Papers in Evolutionary Economic Geography

# 13.24

Do inventors talk to strangers? On proximity and collaborative knowledge creation

Riccardo Crescenzi, Max Nathan and Andrés Rodríguez-Pose

Utrecht University
Urban & Regional research centre Utrecht

DO INVENTORS TALK TO STRANGERS?
ON PROXIMITY AND COLLABORATIVE KNOWLEDGE CREATION

Riccardo Crescenzi*, Max Nathan and Andrés Rodríguez-Pose***

* Department of Geography & Environment, London School of Economics and SERC

** LSE, SERC and National Institute of Economic and Social Research

Abstract

This paper investigates how physical, organisational, institutional, cognitive, social, and ethnic proximities between inventors shape their collaboration decisions. Using a new panel of UK inventors and a novel identification strategy, this paper systematically explores the net effects of all these 'proximities' on co-patenting. The regression analysis allows us to identify the full effects of each proximity, both on choice of collaborator and on the underlying decision to collaborate. The results show that physical proximity is an important influence on collaboration, but is mediated by organisational and ethnic factors. Over time, physical proximity increases in salience. For multiple inventors, geographic proximity is, however, much less important than organisational, social, and ethnic links. For inventors as a whole, proximities are fundamentally complementary, while for multiple inventors they are substitutes.

JEL Classification: O31, O33, R11, R23
Keywords: innovation, patents, proximities, regions, knowledge spillovers, collaboration, ethnicity,

1. Introduction

Over the past few decades, collaboration has become an increasingly important element in innovative activity, whether between firms, universities, public agencies and research teams. As national governments seek to develop innovation 'ecosystems', and as firms internationalise their activities, a growing literature has explored the rise of outsourcing, and the tendency of multinational firms' to couple with local partners (Cantwell, 2005, Yeung, 2009); university-industry joint ventures and the growth of Triple Helix relationships (Leydesdorff and Etzkowitz, 1998, D'Este and Iammarino, 2010; D'Este et al. 2013), as well as global scientific collaborations (Archibugi and Iammarino, 2002, Archibugi and Pietrobelli, 2003)..

Collaboration at individual level has, however, been much less well explored. We know relatively little about what drives researchers to work together and about whether agents talk to strangers or prefer to collaborate with people they know. Collaboration is generally driven by physical closeness to the generation of ideas, and the role of geographical proximity in innovation has been understood since Marshall. Collaboration is, nevertheless, not only about physical proximity. It is by definition a social act and, in addition to personal preferences and circumstances, it is shaped by an individual's position in an organisation, the nature and capacity of those organisations, the type of work they do, and by a range of external circumstances – such as legal and funding frameworks, industry and policy trends. In particular, social networks and institutional links (Agrawal et al., 2006, Breschi and Lissoni, 2006) have come to the fore as drivers of innovation. Besides these 'individual', 'organisational' and 'environmental' factors (Lee and Bozeman, 2005), 'relational' influences – specifically, the closeness of individuals to each other in physical, organisational, social or other space – should also affect collaboration decisions. In this paper we focus on these relational factors, or 'proximities', and how they affect inventors' decisions to patent together or alone.

In a seminal paper, Boschma (2005) brings these perspectives together, arguing that geographic proximity is neither a necessary or sufficient condition for facilitating innovative activity. Rather, five different 'proximities', with different pros and cons, may complement or substitute each other. Geographic, institutional, organisational, social and cognitive proximities are all likely to shape collaboration. It is thus important to understand which are most important for different actors, and how they may or may not interact (Torre and Rallet, 2005).

The ‘globalisation of innovation’ in recent years has added further layers of complexity (Fu and Soete, 2010, Scott and Garofoli, 2007, Mowery, 2001). As organisations’ geographic ‘reach’ extends (Mudambi, 2008), are other forms of proximity becoming important? A number of authors, notably Saxenian (2006), Agrawal et al (Agrawal et al., 2008) and Kerr (Kerr, 2008b, Kerr, 2009) have suggested that co-ethnic communities and diasporic networks play critical enabling roles for international teams and multinational firms.

This paper makes multiple contributions to these literatures. First, we work with all five dimensions of the Boschma framework and add a sixth: ethnicity. We do this by developing a new panel of EPO patents microdata taken from the KITES-PATSTAT dataset. We then use the ONOMAP name classification system to identify inventor ethnicity and thus, ethnic/cultural proximity. Second, our data allows us to work at inventor level, and to look at determinants and incidence of collaborative knowledge creation – both areas under-explored in the field to date (Boschma and Frenken, 2009). Third, working at the inventor pair level, we develop a novel case-control-type identification strategy. We are thus able to look at both single and multiple inventors, across all technology fields and to control for a range of individual, institutional, and macro factors – including individual human capital and preferences / constraints. This helps us identify causal effects.

Our results survive multiple cross-checks and contain a number of important findings. For inventors as a whole, we find that local geographic proximity is an important supporting influence on collaboration – but as other studies have found, it is mediated by organisational proximity, and in some cases, by cultural/ethnic closeness. In contrast to views which have heralded the ‘death of distance’ (e.g. Cairncross, 1997), physical proximity has become more important over time, with organisational and institutional proximity declining in salience. For this group, proximities are fundamentally complementary, with joint effects showing up as robust and positive.

Conversely, for multiple inventors (i.e. those who patent more once in their lifetime), we find that geographic proximity is much less important than organisation, social and cultural/ethnic factors. The analysis also confirms the critical role of social proximity and social networks in mediating collaborative activity. For multiple inventors, proximities also appear to be substitutes, not complements.

Overall, the results highlight important differences between the proximities that help inventors collaborate for the first time, the factors shaping repeat interactions, and the behaviour of serial inventors. Physical proximity is critical to break the ice; once a relationship has been established, however, other forms of proximity become more important. And for multiple inventors, geography disappears almost completely as an influence.

The paper is structured as follows. Section 2 reviews the existing literature on the drivers of collaborative working among inventors and outlines a conceptual framework for the analysis of collaboration decisions. Section 3 introduces our data and sets out our identification and estimation strategies. Section 4 gives some stylised facts. The empirical results with a number of robustness checks are discussed in Section 5. Section 6 concludes.

2. The drivers of collaborative working among inventors

The rise of collaborative working can be seen in the growing number of co-authored scientific publications, both international (Glänzel and Schubert, 2005; Glänzel, 2001) and within specific countries. In the US, for instance, Adams et al. (2005) find a 50% rise in the average number of authors per academic paper during the period 1981-1999. Similar, dramatic shifts can be seen in patenting activity. In the UK ‘co-invented patents’ rose from around 100 per year in 1978 (24.2% of all patents) to over 3,300 in 2007 (66.6% of all patents): over the period as a whole, 57.3% of patents had more than one inventor. Co-patenting increased across all major technology fields; while the share of inventors only working alone fell dramatically. During the period of analysis the mean size of patenting teams rose from under two to over four.

2.1 Is it all about spatial proximity?

In the 1990s a number of economic analyses – most notably Jaffe et al (1993) – suggested that geographical proximity plays an important role in facilitating local knowledge spillovers and innovative activity. Jaffe et al. (1993) was the first study to use patent citations as a way to provide a 'paper trail' for knowledge flows. However, their method and the role of spatial distance has been increasingly criticised, with some research even suggesting that the role of geographical factors may have been overstated (Thompson and Fox-Kean, 2005). Recent research has also shifted its attention to the relevance of different types of distance and to the relative importance of geographic vs. other non-spatial factors in explaining the drivers of knowledge exchange.

Some analyses have re-stated the importance of physical proximity. Lobo and Strumsky (2008), for example, find that spatial agglomeration of inventors in US MSAs is more important than the density of inventor connections (which are negatively correlated to patenting). Similarly, Fleming et al (2007) fail to find links between small world networks and regional innovation, suggesting that spatial proximity might be a crucial enabling factor for the effective transmission of knowledge flows.

By contrast, empirical analyses of innovation drivers suggest that socio-economic factors and proximities might play a crucial role after controlling for the exposure to spatially-mediated knowledge flows both in Europe and in the United States (Crescenzi et al. 2007; Rodriguez-Pose and Crescenzi 2008). Breschi and Lenzi (2011) find that the social networks not only play a crucial role for the development of synergies within agglomerated urban contexts, but also that they ensure knowledge circulation between urban centres and are positively linked to innovation in US metropolitan areas. In the case of the European regions, ideas generation and knowledge exchange seem to be driven more by technological and cognitive congruence between innovative agents than by physical distance, while social and organisational links have a modest effect (Marrocu et al. 2011). The analysis of cluster-level links in France, suggests simultaneously positive effects of relational, cognitive, and geographical proximities on the intensity of interactions (Amisse et al. 2011). In a similar vein, Ponds et al (2010) find that both geographic proximity and university-industry networks help explain the innovative performance of Dutch regions.

Micro-level analyses have reached similar conclusions on the simultaneous interplay of a variety of drivers for knowledge exchange and cooperative innovation projects. The analysis of the spatial spread of world-wide patent citations suggests a long term-decline in the ‘home bias’ effect, i.e. in the tendency of patents to disproportionately cite other patents from their same country of origin (Griffith et al. 2011). This highlights the existence of a gradual decline in the importance of geographical distance. However, the fall in home bias over time varies across sectors. It is weaker for pharmaceuticals and information/communication technology than for other sectors (Lychagin et al. 2010). The role of cognitive and social proximity has also been stressed when explaining the formation of R&D networks in Europe (Paier and Scherngall 2008). In the UK, both geographic proximity and research quality are significant explanatory factors in the frequency of university-industry partnerships (D’Este and Iammarino 2010; D'Este et al. 2013).

2.2 Relational factors, non-spatial proximities and collaborative activities

In cities, the role of physical proximity is likely to be more important. Agents physically close to each other should be able to easily work together, can choose optimal collaborators from a large set of potential team members, and draw ideas from their surroundings via local knowledge flows. However, social network theory emphasises the role of social capital and relational structures in fostering relationships and enabling collaboration (Burt, 1992, Granovetter, 1973). Breschi and Lissoni (2009) argue, contra Marshall, that knowledge is not 'in the air' and accessible to all actors in an area, but rather follows specific channels between linked individuals.

A handful of recent papers have begun to explore the role of non-spatial factors in the formation of relationships and links at the individual and group level. From a social network perspective, Evans et al (2011) find some evidence that homophily explains co-authorship on scientific papers, but institutional and geographic proximity play bigger roles. Similar conclusions are reached by Cassi and Plunket (2010) who study genomics patents in France, and suggest that spatial proximity is highly complementary to social proximity, even if individual partners are in different types of organisations. Singh (2005) confirms that the effect of geography and firm boundaries on knowledge flows diminishes substantially once interpersonal networks have been accounted for. The link between spatial proximity and social networks has been explored be looking at mobility: Agrawal et al (2006) suggest that prior social relationships between inventors help explain current citation patterns even after spatial proximity is altered by mobility decisions. Agrawal et al (2008) conclude that geographic and social proximity operate as substitutes.

From an industrial organisation perspective, Coase (1937) famously shows how organisations arise through the need to co-ordinate multiple market contracts. Those working in firms can take advantage of common organisational rules and culture, enabling easy collaboration and minimising principal-agent problems (Singh, 2005).

Economists have also explored social networks, focusing on the benefits and costs agents face when considering potential connection / collaboration (see Jackson (2006) for a recent review). A number of studies have drawn on principal-agent theory to look at contract formation and partner selection at the individual level (Ackerberg and Botticini, 2002, Sedikes et al., 1999). Building on these insights, Christakis et al. (2010) develop a game-theoretic model of strategic network formation, a sequential process where at each stage, a pair of agents will connect if they perceive this as utility-enhancing. Utility depends on the characteristics of agents, potential partners, and overall network structure.

Boschma (2005) brings these traditions together in a high-level framework. He suggests that various ‘proximities’, by bringing actors closer together in terms of space, knowledge, relationships or contracts, assists innovation by overcoming co-ordination and control problems. He distinguishes five types of proximity – cognitive, organizational, social, institutional and geographical. Boschma suggests first, that these factors may operate as substitutes or complements; and second, that proximities are not always beneficial. Excessive proximity causes forms of ‘lock-in’ – a lack of openness and flexibility that inhibits innovation.

For example, cognitive or ‘technological’ proximity allows agents to communicate in the same research field (Seely Brown and Duguid, 2002). Organisational and social proximity lower transaction costs via (respectively) contracts and social relationships (Kaiser et al., 2011). However, hierarchical organisational structures or supply chain relationships close firms off from the technological opportunities that ‘open innovation’ models provide (Von Hippel, 2005). And social networks based on ‘strong ties’ may be less effective than larger networks of ‘weak ties’, if they do not admit new members or new thinking (Granovetter, 1973; Crescenzi et al. 2013).

