technological opportunities via knowledge of different but related products, solutions or applications prevailing in the (possibly geographically distant but) socially connected market. This intuition is taken up in the work of Breschi and Lenzi (2016), who, although focusing on professional ties between inventors, emphasise the importance of allowing for social connections between different cities as a means to "enriching and renewing a city's knowledge base by facilitating access to fresh external knowledge" (p.66). More recently, Akcigit et al. (2018) also emphasise the pernicious effects of restricted access to external knowledge, which can limit innovation productivity. This is due to the 'proximity paradox', whereby the absence of inflow of new ideas from interaction across clusters results in too much specialisation, cognitive lock-in, and lower idea quality (Miguélez and Moreno, 2015). The conceptual argument made in this paper is similar. The emphasis however is on latent knowledge embedded in informal connections, or 'knowledge in the air' as originally conceptualised by Marshall (1890). According to the proposition of this paper, ideas are not

[8] These figures are particularly high considering that the survey could not distinguish between citations made by the applicant from those included by the patent examiner during the review process. The frequency of inventors who answered '[learned] during patent application process' and 'never before now' (about 60% in total) is consistent with the average incidence of examiner-added citations (about 60% of all patent citations, according to the data used in this paper).

only channelled through one specific network connection but rather permeate the broader informal social context in which innovation occurs. Accordingly, our empirical analysis will attempt to isolate the effect of social connectedness on knowledge flows via light conveyors, as opposed to its influence through heavy channels such as inventor networks and mobility.

3 Data and Empirical Methods

3.1 Variable Definition and Measurement

This paper relies on two main sources of data. Social connectedness is measured using information on friendship links on Facebook, a popular social media platform. Knowledge flows are proxied using patent citation data. Additionally, the analysis also uses data from the 2010 US Decennial Census and the Internal Revenue Service (IRS). What follows gives details on how the key variables of interest are defined.

3.1.1 Informal Social Networks: the Social Connectedness Index

The proposed measure of informal social networks, or social connectedness, relies on an index developed by Bailey et al. (2018b): the Social Connectedness Index (SCI). This index essentially captures the social graph for the universe of *active* US Facebook users as of April 2016, aggregated up to the level of counties.[9] Users are deemed active if they interacted with Facebook in the 30 days prior to the April 2016 snapshot. Geographic location is assigned using the IP address from which users login most frequently. For all users i and j and for each pair of counties c and k, the index is constructed as:

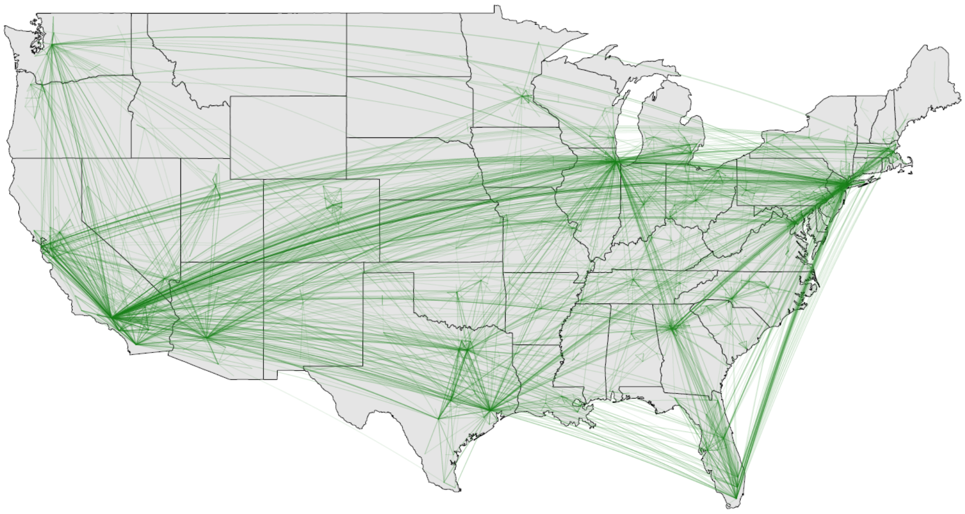
$$SCI_{ck} = \sum_{i \neq j} \sum_{j} \mathbb{1}_{ij}, \text{ for } i \in c \text{ and } j \in k$$

Where $\mathbb{1}_{ij}$ is an indicator variable that takes the value of 1 if two users are friends with each other, and 0 otherwise. Due to confidentiality concerns, Facebook only releases a re-scaled version of these data. The index thus ranges between 0 and 1,000,000, the highest observed value, which is assigned to connections of Los Angeles County to itself (i.e., friendships within the county). The result is a weighted social graph consisting of 3,136 nodes and 9,462,485 edges. Figure 1 visualises the top one percent of edges in the data, assigning darker colours and thicker lines to stronger connections. The concentration of social ties between counties hosting the largest cities in the US is evident.

Nevertheless, there are limitations in the use of the SCI to capture real-life ties. The

[9] In principle it would be more accurate to refer to Facebook *accounts* rather than *users*. However, the same expression as in Bailey et al. (2018b) is used here for consistency.

Figure 1: Top one percent of social connections across US counties



geography of connectedness might be measured imprecisely to the extent that Facebook users do not represent the average American. Because friends are typically added on Facebook rather than deleted, it is also possible that this measure overestimates real-life interaction between people and places, a concern only partly mitigated by the fact that Facebook imposes a limit of 5,000 friendships on personal accounts. However, erroneous measurement is unlikely to bias estimates unless there are reasons to believe that this error is correlated with the outcome of interest.[10]

Another important concern is that using the SCI involves a loss of precision in the measurement of relevant informal social networks, insofar as they are imputed to each inventor on the basis of their neighbourhood, rather than their actual social ties. We conceptualise informal networks as those broad-based relationships individuals entertain in their

[10] Unfortunately, Facebook does not release covariates for these data. However, it is possible to gauge some descriptive facts from secondary sources. At the time the data were extracted, there were over 220 million active monthly Facebook users in the United States and Canada (according to Facebook's 2016 quarterly results report). A Pew Research Center study published in that same year estimates that about 70% of US adults (aged 18 or more) used the social media platform (Greenwood et al., 2016). Women, younger individuals (aged 50 or less), college educated and relatively poorer adults were slightly overrepresented, albeit by small margins. Most Facebook friendships are with people with whom users have ongoing interaction in real life. According to Hampton et al. (2011), ties between Facebook users tend to occur among high school or college peers (31%), immediate or extended family members (20%), co-workers (10%), and neighbours or acquaintances (9%). The remaining ties are with friends of friends, or 'dormant relationships', that may become useful to users in the future. However, only 3% of Facebook friendships are with someone the user has never met in person. Moreover, several studies have shown that Facebook ties are good predictors of real life friendships and friendship strength (Gilbert and Karahalios, 2009; Jones et al., 2013). All this suggest that there is strong potential in these data to be used to study social relationships on a large-scale (Bailey et al., 2018b). A growing literature documents the relevance of these data for explaining socio-economic outcomes, further validating its use in this analysis. Bailey et al. (2020a) consider social interactions in urban areas, Bailey et al. (2020d) examine the European case. Other research considers housing markets (Bailey et al., 2018a, 2019), product adoption (Bailey et al., 2020c), trade and investment flows (Bailey et al., 2020b; Kuchler et al., 2020a; Bali et al., 2019), EITC claiming behaviour (Wilson, 2020), bank lending (Rehbein and Rother, 2019), and the spread of COVID-19 (Kuchler et al., 2020b; Milani, 2020).

personal life beyond work. These include family and friends, but also extend to relationships beyond this inner circle of connections. There are two main ways to define such informal networks empirically. One way is to directly look at each agent's ties (social ties proper, or *interpersonal* networks), restricting these to non-professional relationships (professional ties in this application are inventors' co-patenting networks). Another way is to think of informal connections more generally as the social environment characterising the neighbourhood in which an individual lives or works (*neighbourhood* networks). We adopt the latter definition, which emphasises the value of weak ties (Granovetter, 1973, 1983). The composition of this broader social environment is the aggregate result of choices made by many individuals over many time periods, and therefore represents a potentially richer and more diverse source of knowledge and ideas than strong ties such as family and close friends. Ultimately, neighbourhood ties are simply interpersonal ties aggregated for all agents residing in a given spatial unit. This distinction however matters for at least two reasons. First, even though interpersonal and neighbourhood networks are likely to overlap (most people have friends that live geographically close), some agents in neighbourhood networks may never appear in interpersonal networks (even when considering high-degrees of separation), or enter at a social distance so high that interpersonal networks seem unlikely to matter more than the fact that the same contact can be established due to exposure to the same overall social environment. Second, neighbourhood networks can be considered to be time-invariant over a sufficiently large area and a sufficiently small period of time, due precisely to their aggregate and historically-determined nature. This mitigates endogeneity concerns in the definition of this variable. Moreover, by looking at the overall social environment in which inventors operate this measurement of informal social networks is faithful to Marshall (1890)'s original conceptualisation of spillovers as arising from knowledge 'as it were in the air'.

Importantly, the assumption that the SCI captures informal connections relies on the ability to separately account for interpersonal professional connections, which extant literature finds to significantly influence the exchange of technical knowledge. It is otherwise possible that that social connectedness simply picks up a very noisy estimate of professional ties among inventors. The empirical measurement of such connections is discussed jointly with the patent citation data below.

3.1.2 Knowledge Flows and Professional Networks: USPTO PatentsView

Contrary to the claim that "Knowledge flows [...] leave no paper trail" (Krugman, 1991, p. 53), Jaffe et al. (1993) argue that in fact they sometimes do, for instance in the form of patent citations. Following this intuition, this analysis relies on patent data released by the United States Patent and Trademark Office (USPTO) to measure knowledge transfers. In particular, the USPTO's PatentsView platform offers access to large structured data on over 40 years of patents and patent citations from 1976 until today. From 2001 onwards, these data also include valuable information on who made the citation, the patent's applicant or its examiner. As discussed in Section 3.2, this information is at the core of the proposed identification strategy. There are well known limitations to the use of patent citations as a measure of knowledge flows (Pavitt, 1985; Griliches, 1998; Bessen, 2008). Firstly, patenting is selective, meaning that not all ideas or inventions are observed. In order to be patented, an invention needs to be novel, non trivial and commercially viable. Very often, these criteria make it easier to patent inventions in manufacturing-related activities rather than services, and there is bias within manufacturing industries too. It entails that patents typically represent outcomes of applied, rather than basic research. There is also a strategic component to patenting. Obtaining and maintaining a patent is costly, so that it is likely that only the most valuable ideas are filed for intellectual protection. Similarly, some firms may prefer to maintain their invention entirely secret. Finally, patents necessarily represent a form of knowledge that is relatively structured and that can be codified. This means that the more tacit kinds of knowledge are not captured by this measure. Arguably, however, tacit knowledge is also the kind for which social ties, interpersonal communication and face-to-face contact matter the most.[11]

With these caveats in mind, what follows describes the construction of the estimating sample. We begin by taking the population of citations sent by patents issued in the 2002-2019 period. Each citing and cited patent is mapped onto US counties using the mode of the location of listed inventors residing in the US, breaking ties randomly.[12] We prefer the use of inventor location, rather than the assignee's. Lychagin et al. (2016) show that the geographic location of a firm's researchers better explains cross-firm spillovers than that firm's establishment location. For each citation, we retain its source, whether it was added by the applicant or by the examiner. We then merge in all available patent and county level information, such as issue and application years, technology fields, links over inventors' networks, bilateral geographical distances, social connectedness, and a set of controls based on the 2010 US Census.[13] Technological fields are based on the International Patent Classification (IPC), which provides a hierarchical system of codes.[14] We consider IPC classes (3-digit) and subclasses (4-digit), henceforth IPC3 and IPC4 classes respectively. Moreover, based on this classification, the World Intellectual Property Organisation (WIPO) provides a list of fields that have the advantage of being largely mutually exclusive, with adequate level of differentiation, and appropriate within-field homogeneity (Schmoch, 2008). While a single patent could be associated with multiple

[11] Provided there is sufficient absorptive capacity, especially relevant in the case of technical knowledge.

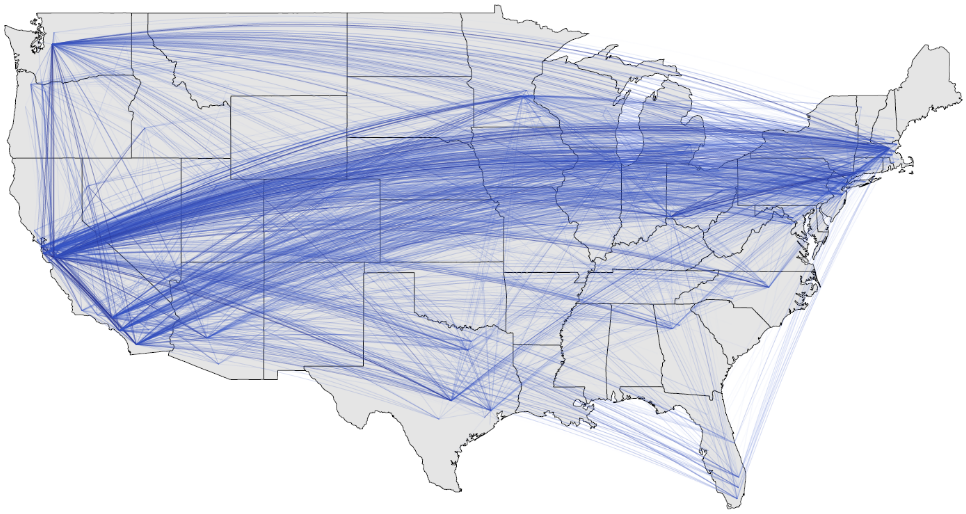
[12] In earlier results, not reported herein, the mapping was also carried out using the location of the first inventor for whom this information was available, with no substantial change in findings.

[13] Appendix Table B.6 provides a list of all variables.

[14] Detailed information on this system is available at this link: `https://www.wipo.int/classifications/ipc/en/`

IPC classes or subclasses, in the vast majority of cases there is only one WIPO field.[15] A complete list is available in Appendix Table B.1. The first listed IPC3 and IPC4 classes are also retained, for robustness checks in the empirical analysis. Some sampling restrictions are applied. First, only national flows of knowledge are considered. Citing and cited patents with no inventors residing in the US at the time the patent was issued are thus dropped. Moreover, citations originating from or received by patents located outside continental US states are also dropped. The sample is then restricted to citing patents whose elapsed time between application and issue date was below the 95th percentile in the distribution because of concerns of unobserved heterogeneity in the top 5% group. Similarly, we drop cited patents whose elapsed time to citation (their 'age' at the time of citing, using differences in application dates) was above the 95th percentile in the distribution. Finally, we restrict our attention to citations originating from patents issued after 2016, as this is the date when the social graph of Facebook was extracted.

Figure 2: Top one percent of knowledge flows (citations) across US counties



The resulting estimating sample consists of 489,230 citing patents and 11,349,396 citations, of which on average about 60% were added by the applicant. Appendix Tables B.2 and B.4 offer descriptive details for citing and cited patents. A large number of patents had all citations made by the applicant (29%) or all citations made by the examiner (22%). This is in line with previous findings (Thompson, 2006; Alcácer and Gittelman, 2006; Alcácer et al., 2009). Still, one might worry that these extreme value patents could bias the analysis. Thus, robustness checks will show that results are unchanged even when these patents are dropped from the sample. Another concern relates to the fact that inventors might cite other patents whose assignee is the same.[16] These citations do

[15] In the few exceptions, the field most frequently associated the listed IPC classes is retained.

[16] An assignee is the legal person to whom ownership of the patent is granted, typically a firm, a university, or another organisation. PatentsView provides assignee-disambiguation. For details, please refer to this webpage: http://www.patentsview.org/community/methods-and-sources.

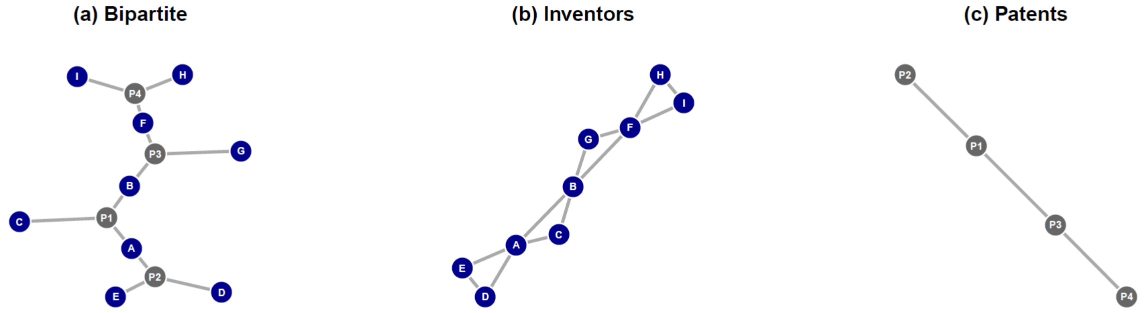
not strictly speaking represent knowledge spillovers since they occur within the boundaries of the same organisation, and are less interesting for the case of knowledge flows. Self-citations so defined represent about 10% of citations in the sample. They are not used in the analysis. Figure 2 maps the top one percent of knowledge flows in the data (aggregate bilateral citation counts for county pairs, irrespective of their direction), assigning darker colours and thicker lines to larger flows. There is a striking overlap between the flows represented in this map, and the social connections in Figure 1. An alternative way of visualising this relationship is proposed in Appendix Figures A.1 and A.2.

Finally, the professional network of inventors is measured in line with existing empirical literature as a co-inventor, or co-patenting, network. To obtain the network, this methodology crucially relies on inventor name disambiguation. Luckily, PatentsView data feature disambiguated inventor identifiers obtained through a discriminative hierarchical coreference algorithm proposed by Nicholas Monath and Andrew McCallum from University of Massachusetts Amherst.[17] We rely entirely on these data and do not attempt to disambiguate inventor names in alternative ways. Using the unique identifiers for *all* listed inventors (not just those located in the US), we construct a dummy for professional networks indicating whether citation patent pairs had a common inventor (self-citation), whether they shared a co-inventor (first-degree connection), and whether they shared the co-inventor of a co-inventor (second-degree connection). Figure 3 illustrates this network. We begin with a bipartite representation of the data in panel (a), where each inventor (blue nodes) is linked to patents (grey nodes). This graph can be converted to a one-mode projection for inventors (panel b), showing co-patenting relationships. In this example, A and B are connected by a first degree tie due to the common authorship of patent P1. F is the co-author of B, a co-author to A, meaning he or she shares a second degree connection with A. Finally, projecting the graph in (a) by patents allows to track whether a citation falls within the inventors' network. As shown in panel (c), patent P1 is connected to P2 and P3 due to a common inventor ('degree zero'). Patent P1 is also connected to P4 via a co-author, F, while patent P2 is linked to P4 via a second degree connection due to F being the co-author of B who is co-author of A.

Just under 20% of citations in the data are linked by a professional connection. Unfortunately, it was not computationally feasible to build higher order network links. Reassuringly, however, Breschi and Lissoni (2009) document that the effect of inventor networks on patent citations drops sharply in the degree of social distance.

[17] Details on the discriminative hierarchical coreference algorithm are available at this webpage: http://www.patentsview.org/data/presentations/UMassInventorDisambiguation.pdf.

Figure 3: Illustration of a professional network based on co-patenting


(a) Bipartite

(b) Inventors

(c) Patents

3.2 Empirical Strategy

Researchers studying the geographic localisation of knowledge exchange face the challenge of controlling for the pre-existing geography of production, that is, the propensity of industries to cluster in space. Firms or workers might exchange knowledge locally within a given industry simply as a result of their co-location due to mechanisms other than learning. That is to say, inventors might disproportionately cite nearby knowledge not because of some spatial friction that limits their access to knowledge produced farther away, but simply because the most relevant knowledge tends to be created locally anyway (and the inventor is located in that cluster precisely for that reason). This would not be a problem if the concentration of relevant activities were entirely driven by spatial frictions, but the literature shows that there are other reasons for the emergence of industrial clusters. Indeed, agglomeration may also arise in the presence of economic externalities due to matching and sharing benefits, such as thicker labour markets or input-output relationships (Duranton and Puga, 2004). In this setting, the correlation between knowledge flows and proximity would be spurious. It is therefore important to empirically isolate learning as a distinct channel other than matching and sharing.

This empirical concern applies analogously to analyses that focus on the social space, rather than the physical one (i.e., social connections). Firstly, because homophily in social relationships typically entails that similar people like each other, thus making it rather likely that the social network measure also reflects the geographic concentration of industries (an example of what Manski, 1993, termed 'correlated effects'). For instance, software engineers are likely to be friends with each other, but also tend to work in the same industries, which cluster around Silicon Valley. At the same time, in Silicon Valley workers might share knowledge independently of these friendship links. Secondly, and more trivially, the clustering of industrial activity matters because in this particular analysis social connectedness is imputed to inventors on the basis of their geographic location. Another possible biasing factor relates to common unobserved environmental factors in the respective locations of each agent, which simultaneously affect their propensity to interact and the possibility of observing a flow of knowledge without the need that interaction is associated to knowledge exchange. This could be the case, for instance, if two large university colleges facilitate interaction between graduate students, thus creating social ties, whilst at the same time being host to important research centres that use knowledge produced by the other university. Yet this knowledge could be sourced autonomously in complete absence of social learning, despite the existence of social links between students.

To identify the effect of social connectedness on knowledge flows using patent citation data, this paper relies on a strategy devised by Thompson (2006). This strategy exploits information available on patents from 2001 onwards about the source of each citation: whether it was the patent's applicant, or if the citation was included by the examiner during the review process.[18] Examiner-added citations are then used as a control group for knowledge flows. In practice, a variable C_{ij} is coded to denote whether a citation of patent j by patent i can be interpreted as a flow of knowledge:

$$C_{ij} = \begin{cases} 1, & \text{if } j \text{ is cited by the applicant} \\ 0, & \text{if } j \text{ is cited by the examiner} \end{cases}$$

Levels of social connectedness between the counties $c(i)$ and $c(j)$ where patents i and j where created are compared for $C_{ij} = 1$ against those for $C_{ij} = 0$, controlling for other possible confounding factors, notably physical geography. This is achieved by means of a Linear Probability Model (LPM) that estimates how physical and social distances influence the likelihood that the citation of patent j by patent i is made by patent i's applicant, as opposed to its examiner.[19] Econometrically, this relationship can be represented as follows:

$$\begin{aligned} C_{ij} = \beta \ln SCI_{c(i)c(j)} + \delta \ln DIS_{c(i)c(j)} + \eta NET_{ij} + X'_{c(i)c(j)}\gamma \\ +\psi_{c(i)} + \psi_{c(j)} + \theta_{t(i)} + \theta_{t(j)} + \mu_{g(i)g(j)} + \pi_i + \pi_j + \epsilon_{z(i)z(j)} \end{aligned} \tag{1}$$

Where $\ln SCI_{c(i)c(j)}$ is the natural log of the Social Connectedness Index between counties $c(i)$ and $c(j)$, $\ln DIS_{c(i)c(j)}$ is the natural log of physical distance (great circle, in thousand kilometres), NET_{ij} is the professional networks dummy, and $X_{c(i)c(j)}$ is a vector of bilateral controls defined at county pair level. Note that controlling for professional networks allows to interpret β as the effect of informal social connections in the inventor's neighbourhood. Additionally, all specifications also include citing and cited counties fixed effects (FEs) $\psi_{c(i)}$ and $\psi_{c(j)}$, citing and cited patents cohort fixed effects $\theta_{t(i)}$ and $\theta_{t(j)}$ (using issue years

[18] Examiners are specialised administrative officers whose job is to deliberate whether or not a patent can be granted. The patent examination process is described in detail in Cockburn et al. (2002).

[19] A linear probability model is preferred over the probit or logit options due to the use of high dimensional fixed effects, which would make probit and logit estimation computationally very demanding. Moreover, this allows for a more straightforward interpretation of coefficients as marginal effects.

t[20]), and a technology-pair fixed effect $\mu_{g(i)g(j)}$ for both patents (using WIPO fields). We also explore the use of citing and cited patents fixed effects π_i and π_j. Finally, $\epsilon_{z(i)z(j)}$ is an error term double-clustered by citing and cited commuting zones $z(i)$ and $z(j)$. This adjustment is required when one clustering dimension is not nested within the other.

In this paper we are interested in estimating β. However, the estimating equation will give biased estimates if $\mathbb{E}[\epsilon_{z(i)z(j)} | \ln SCI_{c(i)c(j)}] \neq 0$. We argue that the use of examiner citations, combined with technology-pair fixed effects, allows to address the main sources of bias discussed above by providing a set of 'control' citations that is orthogonal to the physical and social geographies of the applicant. In addition, county-level fixed effects capture any source of bias deriving from different propensities of counties to generate, patent, or cite ideas, as well as their initial stock of patents, is absorbed. In fact, county fixed effects solve any issue related to observed or unobserved characteristic specific to each county. In selected specifications, citing patent fixed effects control for different propensities of citations to be added by the examiner, which may be correlated with the outcome at patent level. They also capture unobserved examiner characteristics and whether the patent has an institutional assignee or not. Finally, the set of bilateral controls for differences in observable characteristics of counties $c(i)$ and $c(j)$ mitigates issues related to omitted variables specific to each county pair. For instance, we include a dummy coding the presence of a large, leading, research intensive university in both counties.[21] We also include a variable capturing the log of gross migration flows between all county pairs is therefore also included in the analysis. This addresses the concern that the SCI is simply a result of past migration patterns between county pairs.[22] Additional bilateral controls include absolute differences in: the share of adult population with a bachelor degree or higher, the share of children born in 1980-1984 who become inventors in the 2001-2014 period (by CZ where they grew up)[23], population density, median income, unemployment rates, and shares of White, Black, Asian and Hispanic Americans in each county.[24]

Using examiner citations as controls requires two main assumptions (Thompson, 2006). Firstly, this method posits that citations made by the applicant are on average more

[20] Application year cohort fixed effects were also tested in robustness checks, with no change in findings.

[21] The data is obtained from the 2018 THE Ranking of US universities, retaining the top 50 institutions.

[22] This variable is constructed using rolling five-year cumulative counts of yearly county-to-county migration flows. It is assigned to each patent using county and application year information. The data on mobility come from the Statistics of Income Division (SOI) of the US Internal Revenue Service (IRS). They provide one of the most detailed sources of information at this level, based on address changes in the records of all individual income tax forms filed between 1990 and today. The data were retrieved at: https://www.irs.gov/uac/soi-tax-stats-migration-data.

[23] This variable is obtained from Bell et al. (2018), please consult the original paper for further details. The original data can be downloaded from: https://opportunityinsights.org/data/.

[24] Unless otherwise specified, all variables are defined at county-level and are constructed using data from the 2010 US Decennial Census.

likely to represent genuine knowledge transfers than citations by the examiner. Examiner citations are assumed to rather reflect an administrative act required to complete the scope of prior art available for that patent. In other words, this strategy requires that examiner citations can be credibly interpreted as counterfactuals for inventor citations: knowledge that the inventor ought to have had, but did not, and that this knowledge was not in turn added by the examiner as a result of knowledge flows.[25] Note this method does not require that *all* applicant citations reflect knowledge flows. Indeed some citations may have been added by the patent attorney (Jaffe and de Rassenfosse, 2017). However, as long as applicant citations are systematically *more likely* to reflect a knowledge flow than examiner ones, incorrectly attributed citations are simply noise, and the method we propose works. A second identifying assumption is that examiners do not learn via their social connections or geographic location. In other words, the geographic and social locations of examiners must be orthogonal to the predominant knowledge base of the patent being examined, so that examiners cannot learn about prior art from the same localised knowledge flows that are specific to the particular set of technologies of the examined patent. Importantly, this same requirement must also hold for the social space: the position of examiners in the network of social relationships must be exogenous to the predominant technological class of the citing patent so that exposure to the same social networks as the inventors cannot be the reason why examiners cite the patent. These conditions address the well known observation that firms and workers in specialised industries co-locate (sorting), and that people with similar characteristics are more likely to interact socially (homophily). Both these conditions are likely to be met in our data. Cockburn et al. (2002) and Thompson (2006) point out that most examiners work from one office located in Alexandria, VA. Moreover, within a given subject area, patents are assigned to examiners in the order by which applications are filed to the office, which introduces an extra dimension of randomness in case one worries about the place of origin of the examiner before relocating to Virginia.[26] As regards the social space, exogeneity in the physical location of examiners allows to draw the same conclusion for connectedness, to the extent that the latter is defined for geographical units, and that it reflects relationships between the full population of Facebook users, and not just inventors. This is another advantage of using the SCI.[27]

[25] In partial support for this claim, a survey of inventors confirms that applicant citations do represent a measure of knowledge transfers - although noisy - and that when inventors were unaware of citations made in their patent, this was typically due to the citation being added by the examiner (Jaffe et al., 2000).

[26] Note that recent literature has documented the tendency of examiners to specialise by technological areas, see Righi and Simcoe (2019), but technology pair fixed effects address this issue.

[27] Appendix Figure A.3 gives additional credit to our argument. The kernel density plots show the distributions of geographical distance (a) and social connectedness (b) for applicant (in blue) and examiner citations (control, in red), along with a distribution for control citations whose origin was replaced with that of Alexandria, VA, where examiners are actually located (in green). Comparing these fictional distributions to those of applicant citations and observed examiner citations, it is evident

Finally, note that combining technology pair dummies with dummies for citing and cited patent cohorts and patent-level FEs to some extent mimics the case-control matching method first implemented by JTH and often used in this literature (including by Bailey et al., 2018b). In fact, by combining estimates of the within technology-pairs, cohorts, citing and cited patents effects with the examiner-control method, we believe that this analysis imposes stricter constraints on the data.

4 Results and Discussion

4.1 Main Regression Models

We begin by showing in Table 2 the results of baseline models regressing the main variables of interest separately one from the other. The models are estimated using Equation 1, selectively restricting the coefficients β, δ, η and γ to zero. Each coefficient is estimated in its raw form and with key fixed effects. For ease of reading, the outcome is expressed in percentage points. Columns (1), (3) and (5) give raw correlations between citations and social connectedness, geographical distance, and inventors' professional networks, respectively. Columns (2), (4) and (6) restrict the sample to citations across assignees and counties, and introduce the main set of fixed effects used in this analysis: dummies for citing and cited patent counties, cohorts (issue years), and pairs of WIPO technology fields. Restricting the sample is important for two reasons. First, we are interested in studying the impact of social connectedness *across*, rather than within the same region. Second, and most importantly, this allows to implicitly control for inventor mobility, which was one of the heavy knowledge conveyors discussed in Section 2. When the sample excludes within-county and within-assignee citations, inventor self-citations (accounted for by the professional network dummy) necessarily denote instances where the inventor changed employer (at least for that particular patent), favouring one located in a different county. Because the professional network dummy also controls for endogenous inventor networks (another heavy channel), the SCI coefficient in this specification should only capture light conveyors of knowledge such as chance meetings, referrals, or salience of market opportunities (refer to Table 1 for an overview of all mechanisms). Moreover, with respect to fixed effects, note that other than the previously mentioned omitted variable concerns, county dummies also allow to account for the fact that larger county pairs naturally display higher social connectedness.[28]

that examiners tend to draw citations from the social network (and geographic location) of applicants rather than their own, confirming the orthogonality requirement discussed above.

[28] Their inclusion equals to controlling for the natural logarithm of the product of each county's population, which combined with the logarithm of the SCI mimics a measure of logged relative probability of friendship (Bailey et al., 2018b) giving the number of existing connections over the number of total possible connections between two regions.

Table 2: Baseline Regressions

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
ln SCI	0.243	0.452					0.268
	(0.116)^b	(0.0556)^a					(0.108)^b
ln Distance			-0.506	-0.369			-0.00372
			(0.194)^a	(0.0623)^a			(0.0983)
Prof. network					3.058	3.285	2.941
					(0.775)^a	(0.561)^a	(0.555)^a
Counties FEs		●		●		●	●
Years FEs		●		●		●	●
WIPO pairs FEs		●		●		●	●
Other county		●		●		●	●
Other assignee		●		●		●	●
Adj. R^2	0.0006	0.1140	0.0005	0.1139	0.0017	0.1144	0.1145
R^2	0.0006	0.1146	0.0005	0.1145	0.0017	0.1151	0.1152
N	11,288,174	8,803,245	11,274,650	8,791,193	11,288,174	8,803,245	8,791,193

Two-way cluster-robust standard errors for citing and cited CZ pairs (Cameron et al., 2011). Significance levels: $^a p < 0.01$; $^b p < 0.05$; $^c p < 0.1$. The outcome variable is expressed in terms of percentage points.