Two recent trends have added another layer to this type of relational thinking. As innovation systems ‘globalise’ and numbers of international migrants, have grown researchers have become interested in ethnic/cultural proximity, and specifically co-ethnic and diasporic networks (see Docquier and Rapoport (2011) for a recent survey). By increasing trust and lowering transactions costs, co-ethnic group membership may assist collaborative ideas generation and diffusion. In a global context, transnational diasporic networks may accelerate knowledge transfer – shaping the development of high-tech hubs in both ‘host’ and ‘home’ countries (Kerr and Lincoln, 2010, Saxenian and Sabel, 2008, Kapur and McHale, 2005). Just as with other proximities, through, however, co-ethnic groups’ capacity may vary substantially, and is potentially limited by discrimination.

3. Research questions and empirical approach

This review of the literature highlights a number of under-explored areas. First, we still know relatively little about how proximities shape individuals’ behaviour – most studies aggregate outcomes to firms, cities and regions. Second, while many studies explore the effect of collaboration on total innovative activity, or the impact of research, much less work has been done on determinants of knowledge creation (Boschma and Frenken, 2009). Third, due to constraints on data and identification, there are few studies that have been able to explore time periods above a decade. Fourth, the role of cultural and ethnic proximity has been particularly neglected in quantitative analysis outside the US (see Kerr and Kerr (2011) for a review of the American literature). As far as we are aware there is only one study for Europe – Nathan (2011b) for the UK.

These gaps raise three important research questions:

1) What forms of proximity influence the incidence of collaborative knowledge creation at the individual level?
2) What is the interaction between different proximities on inventors’ behaviour?
3) How has the salience of these proximities changed over time?

This paper aims to answer these questions by looking at how inventors' characteristics and relationships to each other influence levels of collaborative knowledge creation, specifically co-invention. We focus on collaborations between individuals by means of a simple model of individual 'collaboration decisions' in order to disentangle the impact of non-spatial/relational factors from spatial proximity and other collaboration drivers identified in the existing literature (i.e. 'individual', 'institutional' and 'environmental')..

How do individuals decide to work together? A useful way to think about the 'collaboration decision' is to think of it as three linked decisions:

1) Should I collaborate or not?
2) Who should I collaborate with?
3) What type of collaboration should I undertake?

Three important consequences follow from this perspective. First, these decisions are – obviously – closely interlinked. Consider a manager working with their employees, a contract between two individuals, and a research team of peers. In each case the *type* of possible collaborations will clearly influence, and be influenced by, the set of *potential collaborators* and the incidence of *actual* collaborations that occur. Second, we suggest that this interconnectedness means that collaboration decisions are generally taken simultaneously. Third, as the example above implies, decisions are not taken in a vacuum – as we discuss in the next section, individuals' opportunities, choices and constraints will likely be shaped by their age, gender, professional role, institutional power structures and wider social, economic and cultural factors.

Ideally we would want to able to observe all three decisions, as well as these wider conditions. In practice it is difficult to think of a dataset (s) that would allow us to do this. In the case of patent data, we are able to observe 1) actual collaborations and 2) the collaborators, but will need to infer potential collaborators not chosen. We have rather less information on 3) the type of collaboration, but are able to identify whether inventors work in a 'pure' partnership, or as part of a larger team.

3.1 Data

Our dataset contains EPO patents microdata from the PATSTAT database, modified by the KITES team at Universita' Bocconi (hence 'KITES-PATSTAT'). The raw data runs from 1978-2010, comprising 116,351 patents with at least one UK-resident inventor. 173,180 inventors are associated with these patents, of whom 133,610 are UK residents. Unlike standard patents data, KITES-PATSTAT has been cleaned to allow robust identification of individual inventors, their spatial location and patenting histories, as well as the usual array of patent and applicant -level characteristics (see Lissoni et al (2006) for details of the cleaning process).[1]

Patent data have a number of advantages for exploring individuals' collaborative behaviour. They provide rich data available over a long time period, as well as detailed information on individual inventors and their past/present collaborators, the type of research they work in (via detailed 'technology field' codes), and the organisations they work for (typically the patent applicant) (OECD, 2009). On the other hand, the data have two inherent limitations. Patents measure *invention* rather than *innovation*; and they tend to only observe some inventions and inventors (for instance, some members of a research team may be left off the patent application). We argue that these issues generate noise, rather than bias. Other limitations which might induce bias, such as patenting's manufacturing focus and vulnerability to policy shocks, are more easily dealt with using appropriate industry controls and time trends, as we discuss below.

We make some basic edits to the data to make it fit for purpose. First, there is typically a lag between applying for a patent and its being granted. This means that in a panel of patents, missing values typically appear in final periods. Following Hall et al (2001), we truncate the dataset by three years to end in 2007. Second, we geo-locate UK-resident inventors in UK Travel to Work Areas (TTWAs). TTWAs are designed to represent functional labour markets, and offer a good proxy for the local spatial economy.

[1] KITES-PATSTAT also provides extensive applicant-level information, particularly for corporate applicants: names/address details are matched to company information from Dun and Bradstreet.

Third, EPO patent data gathers patent applications through EU countries' national patent offices, through European-wide applications to the EPO ('Euro-direct'), and international applications that have reached the European examination stage ('Euro-PCT') (OECD, 2009). As such, our dataset may not include all PCT patent applications, which cover international applications to multiple patent offices. Since PCT applications are increasingly the favoured route for researchers looking to access international markets, there is a risk of selecting out some collaborative activity. However, the use of PCT applications has increased substantially since the early 2000s, so this is likely to affect only part of our sample. Further, by truncating the end of the time series, we minimise the PCT selection issue.

3.2 Identification

In order to build a sample for regression analysis we work at inventor level.[2] In particular, as we are interested in collaborations between inventors, our preferred approach is to make the unit of observation the inventor pair. In this we follow some key studies (e.g. Agrawal et al (2008, 2006) and Ponds et al (2007)). More importantly, we also follow the basic structure of our data – two-inventor patents comprise the single biggest category of co-invented patents. We are fundamentally interested in what influences two inventors to work together (or not), and specifically what makes the incidence of co-invention lesser or greater. This means we are interested in both *possible pairs* – inventors who *might* work together – and *actual pairs* (those who do).

The intuition is that by specifying each inventor's options for collaboration, and by the examining the characteristics of both possible and actual pairs, we identify the factors influencing collaboration *and* control for the underlying incidence of co-invention. Our approach can thus be considered a form of case-control analysis. However, the particularities of inventor activity and research collaboration force some departures from the techniques deployed by Jaffe et al (1993), Thompson and Fox-Kean (2005), Agrawal et al (2006) and others.

[2] To see why inventor-level analysis is appropriate, consider an opposite scenario: working at the patent level. For example, we could estimate a model where the dependent variable would be a dummy taking the value 1 for a co-invented patent; independent variables would cover characteristics for the inventor / set of inventors involved. This has some desirable characteristics – not least, allowing us to look at all the patents in our sample – but discards a lot of inventor-level information. An additional challenge, for patents with more than one inventor, is where to locate the patent in space.

In theory, we might want to observe all possible inventor pairs in our sample, including the subset of inventor pairs who actually collaborate. Specifically, we would observe the set of *all* possible inventor pairs *ij* where $i \neq j$, giving a panel of all inventor pairs * time * area. In practice, we need to impose reasonable restrictions on possible pairs – for example, it is very unlikely that those inventors active in 1978 will be collaborating with those active in 2007.[3]

Our approach is therefore as follows. First, we randomly sample 5% of patents; we stratify the sample by year, 121 three-digit technology fields and inventor team size.[4] Stratification helps ensure that the sample conditions for underlying time and field trends in patenting activity, as well as the underlying distribution of collaborative activity across time, by field and by area. Second, using the sample of patents to create a sample of individual inventors, we create a set of *possible inventor pairs* (pairs who might have co-invented), alongside a much smaller set of *actual inventor pairs* (pairs who really did co-invent). We impose the basic constraint that each member of a possible pair must be active in the same time period.[5] Third, for each year group we create the set of possible pairs, combine with actual pairs, and append each cross-section to create an unbalanced panel of inventor pairs * years * TTWAs. Each inventor is separately coded by address and by patent applicant, allowing us to specify various fixed effects for each member of the pair.

We build a 16-year panel for the years 1992-2007 inclusive. We reserve the period 1978-91 to provide historic information on inventors' patenting activity, and on local patent stocks (see below). This helps us control for otherwise unobserved heterogeneity in individual and area characteristics. We end up with an unbalanced panel of 1,484,074 observations for 1,483,775 possible and actual pairs. Of the inventor pairs, 2,254 are actual pairs, with actual pairs making up between 0.09 and 0.75% of the sample in any given year.

[3] Working with all inventor pairs is also computationally intensive – for example, there are 173,180 inventors in our full sample 1978-2007, which makes for several billion inventor pairs.

[4] Specifically, we want to control for underlying trends in patenting over time and across technology fields, and unobserved macro factors influencing team size. Given the size of the original population we sample without replacement. We seed the sample so that regression results are reproducible. We relax this in robustness checks, to test whether sample construction affects our findings.

[5] In principle we could also restrict possible pairs to those working in the same technology field. However, diagnostics suggest increasing incidence of individuals patenting *across* technology fields. More importantly, we want to explore whether this 'cognitive distance' affects levels of co-inventing – so reserve the variation for regressions rather than build into the matching process.

Our sampling approach is unavoidably noisy, as it only provides information on what our inventor sample did on the *sampled* patents, not on all EPO patents.[6] However, assuming we have both sampled randomly and stratified appropriately, our estimates will not be biased.

We also look at the subset of multiple inventors. In our full sample we have 10,420 multiple inventors (just under 10% of all inventors in the panel). There are two main reasons to isolate this group. First, since the majority of inventors only patent once, the multiple group is, by definition, an unusual minority. The inventor lifecycle literature suggests multiple inventors are likely to be highly productive individuals, often in senior positions in science (Azoulay et al., 2006, Lee and Bozeman, 2005). It is therefore interesting to see if proximities affect their behaviour differently to the pooled sample. Second, looking at multiple inventors allows us to explore the effects of a broader set of proximities for this group. Specifically, social and cognitive proximity measures need to be based on historic behaviour in order to avoid a mechanical link between dependent and independent variables in our model. In turn, this requires that we observe more than one patenting event. Including these proximities thus involves restricting our sample to patenting by multiple inventors (specifically, to the patents with only multiple inventors involved).

Because patenting by multiple inventors is a much smaller share of total patenting activity, we are thus able to sample more of the patents, increasing the precision of our estimates. Specifically, we sample 25% of the patents, build actual and potential inventor pairs as before – then combine this with social network information (see ‘model proximities’ section below). For this panel we have 54,425 observations for 595 actual pairs and a much larger number of potential pairs, with actual pairs making up between 0.03 and 0.13% of the sample.

3.3 Estimated model

We estimate the following model. The left-hand side variable covers various aspects of collaborative activity between an actual or potential inventor pair. The right-hand side variables are dummies for inventor pair characteristics (the various proximities we are

[6] We would have a problem if inventor pairs do not patent in our sub-sample, but do patent in the rest of the dataset, and this process is non-random. Since there is no evident reason to expect this pattern of activity, we treat this issue as generating noise only.

interested in), plus vectors of controls for individual, institutional and environmental factors (see Appendix A.1 for a summary of the variables included in the analysis and in the robustness checks).

For inventor pair *ij* in area *a*, year group *t* and technology field *f*, our basic estimating specification is then:

$$Y_{ijatf} = a + \mathbf{PROX}\boldsymbol{b}_{ijatf} + \mathbf{IND}\boldsymbol{c}_{ij} + \mathbf{INST}\boldsymbol{d}_{ija} + \mathbf{ENV}\boldsymbol{e}_{ija} + e_{ijatf} \quad (1)$$

Dependent variables

We construct two dependent variables, covering different aspects of the collaboration decision. These are:

- Collaboration dummy (DCOINVENT) – for a given inventor pair, this takes the value 1 if they patent together, 0 if not. We interpret this as covering both the decision to collaborate, and the collaborator(s) chosen;

- Collaboration count (#COINVENT) – this is a continuous variable recording the number of collaborations per pair in a given year. We take this as a measure of the "productivity" of an inventor partnership.

Independent variables

Our variables of interest are given by **PROX**, a vector of proximities covering spatial, organisational, institutional, cognitive, social and cultural-ethnic proximities . We interpret the estimated coefficients (*b*s) of these PROX variables as the effect of an inventor pair possessing the relevant relational quality on co-inventing activity of that pair – relative to not possessing that quality, after controlling for individual, institutional and macro factors.

For all inventor pairs we fit:

- Geographic proximity (PROXG_LD) – we calculate the linear distance between TTWA centroids where each inventor is located. For geographical proximity, our basic specification is a linear inverse distance function normalised to take a maximum value of 1 where two inventors are based in the same TTWA. For robustness checking we also construct a more sophisticated inverse linear distance function, with a threshold function to capture knowledge spillovers decay (PROXG_LDT). The mean linear distance in our sample is 197km, so we set the threshold at 200km.

- Organisational proximity (PROXO) – patents contain information on both individual inventors and the patent applicant, which is typically the organisation that the inventor works for. We use this information to build a dummy taking the value 1 if inventors in a pair are based in the same applicant, 0 if not. The variable is set as blank where there is no applicant information, or where the applicant is the individual;[7]

- Institutional proximity (PROXI) – KITES-PATSTAT contains detailed information on applicant types. We use this to build a dummy taking the value 1 if inventors in a pair are based in the same *type* of applicant (coded as business / private research lab; university/ public research lab; foundation / NGO / consortium; or individual, the reference category). The variable is blanked where there is no applicant information;

- Cultural-ethnic proximity (PROXE) – these variables are developed using the ONOMAP system, which using inventor surname and forename information to identify likely ethnicity. ONOMAP is described in more detail in Appendix A.2. PROXE_CEL, PROXE_ETH and PROXE_GEO take the value 1 if inventors in a pair share, respectively, the same ONOMAP cultural-ethnic-linguistic (_CEL) subgroup, likely ONS ethnic group (_ETH) or likely geographical origin (_GEO). Of the three, CEL subgroups are coded across 67 categories, ethnic groups nine categories, and geographical origin 13 categories. CEL is the most precise, and thus preferred, measure; the others are reserved for robustness checks.

[7] Around 3% of observations have no applicant information. 34% of applicants are individuals.

For multiple inventors, we also fit the following additional proximity variables that, as discussed above can be computed only for inventors patenting more than once in their lifetime:

- Cognitive proximity (PROXC) – patents are coded using 'technology field' codes, which can be used at varying levels of detail (OECD, 2009). We use this to measure the cognitive proximity of an inventor pair: for a given technology field, PROXC takes the value 1 if both inventors in a pair have patented previously in that technology field. We build the measure for 121 3-digit IPC fields (PROXC_3) and for 1581 more detailed 6-digit IPC fields (PROXC_6). Following Thompson and Fox-Kean (2004) and Singh (2005), the most conservative six-digit specification is our preferred measure.

- Social proximity (PROXS) – a measure of the inverse social distance between two inventors in a pair, specifically whether they have co-invented in the past, have co-authors in common, or more indirect links to actual / possible partners. We take a simple measure of social distance: given the very large sample, it makes sense to choose the simplest possible way of capturing social relationships between inventors (Singh, 2005).[8] We assume that ties decay after five years.

  For a given year, we then measure the number of 'steps' between inventors *i* and *j* based on their activity in the previous five-year period. This is the social distance between *i* and *j*. We then take the inverse distance to generate the social proximity between the two. For example, if *i* and *j* have co-invented together in the past, the number of steps between them is 0. If *i* and *j* have not collaborated directly, but have both worked with *k*, then there is 1 step between them; if *i* is connected to *j* through *k* and *l*, there are two steps; and so on. Respective degrees of social proximity are then 0, -1 ,and -2, through to minus infinity (no link). We then specify PROXS, a continuous measure of social proximity.

[8] We are trading off some finer-grained information (density of relationships, hierarchy etc) for a more tractable solution.

For robustness checking we also build PROXS2, social proximity rescaled into three categories (3 = direct / 2 = indirect / 1 = no link); and PROXS2D1-3, dummies for these three categories. In regressions we take no link (PROXS2D1) as the reference category.

Control variables

In order to identify causal effects, our model also needs to control for wider factors affecting the likelihood, type and incidence of collaboration. Drawing on Lee and Bozeman (2005), Azoulay et al (2006) and others, we group these influences into 'individual' factors, such as individual inventor characteristics; 'institutional' factors, such as differences between universities and private companies; and a series of 'macro / environmental' factors, including macro shocks in a given time period, industry-specific shocks and trends, and geographical concentration of innovative activity and of specific sectors.

IND is a vector of controls for individual characteristics, covering human capital, inventor patenting preferences and status (Lee and Bozeman, 2005, Bozeman and Gaughan, 2011, Evans et al., 2011). Controlling for individual-level factors is difficult using patents data, because so little individual-level information is directly observed. We use inventors' historic patenting activity to develop individual-level controls for human capital/status, and for individual preferences. We follow the approach developed by Blundell et al (Blundell et al., 1995) in a seminal study of innovation in firms. Blundell and co-authors argue that historic patenting information represents an accumulation of knowledge and thus, human capital. Importantly, they argue that this information approximates an individual fixed effect. We argue that this approach also works at individual level: by dividing our sample into two time periods, we suggest that looking at inventors' patenting behaviour in the 'historic' period provides some information on human capital endowments in the 'present' period.

Specifically, for each inventor in a pair we first set the historic period as 1978 to 1991 and model human capital as the inventor's average patenting during this period. For inventors who do not patent in the historic period, we zero the figure. Next, for each inventor we create a dummy variable which takes the value one for inventors who do not patent in the historic period.

We also suggest that exploring the *type of patenting* in a past period (solo, collaborative or mixed) provides similar information on inventors' inherent preferences, as well as indirectly indicating other individual factors such as age and status. Again for each inventor in the pair, we control for individuals' patenting preferences by creating dummy variables for each inventor patenting 'style' in the historic period – always filing solo patents, always co-inventing, or combining solo and co-inventing. We create an additional dummy taking the value one if an inventor in a pair does not patent in the historic period, which we treat as the reference category.[9] Second, as a robustness check we use standard fixed effects at inventor level (this is very computationally intensive, so is not our preferred strategy).

INST is a vector of applicant-level controls, covering institutional conditions. Institutional factors are likely to include institutional type, quality, culture and capacity. Staff at more prestigious research institutions, where the overall quality of research is high, are likely to receive more offers of collaboration (D'Este and Iammarino, 2010). Cutting across this, some research institutions may seek to foster a culture of active collaboration; others may discourage it (Azoulay et al., 2006, Ponds et al., 2010). Regardless of institutional strategy, the capacity to foster collaborative activity may vary substantively across organisations. For academics, in particular, the quality of a University's Technology Transfer Office could make a substantive difference to the pool of potential non-academic collaborators (Lee and Bozeman, 2005). To control for these issues, we fit dummies for public sector, private sector and 'other' applicant types, with individual applicants the reference category. Again, two sets of dummies are fitted.

ENV covers macro / environmental factors, including technology or subject field differences, local area context, and shifting policy frameworks. The increasing costs of research equipment in some sectors, notably hard sciences, are likely to increase incentives to collaborate; as a result, in recent years a number of interdisciplinary fields have emerged, such as biotechnology (Lee and Bozeman, 2005). And a sequence of policy decisions both at national and European level – notably the structure of EU research funding – also increasingly encourage collaboration (Ponds et al., 2007). In **ENV** we thus include 16 year dummies; a grouping variable for 1581 technology fields, following Thompson and Fox-

[9] Fitting an ordinal variable would be more parsimonious, but there is no obvious scaling for the different types of patenting activity.

Kean (2005); 243 TTWA dummies for each inventor in the pair, and historic area weighted patent stocks for each inventor's TTWA.

Choice of estimator

Our choice of estimator comes down to whether or not to fit a linear model. Our preferred strategy to deal with individual-level unobservables uses two sets of 'levels effects' rather than conventional fixed effects; we also run a large number of dummies at technology field, applicant and area levels. This presents us with a 'high-level fixed effects' issue, where conventional linear / non-linear estimators may be extremely inefficient and may not be computationally possible (McCaffrey et al., 2010). We experiment with a number of potential high-level fixed effects estimators, and explore alternative specifications in robustness tests.[10]

Given our binary / count data structure, some would argue that a non-linear specification is preferred to deliver efficient estimates. Running non-linear estimators with so many fixed effects is not straightforward. Angrist and Pischke (2009) also convincingly argue that once raw coefficients from non-linear estimators are converted to marginal effects, they offer little efficiency or precision gains over linear specifications. In robustness tests, we check whether linear/non-linear specifications make a difference.

3.4 Wider endogeneity challenges

In order to identify causal effects, we also need to tackle a series of other endogeneity issues. We deal with each in turn.

The first issue is endogenous partner selection. Consider a contract decision between a principal P and an agent A. Ideally, P and A observe everything about each other, reaching the optimal contract. In reality, there are unobservable qualities of P and A which affect type of contract chosen. This is the 'endogenous matching problem'. Typically, models of contract choice use proxies for aspects of P and A that will affect contract choice. However, this does

[11] This classification is used for illustrative purposes only. In the regression analysis we use this 30-fold typology, alongside more detailed typologies of 121 IPC three-digit sub-classes and 1851 six-digit main classes to generate technology field fixed effects.

not solve the endogenous matching problem – unobservables remain, so proxies are correlated with the error term and coefficients of P and A characteristics are biased.

We might face a version of this problem, since our pair-level controls are proxies which are unlikely to capture every salient factor shaping collaboration decisions. As set out by Ackerberg and Botticini (2002), however, we can use our individual-level controls to operate as proxy individual fixed effects.

A second issue concerns the presence of third parties. So far we have assumed that A's decision to co-invent with B (or not) is not affected by the presence of C or D. But as Sedikes et al (1999) point out, this assumption may not hold. My decision to partner with A rather than B may be affected by the presence of C, which shifts relative positions of A-B-C on specific decision axes. This relates to our third collaboration decision, namely how co-invention activity is affected by the *structure* of collaboration chosen (see section 3). In this case, patent data gives us a limited view on collaboration structure by allowing us to see inventor team size; we use this to generate a TEAM dummy which takes the value 1 if the pair is part of a co-inventing team larger than two individuals. Since our starting point is that inventor team structure is determined simultaneously with the decision to patent, this implies TEAM cannot be fitted with models with the co-invention dummy on the left hand side. However, in robustness checks we explore whether inventor teams affect the *number* of co-invented patents filed by a given inventor pair.

Working at inventor level raises two further identification issues. First, we are interested in the number of co-invented patents which inventor pairs produce in a given year, and we set inventor activity to zero in cells when they are not patenting. But our data structure means that we do not actually *observe* inventors when they are not filing patents: they might be working on other inventions, or might be inactive. If there are structural patterns here, this may lead to omitted variable bias – in which case a more conservative specification would be to *blank* all cells in which inventors are not active. In a related paper, Nathan (2011) tests both approaches on a subset of multiple inventors – finding both deliver identical results. We therefore feel confident with a zero-basing assumption for our analysis.

Second, we face potential simultaneity / reverse causality problems if inventors tend to move into TTWAs that enable collaboration (or to larger / higher-capacity applicants). Resolving this issue is harder because it is impossible to definitively identify movers using patents information. Following the procedure in Agrawal et al (2006), we identify 14.2% of inventors as *likely* movers across TTWAs. Using a more cautious identification strategy on the same data, Crescenzi and Gagliardi (forthcoming) find around 5% of inventors as likely movers. On this basis, we suggest that potential moving inventors are unlikely to affect results.

4. Descriptive analysis

Collaborative research and invention has progressively increased its importance over time and across technology fields. Figure 1 provides the time trend for the count of co-invented patents. Over the whole period, 57.3% of patents were 'co-invented'. Co-invention is the norm: 15.9 % of inventors only work alone (i.e. never co-invent); 4.7% sometimes co-invent; 79.4% only co-invent. Patents with five or fewer inventors comprise over 95% of the sample, of which over half are co-invented. Most co-invented patents have two or three inventors, with two being – by some way – the single largest group (26.2% of patents, versus 14.7% of patents with three inventors).We can see three distinct phases within the sample period: from the late 1970s to the late 1980s; the 1990s, with a peak in co-inventing in 2000; and then a plateau period, with a slight decline at the end of the panel (probably reflecting fewer granted patents).

Figure 1 about here

Co-invention trends vary substantially across patent fields. Figure 2 shows the trends in co-invented patents across seven major technology fields (generated using the OST reclassification)[11]. At the start of the sample period, shares of co-inventing are low (from 0.14 in consumer goods to 0.22 in electrical engineering). By the end of the period shares are higher but there is also much more variation, from 0.49 in consumer goods to 0.83 in chemicals and materials. In six out of seven cases, co-invention has shifted from minority to majority type.

[11] This classification is used for illustrative purposes only. In the regression analysis we use this 30-fold typology, alongside more detailed typologies of 121 IPC three-digit sub-classes and 1851 six-digit main classes to generate technology field fixed effects.

Insert Figure 2 about here

Inventors' behaviour has also altered over time, in line with patenting shifts. Figure 3 shows that these aggregates hide some quite large movements within the sample. Counts and shares of 'only co-inventing inventors' have risen substantively; in contrast, while counts of 'only solo' inventors have risen slightly, their relative shares have declined a lot.

Insert Figure 3 about here

Finally, we look at inventor team composition. Figure 4 shows the trend in average team size. The trend line is a lot spikier than co-invention trend, but the general shape is the same. We can see that the average patent in 1978 had 1.73 inventors; in 2007 this had risen to just over four inventors.