In column (2), the positive and significant coefficient of 0.45 suggests that a one percent change in social connectedness leads to a 0.0045 percentage points increase in the probability of citation. Equivalently, it means that doubling the SCI yields a 0.31 percentage point increase in citation likelihood ($\beta \times \ln 2$). This is more than ten times smaller than the 4.37 percentage points estimated by Bailey et al. (2018b) for the same change using JTH's case-control matching method. Interestingly, Column (4) shows that physical distance displays a very similar effect, although with opposite sign. A county twice as far to where knowledge is produced is a quarter of a percentage point less likely to cite that knowledge, compared to a another located only half that distance away. Column (6) shows the effect of professional networks. Being the co-author of a patent, having co-authored with an author of that patent, or sharing a co-author with an author of that patent increases the probability of citation by just over 3 percentage points. These effects are all statistically significant at the highest conventional level. By contrast, Column (7) shows that when estimating all parameters simultaneously and controlling for the same variables mentioned above, the coefficient on distance becomes insignificant. Social connectedness, about 60% of the original magnitude, is only significant at the 5% level. Although slightly reduced, the coefficient on professional networks also remains statistically significant. This specification represents the basis on which all other main models in this paper are estimated.

The main results of the present analysis are reported in Table 3. For reference, the first column in this table copies the estimates of Column (7), Table 2. Column (2) introduces a vector of controls for citing patents and for all county pairs. We rule out that the citing patent's team size and geographical diversity (proxied by the number of different

US counties listed for all patent's inventors) are driving the effects of social connectedness. Moreover, the log of gross migration flows across counties over the previous five years addresses the concern that social connectedness does nothing more than to proxy population mobility between regions. Similarly, a dummy coding the presence of a major college in both citing and cited counties addresses the concern of spurious correlation due to the co-presence of students and researchers, with the former affecting friendship links and the latter generating citations, without any actual relationship between the two. Other bilateral controls include differences in education attainment, inventor and population densities, incomes, and ethnicities. Reassuringly, even after controlling for all the above, the coefficient on the SCI remains significant, even increasing in magnitude. Details on the marginal effects of each bilateral control are available in Appendix Table B.7. Columns (3) and (4) introduce fixed effects for citing and cited patents. In both specifications, we restrict our sample to patents sending or receiving at least 10 citations, to avoid bias due to too few observations within each absorbed unit. The marginal effect of social connectedness is much smaller in both cases. It remains significant when controlling for unobserved characteristics specific to each citing patent, but cannot be distinguished from zero when effects for each cited patent are also added. However, we argue that it is excessive to restrict our estimates to within-citing, or within-citing *and* cited patents effects. In the latter case, the identifying variation effectively would only come from the list of cited patents within each citing patent, when the cited patent is also cited by other patents, net of all other fixed effects.[29] Given our identification strategy, our main concern at the citing patent level is that results could be driven by unobserved examiner characteristics. To further reassure ourselves that this is not the case, in Column (5) we introduce fixed effects for nearly 600 examination art units and groupings of examiners, obtained directly from PatentsView.[30] The resulting coefficients are only slightly smaller than those in Column (2). Column (5) is our preferred specification. It suggests that two counties at the 75th percentile of the SCI are 1.1 percentage points more likely to cite one another than a pair of counties at the 25th percentile (see the Appendix, Table B.6, for summary statistics). Finally, Column (6) reports the same specification in (5) but includes main effects for several dimensions of heterogeneity that we intend to explore using this estimating sample: different spatial boundaries (same county, same state, other state), cited patent age in years, maximum age of citing patent assignees, and technological distance (deciles of distance across IPC4 classes). This specification is included here for reference as it represents the baseline for all models that include heterogeneous SCI effects. This ensures that the intercept is the same across specifications even as the

[29] In addition, we show in Appendix Table B.7 that much of the reduction in magnitude is due to a change in the estimating sample, as opposed to cited patent fixed effects.

[30] Unfortunately, we do not have disambiguated identifiers for each examiner, but within citing patent estimates in Column (3), albeit perhaps too restrictive, also reassure us that unobserved examiner characteristics are unlikely to be driving our results.

coefficient on connectedness is broken down by different variables, allowing like-for-like comparison (see Section 4.3 for further details). Despite this change, all coefficients are comparable in magnitude to those in Column (5).

Table 3: Main Regressions

	(1)	(2)	(3)	(4)	(5)	(6)
ln SCI	0.268	0.446	0.104	0.0109	0.389	0.393
	(0.108)^{b}	(0.106)^{a}	(0.0238)^{a}	(0.0233)	(0.0913)^{a}	(0.0953)^{a}
ln Distance	-0.00372	0.0657	0.0169	-0.0253	0.0227	0.00492
	(0.0983)	(0.0850)	(0.0190)	(0.0181)	(0.0856)	(0.0956)
Prof. network	2.941	2.457	0.359	0.118	1.960	2.051
	(0.555)^{a}	(0.495)^{a}	(0.120)^{a}	(0.0579)^{b}	(0.387)^{a}	(0.381)^{a}
WIPO pairs FEs	●	●	●	●	●	●
Controls		●	●	●	●	●
Within citing			●	●		
Within cited				●		
Art unit FEs					●	●
Interaction samp.						●
Adj. R^2	0.1145	0.1174	0.5173	0.4630	0.1413	0.1477
R^2	0.1152	0.1180	0.5279	0.4956	0.1420	0.1484
N	8,791,193	8,787,417	7,882,961	6,054,214	8,787,348	8,785,291

Two-way cluster-robust standard errors for citing and cited CZ pairs (Cameron et al., 2011). Significance levels: $^{a}p < 0.01$; $^{b}p < 0.05$; $^{c}p < 0.1$. The outcome variable is expressed in terms of percentage points. All specifications use citing and cited year and county fixed effects. The sample excludes citations within same assignee or same county. Within citing and cited patent specifications restrict to at least 10 citations. Controls: citing team size and geography (no. of US counties), gross migration, top 50 college, diff. in education, inventors, density, income, ethnicity. Interaction controls: main effects for own CZ or state, other state, elapsed time, assignee age, IPC4 technological distance.

4.2 Robustness Checks

Before investigating heterogeneous effects, what follows explores the robustness of estimates in Column (5), Table 3, to changes in model specifications and in the sample. Table 4 summarises the findings. Column (1) simply copies the estimates of the preferred specification (5) in Table 3, for reference. Column (2) shows that the estimates are robust to including fixed effects for application year cohorts, rather than issue year, for both citing and cited patents. Column (3) addresses the possibility of omitted variable bias due to assignment of location as the most frequently observed one among all inventors on the patent. Bias could arise if the other locations of co-inventors are also likely to be the most socially connected ones to the modal county of the patent. Knowledge flows from these counties would then be erroneously attributed to connectedness, while in reality they can be explained by (unobserved) co-location of one of the inventors. To address this, we restrict the estimating sample to citations made and received by patents with a single inventor. In such instances, location is necessarily assigned correctly with our method and there is no omitted variable bias of this kind. Doing so significantly reduces the size of the estimating sample, which falls to roughly 700,000 citations. Despite this very restrictive test, the coefficients on social connectedness and professional networks remain statistically significant. In fact, both increase somewhat in magnitude (especially the latter), suggesting that inventors who patent alone might disproportionately rely on informal and professional ties for access to knowledge (or perhaps this is simply due to more accurate measurement of location, further research may wish to explore this claim more in detail).

Table 4: Robustness checks

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
ln SCI	0.389	0.302	0.403	0.328	0.301	0.399	0.372
	(0.0913)^a	(0.0971)^a	(0.200)^b	(0.0834)^a	(0.0812)^a	(0.134)^a	(0.0800)^a
ln Distance	0.0227	-0.0362	-0.229	0.0226	0.0202	-0.110	0.00308
	(0.0856)	(0.0896)	(0.185)	(0.0776)	(0.0731)	(0.177)	(0.0872)
Prof. network	1.960	1.904	5.971	1.942	1.805	1.826	1.713
	(0.387)^a	(0.361)^a	(0.818)^a	(0.381)^a	(0.367)^a	(0.580)^a	(0.388)^a
Tech. pairs FEs	WIPO	WIPO	WIPO	IPC3	IPC4	WIPO	WIPO
Controls	●	●	●	●	●	●	●
Appl. year FEs		●					
Single-authored			●				
Non coastal						●	
Trimmed							●
Adj. R^2	0.1413	0.1406	0.2331	0.1465	0.1695	0.1594	0.1091
R^2	0.1420	0.1413	0.2400	0.1481	0.1762	0.1606	0.1101
N	8,787,348	8,787,347	715,733	9,042,076	9,016,933	5,022,152	5,787,251

Two-way cluster-robust standard errors for citing and cited CZ pairs (Cameron et al., 2011). Significance levels: $^a p < 0.01$; $^b p < 0.05$; $^c p < 0.1$. The outcome variable is expressed in terms of percentage points. All specifications use citing and cited year and county fixed effects, and citing art unit effects. The sample excludes citations within same assignee or same county. Controls: citing team size and geography (no. of US counties), gross migration, top 50 college, diff. in education, inventors, density, income, ethnicity. The single-authored sample drops citations sent or received by patents with multiple authors. The non coastal sample drops citations originating or received in Census Divisions bordering the Atlantic and Pacific coasts. The trimmed sample drops patents with citations added exclusively by the applicant or the examiner.

Columns (4) and (5) replace fixed effects for WIPO technology field pairs with fixed effects at IPC class (3-digit) and subclass (4-digit) levels. This entails moving from a set of just under 1,200 possible combinations to over 300,000 and 13 millions respectively, since there are more than 550 IPC classes and 3,700 subclasses. Despite this demanding change, the coefficient on connectedness is only sightly reduced and remains significant at the 99% level. Column (6) restricts the sample to non coastal areas only, dropping all citations originating from or destined to Census Divisions not bordering the Atlantic and Pacific coasts. It addresses the concern that population and economic activity naturally cluster along the coasts, and so does innovation activity. As a result, more interaction is to be expected between coastal areas, as well as greater exchange of knowledge (more coast-to-coast citations), without the two being necessarily causally related to each other (essentially, an omitted variable bias due to an unobserved 'coast effect'). The size of the estimating sample is significantly reduced, but results are not affected by this restriction either. We infer that our findings are not limited to coastal locations but apply throughout the US territory. Finally, Column (7) trims the data by excluding patents whose citations were added exclusively by the applicant or by the examiner. As discussed, these represent a large group in our sample, and there is a concern that results are mainly driven by these patents. Reassuringly, this trimming does not alter findings.

Appendix Table B.8 repeats the same exercise but includes fixed effects for citing patents across all models, despite concerns that this specification might be too restrictive. Once again, there does not seem to be any sizeable change in the coefficients compared to the original estimates, with the exception of single-authored patents, where the sample is likely too small to allow precise estimate of within-citing effects (indeed, the coefficient magnitude is stable, but standard errors are inflated).[31]

4.3 Heterogeneous Effects

This section explores possible dimensions of heterogeneity in the marginal effects of social connectedness. In line with the literature and with the conceptual framework outlined in Section 2, we investigate three main drivers, described separately below. To empirically test for heterogeneous effects, we estimate models that take the following general form:

$$\begin{aligned} C_{ij} = \sum_h \beta_h \ln SCI_{c(i)c(j)} \times INT_h + \sum_h \delta_h \ln DIS_{c(i)c(j)} \times INT_h + \eta NET_{ij} \\ + X'_{c(i)c(j)}\gamma + \xi_{ij} + \psi_{c(i)} + \psi_{c(j)} + \theta_{t(i)} + \theta_{t(j)} + \mu_{g(i)g(j)} + \epsilon_{z(i)z(j)} \end{aligned} \qquad (2)$$

Where all variables are defined as in Equation (1), with FEs for citing and cited counties, issue year cohorts, and WIPO technology pairs. In addition, the interaction term INT_h takes different values depending on the heterogeneous margin of interest: We consider heterogeneity over discrete geographical boundaries $GEO_{c(i)c(j)}$, cited patent age AGE_{ij}, citing assignee age ENT_i (elapsed time since first patent), and quintiles of technological distance $TDS_{g(i)g(j)}$. Note that all interaction terms are categorical variables, so that β_h and δ_h capture the marginal effect of social connectedness and distance for category h of the interacted variable. At the same time, we always include main effects for all interaction variables, captured by ξ_{ij}. The sample restrictions discussed in Section 4.1 are always applied: we drop within county and assignee citations. In the absence of any interaction term, thus, the baseline model reported in Table 3, Column (5), is estimated. An additional driver of heterogeneity we examine is the issue year cohort of the citing patent. In this particular case, however, we construct a new estimation sample dating

[31] In unreported results, we also confirm that our findings are robust to controlling for citing and cited patent lawyer dummies, and for citing and cited patents sharing the same lawyer. In both instances, we exploited lawyer disambiguation available from PatentsView.

back to patents issued in 2002, to consider a longer time-span. As the overall estimate of β is not directly comparable to that discussed so far anyway, the main effects term ξ_{ij} for all interaction variables is omitted (citing patent year cohort dummies are absorbed anyway). We also do not consider heterogeneity in the geographical distance coefficient δ here. The model specification is otherwise the same as in (2). We begin by discussing this last case of heterogeneity.

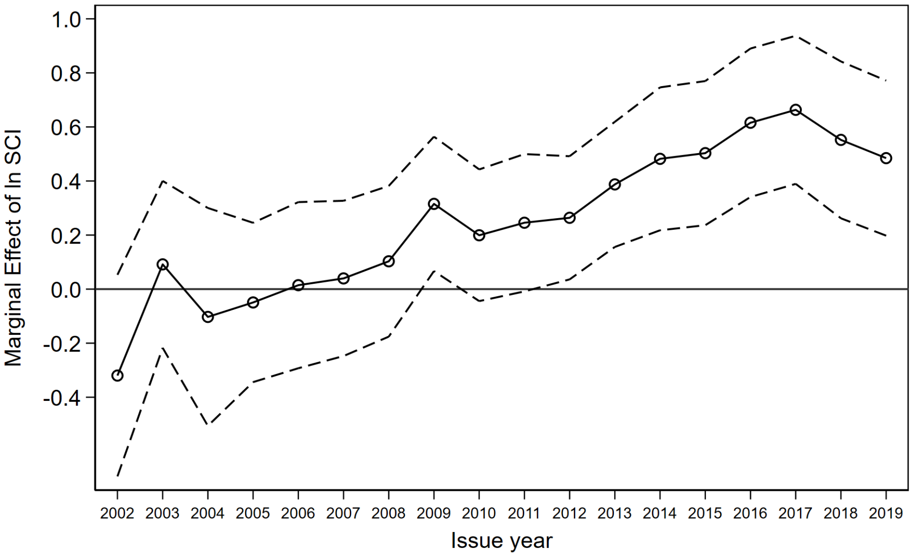
4.3.1 Time Trends

Sonn and Storper (2008) show that geographical proximity has become more important for knowledge production over time, despite advances in information and communication technologies. Using the JTH case-control matching method, the authors reveal a greater likelihood of observing US citations to the same state or city in 1997 compared to 1975. The propensity to rely on local knowledge increases almost monotonically over this period. The underlying causes for this trend, the authors argue, have to do with greater reliance on tacit and non-codified knowledge on the technological frontier, faster product lifecycles requiring more rapid innovation rates, and more complex organisational strategies in knowledge production. More recently, Bloom et al. (2020) document a progressive decline in the productivity of research, defined as the ratio of total factor productivity (TFP) growth and the effective number of researchers. The authors thus conclude that "ideas are getting harder to find". Their result aligns with previous evidence by Jones (2009) on the changing nature of innovation, which he argues is becoming increasingly difficult and requires greater collaborative efforts. In keeping with these findings, we formulate the hypothesis that social connectedness may have also become more relevant over time, as a means to compensate for the increasingly demanding task of accessing relevant knowledge.

We test whether the effect of connectedness changes over time by allowing β to vary depending on the issue year cohort of the citing patent. To this end, we introduce a new sample that includes all patents issued from 2002 onwards. Information on the source of citation, crucial for the identification strategy, was unavailable before this period. The sample construction follows the same method described in Section 3.1.2, with the exception that the size of the resulting list of citations is too large to work with, so a stratified random sample of 20% is drawn. Randomisation is performed at the level of citing patents to ensure that the drawn sample does not over-represent patents with many citations. The resulting estimating sample consists of 364,372 citing patents and 7,212,370 citations, of which about 60% on average are made by applicants. Appendix Tables B.3 and B.5 offer descriptive statistics for citing and cited patents.