Insert Figure 4 about here

Taken together, these stylised facts suggest a substantial rise in co-patenting between the late 1970s and the late 2000s, across technology fields, and involving significant changes in inventor behaviour.

5. Results

The regression analysis explores the drivers of these changes, and is organised into three sections. The first section looks at the results for all inventors (larger sample but more limited set of explanatory variables in terms of proximities), while the second section looks at the sub-sample of multiple inventors (for these more established and 'regular' inventors we can compute a broader set of indicators, including position in social networks). The third section includes a number of robustness checks.

5.1 All inventors

For the full sample of inventors it is possible to test the effect of geographical, organisation, institutional and cultural/ethnic proximities. Results for the collaboration dummy are given in Table 1, and for co-invention counts in Table 2. Columns 1 includes the basic specification with all proximities and controls (human capital, preferences, applicant type, and area-level and technology field characteristics as discussed in section 3). Columns 2-3 look at the time split for the 1990s and 2000s while in columns 4 to 7 all non-spatial proximities are progressively interacted with spatial distance. Columns 8-9 include the time splits for these interaction terms. In the interpretation, we focus on the relative sign, size and significance of the proximities variables, rather than trying to unpack specific point estimates.[12]

Collaboration

The results for the full specification are reported in Table 1. [13] Column 1 suggests a robust positive influence of local geographic proximity on inventor pairs' tendency to co-invent, with the point estimate significant at 1%. Echoing other literature, we also find a much stronger influence of organisational proximity, also significant at 1%. The pairwise correlation between these two measures is about 0.221, indicating that some 'hyper-local' proximity is captured, but largely an organisational proximity dynamic. It is also likely that two omitted proximities, social and cultural closeness, are partly reflected in this result – something confirmed in our multiple inventors analysis (see next section). Institutional proximity, perhaps surprisingly, has a negative significant effect, at 5%, but point estimates are very small and close to zero. In this model, cultural/ethnic proximity is insignificant.

[Insert Table 1 here]

Columns 2 and 3 split the panel into two periods, 1992-1999 (column 2) and 2000-2007 (column 3). The results point to an increasing salience of geographic proximity from the 1990s to the 2000s, with a decreasing influence of organisational proximity (although the

[12] For the full sample, Table B-1 in Appendix B gives summary statistics (first panel) and correlations matrices of the proximities variables (second panel). As we do not have local area information for all inventors, sample sizes for panel regressions drop to around 844,000 observations once controls are added. Correlation matrices indicate that our variables of interest are free of collinearity problems, and the results are confirmed in VIF tests. Other model fit statistics are also satisfactory, with R^2 around 0.38.

[13] When controls for individual, institutional, and environmental conditions are fitted progressively before the full specification presented in column 1, the basic specification survives the addition of the full set of controls more or less unscathed, although technology field controls do shift point estimates substantially.

effect of the latter is still much larger than the former). The effect of institutional proximity also wanes; ethnic-cultural proximity coefficients change little between the time periods. As expected, these results reflect some of the global trends in the organisation of research activity over the same period, with a growth in cross-organisation and public-private working.

Column 4 fits the interaction of geographic proximity with the other proximity measures, in turn (columns 4-6) and together (column 7). The results show that geographical proximity is complementary to the other proximities: with the exception of cultural/ethnic closeness, joint effects are all positive significant at 1%. In the full specification, however, only geography * organisational proximity remains significant, and the joint effect of geography * institutional proximity turns negative (although insignificant). Any separate effect of local geography disappears in the complete model, as it does for both time periods.

<u>Inventor pair activity</u>

Table 2 looks at the counts of co-invented patents for actual and possible inventor pairs.

[Insert Table 2 here]

Column 1 shows that results for the full specification are broadly similar to the co-invention dummy models. Local geographic proximity matters for the amount of collaboration between inventor pairs, but organisational proximity matters more. The basic difference with the previous results is that ethnic/cultural proximity is now marginally significant and positive. This suggests that the salience of proximities also varies depending on the type of activity observed. Columns 2 and 3 run time splits. Unlike the collaboration dummy models, both local geography and organisation proximity become less important over time, although they remain statistically significant in the 2000s, at 5% and 1% respectively. Splitting the sample removes the significance of ethnic/cultural proximity, although coefficients in the 1990s are a lot bigger than the 2000s. Columns 4 to 6 show results for interactions, which are very similar to those for the collaboration dummy: joint effects dominate, especially the joint effect of physical and organisational proximity.

5.2 Multiple inventor analysis

We now look at the subset of multiple inventors in more detail. As mentioned in Section 3, multiple inventors make up around 10% of our inventor sample. Focusing on multiple inventors allows us to explore the behaviour of this distinct group and to estimate the effect of a wider set of proximities on their patenting behaviour.[14] The results are given in Tables 3 and 4.

Collaboration

Collaboration results are covered in Table 3. Column 1 fits a specification for multiple inventors, with the same four proximities as the full sample. Column 2 adds social and cognitive proximity measures. Columns 3 and 4 split the sample into the 1990s and 2000s time periods. Columns 5 to 10 interact each proximity with geographical distance.

[Insert Table 3 here]

The base specification (column 1) confirms that multiple inventors are differently affected by proximities than occasional ones. Most strikingly, for those involved in more than one patenting project, local geographic proximity is no longer significant, at least as measured by linear distance, and point estimates are close to zero (column 1). By contrast, organisational proximity is a very powerful determinant of repeat collaborative activity. Note that some of this result reflects ‘hyper-local’ geographic proximity – for instance, people working in the same building – but given the pairwise correlation of 0.43, we are also picking up a distinct organisational proximity effect.

As expected, exploring the full set of proximities also reveals some important differences (column 2). Specifically, once we add social and cultural proximity measures, the magnitude of organisational proximity declines markedly. Social proximity exhibits a much stronger

[14] Table B-2 in Appendix B gives summary statistics (first panel) and correlations matrices of the proximities variables (second panel). In this case, because the social and cognitive proximity variables are based on past inventor behaviour, pairwise correlations between these proximities and dependent variables are higher than we might like (between 0.5 and 0.54), but do not indicate fatal collinearity problems. With the patent sampling base increased from 5 to 25%, model fit statistics are substantially higher than for the full panel, with the R^2 rising to about 0.8 with controls.

effect, significant at 1% – even taking into account potential collinearity issues, this is a strong result. In the pooled sample ethnic/cultural proximity also becomes marginally significant.

In the (time) split samples (columns 3 and 4), organisational proximity becomes weaker over time, and social proximity becomes stronger; institutional proximity is 5% significant in the 1990s, but insignificant in the 2000s. Unlike the full sample of all inventors, ethnic/cultural proximity effects increase in time for multiple inventors. The only puzzling result is for cognitive proximity, which is insignificant or slightly negative depending on the model fitted. Given the wider literature, this may reflect the way we have constructed the variable.

The analysis of interaction effects highlights the distinct effects of proximities on multiple inventors. The major difference to the pooled inventor sample is that while social, cultural/ethnic and organisational proximities have strong positive effects on collaboration, joint effects generally are not significant and often take a negative sign (columns 5-9). The full specification (column 10) shows that cultural/ethnic proximity is positive significant at 5%, while its joint effect with geography is insignificant. Column 10 also highlights negative significant joint effects for institutional and cognitive proximities with local geographic closeness. For multiple inventors, these results suggest that proximities are fundamentally substitutes, not complements.

Multiple inventor pair activity

Table 4 covers counts of co-inventive activity. As before, fitting the base specification to the subset of multiple inventors (Table 4, column 1) indicates that a different configuration of proximities are in play when looking at the frequency of collaborations by multiple inventors. Local geographic proximity is insignificant, and in this case is dominated by organisational and institutional proximity, the latter now significant at 5%.

[Insert Table 4 here]

When the full set of proximities is fitted (column 2), organisational proximity effects weaken substantially, becoming marginally significant; geographic proximity becomes negative and marginally significant. Counts of co-invention activity for multiple inventors are largely driven by social, institutional and cultural/ethnic proximity. Time splits (column 3 and 4) indicate that institutional and organisation effects are most prevalent in the 1990s. By contrast, social and cultural/ethnic proximities take on an increasingly important role.

Interaction effects again suggest that for multiple inventors, proximities are fundamentally substitutes, not complements. In this case, the joint effects of geography with social, organisational and cognitive proximity are all negative significant, while the individual coefficients are almost all robust positive effects.

5.3 Robustness checks

Section 3 highlights a number of identification challenges. To deal with these, we subject our model to a series of robustness checks.

Omitted / mis-specified variables

The way we measure proximities affects the results raising, perhaps, mis-measurement problems. In order to address this issue, Table C-1 in Appendix C refits the collaboration dummy and count models with alternative measures for geographic and ethnic proximity, using a linear distance threshold (columns 1 and 5), a TTWA dummy for inventors in a pair (columns 2 and 6), and alternative cultural/ethnic dummies (columns 3-4 and 7-8). Although point estimates change slightly, the overall pattern of the results does not change.

We might also worry that the *type* of collaboration chosen affects the number of collaborations. Column 9 therefore refits the co-invention counts model with a dummy that takes the value of 1 if the inventor pair is part of team larger than two individuals. The results are robust to changes in the measurement of distances, with the team dummy wholly dominating the results. The correlations matrix in Table B-1 (Appendix B) confirms simultaneity, with pairwise correlations of 0.76 between the team dummy and co-invention counts (and 0.99 with the collaboration dummy).

Table C-2 in Appendix C repeats this exercise for multiple inventors, focusing on the collaboration dummy model (results for collaboration counts). Alongside the existing alternative proximity measures already introduced, we also fit alternative specifications of social proximity (columns 1, 2 and 4), cognitive proximity (column 3), and organisational proximity (column 5). Once again, our main results remain robust to these changes.

The results also suggest that the six-digit technology field control is a powerful conditioning influence on proximity effects. In Table C-3 in Appendix C, we fit an alternative specification of the collaboration dummy model to see if results survive. Specifically, in column 2 we introduce a 'technology regime' control, consisting of 1581 six-digit technology field group * year fixed effects, designed to capture the underlying conditions shaping knowledge creation. This specification shifts up point estimates for geographic proximity and organisation proximity, but does not change their relative magnitudes or the overall pattern of results.

More seriously, we might worry that we have mis-specified individual-level controls. Table C-4 in Appendix C includes two alternative individual fixed effects specifications, a single inventor-pair level fixed effect (column 2) and two compound fixed effects, based on each vector of individual, institutional, area and historical specificities (column 3). These specifications both marginally shift coefficients for institutional and cultural/ethnic proximity, but do not change their statistical significance. Other variables of interest are unaffected.

Estimation

A second set of concerns centres on estimation issues. Our main regressions are fitted with robust standard errors; Table C-5 in Appendix C strengthens the specification, with HAC standard errors clustered on inventor pairs. Main results survive essentially unchanged; model fit as measured by the F-statistic decreases marginally.

We fit our main models with a linear estimator, arguing that once converted to marginal effects non-linear models offer little extra precision. The very high number of zeroes in the collaboration dummy makes a logit model hard to converge; for the co-invention count variable, diagnostics indicate excess zeroes and over-dispersion. We therefore test a reduced form version of the co-invention count model in OLS and as a negative binomial, running the latter with and without marginal effects. Table C-6 in Appendix C gives results. Once converted to marginal effects, the non-linear model has some noticeable differences in point estimates. However, as expected relative magnitudes are about the same.

Sample construction

Finally, we might worry that our results are simply the product of a particular sub-sample of patents and inventors. Since sampling patents and inventors is the building block of our identification strategy, it is important to test this. To do so, we rebuild the main panel five times using different samples of patents. We then re-run the collaboration dummy model on each panel in sequence. Results are given in Table C-7 (Appendix C), and show very little variation. We conclude that our identification strategy is robust to sample choice.

6. Conclusions

Innovation has become an increasingly collaborative activity in recent years, with a range of studies suggesting that geographic and other proximities have an important influence on ideas generation and diffusion.

This paper explores the role of a range of proximities on knowledge creation. Using a rich microdata set and a novel identification strategy, we have been able to focus on individual inventors across the full range of technology fields and in different time periods, examine geographic, organisational, social, institutional, cognitive and ethnic/cultural factors singly and in combination, and identify causal effects by controlling for a range of other individual, institutional and macro influences. In doing so, we address empirically a number of issues which have tended to be considered from a more theoretical perspective in the literature.

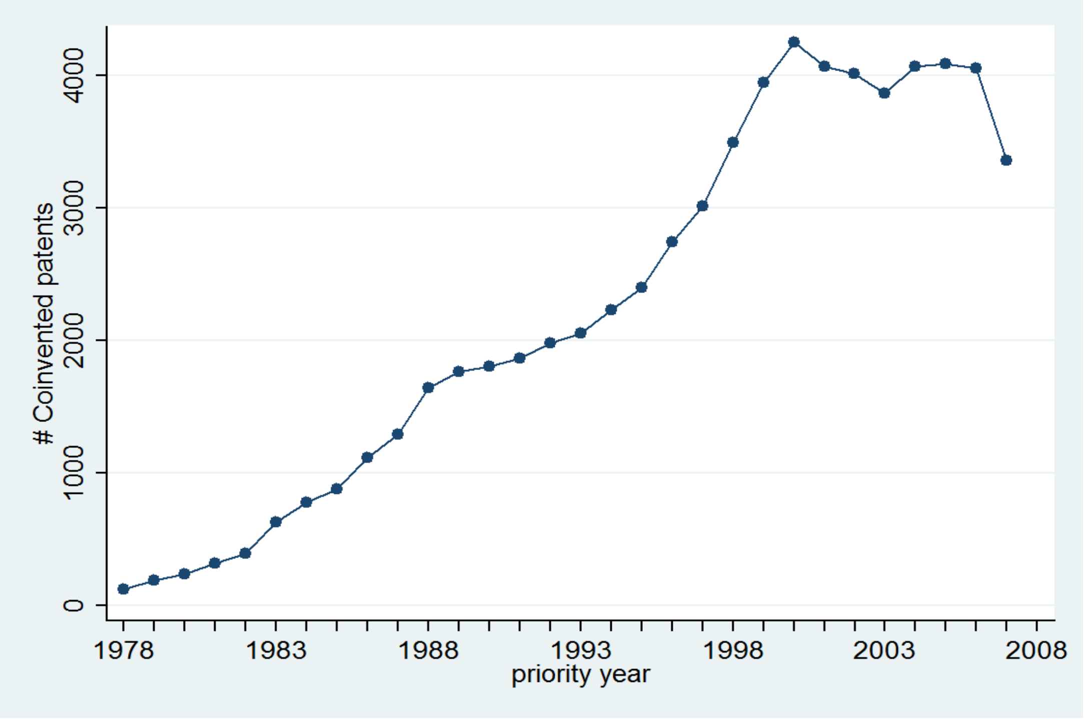
Our results are robust to multiple cross-checks, and contain a number of important findings. For inventors as a whole, we find that local geographic proximity is an important supporting influence on collaboration – but as other studies have found, it is mediated by organisational proximity, and in some cases, by cultural/ethnic closeness. Physical proximity has become more important over time, with organisational and institutional proximity declining in salience – shifts which match with the stylised facts about the globalisation of innovative activity. For this group, proximities are fundamentally complementary, with joint effects showing up as robust and positive. For multiple inventors, we find that geographic proximity is much less important than organisation, social and cultural/ethnic factors. Our results confirm that multiple inventors behave in a clearly distinct way to more occasional inventors. The analysis also confirms the critical role of social proximity and social networks in mediating collaborative activity. For multiple inventors, proximities also appear to be substitutes, not complements.

Overall, the results highlight important differences between the proximities that help inventors collaborate for the first time, the factors shaping repeat interactions, and the behaviour of serial inventors. Physical proximity is critical to break the ice. Once a relationship has been established, however, other forms of proximity become more important. And for multiple inventors, geography disappears almost completely as an influence.

Our analysis contributes to open up a number of avenues for future research. Our results for cognitive proximity are unusual, and point to a need for follow-up work testing out alternative specifications of technological closeness. We have also chosen deliberately simple social proximity measures. Further work could use more complex social proximity metrics, or focus on hub inventors' ego-networks rather than inventor pairs. Finally, given the scope of the paper, we have been unable to delve into other interesting aspects, such as age or gender differences, which intuition suggests may be important influences on what inventors do. Future analysis could explore these issues in detail.

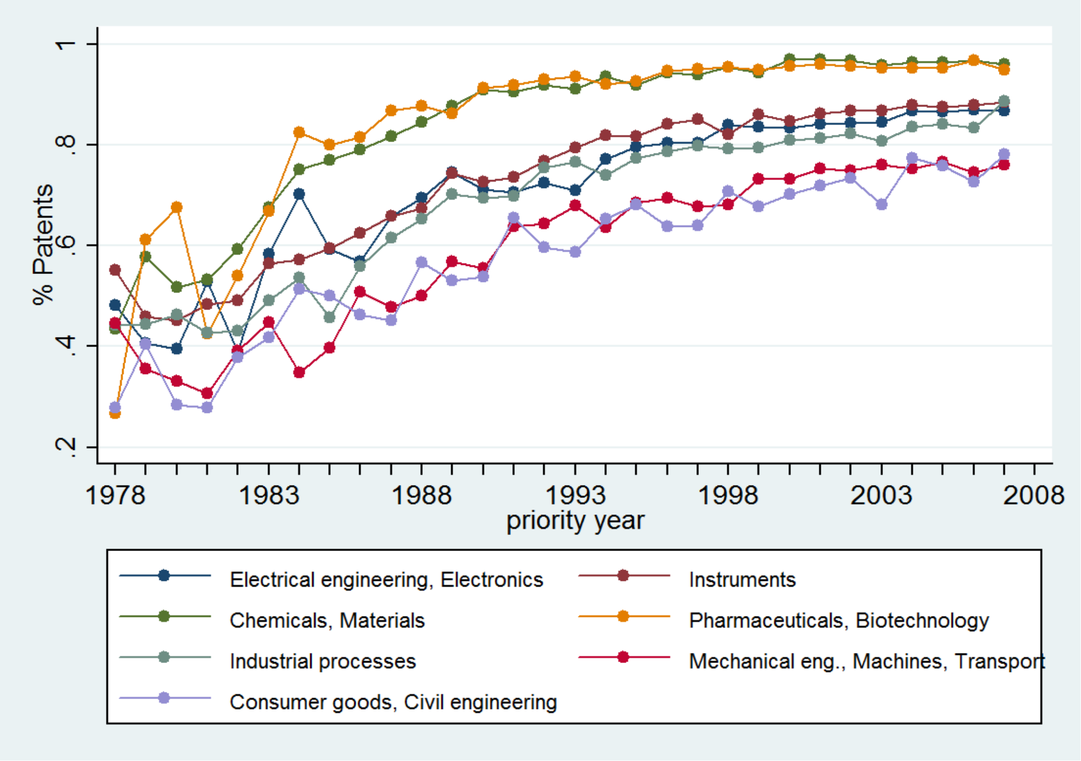
List of figures

Figure 1. Co-invented patents, 1978-2007.



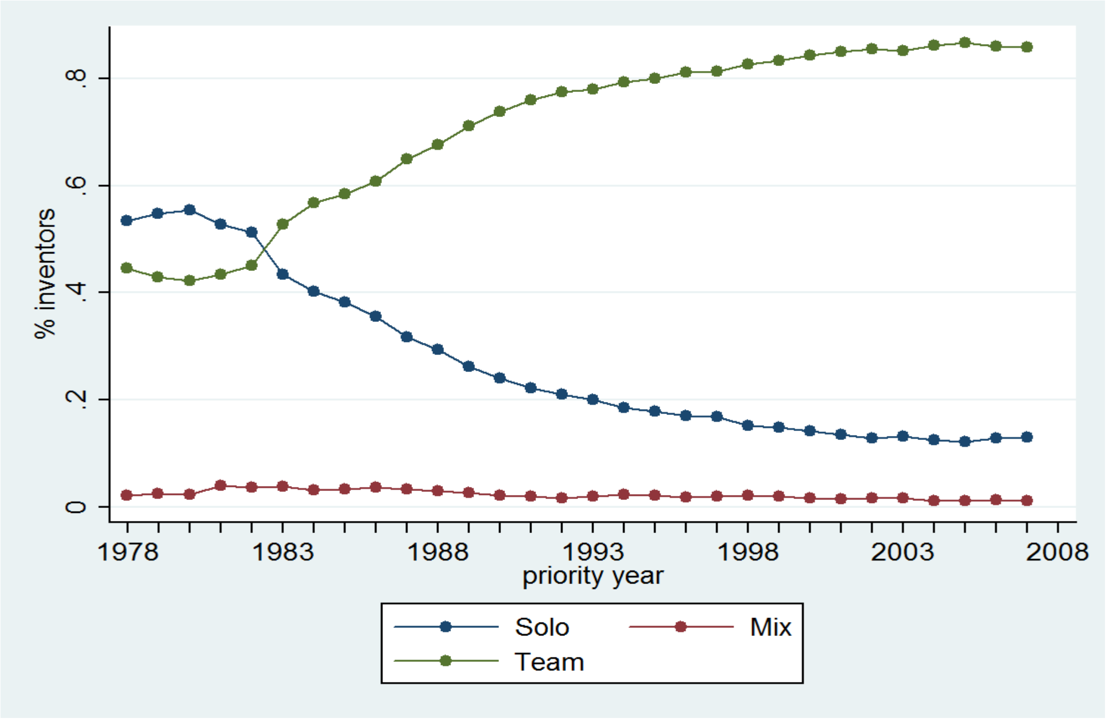
Source: KITES-PATSTAT.

Figure 2. Co-invented patents by technology field, 1978-2007.



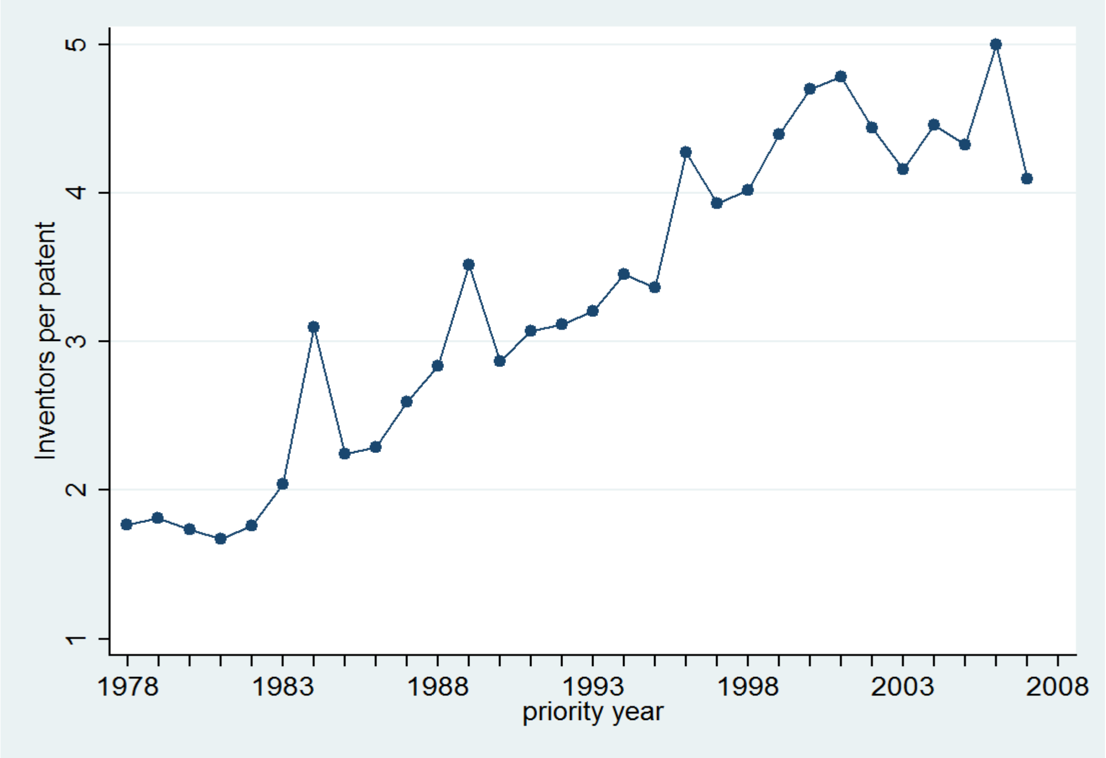
Source: KITES-PATSTAT.

Figure 3. Inventor behaviour, 1978-2007.



Source: KITES-PATSTAT.

Figure 4. Inventor team size, 1978-2007.



Source: KITES-PATSTAT.

List of tables

Table 1 - All inventors,Co-invention dummy, 1978-2007.