Our results are summarised by the coefficient plot in Figure 4, which reports marginal effects of social connectedness over time. All coefficients were obtained from the same regression that interacts the log of the connectedness index with citing patent issue year

Figure 4: Marginal effects by citing patent issue year



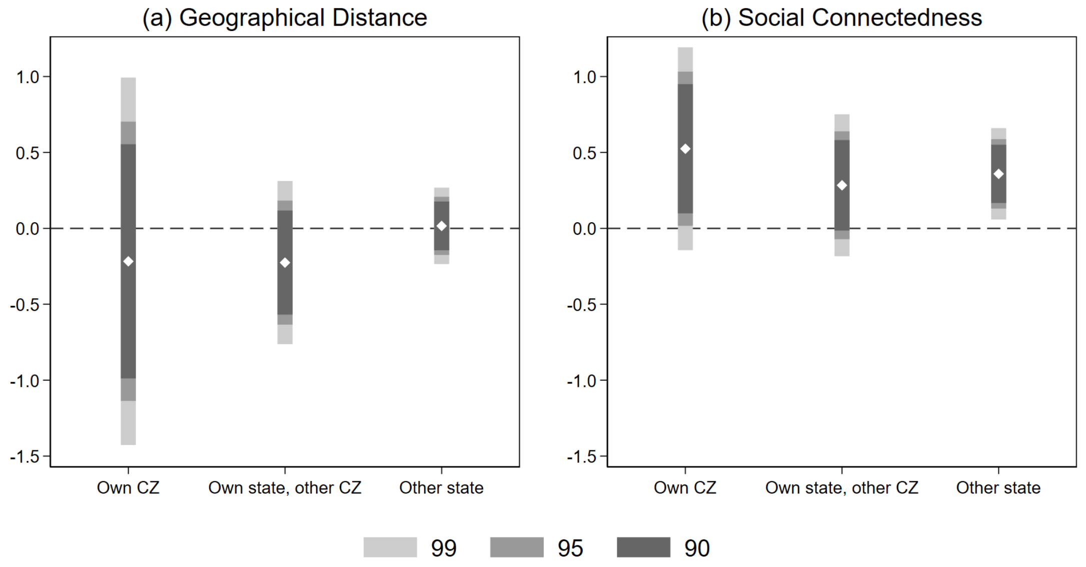
The graph displays coefficients obtained from a regression where ln SCI is interacted with the patent's issue year, conditional on issue year main effects, ln Distance, inventor networks, and the full set of controls. Dashed lines are 95% CIs.

dummies, controlling for issue year main effects, geographical distance, inventor networks, and differences in county-level observables. The results support our hypothesis. Not only are point estimates significantly higher in recent years compared to the early 2000s, marginal effects are mostly not statistically different from zero at the 95% level before 2012. As a robustness check, an alternative regression is run where the application year of the citing patent is used, rather than the issue year. Results, reported in Appendix Figure A.4, are largely unaffected. This finding also provides an additional reason for the decision to restrict this analysis to knowledge flows occurring in the 2016-2019 period. It should be noted, however, that the increasing magnitude of the effects could potentially also be related to the measure of social connectedness becoming more accurate over time, as it reflects a snapshot taken in 2016.

4.3.2 Spatial Boundaries

In this instance, we explore whether social connectedness becomes more important at greater distances. This would be consistent with the notion that connectedness allows to substitute for informal interaction that would otherwise occur locally due to geographic proximity (of two different counties, as we consider cross-county flows only - we do not test for substitution of co-location in the same county). For the same reason, we are also interested in comparing these results with what would happen if we only used physical distance as a proxy for this kind of interaction. As argued in Section 2, distance is likely to be inadequate in capturing this effect. To validate this, we would expect the coefficient on physical distance to be insignificant across discrete spatial boundaries capturing progressively larger areas.

Figure 5: Marginal effects by spatial boundaries



Each graph displays coefficients obtained from the same regression, where ln Distance and ln SCI are interacted with dummies for spatial boundaries. Main effects for spatial boundaries are also included. Dashed lines are 95% CIs.

Results reported graphically in Figure 5 cannot confirm this conjecture. The plot displays coefficients on physical distance (a) and connectedness (b) broken down by three discrete spatial boundaries: citations within CZ, within own state (but not own CZ), and across states. They are all obtained from the same regression, as in Equation (2). There is no evidence that the importance of connectedness increases as one considers progressively more (physically) distant interactions. Similarly, whether looking at county pairs within a cummuting zone, within the same state, or even across states, the marginal effect of SCI is the same. There is thus no evidence supporting the hypothesis that social and geographical proximity are strictly speaking substitutes. This contrasts with the findings by Agrawal et al. (2008), who study the interaction effect between geographical distance and co-ethnicity of inventors on citation likelihood.

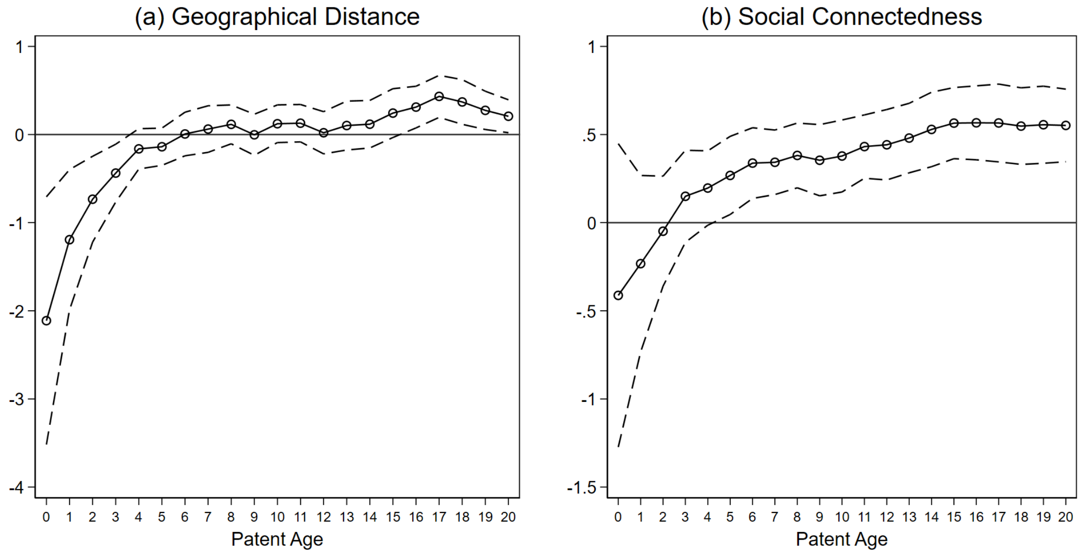
4.3.3 Patent Age

This section considers the role of elapsed time to citation in mediating the effect of social connectedness and geographical distance. Elapsed time to citation can be though of as the 'age' of patent j when it was cited by i at time t, measured as:

$$AGE_{ij} = t_i^{app} - t_j^{app} + 18\ months$$

Where t_i^{app} is the application date of citing patent i, and t_j^{app} is the application date of cited patent j. Since November 29, 2000, all applications received by the USPTO are published 18 months after being filed irrespective of whether or not they are granted. We thus consider this to be the relevant 'birth date' for cited patents. Patent age, initially expressed in months, is then discretised into years using a floor function that assigns the greatest integer less than or equal to the value in months divided by twelve. We conjecture that the impact of social connectedness might change over the interval defined by AGE_{ij}. The pattern of heterogeneity, however, is uncertain a priori. It is possible that social and geographic proximity matter most for the citation of young patents, when frictions in knowledge flows are highest. For geographic proximity, this effect is documented in JTH, where it is shown that localisation of citations decreases as the cited patent becomes 'older'. In the case of social connectedness, analogously, stronger informal ties might be especially relevant for the exchange of knowledge that is yet to become common domain. By contrast, it is also possible that once a patent does become common knowledge, its citation depends increasingly on the presence of some linkage, whether of geographical or social nature, which nudges inventors to tap into that pool of ideas as opposed to another. Older patents, for instance, might have been 'forgotten'. Making predictions about the direction of heterogeneity is further complicated by the fact that geographic proximity and social connectedness are not independent from each other, so that at different points in time the effect of one might influence that of the other. Ultimately, thus, this is an empirical question.

Figure 6: Marginal effects by cited patent age



Each graph displays coefficients obtained from the same regression, where ln Distance and ln SCI are interacted with cited patent age (app. - app. + 18 months). Main effects for patent age are also included. Dashed lines are 95% CIs.

Figure 6 graphically reports the marginal effects on geographical distance (a) and the SCI (b), allowing the coefficients to vary across the age of cited patents.[32] Dashed lines denote 95% confidence intervals. Controlling for the effect of social connectedness, geographic

[32] An unreported coefficient controls for the effect on all patents older than 20 years.

proximity matters most for the citation of young patents, confirming previous results by JTH. Greater distance between two counties decreases the probability of patents produced in one to cite those produced in the other during their first five years of circulation. The friction imposed by geographical distance is strongest for very young patents, then falls sharply and becomes largely insignificant. By contrast, controlling for the effect of geographical distance, social connectedness displays the opposite pattern. The marginal effect of the SCI is insignificant for cited patents aged five years or younger, but increases almost monotonically after that. The synchrony in the fading effect of physical geography as that of social connections becomes relevant is striking. It suggests that there is some degree of interaction between the two effects over time. It is difficult to interpret the graph unambiguously, however. It appears that as patents become common domain in a spatial sense, their likelihood of being cited depends increasingly on social connections. This could reflect a degree of bias in the sources of available knowledge inventors tap into, whereby they disproportionately rely on knowledge produced in places with stronger informal ties to their location. It could also show that social connectedness mitigates a propensity for older patents to become forgotten (without necessarily being obsolete).

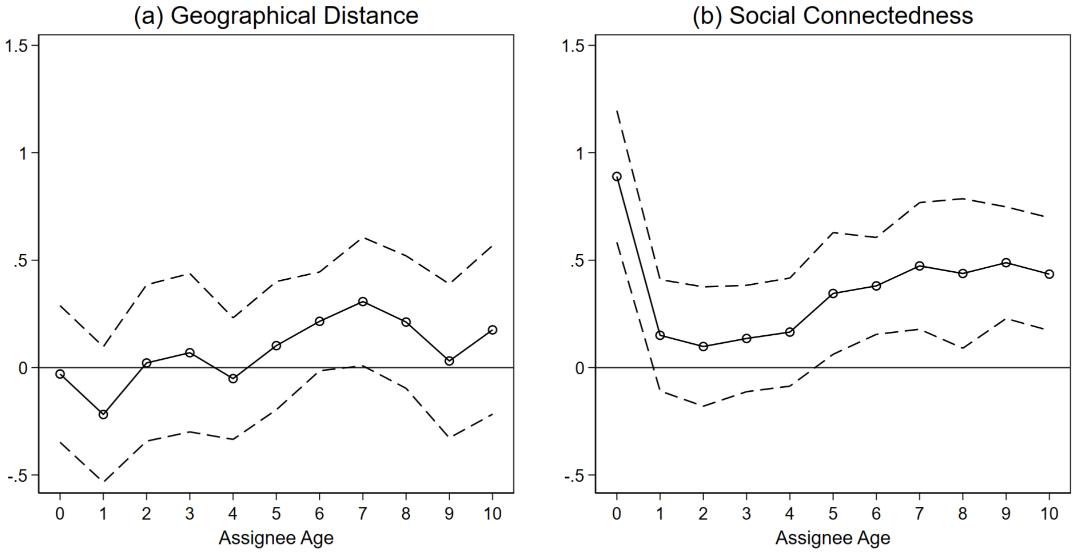
4.3.4 Entrepreneurship and Garage Inventors

Does social connectedness matter differentially for organisations at different stages in their life? In particular, are entrepreneurs and garage inventors disproportionately reliant on their informal social environment as a source of ideas and innovation? In many organisations, inventors ‘work for hire’ with little flexibility in terms of process, and relatively strict guidance with respect to expected outputs. This is likely to be the case especially for more junior inventors in established teams, and generally in larger firms. For instance, Agrawal et al. (2010) find that inventors employed by large firms in company towns (places where innovation is concentrated in a single organisation) are more likely to draw on knowledge produced within the firm’s institutional boundaries. The type of ‘light’ contributions channelled by social connectedness, such as salience of market opportunities, experimental ideas, or chance meetings, are perhaps of secondary relevance for this group. By contrast, they should matter most for more independent types of organisations, such as smaller and younger firms, entrepreneurs, and garage inventors (that is, inventors who work independently, on their own, often at the early stages of a new idea). Duranton and Puga (2001) introduce the concept of ‘nursery cities’ to highlight the role that access to diversified knowledge observed in large urban agglomerations plays in fostering innovation and entrepreneurship. Analogously, we test the hypothesis that social connectedness provides a similar source of advantage in the early stages of a firm’s economic life. [33]

[33] Consistent with this hypothesis, Percoco (2012) shows that local social capital is positively associated with entrepreneurship in Italian cities, not least because of a possible effect on information transfers.

A variable ENT_i is created, which tracks the maximum age of all assignees listed for citing patent i. Assignee age is defined by exploiting disambiguated identifiers on organisations owning each patent. Organisations are assumed to have been established at the time they were issued their first patent. Subsequently, for any patent i, assignee age is the elapsed time between the issue year of the citing patent, and the issue year of the first observed patent for that same assignee. By construction, therefore, year zero is when none of the assignees owning the invention had previously patented. We think of them as entrepreneurs, or garage inventors. Equation (2) is then estimated for $h = 3$, allowing the coefficients on geographical distance and social connectedness to vary over assignee age. Results are reported in Figure 7.

Figure 7: Marginal effects by maximum age of citing assignee(s)



Each graph displays coefficients obtained from the same regression, where ln Distance and ln SCI are interacted with the maximum assignee age since first patent. Main effects for assignee age are also included. Dashed lines are 95% CIs.

While the marginal effect of geographical distance in (a) is mostly indistinguishable from zero across all values of assignee age, the effect of social connectedness in (b) is at least twice, and up to four times, as large for garage inventors and start-up firms (year zero), than it is for older organisations. This difference is statistically significant compared to coefficient values for firms that are up to three years older. During this period, in fact, social connectedness does not matter for citation probability. From year four onwards, then, stronger informal ties matter again, although with reduced strength compared to garage inventors. This pattern is consistent with demographic studies of firms. Bartelsman et al. (2005) find that in the US firms enjoy a honeymoon phase in their first year of life, with the probability of exiting the market increasing significantly in the second year before settling at a constant rate. By year three, about 30% of newly established firms will have exited the market. Interpreting our results through this lens would suggest that while social connectedness is strongest for start-up firms, it is also lowest among firms that are more likely to fail. Perhaps, then, firms that survive this high risk phase and are still observed patenting as they age were somewhat advantaged by their greater social connectivity. This interpretation, however, is largely speculative and cannot be tested within the scope of this analysis. It is also possible, in fact even likely, that the proposed garage inventor measure correlates with the size of the patenting firm, with smaller firms (indeed potentially also young firms) disproportionately relying on external sources of knowledge. Appendix Figure A.5 shows that our findings also hold when replacing assignee age with the assignee's cumulative patents at the time of citation. The strength of social connectedness effects decays rapidly in the number of patents owned by assignees, becoming insignificant after the third one is granted.

4.3.5 Technological Distance

In this section, we explore the possibility that social connectedness matters differentially for the flow of ideas depending on the type of knowledge that is exchanged. It is well known that higher density leads to more innovation (Carlino et al., 2007). However, this relationship is non monotonic, since patenting rates are highest at medium levels of population density (Carlino et al., 2007; Henderson, 2007). Building on this finding, Berkes and Gaetani (2020) propose a model where informal interaction spurred by high density living sustains knowledge exchange across *distant* technologies. In other words, while overall innovation occurs in medium-sized specialised clusters, it would appear that 'unconventional innovation', as the authors call it, builds on informal interactions made possible by very dense urban agglomerations. Following this intuition, we investigate whether informal interaction fostered by stronger connectedness, rather than spatial proximity, can play a similar role in bridging gaps between different communities of inventors across the US. According to this hypothesis, social proximity would allow the diversity of knowledge bases typical of large urban agglomerations to exist beyond the constraints of geography. Feldman and Audretsch (1999) show that greater diversity in the industrial composition of a region is associated with higher rates of local innovation. One can think of informal social connectedness as a way to tap into a broader pool of knowledge. This hypothesis is consistent with research suggesting that a city with strong connections to other clusters benefits from the renewal and enrichment of its knowledge base by gaining access to new external ideas (Bathelt et al., 2004; Breschi and Lenzi, 2016; Akcigit et al., 2018), conditional on having sufficient absorptive capacity to do so (Miguélez and Moreno, 2015).