Depvar = co-invention dummy	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**
	all	**90s**	**00s**	**all**	**all**	**all**	**all**	**90s**	**00s**
inverse linear distance between i and j	0.00159***	0.00122**	0.00187***	0.000300***	0.000508***	0.00118***	0.000114	0.000766	-0.000207
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)
inventor pair same applicant	0.0276***	0.0375***	0.0229***	0.0181***	0.0271***	0.0276***	0.0180***	0.0309***	0.0109***
	(0.002)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.005)	(0.003)
inventor pair same type of applicant	-0.0000610***	-0.0000903*	-0.0000552**	-0.0000610***	-0.000145***	-0.0000603***	-0.0000511**	-0.0000618	-0.0000558***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
inventor pair same CEL subgroup	0.0000186	-0.0000323	0.0000301	0.0000229	0.0000199	-0.0000218	-0.00000680	-0.00000949	-0.0000225
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
distance * same applicant				0.0208***			0.0209***	0.0161**	0.0250***
				(0.004)			(0.004)	(0.008)	(0.005)
distance * same institution type					0.00213***		-0.000239	-0.000607**	-0.0000510
					(0.000)		(0.000)	(0.000)	(0.000)
distance * same CEL subgroup						0.000948	0.000692	-0.000291	0.00121
						(0.001)	(0.001)	(0.001)	(0.001)
Human capital effects	Y	Y	Y	Y	Y	Y	Y	Y	Y
Preferences effects	Y	Y	Y	Y	Y	Y	Y	Y	Y
Applicant type dummies for inventors *i*, *j*	Y	Y	Y	Y	Y	Y	Y	Y	Y
Historic area patent stocks	Y	Y	Y	Y	Y	Y	Y	Y	Y
TTWA dummies for inventors *i*, *j*	Y	Y	Y	Y	Y	Y	Y	Y	Y
IPC six-digit main group dummies	Y	Y	Y	Y	Y	Y	Y	Y	Y
Observations	843894	307745	536149	843894	843894	843894	843894	307745	536149
F	28.537	18.459	17.992	28.778	28.498	28.427	28.615	18.227	18.438
r2	0.767	0.772	0.764	0.767	0.767	0.767	0.767	0.772	0.765

Source: KITES-PATSTAT. - Notes: All models use time dummies. Robust standard errors in parentheses. Constant not shown - * p<0.1, ** p<0.05, *** p<0.01

Table 2 - All inventors, Co-invention counts, 1978-2007.

depvar = #co-invented patents	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**
	all	**90s**	**00s**	**all**	**all**	**all**	**all**	**90s**	**00s**
inverse linear distance between i and j	0.0131***	0.0266***	0.00804**	0.00232*	0.00251	0.00709	-0.00326	- 0.000775	-0.00415
	(0.004)	(0.008)	(0.004)	(0.001)	(0.002)	(0.005)	(0.005)	(0.012)	(0.004)
inventor pair same applicant	0.203***	0.323***	0.146***	0.123***	0.199***	0.203***	0.123***	0.163***	0.0922***
	(0.017)	(0.042)	(0.016)	(0.027)	(0.017)	(0.017)	(0.027)	(0.056)	(0.029)
inventor pair same type of applicant	0.0000976	0.000389	-0.000126	0.0000979	-0.000721*	0.000109	0.0000622	0.000667	-0.000305
	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)
inventor pair same CEL subgroup	0.000776*	0.00118	0.000407	0.000812*	0.000789*	0.000188	0.000317	0.000773	0.0000943
	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)
distance * same applicant				0.174***			0.173***	0.386***	0.110**
				(0.054)			(0.053)	(0.131)	(0.051)
distance * same institution type					0.0209***		0.00115	-0.00383	0.00425
					(0.006)		(0.002)	(0.003)	(0.003)
distance * same CEL subgroup						0.0138	0.0116	0.0152	0.00727
						(0.010)	(0.010)	(0.022)	(0.008)
Human capital effects	Y	Y	Y	Y	Y	Y	Y	Y	Y
Preferences effects	Y	Y	Y	Y	Y	Y	Y	Y	Y
Applicant type dummies for inventors *i*, *j*	Y	Y	Y	Y	Y	Y	Y	Y	Y
Historic area patent stocks	Y	Y	Y	Y	Y	Y	Y	Y	Y
TTWA dummies for inventors *i*, *j*	Y	Y	Y	Y	Y	Y	Y	Y	Y
IPC six-digit main group dummies	Y	Y	Y	Y	Y	Y	Y	Y	Y
Observations	843894	307745	536149	843894	843894	843894	843894	307745	536149
F	8.539	4.843	5.927	9.149	8.582	8.513	9.123	4.989	6.134
r2	0.376	0.354	0.424	0.377	0.376	0.376	0.377	0.356	0.424

Source: KITES-PATSTAT.

Notes: All models use time dummies. Robust standard errors in parentheses. Constant not shown

* p<0.1, ** p<0.05, *** p<0.01

Table 3 - Multiple inventors - Co-invention dummy, interactions and time splits.

Depvar = co-invention dummy	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**	**(10)**
	base	**all**	**90s**	**00s**	**all**	**all**	**all**	**all**	**all**	**all**
Inv. linear distance between i and j	-0.000103	-0.00161	0.00139	-0.00300	-0.000972**	-0.000170	-0.00135	-0.000530	0.000147	0.00268
	(0.002)	(0.002)	(0.002)	(0.004)	(0.000)	(0.000)	(0.004)	(0.001)	(0.002)	(0.003)
inventor pair same applicant	0.0555***	0.0194***	0.0166**	0.0220*	0.0237	0.0201***	0.0195***	0.0197***	0.0209***	0.0154
	(0.008)	(0.007)	(0.007)	(0.013)	(0.016)	(0.007)	(0.007)	(0.007)	(0.007)	(0.011)
inventor pair same type of applicant	0.000351	0.000436	0.000890**	-0.000136	0.000432	0.000530	0.000434	0.000427	0.000451	0.000526*
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
inventor pair same CEL subgroup	0.000372	0.000500*	0.0000866	0.000597	0.000494**	0.000493*	0.000521**	0.000476*	0.000486*	0.000522**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
continuous social proximity		0.0797***	0.0342**	0.132***	0.0803***	0.0799***	0.0797***	0.0951***	0.0804***	0.0993***
		(0.017)	(0.016)	(0.032)	(0.018)	(0.017)	(0.017)	(0.036)	(0.017)	(0.035)
inventor pair share past IPC6 field		-0.00102	-0.000747*	-0.00186	-0.00101	-0.00103	-0.00102	-0.00105	0.00104	0.000968
		(0.001)	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
distance * same applicant					-0.00671					0.0103
					(0.022)					(0.016)
distance * same applicant type						-0.00256				-0.00236**
						(0.004)				(0.001)
distance * same CEL subgroup							-0.000461			-0.00139
							(0.005)			(0.005)
distance * social proximity								-0.0220		-0.0284
								(0.039)		(0.040)
distance * same past IPC6 field									-0.0361***	-0.0360***
									(0.009)	(0.009)
Observations	43973	43973	24892	19081	43973	43973	43973	43973	43973	43973
F	17.781	17.223	18.563	12.137	16.980	17.076	17.102	16.660	17.070	16.172
r2	0.801	0.807	0.873	0.747	0.807	0.807	0.807	0.807	0.808	0.808

Source: KITES-PATSTAT. - Notes: All models use time dummies + individual + inst + area + hist + techfield fixed effects. Robust standard errors in parentheses. Constant not shown - * p<0.1, ** p<0.05, *** p<0.01

Table 4 - Multiple inventors - Co-invention counts, interactions and time splits, 1978-2007.

depvar = # co-invented patents	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**	**(10)**
	base	**all**	**90s**	**00s**	**all**	**all**	**all**	**all**	**all**	**all**
inverse linear distance between i and j	-0.00986	-0.0142*	-0.00225	-0.0184	-0.00206	-0.0000820	-0.0195	-0.0000219	-0.00921	0.00318
	(0.008)	(0.008)	(0.008)	(0.013)	(0.001)	(0.001)	(0.012)	(0.003)	(0.008)	(0.010)
inventor pair same applicant	0.155***	0.0514*	0.0423**	0.0537	0.131**	0.0582**	0.0509*	0.0549**	0.0555**	0.0410
	(0.027)	(0.027)	(0.019)	(0.052)	(0.059)	(0.027)	(0.027)	(0.027)	(0.027)	(0.040)
inventor pair same type of applicant	0.00216**	0.00241**	0.00444***	0.000150	0.00235**	0.00333***	0.00245**	0.00230**	0.00245**	0.00252**
	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
inventor pair same CEL subgroup	0.00169*	0.00205**	0.000409	0.00313*	0.00195**	0.00198**	0.00163**	0.00174**	0.00202**	0.00155**
	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
continuous social proximity		0.228***	0.0900**	0.406***	0.238***	0.230***	0.229***	0.428***	0.230***	0.442***
		(0.060)	(0.040)	(0.121)	(0.063)	(0.061)	(0.061)	(0.132)	(0.060)	(0.129)
inventor pair share past IPC6 field		-0.00158	-0.00297***	-0.00335	-0.00147	-0.00167	-0.00159	-0.00202	0.00421	0.00317
		(0.003)	(0.001)	(0.006)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
distance * same applicant					-0.126					0.0298
					(0.081)					(0.057)
distance * same applicant type						-0.0249*				-0.00430
						(0.014)				(0.003)
distance * same CEL subgroup							0.00923			0.00320
							(0.016)			(0.016)
distance * social proximity								-0.288**		-0.307**
								(0.143)		(0.146)
distance * same past IPC6 field									-0.102***	-0.0926***
									(0.029)	(0.027)
Observations	43973	43973	24892	19081	43973	43973	43973	43973	43973	43973
F	15.378	14.066	28.637	6.754	12.655	13.792	13.983	13.414	13.992	13.317
r2	0.685	0.691	0.798	0.618	0.693	0.692	0.691	0.696	0.692	0.697

Source: KITES-PATSTAT.

Notes: All models use time dummies + individual + inst + area + hist + techfield fixed effects. Robust standard errors in parentheses. Constant not shown - * p<0.1, ** p<0.05, *** p<0.01

References

ACKERBERG, D. & BOTTICINI, M. 2002. Endogenous Matching and the Empirical Determinants of Contract Form. *Journal of Political Economy,* 110, 564-591.

ADAMS, J. D., BLACK, G. C., CLEMMONS, J. R. & STEPHAN, P. E. 2005. Scientific teams and institutional collaborations: Evidence from U.S. universities, 1981–1999. *Research Policy,* 34, 259-285.

AGRAWAL, A., COCKBURN, I. & MCHALE, J. 2006. Gone but not forgotten: knowledge flows, labor mobility, and enduring social relationships. *Journal of Economic Geography,* 6, 571-591.

AGRAWAL, A., KAPUR, D. & MCHALE, J. 2008. How do spatial and social proximity influence knowledge flows? Evidence from patent data. *Journal of Urban Economics,* 64, 258-269.

AMISSE, S., HUSSLER, C., MULLER, P. & RONDÉ, P. 2011. Do birds of a feather flock together? Proximities and interclusters network.

ARCHIBUGI, D. & IAMMARINO, S. 2002. The globalization of technological innovation: definition and evidence. *Review of International Political Economy,* 9, 98–122.

ARCHIBUGI, D. & PIETROBELLI, C. 2003. The globalisation of technology and its implications for developing countries: windows of opportunity or further burden? . *Technological Forecasting and Social Change,* 70, 861-883.

AZOULAY, P., DING, W. & STUART, T. 2006. The Determinants of Faculty Patenting Behaviour: Demographics or Opportunities? *Academic Science and Entrepreneurship: Dual Engines of Growth?* Santa Fe, NM: NBER.

BLUNDELL, R., GRIFFITH, R. & VAN REENEN, J. 1995. Dynamic Count Data Models of Technological Innovation. *The Economic Journal,* 105, 333-344.

BOSCHMA, R. 2005. Proximity and Innovation: A Critical Assessment. *Regional Studies,* 39, 61 - 74.

BOSCHMA, R. & FRENKEN, K. 2009. The Spatial Evolution of Innovation Networks: A proximity perspective. *In:* BOSCHMA, R. & MARTIN, R. (eds.) *Handbook of Evolutionary Economic Geography.* Cheltenham: Edward Elgar.

BOZEMAN, B. & GAUGHAN, M. 2011. How do men and women differ in research collaborations? An analysis of the collaborative motives and strategies of academic researchers. *Research Policy,* 40, 1393-1402.

BRESCHI, S. & LENZI, C. 2011. Net City: How Co-Invention Networks Shape Inventive Productivity in US Cities Milan: Universita Bocconi.

BRESCHI, S. & LISSONI, F. 2006. Mobility of inventors and the geography of knowledge spillovers. New evidence on US data. *KITeS Working Papers 184.* Milan: Universita' Bocconi,.

BRESCHI, S. & LISSONI, F. 2009. Mobility of skilled workers and co-invention networks: an anatomy of localized knowledge flows. *Journal of Economic Geography*, 9, 439-468.

BURT, R. 1992. *Structural Holes: The Social Structure of Competition,* Cambridge, MA, Harvard University Press.

CANTWELL, J. 2005. MNCs, local clustering and science-technology relationships. *In:* SANTANGELO, G. (ed.) *Technological Change and Economic Catch-Up: The Role of Science and Multinationals.* Cheltenham: Edward Elgar.

CASSI, L. & PLUNKET, A. 2010. The determinants of co-inventor tie formation: proximity and network dynamics. *Papers in Evolutionary Economic Geography # 10.15.* Utrecht: Utrecht University.

CHRISTAKIS, N., FOWLER, J., IMBENS, G. & KALYANARAMAN, K. 2010. An Empirical Model for Strategic Network Formation. *NBER Working Paper No. 16039.* Cambridge, MA: NBER.

COASE, R. 1937. The Nature of the Firm. *Economica*, 4, 386-405.

CRESCENZI R., GAGLIARDI L. & PERCOCO M. "Social capital and the innovative performance of Italian provinces", *Environment and Planning A*, 45(4), 908-929, 2013

CRESCENZI R., RODRÍGUEZ-POSE A. & STORPER M. 2007. The territorial dynamics of innovation: a Europe-United States Comparative Analysis, *Journal of Economic Geography*, 7(6), 673-709

D'ESTE, P. & IAMMARINO, S. 2010. The spatial profile of university-business research partnerships. *Papers in Regional Science,* 89, 335-350.

D’ESTE P., GUY, F. and IAMMARINO, S. (2013). Shaping the formation of university-industry research collaborations: what type of proximity does really matter? *Journal of Economic Geography, 13 (4): 537-558.*

DOCQUIER, F. & RAPOPORT, H. 2011. Globalisation, Brain Drain and Development. *IZA DP 5590.* Bonn: IZA.

EVANS, T. S., LAMBIOTTE, R. & PANZARASA, P. 2011. Community Structure and Patterns of Scientific Collaboration in Business and Management. *Scientometrics,* forthcoming.

FLEMING, L., KING, C. & JUDA, A. I. 2007. Small Worlds and Regional Innovation. *Organization Science*, 18, 938-954.

FU, X. & SOETE, L. (eds.) 2010. *The Rise of Technological Power in the South,* Basingstoke: Palgrave MacMillan.

GLÄNZEL, W. 2001. National characteristics in international scientific co-authorship relations. *Scientometrics,* 51, 69-115.

GLÄNZEL, W. & SCHUBERT, A. 2005. Analysing Scientific Networks Through Co-Authorship. *In:* MOED, H., GLÄNZEL, W. & SCHMOCH, U. (eds.) *Handbook of Quantitative Science and Technology Research.* Springer Netherlands.

GRANOVETTER, M. S. 1973. The Strength of Weak Ties *The American Journal of Sociology,* 78, 1360-1380.

GRIFFITH, R., LEE, S. & VAN REENEN, J. 2011. Is distance dying at last? Falling home bias in fixed effects models of patent citations. *cemmap working paper CWP18/11.* London The Institute for Fiscal Studies / Department of Economics, UCL.

HALL, B., JAFFE, A. & TRAJTENBERG, M. 2001. The NBER Patent Citations Data File: Lessons, Insights and Methodological Tools. Cambridge, Mass.: NBER.

JACKSON, M. 2006. The Economics of Social Networks. *In:* BLUNDELL, R., NEWEY, W. & PERSSON, T. (eds.) *Advances in Economics and Econometrics: Ninth World Congress of the Econometric Society.* Cambridge: Cambridge University Press.

JAFFE, A. B., TRAJTENBERG, M. & HENDERSON, R. 1993. Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations. *The Quarterly Journal of Economics,* 108, 577-598.

KAISER, U., KONGSTED, H. C. & RØNDE, T. 2011. Labor Mobility, Social Network Effects, and Innovative Activity. *In:* 5654, I. D. N. (ed.). Bonn: IZA.

KAPUR, D. & MCHALE, J. 2005. Sojourns and Software: Internationally mobile human capital and high tech industry development in India, Ireland and Israel. *In:* ARORA, A. & GAMBARDELLA, A. (eds.) *From Underdogs to Tigers: The Rise and Growth of the Software Industry in Brazil, China, India, Ireland and Israel.* Oxford: OUP.

KERR, W. 2008a. The Agglomeration of US Ethnic Inventors. *HBS Working Paper 09-003.* Boston, MA: Harvard Business School.

KERR, W. 2008b. Ethnic Scientific Communities and International Technology Diffusion. *Review of Economics and Statistics,* 90, 518-537.

KERR, W. 2009. Breakthrough Innovations and Migrating Clusters of Innovation. *NBER Working Paper 15443.* Cambridge, MA: NBER.

KERR, W. & LINCOLN, W. 2010. The Supply Side of Innovation: H-1b Visa Reforms and US Ethnic Invention *NBER Working Paper 15768.* Cambridge, Mass.: NBER

LEE, S. & BOZEMAN, B. 2005. The Impact of Research Collaboration on Scientific Productivity. *Social Studies of Science,* 35, 673-702.

LEYDESDORFF, L. & ETZKOWITZ, H. 1998. The triple helix as a model for innovation studies. *Science and public policy,* 25, 195-203.

LISSONI, F., TARASCONI, G. & SANDITOV, B. 2006. The KEINS Database on Academic Inventors: Methodology and Contents. *CESPRI Working Paper 181.* Milan: Universita' Bocconi.

LOBO, J. & STRUMSKY, D. 2008. Metropolitan patenting, inventor agglomeration and social networks: A tale of two effects. *Journal of Urban Economics,* 63, 871-884.

LYCHAGIN, S., PINKSE, J., SLADE, M. E. & VAN REENEN, J. 2010. Spillovers in Space: Does Geography Matter? *NBER Working Paper 16188.* Cambridge, MA: NBER.

MARROCU, E., PACI, R. & USAI, S. 2011. Proximity, Networks and Knowledge Production in Europe. *CRENoS Working Paper 2011_09.* Cagliari: CRENoS.

MCCAFFREY, D., LOCKWOOD, J., MIHALY, K. & SASS, T. 2010. A Review of Stata Routines for Fixed Effects Estimations in Normal Linear Models. *The Stata Journal,* vv, 1-22.

MOWERY, D. C. 2001. Technological Innovation in a Multipolar System: Analysis and Implications for U.S. Policy. *Technological Forecasting and Social Change,* 67, 143-157.

MUDAMBI, R. 2008. Location, control and innovation in knowledge-intensive industries. *Journal of Economic Geography,* 8, 699-725.

NATHAN, M. 2011a. The Economics of Super-Diversity: Findings from British Cities, 2001-2006. *SERC Discussion Papers SERCDP0068.* London: SERC.

NATHAN, M. 2011b. Ethnic Inventors, Diversity and Innovation in the UK: Evidence from Patents Microdata. *Spatial Economics Research Centre SERCDP0092.* London: SERC.

OECD 2009. OECD Patent Statistics Manual. Paris: OECD.

OTTAVIANO, G., BELLINI, E. & MAGLIETTA, A. 2007. Diversity and the Creative Capacity of Cities and Regions. *SUSDIV Paper 2.2007.* Bologna: FEEM.

PAIER, M. & SCHERNGELL, T. 2008. Determinants of Collaboration in European R&D Networks: Empirical evidence from a binary choice model perspective. *NEMO Working Ppaer #10.* Vienna: ARC.

PONDS, R., VAN OORT, F. & FRENKEN, K. 2007. The geographical and institutional proximity of research collaboration*. *Papers in Regional Science,* 86, 423-443.

PONDS, R., VAN OORT, F. & FRENKEN, K. 2010. Innovation, Spillovers and university-industry collaboration: an extended knowledge production function approach. *Journal of Economic Geography,* 10, 231-255.

RODRÍGUEZ-POSE A. & CRESCENZI R. 2008. R&D, spillovers, innovation systems and the genesis of regional growth in Europe *Regional Studies*, 42(1), 51-67

SAXENIAN, A.-L. 2006. *The New Argonauts: Regional Advantage in a Global Economy,* Cambridge, MA, Harvard University Press.

SAXENIAN, A.-L. & SABEL, C. 2008. Venture Capital in the 'Periphery': The New Argonauts, Global Search and Local Institution-Building. *Economic Geography*, 84, 379-394.

SCOTT, A. & GAROFOLI, G. (eds.) 2007. *Development on the Ground: Clusters, Networks and Regions in Emerging Economies,* Oxford: Routledge.

SEDIKES, C., ARIELY, D. & OLSEN, N. 1999. Contextual and Procedural Determinants of Partner Selection: Of Asymmetric Dominance and Prominence. *Social Cognition,* 17, 118-139

SEELY BROWN, J. & DUGUID, P. 2002. *The Social Life of Information,* Cambridge, MA, Harvard Business School Press.

SINGH, J. 2005. Collaborative Networks as Determinants of Knowledge Diffusion Patterns. *Management Science*, 51, 756-770.

THOMPSON, P. & FOX-KEAN, M. 2005. Patent Citations and the Geography of Knowledge Spillovers: A Reassessment. *American Economic Review,* 95, 450-460.

TORRE, A. & RALLET, A. 2005. Proximity and Localization. *Regional Studies,* 39, 47 - 59.

VON HIPPEL, E. 2005. *Democratising Innovation,* Cambridge, MA, MIT Press.

YEUNG, H. 2009. Regional Development and the Competitive Dynamics of Global Production Networks: An East Asian Perspective. *Regional Studies,* 43, 325-351.

APPENDIX A.1 – Description of the variables

Variable name	**Definition**	**Source**
Dependent variables		
DCOINVENT	Dummy variable for inventor pairs, coded as 1 if pair patent together in a given year, 0 if not	KITES-PATSTAT
#COINVENT	Continuous variable for inventor pairs, recording the count of collaborations in a given year	KITES-PATSTAT
Independent variables		
PROXG_LD	Inverse linear distance in km between Travel to Work Area (TTWA) centroids occupied by each inventor in a pair, based on inventor address. Normalised to take maximum value 1	KITES-PATSTAT, UK Office of National Statistics
PROXG_LDT*	Inverse linear distance in km between TTWA centroids occupied by each inventor in a pair, with threshold set at 200km	KITES-PATSTAT, UK Office of National Statistics
PROXG*	Dummy for inventor pairs, coded as 1 if both are based in same TTWA, 0 if not, blank if unknown	KITES-PATSTAT
PROXO	Dummy for inventor pairs, coded as 1 if both are based in same applicant, 0 if not, blank if unknown	KITES-PATSTAT
PROXI	Dummy for inventor pairs, coded as 1 if both are based in same applicant type, 0 if not, blank if unknown. Types are coded as i) business / private research lab, ii) university / public research lab; iii) foundation / NGO / consortium; iv) individual.	KITES-PATSTAT

PROXE_CEL	Dummy for inventor pairs, set as 1 if both are in the same ONOMAP ‘cultural-ethnic-linguistic’ (CEL) subgroup, 0 if not, blank if unknown. ONOMAP coding is based on inventor name information (see Appendix A).	KITES-PATSTAT, ONOMAP
PROXE_ETH*	Dummy for inventor pairs, set as 1 if both are in the same ONOMAP ‘geographical origin’ subgroup, 0 if not, blank if unknown. ONOMAP coding is based on inventor name information (see Appendix A).	KITES-PATSTAT, ONOMAP
PROXE_GEO *	Dummy for inventor pairs, set as 1 if both are in the same ONS ethnic group, 0 if not, blank if unknown. Coding is via ONOMAP, based on inventor name information (see Appendix A).	KITES-PATSTAT, ONOMAP
PROXC_6	Dummy for multiple inventor pairs, set as 1 if both have previously patented in the same 6-digit IPC technology field, 0 if not.	KITES-PATSTAT
PROXC_3*	Dummy for multiple inventor pairs, set as 1 if both have previously patented in the same 3-digit IPC technology field, 0 if not.	KITES-PATSTAT
PROXS	Inverse social distance between inventors in a pair. For a given year, social distance is defined as the number of steps between pair members in the previous five years, from 0 (collaboration) to minus infinity (no connection).	KITES-PATSTAT, University of Greenwich
PROXS2*	Inverse social distance between pair members, defined as PROXS but rescaled into 3 categories (3 = direct link, 2 = indirect link, 1 = no link).	KITES-PATSTAT, University of Greenwich

PROXS2D*	Dummies for social distance between inventors in a pair, defined as PROXS2. PROXS2D1 (no link) is taken as the reference category.	KITES-PATSTAT, University of Greenwich
Control variables		
IND	Vector of individual characteristics controls for each inventor in a pair: 1/ Dummy taking the value 1 if inventor is active in the pre-sample period 1978-1991, 0 if not; 2/ Inventor's average patenting in the pre-sample period 1978-1991, zeroed if inventor is inactive pre-sample; 3/ Dummies for inventor's type of patenting activity in pre-sample period: i) always solo ii) always co-inventing iii) mix solo and co-inventing iv) inactive. Inactive is set as the reference category.	KITES-PATSTAT
INST	Vector of institutional characteristics controls for each inventor in a pair: 1/ Dummies for type of inventor's applicant type, coded as i) business / private research lab, ii) university / public research lab; iii) foundation / NGO / consortium; iv) individual. Individual is the reference category.	KITES-PATSTAT
ENV	Vector of macro / area characteristics controls for each inventor in a pair: 1/ Year dummies; 2/ Grouping variable for 6-digit IPC technology fields, zeroed for potential pairs; 3/ TTWA dummies; 4/ Pre-sample (1978-1991) weighted patent stocks for each TTWA.	KITES-PATSTAT

*Variable used only in robustness checks

APPENDIX A.2 - Identifying cultural-ethnic categories with ONOMAP

Following Nathan (2011b), Agrawal et al (2008) and Kerr (2008a), we use a name classification system to generate culture-ethnicity-linguistic information for individual inventors. ONOMAP was originally designed for mining patient data for the UK National Health Service, and classifies individuals according to most likely cultural, ethnic and linguistic characteristics identified from forenames, surnames and forename-surname combinations. ONOMAP is built from a very large names database drawn from UK Electoral Registers plus a number of other contemporary and historical sources, covering 500,000 forenames and a million surnames across 28 countries. These are then algorithmically grouped together, combining information on geographical area, religion, language and language family. Separate classifications of surnames, forenames and surname-forename combinations are produced.