Technological distance is measured as the cosine dissimilarity in the reference set of each pair of technologies (Yan and Luo, 2017), using IPC technology classes (IPC3) or subclasses (IPC4), and the complete list of citations made by patents issued over the

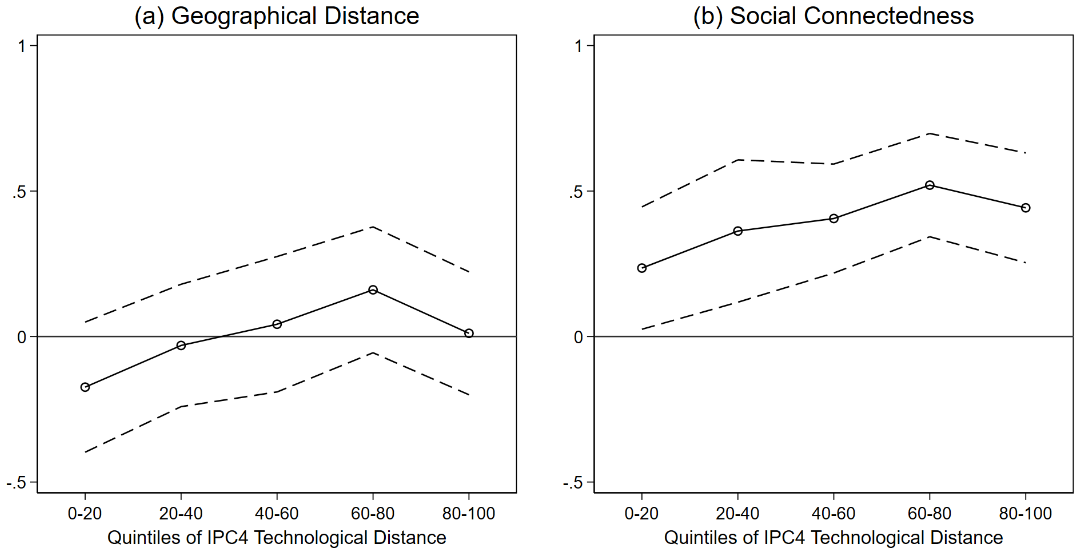
2002-2019 period. Because each citing and cited patent can belong to multiple classes or subclasses, we consider a weighted average measure. We proceed as follows (we discuss classes only, the method is the same for subclasses). First, we inflate the citation list by assigning to each citing and cited patent all the classes they are associated with. We then assign a citation of each patent pair proportionally to the number of citing and cited classes for that pair. For instance, if citing patent i belongs to two classes and cited patent j belongs to four classes, each class pair is assigned one eighth of that citation. The resulting dataset is then collapsed summing up weighted citations by citing and cited classes. This is used to compute the cosine dissimilarity measure. In particular, for every pair of citing $g(i) = \mathcal{A}$ and cited $g(j) = \mathcal{B}$ classes, technological distance is measured as:

$$TDS_{g(i)g(j)} = 1 - \frac{\sum_k C_{\mathcal{A}k} C_{\mathcal{B}k}}{\sqrt{\sum_k C^2_{\mathcal{A}k}}\sqrt{\sum_k C^2_{\mathcal{B}k}}} \tag{3}$$

Where $C_{\mathcal{A}k}$ and $C_{\mathcal{B}k}$ denote the weighted number of citations sent from patents in technology class $\mathcal{A}$ and technology class $\mathcal{B}$ to patents in technology class k, with k indexing all available classes. Intuitively, the fraction in (3) gives the similarity in the two vectors representing the distribution of citations of each class to all classes (the cosine of their angle), which is bounded in the $[0, 1]$ interval. Subtracting this value from one thus gives a measure of dissimilarity, or distance, based on how different the knowledge bases of the two classes are. Finally, we assign a weighted average of this measure to each patent pair in the estimating sample, based on all the technology classes associated with the citing and cited patents. We also recode the variable in terms of quintiles over the distribution in 2016-2019 (we retain the same variable name for simplicity). Equation (2) is then estimated for $h = 4$.

Figure 8 graphically reports the marginal effects of geographical distance (a) and social connectedness (b), allowing the coefficients to vary across quintiles of technological distance between citing and cited patents (IPC4 level). Vertical bars denote 95% confidence intervals. These estimates appear to give some credit to our hypothesis with respect to social connectedness, but the relationship is very noisy. The coefficients display a positive sloping trend, with the SCI being nearly statistically indistinguishable from zero for the bottom quintile of technological distance. Moreover, the point estimates on the most technologically distant groups of citations are almost twice as large as that measured for the first quintile. Yet we cannot argue that the coefficients are different from each other in a statistical sense. By contrast, there does not seem to be any statistically significant relationship between geographical distance and citation irrespective of which quintile is considered. These results broadly hold also if distance between classes (IPC3), rather than subclasses, is considered (Appendix Figure A.6).

Figure 8: Marginal effects by technological distance (IPC4)



Each graph displays coefficients obtained from the same regression, where ln Distance and ln SCI are interacted with quintiles of technological distance. Main effects for each quintile are also included. Dashed lines are 95% CIs.

5 Conclusions

This paper explored the role of informal social interaction, defined in terms of social connectedness, in the transfer of knowledge as captured by patent citation data. Using an index of aggregate Facebook ties to measure social connectedness between places, it finds that social proximity does seem to matter, positively influencing the probability of observing a citation between two places. This is robust to controlling for physical distance, the pre-existing geography of production (e.g., clustering due to other Marshallian forces such as matching or sharing), and the existence of professional links between any inventor involved in creating the citing or the cited patent (up to two degrees of distance). Interestingly, these effects seem to explain away the statistical significance of physical proximity. This suggests that informal social connectedness, despite its likely correlation with geographical distance, offers perhaps a more accurate measure to study knowledge flows. By this we do not mean to say that being socially connected can replace the importance of being co-located. Our analysis did not directly test for substitution of co-location in the *same* county, nor was it conclusive with respect to substitution between social and geographical proximity *across* counties. Rather, we note that physical proximity and social connectedness appear to be two ways by which inventors can access existing knowledge. In practice, most inventors will rely on both, especially to the extent that physically proximate places are also likely to be strongly connected socially. We document that the age of the cited patent might play a role in explaining the relevance of geographical, as opposed to social proximity. In the early stages of knowledge creation, spatial frictions are strong and spatial proximity facilitates access to knowledge. However, as knowledge becomes common domain in a geographical sense, the informal social environment in which inventors operate is increasingly important in shaping knowledge flows across counties, irrespective of physical distance. We also show that social connectedness matters most for entrepreneurs and garage inventors, and that it contributes bridging gaps between technologically distant knowledge areas. Our key takeaway is that no inventor is an island, as knowledge creation is inherently a social process. This is not just true for interactions with colleagues in the profession, but also with respect to informal ties in the inventors' social environment.

In terms of magnitude, the effect of informal interaction is quite small. According to our preferred specification, doubling social connectedness increases citation likelihood by about a third of a percentage point. Social connectedness, however, can be economically meaningful. Two counties at the 75th percentile of social connectedness are on average 1.1 percentage points more likely to cite one another than a pair of counties at the 25th percentile. To be more concrete, consider the following example. The counties of Colleton and Dorchester in South Carolina neighbour each other geographically. The latter, however, has a connectedness strength to Santa Clara County in California (one of the top patenting counties in the US) at the 75th percentile of the overall distribution for county-pairs, while the former is only at the 25th percentile. Between 2016 and 2019, there were fourteen times as many applicant citations between Santa Clara and Dorchester, than between Santa Clara and Colleton.[34] This difference is striking considering that the two counties are contiguous and certainly within commuting distance from each other. Moving inventors from one to the other can potentially have implications for their exposure to ideas. While admittedly anecdotal, and granted that it is hard to imagine that there is actually a sharp discontinuity in connectedness at the county border, this example helps illustrate the local variation existing in this measure, and the tangible difference that social connectedness can make for knowledge flows. There are several other instances where this type of change can be achieved by moving relatively close in space. Appendix Figure A.7 shows counties connected to Santa Clara, CA, with strength at least as strong as the upper quartile (in blue), or at least as weak as the lower quartile (in green). Evidently, green and blue counties are frequently located in close proximity.[35]

There are several limitations to the present work. The most important concern relates to measurement. What is the SCI capturing in practice? With the level of aggregation used in this analysis, we can only gauge an indirect picture. Ideally, one would observe the entire social graph of inventors, allowing to explicitly account for the nature and strength

[34] In terms of propensity, the likelihood of observing a citation by an applicant, compared to all citations, is 20% greater between Santa Clara and Dorchester, than between Santa Clara and Colleton.

[35] More systematically, Appendix Figure A.8 shows that of all county pairs strongly and weakly connected to the same third county, over 5% are within 400 kilometres of distance from each other, and over 20% are within a 1000 kilometres catchment area.

of connections, as well as more generally to study the topography of this graph. The SCI, however, also has some advantages over analyses of this kind. To our knowledge, for instance, this index represents the most comprehensive measure of revealed social interaction available yet for the entire geography of the US. Moreover, failing to observe the full network of inventors, we align to previous work by measuring the professional network of inventors as proxied by co-patenting links. Future work could consider focusing on a subset of the data to construct higher-order connections, which could not be done in this paper due to computational constraints. Another problem relates to the possible endogeneity of the SCI measure. Omitted variable bias, for instance, could arise to the extent that people and economic activity tend to cluster around certain areas in response to natural comparative advantages and history. Our estimating framework has attempted to mitigate this concern, along with robustness checks that restricted the sample of citations to exchanges between non-coastal regions. Admittedly, however, this strategy is incomplete. The ideal experiment would randomly re-wire the social connectivity of all US citizens and measure the resulting effects on knowledge exchange. Finally, a reminder that all results depend on the identifying assumptions underlying the use of examiner citations as a control group. The literature is yet to form a clear view regarding the nature of these citations and possible biases they may cause (Alcácer and Gittelman, 2006; Alcácer et al., 2009; Righi and Simcoe, 2019). In the ideal picture, the examiners simply fill in all technological connections to a patent that the applicant was not aware of. In practice, however, citations are potentially also added by patent attorneys, and examiners might be limited by their own imperfect search process. As such, results should be interpreted as the relative effect of knowledge flows to the applicant, above and beyond any bias accruing to the examiner (rather than relative to an ideal omniscient actor). This, however, is likely to work against the detection of any effect. A comparison of our estimates to those of Bailey et al. (2018b), who use a case-control matching approach and estimate stronger effects, would indeed suggest that any distortion in our method biases results downward. The estimate we provide is thus conservative. We also express a word of caution in terms of the way knowledge flows are measured in this paper. We relied on patents due to the ease of tracking exchanges via citations and to the availability of structured data, but these data have well-known limitations (see Section 3.1.2). Future work could investigate other types of knowledge exchange that would be more likely to be channelled over informal ties.

There are also ways in which this work can be refined and expanded. One possibility is to investigate whether stronger social connectedness is significantly associated to weaker industrial agglomeration locally. Similarly, it would be interesting to study what types of clusters rely more on this resource. Could it be that large diversified urban agglomerations draw on this connectedness, or is it smaller, more specialised clusters that reap most benefits from stronger informal ties to actors elsewhere? Another important although more challenging question would be to distinguish SCI-mediated knowledge flows from pure spillovers. Indeed, observing that knowledge is more likely to flow from one place to another does not necessarily entail that it causes productivity-enhancing spillovers, or that the exchange took place outside market boundaries. In its simplest form, this analysis would investigate whether stronger social connectedness is associated to the production of higher-quality ideas holding inputs constant, where quality can be approximated using counts of downstream citations. This could be additionally integrated with the study of spillovers between specific industries, contributing to the understanding of how different 'trees of knowledge' emerge. Finally, another line of inquiry could take a closer look at the nature of populations and their social ties, exploring how and why people in different places are interconnected.

In conclusion, while this paper has attempted to set the ground for a sound investigation into the physical and social geographies of knowledge exchange, evidently a great amount of work still lies ahead.

References

Aghion, P. and Howitt, P. (1992). A Model of Growth Through Creative Destruction. *Econometrica*, 60(2):323.

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

Agrawal, A. K., Cockburn, I., and Rosell, C. (2010). Not Invented Here? Innovation in Company Towns. *Journal of Urban Economics*, 67(1):78–89.

Agrawal, A. K., 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.

Akcigit, U., Caicedo, S., Miguelez, E., Stantcheva, S., and Sterzi, V. (2018). Dancing with the Stars: Innovation Through Interactions. *National Bureau of Economic Research Working Paper Series*, No. 24466.

Alcácer, J. and Gittelman, M. (2006). Patent Citations as a Measure of Knowledge Flows: The Influence of Examiner Citations. *The Review of Economics and Statistics*, 88(4):774–779.

Alcácer, J., Gittelman, M., and Sampat, B. (2009). Applicant and examiner citations in U.S. patents: An overview and analysis. *Research Policy*, 38(2):415–427.

Atkin, D., Chen, K., and Popov, A. (2020). The Returns to Serendipity: Knowledge Spillovers in Silicon Valley. *Working Paper.*

Audretsch, D. B. and Feldman, M. P. (1996). R&D Spillovers and the Geography of Innovation and Production. *The American Economic Review*, 86(3):630–640.

Audretsch, D. B. and Feldman, M. P. (2004). Knowledge Spillovers and the Geography of Innovation. In Henderson, J. V. and Thisse, J.-F. B. T., editors, *Handbook of Regional and Urban Economics: Cities and Geography*, volume 4, chapter 61, pages 2713–2739. Elsevier.

Bailey, M., Cao, R., Kuchler, T., and Stroebel, J. (2018a). The Economic Effects of Social Networks: Evidence from the Housing Market. *Journal of Political Economy*, 126(6):2224–2276.

Bailey, M., Cao, R., Kuchler, T., Stroebel, J., and Wong, A. (2018b). Social Connectedness: Measurement, Determinants, and Effects. *Journal of Economic Perspectives*, 32(3):259–280.

Bailey, M., Dávila, E., Kuchler, T., and Stroebel, J. (2019). House Price Beliefs And Mortgage Leverage Choice. *The Review of Economic Studies*, 86(6):2403–2452.

Bailey, M., Farrell, P., Kuchler, T., and Stroebel, J. (2020a). Social connectedness in urban areas. *Journal of Urban Economics*, 118:103264.

Bailey, M., Gupta, A., Hillenbrand, S., Kuchler, T., Richmond, R., and Stroebel, J. (2020b). International Trade and Social Connectedness. *Working Paper.*

Bailey, M., Johnston, D., Kuchler, T., Stroebel, J., and Wong, A. (2020c). Peer Effects in Product Adoption. *Working Paper.*

Bailey, M., Kuchler, T., Russel, D., State, B., and Stroebel, J. (2020d). The Determinants and Effects of Social Connectedness in Europe. *CESifo Working Papers*, No. 8310.

Bala, V. and Goyal, S. (2000). A Noncooperative Model of Network Formation. *Econometrica*, 68(5):1181–1229.

Bali, T. G., Hirshleifer, D. A., Peng, L., and Tang, Y. (2019). Attention, Social Interaction, and Investor Attraction to Lottery Stocks. *SSRN Electronic Journal.*

Bartelsman, E., Scarpetta, S., and Schivardi, F. (2005). Comparative analysis of firm demographics and survival: evidence from micro-level sources in OECD countries. *Industrial and Corporate Change*, 14(3):365–391.

Bathelt, H., Malmberg, A., and Maskell, P. (2004). Clusters and knowledge: local buzz, global pipelines and the process of knowledge creation. *Progress in Human Geography*, 28(1):31–56.

Bell, A., Chetty, R., Jaravel, X., Petkova, N., and Van Reenen, J. (2018). Who Becomes an Inventor in America? The Importance of Exposure to Innovation. *The Quarterly Journal of Economics*, 134(2):647–713.

Berkes, E. and Gaetani, R. (2020). The Geography of Unconventional Innovation. *Working Paper - Conditionally accepted at The Economic Journal.*

Bessen, J. (2008). The value of U.S. patents by owner and patent characteristics. *Research Policy*, 37(5):932–945.

Bloom, N., Jones, C. I., Van Reenen, J., and Webb, M. (2020). Are Ideas Getting Harder to Find? *American Economic Review*, 110(4):1104–1144.

Breschi, S. and Lenzi, C. (2016). Co-invention networks and inventive productivity in US cities. *Journal of Urban Economics*, 92:66–75.

Breschi, S. and Lissoni, F. (2001). Knowledge Spillovers and Local Innovation Systems: A Critical Survey. *Industrial and Corporate Change*, 10(4):975–1005.

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

Buechel, K., von Ehrlich, M., Puga, D., and Viladecans-Marsal, E. (2019). Calling from the Outside: The Role of Networks in Residential Mobility. *CEPR Discussion Paper Series*, No. DP1361.

Buera, F. J. and Oberfield, E. (2020). The Global Diffusion of Ideas. *Econometrica*, 88(1):83–114.

Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2011). Robust Inference With Multiway Clustering. *Journal of Business & Economic Statistics*, 29(2):238–249.

Carlino, G. A., Chatterjee, S., and Hunt, R. M. (2007). Urban density and the rate of invention. *Journal of Urban Economics*, 61(3):389–419.

Carlino, G. A. and Kerr, W. R. (2015). Agglomeration and Innovation. In Duranton, G., Henderson, J. V., and Strange, W. C., editors, *Handbook of Regional and Urban Economics*, volume 5, chapter 6, pages 349–404. Elsevier.

Catalini, C. (2018). Microgeography and the Direction of Inventive Activity. *Management Science*, 64(9):4348–4364.

Cockburn, I. M., Kortum, S., and Stern, S. (2002). Are All Patent Examiners Equal? The Impact of Examiner Characteristics on Patent Statistics and Litigation Outcomes. *National Bureau of Economic Research Working Paper Series*, No. 8980.

Cohen, W. M. and Levinthal, D. A. (1990). Absorptive Capacity: A New Perspective on Learning and Innovation. *Administrative Science Quarterly*, 35(1):128–152.

Comin, D. A., Dmitriev, M., and Rossi-Hansberg, E. (2012). The Spatial Diffusion of Technology. *National Bureau of Economic Research Working Paper Series*, No. 18534.

Crescenzi, R., Filippetti, A., and Iammarino, S. (2017). Academic inventors: collaboration and proximity with industry. *The Journal of Technology Transfer*, 42(4):730–762.

Crescenzi, R., Nathan, M., and Rodríguez-Pose, A. (2016). Do inventors talk to strangers? on proximity and collaborative knowledge creation. *Research Policy*, 45(1):177–194.

Duranton, G. and Puga, D. (2001). Nursery Cities: Urban Diversity, Process Innovation, and the Life Cycle of Products. *American Economic Review*, 91(5):1454–1477.

Duranton, G. and Puga, D. (2004). Micro-Foundations of Urban Agglomeration Economies. In Henderson, J. V. and Thisse, J.-F., editors, *Handbook of Regional and Urban Economics*, volume 4, chapter 48, pages 2063–2117. Elsevier B.V.

Feldman, M. P. (2002). The Internet revolution and the geography of innovation. *International Social Science Journal*, 54(171):47–56.

Feldman, M. P. and Audretsch, D. B. (1999). Innovation in cities: Science-based diversity, specialization and localized competition. *European Economic Review*, 43(2):409–429.

Gilbert, E. and Karahalios, K. (2009). Predicting tie strength with social media. In *Conference on Human Factors in Computing Systems - Proceedings*, pages 211–220, New York, New York, USA. ACM Press.

Glaeser, E. L. (1999). Learning in Cities. *Journal of Urban Economics*, 46(2):254–277.

Granovetter, M. (1973). The Strength of Weak Ties. *The American Journal of Sociology*, 78(6):1360–1380.

Granovetter, M. (1983). The Strength of Weak Ties: A Network Theory Revisited. *Sociological Theory*, 1:201–233.

Greenwood, S., Perrin, A., and Duggan, M. (2016). Social Media Update 2016. Technical report, Pew Research Center, Washington, DC.

Griliches, Z. (1998). Patent Statistics as Economic Indicators: A Survey. In *R&D and Productivity: The Econometric Evidence*, pages 287–343. University of Chicago Press, Chicago, IL.

Hampton, K. N., Goulet, L. S., Rainie, L., and Purcell, K. (2011). Social Networking Sites and Our Lives. Technical report, Pew Research Center, Washington, DC.

Henderson, J. V. (2007). Understanding knowledge spillovers. *Regional Science and Urban Economics*, 37(4):497–508.

Jaffe, A. B. and de Rassenfosse, G. (2017). Patent citation data in social science research: Overview and best practices. *Journal of the Association for Information Science and Technology*, 68(6):1360–1374.

Jaffe, A. B., Trajtenberg, M., and Fogarty, M. S. (2000). Knowledge Spillovers and Patent Citations: Evidence from a Survey of Inventors. *The American Economic Review*, 90(2):215–218.

Jaffe, A. B., Trajtenberg, M., and Fogarty, M. S. (2002). The Meaning of Patent Citations: Report on the NBER/Case-Western Reserve Survey of Patentees. In Jaffe, A. B. and Trajtenberg, M., editors, *Patents, citations, and innovations: A window on the knowledge economy*, chapter 12. MIT press, Cambridge, MA.

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):577–598.

Jones, A. (2009). Redisciplining generic attributes: the disciplinary context in focus. *Studies in Higher Education*, 34(1):85–100.

Jones, J. J., Settle, J. E., Bond, R. M., Fariss, C. J., Marlow, C., and Fowler, J. H. (2013). Inferring Tie Strength from Online Directed Behavior. *PLOS ONE*, 8(1):1–6.

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

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

Kuchler, T., Li, Y., Peng, L., Stroebel, J., and Zhou, D. (2020a). Social Proximity to Capital: Implications for Investors and Firms. *SSRN Electronic Journal*, (646).

Kuchler, T., Russel, D., and Stroebel, J. (2020b). The geographic spread of COVID-19 correlates with the structure of social networks as measured by Facebook. *Working Paper.*

Lissoni, F. (2018). International migration and innovation diffusion: an eclectic survey. *Regional Studies*, 52(5):702–714.

Lucas, R. E. (1988). On the mechanics of economic development. *Journal of Monetary Economics*, 22(1):3–42.

Lucas, R. E. and Moll, B. (2014). Knowledge growth and the allocation of time. *Journal of Political Economy*, 122(1):1–51.

Lychagin, S., Pinkse, J., Slade, M. E., and Van Reenen, J. (2016). Spillovers in Space: Does Geography Matter? *The Journal of Industrial Economics*, 64(2):295–335.

Manski, C. F. (1993). Identification of Endogenous Social Effects: The Reflection Problem. *The Review of Economic Studies*, 60(3):531–542.

Marshall, A. (1890). *Principles of Economics.* Macmillan and Co., Ltd., London.

Miguélez, E. and Moreno, R. (2015). Knowledge flows and the absorptive capacity of regions. *Research Policy*, 44(4):833–848.

Milani, F. (2020). COVID-19 Outbreak, Social Response, and Early Economic Effects: A Global VAR Analysis of Cross-Country Interdependencies. *medRxiv*, page 2020.05.07.20094748.

Murata, Y., Nakajima, R., Okamoto, R., and Tamura, R. (2013). Localized Knowledge Spillovers and Patent Citations: A Distance-Based Approach. *The Review of Economics and Statistics*, 96(5):967–985.

Pavitt, K. (1985). Patent statistics as indicators of innovative activities: Possibilities and problems. *Scientometrics*, 7(1-2):77–99.

Percoco, M. (2012). Entrepreneurship, Social Capital and Institutions: Evidence from Italy. *Spatial Economic Analysis*, 7(3):339–355.

Powell, W. W. and Grodal, S. (2005). Networks of Innovators. In Fagerberg, J., Mowery, D. C., and Nelson, R. R., editors, *The Oxford Handbook of Innovation*, chapter 3, pages 56–85. Oxford University Press.

Rehbein, O. and Rother, S. (2019). Distance in Bank Lending : The Role of Social Networks. *Working Paper.*

Righi, C. and Simcoe, T. (2019). Patent examiner specialization. *Research Policy*, 48(1):137–148.

Roche, M. P. (2019). Taking Innovation to the Streets: Microgeography, Physical Structure and Innovation. *The Review of Economics and Statistics*, pages 1–47.

Romer, P. M. (1986). Increasing Returns and Long-Run Growth. *Journal of Political Economy*, 94(5):1002–1037.

Romer, P. M. (1990). Endogenous Technological Change. *Journal of Political Economy*, 98(5, Part 2):S71–S102.

Saxenian, A. (1996). *Regional Advantage.* Harvard University Press, Cambridge, MA.

Schmoch, U. (2008). Concept of a Technology Classification for Country Comparisons. Technical report, World Intellectual Property Organization, Geneva, Switzerland.

Sonn, J. W. and Storper, M. (2008). The Increasing Importance of Geographical Proximity in Knowledge Production: An Analysis of US Patent Citations, 1975–1997. *Environment and Planning A: Economy and Space*, 40(5):1020–1039.

Storper, M. and Venables, A. J. (2004). Buzz: Face-to-face contact and the urban economy. *Journal of Economic Geography*, 4(4):351–370.

Thompson, P. (2006). Patent Citations and the Geography of Knowledge Spillovers: Evidence from Inventor- and Examiner-added Citations. *The Review of Economics and Statistics*, 88(2):383–388.

Thompson, P. and Fox-Kean, M. (2005). Patent Citations and the Geography of Knowledge Spillovers: A Reassessment. *The American Economic Review*, 95(1):450–460.

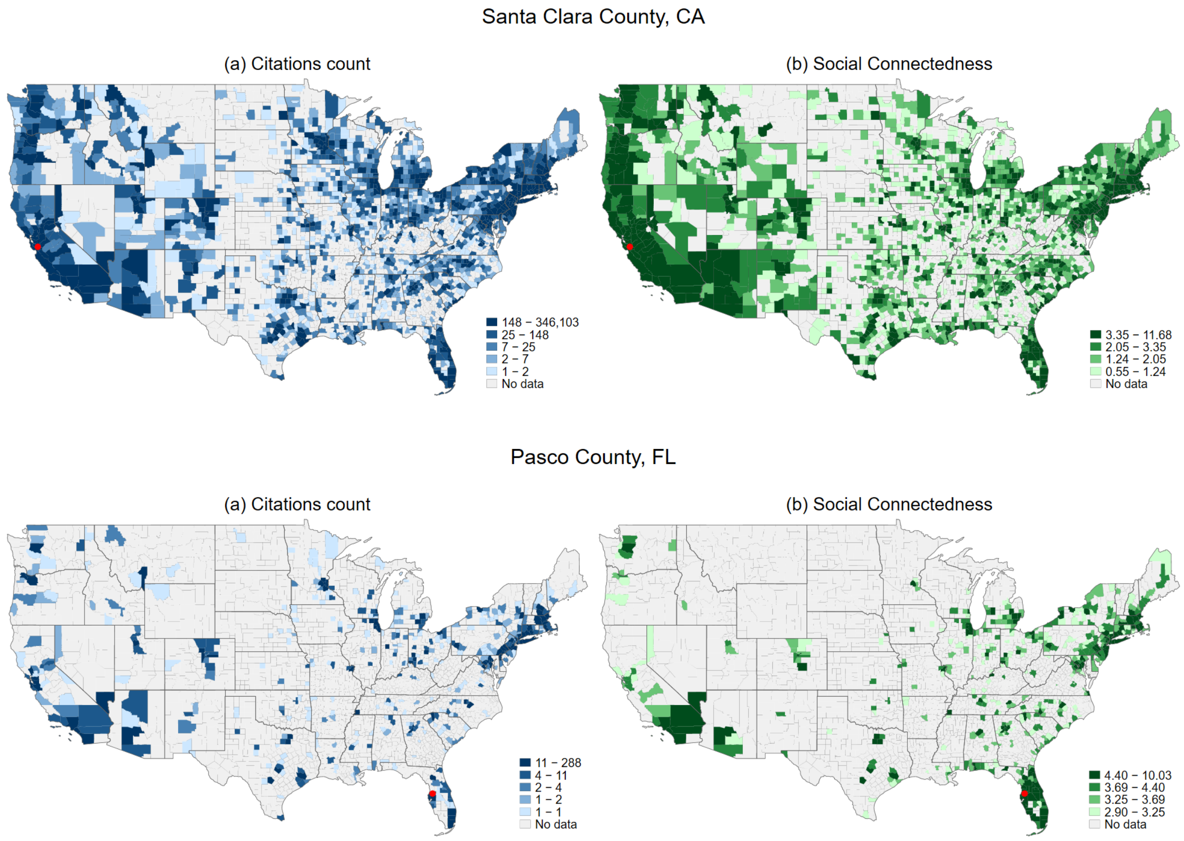
Wilson, R. (2020). The Impact of Social Networks on EITC Claiming Behavior. *Working Paper.*

Yan, B. and Luo, J. (2017). Measuring technological distance for patent mapping. *Journal of the Association for Information Science & Technology*, 68(2):423–437.

Appendices

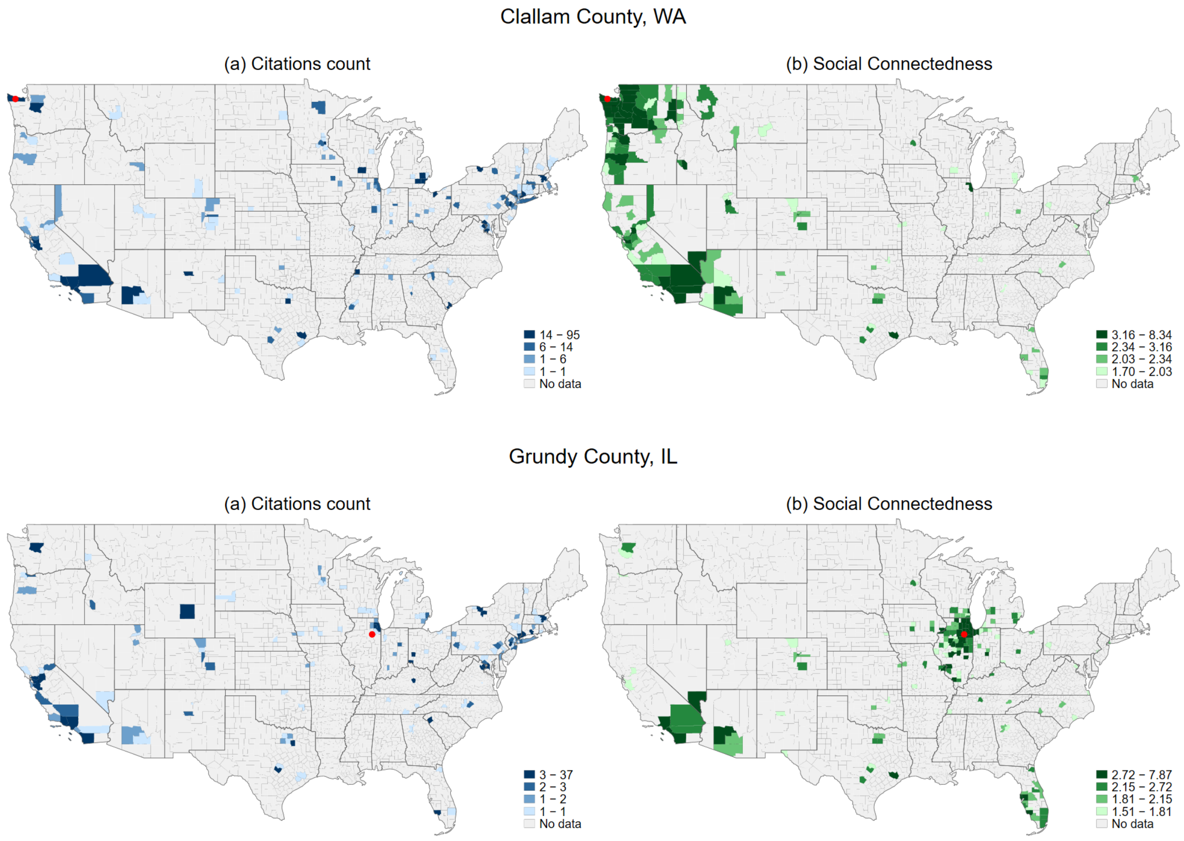
A Figures

Figure A.1: Network Maps of US Counties by Quartiles



Notes: Panel (a) in each map shows, for a given citing county, all counties that receive citations by patents issued in the 2016-2019 period. Polygons are coloured proportional to quartiles of received citation counts. Panel (b) shows the log of social connectedness for counties most strongly connected to the citing one, limiting the sample to the same number of counties as those receiving at least one citation in panel (a). Polygons are coloured proportional to quartiles of connection strength. The similarity in panels (a) and (b) for each citing county suggests that there is a correlation between knowledge flows and social connectedness. Citing counties were selected to represent respectively the 99th, 75th, 50th, and 25th percentiles in the distribution of sent citations, conditional on citing at least 100 different counties.

Figure A.2: Network Maps of US Counties by Quartiles (continued)


Clallam County, WA

(a) Citations count

(b) Social Connectedness

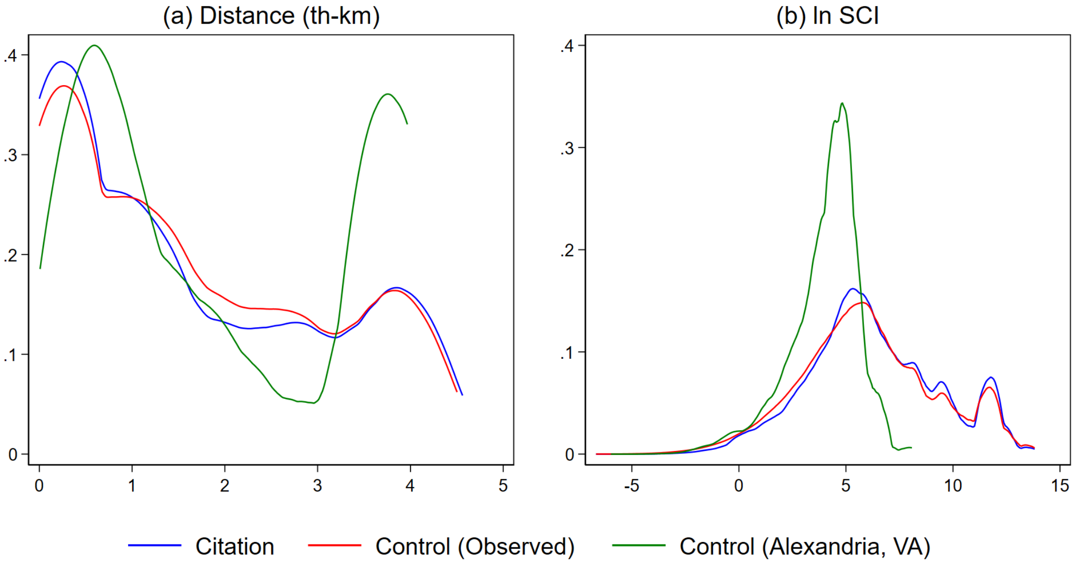
Grundy County, IL

(a) Citations count

(b) Social Connectedness

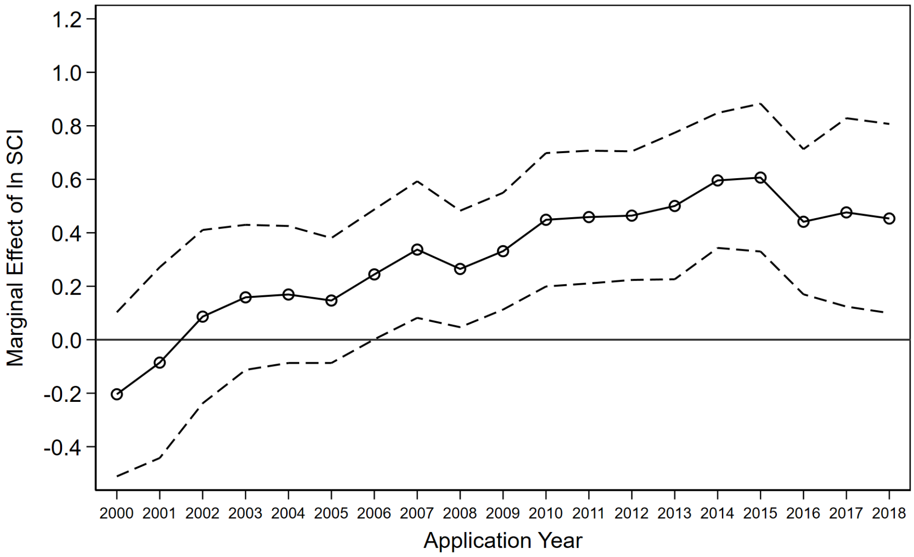
Notes: Continued from previous page. See notes on previous page for details on the interpretation of these maps.

Figure A.3: Distribution of distance and ln SCI for citation and control knowledge flows



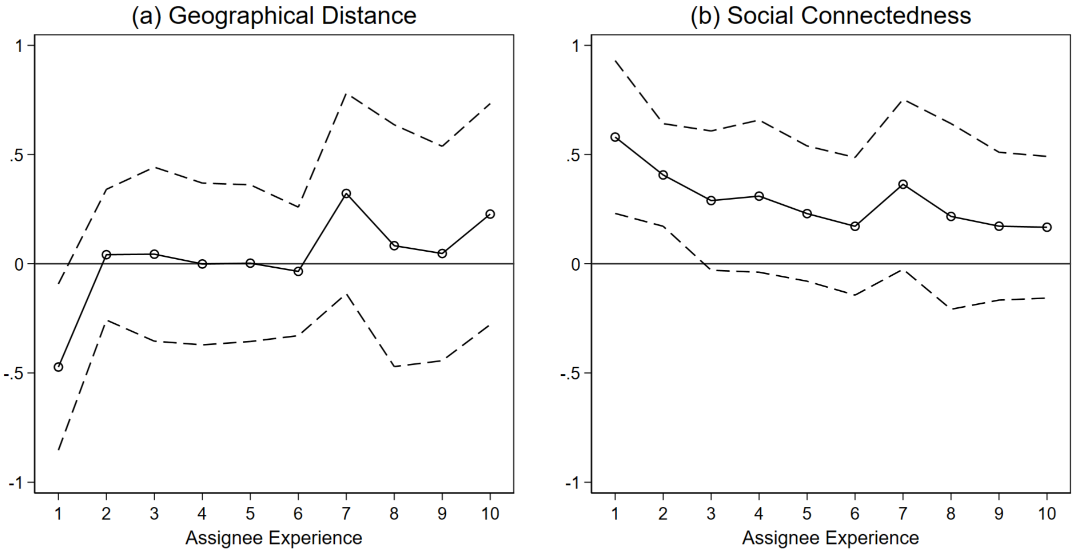
Kernel density plots, epanechnikov kernel with 0.3 bandwidth. Observed control citations are assigned to the county of the applicant. The Alexandria, VA control assigns citations to the examiners' county.

Figure A.4: Marginal effects by citing patent application year



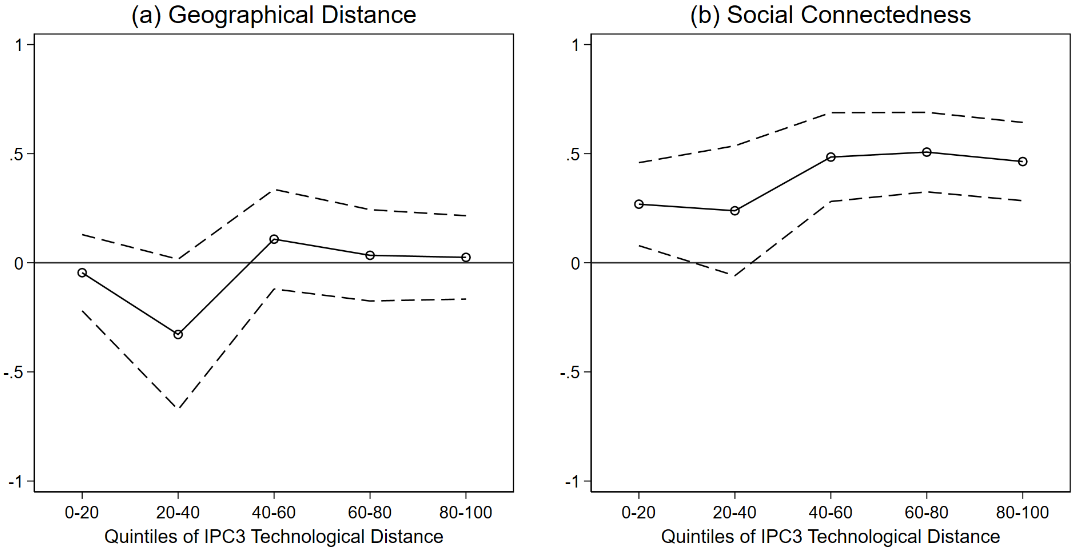
The graph displays coefficients obtained from a regression where ln SCI is interacted with the patent's application year, conditional on application year main effects, ln Distance, inventor networks, and the full set of controls. Dashed lines are 95% CIs.

Figure A.5: Marginal effects by maximum age of citing assignee(s)



Each graph displays coefficients obtained from the same regression, where ln Distance and ln SCI are interacted with the maximum assignee cumulative patent count. Main effects for experience are also included. Dashed lines are 95% CIs.

Figure A.6: Marginal effects by technological distance (IPC3)



Each graph displays coefficients obtained from the same regression, where ln Distance and ln SCI are interacted with quintiles of technological distance. Main effects for each quintile are also included. Dashed lines are 95% CIs.

Figure A.7: Strongly and weakly connected counties to Santa Clara, CA

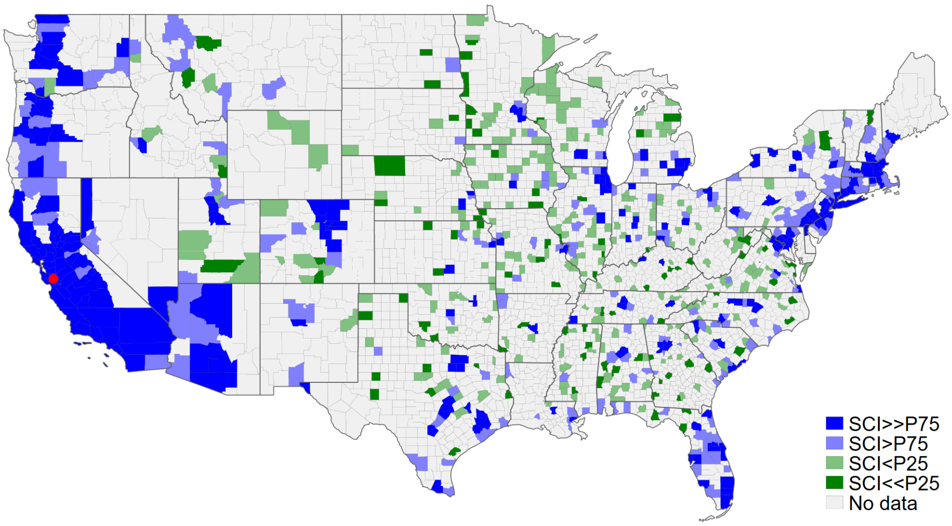
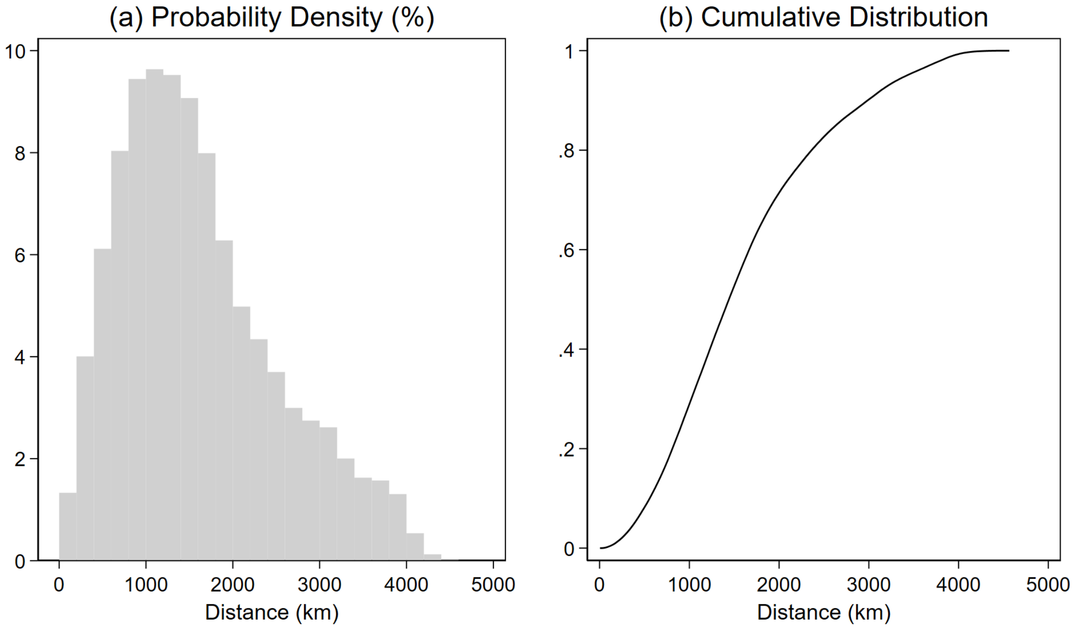


Figure A.8: PDF and CDF of strongly and weakly connected county pairs



PDF and CDF for geographical distance between county pairs below the lower and above the upper quartiles of social connectednes strength with the same third county. In (a), bin width is 200 km.

B Tables

Table B.1: Complete list of WIPO technology fields

Code	Field Title
1	Electrical engineering: Electrical machinery, apparatus, energy
2	Electrical engineering: Audio-visual technology
3	Electrical engineering: Telecommunications
4	Electrical engineering: Digital communication
5	Electrical engineering: Basic communication processes
6	Electrical engineering: Computer technology
7	Electrical engineering: IT methods for management
8	Electrical engineering: Semiconductors
9	Instruments: Optics
10	Instruments: Measurement
11	Instruments: Analysis of biological materials
12	Instruments: Control
13	Instruments: Medical technology
14	Chemistry: Organic fine chemistry
15	Chemistry: Biotechnology
16	Chemistry: Pharmaceuticals
17	Chemistry: Macromolecular chemistry, polymers
18	Chemistry: Food chemistry
19	Chemistry: Basic materials chemistry
20	Chemistry: Materials, metallurgy
21	Chemistry: Surface technology, coating
22	Chemistry: Micro-structural and nano-technology
23	Chemistry: Chemical engineering
24	Chemistry: Environmental technology
25	Mechanical engineering: Handling
26	Mechanical engineering: Machine tools
27	Mechanical engineering: Engines, pumps, turbines
28	Mechanical engineering: Textile and paper machines
29	Mechanical engineering: Other special machines
30	Mechanical engineering: Thermal processes and apparatus
31	Mechanical engineering: Mechanical elements
32	Mechanical engineering: Transport
33	Other fields: Furniture, games
34	Other fields: Other consumer goods
35	Other fields: Civil engineering

Table B.2: Summary statistics for citing patents

	Mean	Median	Std. Dev.	Min.	Max.
Issue year	2017.39	2017	1.07	2016	2019
Application year	2014.69	2015	1.76	2008	2019
Citations per patent	23.36	6	91.15	1	4154
Share of applicant citations	0.62	0.79	0.40	0	1
Cited WIPO	2.83	2	2.72	1	34
Cited IPC3 (first)	3.16	2	3.60	1	65
Cited IPC4 (first)	4.49	3	6.52	1	169
Team size	1.96	1	1.43	1	37
Team US geog.	1.75	1	1.02	1	17
Assignee age (max)	19.11	15	15.06	0	43
Assignee experience (max)	10025.53	660	25651.18	1	131150
Number of citing patents					483,183
Share of citing patens with only applicant citations					0.29
Share of citing patens with only examiner citations					0.22

Table B.3: Summary statistics for citing patents, 20% random sample over 2002-2019

	Mean	Median	Std. Dev.	Min.	Max.
Issue year	2011.37	2012	5.11	2002	2019
Application year	2008.40	2009	5.34	1994	2019
Citations per patent	19.78	7	60.40	1	3303
Share of applicant citations	0.57	0.67	0.40	0	1
Cited WIPO	2.83	2	2.48	1	33
Cited IPC3 (first)	3.16	2	3.20	1	65
Cited IPC4 (first)	4.37	3	5.47	1	141
Team size	1.84	1	1.32	1	37
Team US geog.	1.67	1	0.96	1	21
Assignee age (max)	17.41	15	13.36	0	43
Assignee experience (max)	6963.34	483	18169.60	1	131136
Number of citing patents					362,398
Share of citing patens with only applicant citations					0.21
Share of citing patens with only examiner citations					0.24

Table B.4: Summary statistics for cited patents

	Applicant		Examiner		Total	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
SCI	20,674.50	80,388.83	18,580.43	81,191.61	20,478.16	80,466.76
ln SCI	6.21	3.03	5.80	3.12	6.17	3.04
Distance (km)	1,573.65	1,426.60	1,630.55	1,375.90	1,578.99	1,422.02
ln Distance	6.06	2.63	6.27	2.48	6.08	2.62
Prof. network	0.19	0.39	0.12	0.33	0.19	0.39
Same inventor	0.07	0.26	0.07	0.26	0.07	0.26
Co-authored	0.07	0.25	0.03	0.17	0.07	0.25
Shared co-author	0.05	0.22	0.02	0.15	0.05	0.22
Same assignee	0.11	0.31	0.09	0.29	0.11	0.31
Issue year	2004.01	6.74	2004.50	7.34	2004.06	6.80
Application year	2001.08	6.27	2001.68	6.86	2001.14	6.33
Patent age (since app. +18m)	11.64	6.12	10.93	6.61	11.57	6.17
Patent age (since issue)	10.21	6.59	9.62	7.11	10.16	6.64
Same county	0.12	0.33	0.11	0.31	0.12	0.33
Same CZ	0.05	0.23	0.04	0.20	0.05	0.23
Other state	0.72	0.45	0.77	0.42	0.73	0.45
Number of applicant citations						10,129,600
Number of examiner citations						1,158,574
Total number of citations						11,288,174

Table B.5: Summary statistics for cited patents, 20% random sample over 2002-2019

	Applicant		Examiner		Total	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
SCI	19,926.61	81,609.65	17,562.95	81,970.22	19,607.66	81,662.38
ln SCI	6.13	2.99	5.69	3.07	6.07	3.01
Distance (km)	1,606.18	1,416.38	1,642.18	1,369.14	1,611.04	1,410.15
ln Distance	6.13	2.60	6.33	2.42	6.15	2.58
Prof. network	0.16	0.37	0.10	0.31	0.16	0.36
Same inventor	0.07	0.25	0.06	0.24	0.06	0.25
Co-authored	0.06	0.23	0.03	0.16	0.05	0.23
Shared co-author	0.04	0.20	0.02	0.13	0.04	0.19
Same assignee	0.10	0.30	0.08	0.27	0.10	0.30
Issue year	2000.19	7.13	1999.51	7.51	2000.10	7.19
Application year	1997.59	6.66	1997.03	7.07	1997.52	6.72
Patent age (since app. +18m)	10.15	6.00	8.73	6.32	9.96	6.06
Patent age (since issue)	9.05	6.28	7.77	6.59	8.87	6.34
Same county	0.12	0.33	0.10	0.30	0.12	0.32
Same CZ	0.05	0.22	0.04	0.20	0.05	0.22
Other state	0.74	0.44	0.79	0.41	0.75	0.44
Number of applicant citations						6,047,894
Number of examiner citations						1,118,846
Total number of citations						7,166,740

Table B.6: Overview of all variables used in the analysis

For county pairs:	Mean	Std. Dev.	Min.	25th Pct.	Median	75th Pct.	Max.
SCI	146.75	3,226.25	0.00	1.91	7.71	30.97	1,000,000.00
ln SCI	2.05	2.17	-6.67	0.64	2.04	3.43	13.82
Distance (km)	1,531.01	1,055.81	0.00	714.72	1,282.30	2,189.93	4,561.70
ln Distance	7.00	1.01	0.00	6.57	7.16	7.69	8.43
Gross mig. flow	217.59	2,875.36	0.00	0.00	0.00	0.00	325,606.00
ln Gross mig. flow	0.99	2.19	0.00	0.00	0.00	0.00	12.69
D Bachelor (%)	12.87	9.58	0.00	5.20	11.00	18.70	63.20
D Inventors (%)	0.00	0.00	0.00	0.00	0.00	0.00	0.01
D Density	1,847.29	6,048.23	0.00	203.52	596.08	1,475.67	69,467.53
D Median income	7,936.53	6,442.00	0.00	2,836.00	6,334.00	11,596.00	47,098.00
D Unemployment (%)	2.60	2.09	0.00	1.00	2.10	3.70	24.70
D White (%)	21.14	16.64	0.00	7.57	17.26	31.30	94.65
D Black (%)	11.03	11.88	0.00	2.44	7.03	15.65	81.53
D Asian (%)	4.17	5.43	0.00	0.96	2.36	4.80	33.00
D Hispanic (%)	11.65	12.67	0.00	2.63	7.07	16.22	95.06
For patent pairs:	Mean	Std. Dev.	Min.	25th Pct.	Median	75th Pct.	Max.
Citation	0.91	0.29	0.00	1.00	1.00	1.00	1.00
Prof. network	0.19	0.39	0.00	0.00	0.00	0.00	1.00
Same inventor	0.07	0.26	0.00	0.00	0.00	0.00	1.00
Co-authored	0.07	0.25	0.00	0.00	0.00	0.00	1.00
Shared co-author	0.05	0.22	0.00	0.00	0.00	0.00	1.00
Same assignee	0.11	0.31	0.00	0.00	0.00	0.00	1.00
Same county	0.12	0.33	0.00	0.00	0.00	0.00	1.00
Same CZ	0.05	0.23	0.00	0.00	0.00	0.00	1.00
Other state	0.73	0.45	0.00	0.00	1.00	1.00	1.00
Issue year	2004.06	6.80	1982.00	1999.00	2004.00	2010.00	2019.00
Application year	2001.14	6.33	1981.00	1997.00	2001.00	2006.00	2017.00
Patent age (since app. +18m)	11.57	6.17	0.00	7.00	12.00	16.00	26.00
Patent age (since issue)	10.15	6.64	0.00	4.00	10.00	15.00	27.00
Tech. distance (IPC3)	0.32	0.32	0.00	0.00	0.27	0.56	1.00
Tech. distance (IPC4)	0.40	0.32	0.00	0.03	0.39	0.65	1.00
For the estimation sample:	Mean	Std. Dev.	Min.	25th Pct.	Median	75th Pct.	Max.
ln SCI	5.36	2.41	-6.67	3.89	5.43	7.01	11.37
ln Distance	7.02	1.28	1.43	6.52	7.32	8.03	8.43
Prof. network	0.08	0.27	0.00	0.00	0.00	0.00	1.00

Table B.7: Main regressions with details on bilateral controls

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
ln SCI	0.268	0.347	0.361	0.374	0.446	0.104	0.371	0.0109	0.0161	0.389	0.393
	(0.108)^b	(0.111)^a	(0.110)^a	(0.105)^a	(0.106)^a	(0.0238)^a	(0.0875)^a	(0.0233)	(0.0212)	(0.0913)^a	(0.0953)^a
ln Distance	-0.00372	-0.00839	0.00537	0.0119	0.0657	0.0169	0.0885	-0.0253	-0.00695	0.0227	0.00492
	(0.0983)	(0.0986)	(0.0984)	(0.0955)	(0.0850)	(0.0190)	(0.0673)	(0.0181)	(0.0170)	(0.0856)	(0.0956)
Prof. network	2.941	2.940	2.936	2.480	2.457	0.359	1.386	0.118	0.112	1.960	2.051
	(0.555)^a	(0.556)^a	(0.555)^a	(0.496)^a	(0.495)^a	(0.120)^a	(0.320)^a	(0.0579)^b	(0.0588)^c	(0.387)^a	(0.381)^a
ln Gross mig. flow		-0.0686	-0.0637	-0.0553	-0.0345	0.00968	-0.0335	0.00770	0.00636	-0.0414	-0.0537
		(0.0538)	(0.0535)	(0.0539)	(0.0627)	(0.0158)	(0.0449)	(0.0125)	(0.0117)	(0.0574)	(0.0537)
Top 50 colleges=1			-0.386	-0.397	-0.506	-0.0370	-0.395	-0.0700	-0.0256	-0.377	-0.314
			(0.236)	(0.232)^c	(0.252)^b	(0.0377)	(0.178)^b	(0.0450)	(0.0377)	(0.208)^c	(0.213)
Team size				0.134	0.135	0	0.0782	0	0	0.142	0.145
				(0.0480)^a	(0.0482)^a	(6.36e-18)	(0.0342)^b	(5.56e-18)	(4.64e-11)	(0.0469)^a	(0.0488)^a
Team US geog.				1.328	1.327	0	0.778	0	0	1.217	1.188
				(0.164)^a	(0.164)^a	(4.34e-18)	(0.116)^a	(3.35e-18)	(1.60e-11)	(0.144)^a	(0.137)^a
D Bachelor (%)					0.0283	0.00120	0.0238	0.000350	0.00178	0.0230	0.0221
					(0.0105)^a	(0.00196)	(0.00673)^a	(0.00202)	(0.00194)	(0.00932)^b	(0.00927)^b
D Inventors (%)					158.2	2.223	141.4	-17.46	-8.767	126.8	147.1
					(69.36)^b	(17.59)	(46.67)^a	(11.32)	(15.29)	(58.17)^b	(68.70)^b
D Density					0.0000140	-0.00000437	0.00000371	-0.00000563	-0.00000441	0.00000960	0.0000165
					(0.0000135)	(0.00000449)	(0.0000104)	(0.00000411)	(0.00000174)^b	(0.0000110)	(0.0000102)
D Median income					-0.0000105	-0.00000255	-0.00000686	-0.000000815	-0.00000229	-0.0000109	-0.0000103
					(0.0000129)	(0.00000279)	(0.00000914)	(0.00000279)	(0.00000234)	(0.0000123)	(0.0000128)
D Unemployment (%)					-0.0275	-0.0219	-0.0322	-0.0160	-0.0217	-0.0299	-0.0331
					(0.0392)	(0.00657)^a	(0.0271)	(0.00887)^c	(0.00670)^a	(0.0374)	(0.0369)
D White (%)					-0.00845	-0.0000242	-0.00556	-0.000188	0.000456	-0.00689	-0.00787
					(0.00399)^b	(0.000934)	(0.00277)^b	(0.000963)	(0.000738)	(0.00348)^b	(0.00321)^b
D Black (%)					0.0111	0.00316	0.00840	0.00226	0.000969	0.0131	0.0126
					(0.00704)	(0.00123)^b	(0.00439)^c	(0.00170)	(0.00147)	(0.00616)^b	(0.00599)^b
D Asian (%)					0.00932	0.00203	0.00828	-0.00308	-0.000430	-0.000000376	-0.00584
					(0.0105)	(0.00256)	(0.00665)	(0.00208)	(0.00181)	(0.0101)	(0.0100)
D Hispanic (%)					0.00209	-0.0000972	-0.000348	0.00189	-0.000577	0.00708	0.00648
					(0.00573)	(0.00130)	(0.00412)	(0.00169)	(0.00130)	(0.00540)	(0.00542)
WIPO pairs FEs	●	●	●	●	●	●	●	●	●	●	●
Within citing						●		●	●		
Within cited								●			
Art unit FEs										●	●
Interaction samp.											●
Adj. R^2	0.1145	0.1145	0.1145	0.1174	0.1174	0.5173	0.0914	0.4630	0.4253	0.1413	0.1477
R^2	0.1152	0.1152	0.1152	0.1180	0.1180	0.5279	0.0921	0.4956	0.4391	0.1420	0.1484
N	8,791,193	8,787,610	8,787,610	8,787,610	8,787,417	7,882,961	7,882,961	6,054,214	6,054,214	8,787,348	8,785,291

Two-way cluster-robust standard errors for citing and cited CZ pairs (Cameron et al., 2011). Significance levels: $^a p < 0.01$; $^b p < 0.05$; $^c p < 0.1$. The outcome variable is expressed in terms of percentage points. All specifications use citing and cited year and county fixed effects. The sample excludes citations within same assignee or same county. Within citing and cited patent specifications restrict the sample to patents with at least 10 citations. Interaction controls: main effects for own CZ or state, other state, elapsed time, assignee age, IPC4 technological distance. Column (7) estimates the same model as (5), restricting the sample to that in (6). Similarly, (9) estimates the model in (6) on the sample used in (8). These restrictions allow to compare coefficient changes due to changes in the specification, as opposed to changes in the sample. The reduced ln SCI coefficient in (6) can be largely attributed to the effect of citing patent dummies. By contrast, large part of the fall in the magnitude of ln SCI effects in (8) is due to a change in the sample, as opposed to the use of cited patent dummies.

Table B.8: Robustness checks with citing patent dummies

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
ln SCI	0.104	0.0812	0.100	0.0924	0.0864	0.120	0.153
	$(0.0238)^a$	$(0.0237)^a$	(0.0997)	$(0.0236)^a$	$(0.0244)^a$	$(0.0378)^a$	$(0.0320)^a$
ln Distance	0.0169	0.00358	0.00188	0.00994	0.00482	0.0358	0.0172
	(0.0190)	(0.0188)	(0.0635)	(0.0202)	(0.0195)	(0.0398)	(0.0280)
Prof. network	0.359	0.347	0.666	0.365	0.362	0.246	0.522
	$(0.120)^a$	$(0.106)^a$	$(0.360)^c$	$(0.121)^a$	$(0.123)^a$	$(0.137)^c$	$(0.197)^a$
Tech. pairs FEs	WIPO	WIPO	WIPO	IPC3	IPC4	WIPO	WIPO
Controls	●	●	●	●	●	●	●
Whithin citing	●	●	●	●	●	●	●
Appl. year FEs		●					
Single-authored			●				
Non coastal						●	
Trimmed							●
Adj. R^2	0.5173	0.5167	0.6259	0.5177	0.5225	0.5404	0.3892
R^2	0.5279	0.5274	0.6539	0.5288	0.5365	0.5564	0.4028
N	7,882,894	7,882,893	590,607	8,112,657	8,091,381	4,502,626	5,316,011

Two-way cluster-robust standard errors for citing and cited CZ pairs (Cameron et al., 2011). Significance levels: $^a p < 0.01$; $^b p < 0.05$; $^c p < 0.1$. The outcome variable is expressed in terms of percentage points. All specifications use citing and cited year and county fixed effects, and citing patent fixed effects, restricting to patents with at least 10 citations. The sample excludes citations within same assignee or same county. Controls: citing team size and geography (no. of US counties), gross migration, top 50 college, diff. in education, inventors, density, income, ethnicity. The single-authored sample drops citations sent or received by patents with multiple authors. The non coastal sample drops citations originating or received in Census Divisions bordering the Atlantic and Pacific coasts. The trimmed sample drops patents with citations added exclusively by the applicant or the examiner.

CENTRE FOR ECONOMIC PERFORMANCE
Recent Discussion Papers

1730	Hanming Fang Chunmian Ge Hanwei Huang Hongbin Li	Pandemics, global supply chains, and local labor demand: evidence from 100 million posted jobs in China
1729	Ria Ivandić Tom Kirchmaier Ben Linton	Changing patterns of domestic abuse during COVID-19 lockdown
1728	Jonathan Colmer Ralf Martin Mirabelle Muûls Ulrich J. Wagner	Does pricing carbon mitigate climate change? Firm-level evidence from the European Union emissions trading scheme
1727	Tony Beatton Michael P. Kidd Matteo Sandi	School indiscipline and crime
1726	Maximilian v. Ehrlich Henry G. Overman	Place-based policies and spatial disparities across European cities
1725	Gabriel M. Ahlfeldt Thilo N. H. Albers Kristian Behrens	Prime Locations
1724	Benjamin Handel Jonathan Kolstad Thomas Minten Johannes Spinnewijn	The Social Determinants of Choice Quality: Evidence from Health Insurance in the Netherlands
1723	Claudia Hupkau Barbara Petrongolo	Work, Care and Gender During the Covid-19 Crisis
1722	Ross Levine Yona Rubinstein	Selection Into Entrepreneurship and Self-Employment
1721	Sandra McNally	Gender Differences in Tertiary Education: What Explains STEM Participation?

1720	Edoardo di Porto Paolo Naticchioni Vincenzo Scrutinio	Partial Lockdown and the Spread of Covid-19: Lessons From the Italian Case
1719	Swati Dhingra Stephen Machin	The Crisis and Job Guarantees in Urban India
1718	Stephen J. Redding	Trade and Geography
1717	Arun Advani Felix Koenig Lorenzo Pessina Andy Summers	Importing Inequality: Immigration and the Top 1 Percent
1716	Pol Antràs Stephen J. Redding Esteban Rossi-Hansberg	Globalization and Pandemics
1715	Davin Chor Kalina Manova Zhihong Yu	Growing Like China: Firm Performance and Global Production Line Position
1714	Luna Bellani Anselm Hager Stephan E. Maurer	The Long Shadow of Slavery: The Persistence of Slave Owners in Southern Law-Making
1713	Mathias Huebener Nico A. Siegel C. Katharina Spiess Gert G. Wagner Sevrin Waights	Parental Well-Being in Times of Covid-19 in Germany
1712	Alan Manning Graham Mazeine	Subjective Job Insecurity and the Rise of the Precariat: Evidence from the UK, Germany and the United States

The Centre for Economic Performance Publications Unit
Tel: +44 (0)20 7955 7673 Email info@cep.lse.ac.uk
Website: http://cep.lse.ac.uk Twitter: @CEP_LSE