ONOMAP has the advantage of providing objective information at several levels of detail and across several dimensions of identity. It is also able to deal with Anglicisation of names, and names with multiple origins, giving it additional granularity and validity. Like Kerr's similar work on US patents data (Kerr, 2008a), ONOMAP is unable to observe immigrants, and only observing objective characteristics of identity – the most conservative interpretation is that it provides information on *most likely* cultural identity. However, unlike the MELISSA commercial database used by Kerr, which only identifies high-level ethnicities, ONOMAP allows us to examine inventor characteristics from several angles and at several levels of detail. ONOMAP also matches 99% of inventor names (compared with Kerr's 92-98% success rates).[15]

We use three 'identity bases' from ONOMAP: nine ONS ethnic groups, 13 'likely geographical origin' zones and ONOMAP's 67 bespoke 'cultural-ethnic-linguistic' (CEL) subgroups.[16] These groupings offer different levels of detail, and cover different salient aspects of cultural-ethnic identity (see Nathan (2011a) and Ottaviano et al (2007) for reviews of the debate). Descriptive statistics and correlation matrices for these variables are given in Tables 5 (full sample) and 6 (multiple inventors).

[15] We remove all conflict cases from the sample.

[16] The full set of ONS 1991 groups is White, Black Caribbean, Black African, Indian, Pakistani, Bangladeshi, Chinese and Other. The full set of twelve geographical origin zones is Africa, Americas, British Isles, Central Asia, Central Europe, East Asia, Eastern Europe, Middle East, Northern Europe, South Asia, Southern Europe and Rest of the World. See Nathan (2011) for the full classification of 67 CEL subgroups.

Appendix B - Summary statistics and correlation matrices

Table B.1 - Summary statistics and correlation matrices: full sample

variable	**N**	**mean**	**sd**	**min**	**max**
#co-invented patents per inventor pair *ij* (p_coinvents)	1508248	0.005	0.216	0	46
co-invented patent dummy (p_coinventsd)	1508248	0.001	0.029	0	1
inventors per patenting team (ipcount)	1508248	0.004	0.17	0	16
inventor pair is part of a team (pteamd)	1508248	0.001	0.028	0	1
inventor pair share same TTWA (proxg_l)	970555	0.034	0.181	0	1
inverse linear distance between inventors *i* and *j* (proxg_ld)	967969	0.042	0.18	0.001	1
inverse linear distance between *i* and *j* (200km distance threshold) (proxg_ldt)	970555	0.287	0.309	0	1
inventor pair share same applicant type (proxi)	1358652	0.449	0.497	0	1
inventor pair share same applicant (proxo)	1358652	0.007	0.086	0	1
inventor pair share same cultural-ethnic-linguistic (CEL) subgroup (proxe_cel)	1257109	0.401	0.49	0	1
inventor pair share same geographical origin (proxe_geo)	1257109	0.644	0.479	0	1
inventor pair share same ONS ethnic group (proxe_eth)	1257109	0.743	0.437	0	1

variable	proxg_l	proxg_ld	proxg_ldt	proxi	proxo	proxe_cel	proxe_geo	proxe_eth
Proxg_l	1							
Proxg_ld	0.9992	1						
Proxg_ldt	0.4295	0.4601	1					
proxi	0.0261	0.027	0.0292	1				
proxo	0.2188	0.2211	0.1347	0.0931	1			
Proxe_cel	-0.0167	-0.0162	0.0004	0.0397	0.0064	1		
Proxe_geo	-0.0285	-0.0293	-0.0479	0.031	0.0096	0.525	1	
Proxe_eth	-0.0276	-0.0283	-0.04	0.0162	0.0064	0.3473	0.6576	1

Source: KITES-PATSTAT.

Table B.2 Summary statistics and correlation matrices: multiple inventors

variable	**N**	**mean**	**sd**	**min**	**max**
#co-invented patents per inventor pair *ij* (p_coinvents)	52425	0.002	0.074	0	8
co-invented patent dummy (p_coinventsd)	52425	0.001	0.025	0	1
inventors per patenting team (ipcount)	52425	0.002	0.097	0	9
inventor pair is part of a team (pteamd)	52425	0.001	0.023	0	1
inventor pair share same TTWA (proxg_l)	52425	0.03	0.169	0	1
inverse linear distance between inventors *i* and *j* (proxg_ld)	52425	0.038	0.169	0.001	1
inverse linear distance between *i* and *j* (200km distance threshold) (proxg_ldt)	52425	0.279	0.307	0	1
inventor pair share same applicant type (proxi)	52425	0.555	0.497	0	1
inventor pair share same applicant (proxo)	52425	0.006	0.075	0	1
inventor pair share same cultural-ethnic-linguistic (CEL) subgroup (proxe_cel)	52425	0.54	0.498	0	1
inventor pair share same geographical origin (proxe_geo)	52425	0.817	0.386	0	1
inventor pair share same ONS ethnic group (proxe_eth)	52425	0.904	0.295	0	1
inverse social distance between inventors *i* and *j* (proxs)	52425	0.008	0.084	0	1
inverse social distance between inventors *i* and *j* (scaled 1-3) (proxs2)	52425	1.016	0.170	1	3
inventor pair has patented in same six-digit tech field in the past (proxc_6)	52425	0.012	0.110	0	1
inventor pair has patented in same three-digit tech field in the past (proxc_3)	52425	0.012	0.110	0	1

variable	proxg_l	proxg_ld	proxg_ldt	proxi	proxo	proxe_cel	proxc_6	proxc_3	proxs	proxs2
Proxg_l	1									
Proxg_ld	0.9993	1								
Proxg_ldt	0.4322	0.4627	1							
Proxi	0.0271	0.0291	0.0637	1						
Proxo	0.4329	0.4347	0.2225	0.1015	1					
Proxe_cel	-0.0042	-0.0053	-0.0336	0.018	0.0014	1				
Proxc_6	0.0147	0.0149	0.0098	-0.0012	0.031	-0.0011	1			
Proxc_3	0.0152	0.0154	0.01	-0.0011	0.0319	-0.001	0.9997	1		
Proxs	0.3487	0.3508	0.1868	0.0768	0.7351	-0.0114	0.0325	0.0338	1	
Proxs2	0.3639	0.3659	0.1927	0.079	0.7585	-0.0098	0.0336	0.0348	0.993	1

Source: KITES-PATSTAT.

Appendix C – Robustness Checks

Table C-1. Robustness checks: omitted variables, full sample.

	Co-invents dummy				# co-invented patents				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(7)
inverse linear distance			0.00159***	0.00159***			0.0131***	0.0131***	0.00186
between i and j			(0.000)	(0.000)			(0.004)	(0.004)	(0.003)
inverse linear distance	0.000258***				0.00153				
between i and j (200km)	(0.000)				(0.001)				
inventor pair share local area		0.00159***				0.0132***			
		(0.000)				(0.004)			
inventor pair same applicant	0.0282***	0.0276***	0.0276***	0.0276***	0.208***	0.203***	0.203***	0.203***	0.0142
	(0.002)	(0.002)	(0.002)	(0.002)	(0.018)	(0.017)	(0.017)	(0.017)	(0.011)
inventor pair same type of	-0.0000558***	-0.0000611***	-0.0000607***	-0.0000609***	0.000143	0.0000972	0.000105	0.000101	0.000434*
applicant	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
inventor pair same CEL	0.0000192	0.0000185			0.000777*	0.000772*			0.000648*
subgroup	(0.000)	(0.000)			(0.000)	(0.000)			(0.000)
inventor pair same ONS			0.0000359				0.000775		
ethnic group			(0.000)				(0.001)		
inventor pair same geog				0.0000249				0.000237	
origin zone				(0.000)				(0.001)	
							0.0131***	0.0131***	
							(0.004)	(0.004)	
ipair is part of bigger team									7.023***
									(0.501)
Observations	845948	845948	843894	843894	845948	845948	843894	843894	843894
F	28.099	28.396	28.501	28.503	8.415	8.500	8.540	8.545	25.967
r2	0.767	0.767	0.767	0.767	0.376	0.376	0.376	0.376	0.558

Source: KITES-PATSTAT. Notes: All models use time dummies + individual + inst + area + hist + techfield fixed effects. Robust standard errors in parentheses. Constant not shown. * p<0.1, ** p<0.05, *** p<0.01

Table C-2. Robustness checks: omitted variables, multiple inventors. [co-invents dummy only, co-invents count on request]

Co-invents dummy	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**
inverse linear distance between i and j	-0.00179 (0.002)	-0.00116 (0.002)	-0.00162 (0.002)	0.000189 (0.002)	-0.00143 (0.002)			-0.00162 (0.002)	-0.00162 (0.002)
inverse linear distance between i and j (200km)						-0.00115* (0.001)			
inventor pair share local area							-0.00146 (0.002)		
inventor pair same applicant	0.0177** (0.008)	0.0251*** (0.008)	0.0194*** (0.007)	0.0205*** (0.007)	0.0206*** (0.007)	0.0195*** (0.007)	0.0198*** (0.007)	0.0195*** (0.007)	0.0195*** (0.007)
inventor pair same type of applicant	0.000455 (0.000)	0.000426 (0.000)	0.000435 (0.000)	0.000469 (0.000)	0.000436 (0.000)	0.000441 (0.000)	0.000430 (0.000)	0.000432 (0.000)	0.000431 (0.000)
inventor pair same CEL subgroup	0.000483* (0.000)	0.000502* (0.000)	0.000501* (0.000)	0.000300 (0.000)	0.000477* (0.000)	0.000481* (0.000)	0.000480* (0.000)		
inventor pair same ONS ethnic group								-0.000320 (0.001)	
inventor pair same geog origin zone									-0.000234 (0.000)
scaled social proximity	0.0392*** (0.009)								
direct social link		0.0793*** (0.017)							
indirect social link		0.00740 (0.013)							
continuous social proximity			0.0798*** (0.017)	0.0606*** (0.015)	0.0752*** (0.017)	0.0781*** (0.017)	0.0782*** (0.017)	0.0797*** (0.017)	0.0797*** (0.017)
inventor pair patent in same IPC6 field in the past	-0.00102 (0.001)	-0.00122 (0.001)		-0.00160** (0.001)	-0.00124 (0.001)	-0.000983 (0.001)	-0.000978 (0.001)	-0.00101 (0.001)	-0.00102 (0.001)
inventor pair patent in same IPC3 field in the past			-0.00125 (0.001)						

rough: inventor pair has co-invented previously				0.349*** (0.073)					
rough: inventor pair same applicant previously					0.157* (0.091)				
Observations	43973	43973	43973	43973	43973	44829	44829	43973	43973
F	17.262	17.075	17.362	15.843	19.560	17.067	17.117	17.215	17.205
r2	0.807	0.808	0.807	0.829	0.810	0.807	0.807	0.807	0.807

Source: KITES-PATSTAT. Notes: All models use time dummies + individual + inst + area + hist + techfield fixed effects. Robust standard errors in parentheses. Constant not shown. * p<0.1, ** p<0.05, *** p<0.01

Table C-3. Robustness checks: technology field / regime controls.

Depvar = co-invents dummy	**(1)**	**(2)**
inverse linear distance between i and j	0.00159*** (0.000)	0.00648*** (0.001)
inventor pair same applicant	0.0276*** (0.002)	0.107*** (0.004)
inventor pair same type of applicant	-0.0000610*** (0.000)	-0.0000859** (0.000)
inventor pair same CEL subgroup	0.0000186 (0.000)	0.0000314 (0.000)
Six-digit IPC group controls	Y	N
Six-digit IPC group * year effects	N	Y
Observations	843894	843894
F	28.537	2.514
r2	0.767	0.107

Source: KITES-PATSTAT.
Notes: All models use individual + inst + area + hist + fixed effects. Robust standard errors in parentheses. Constant not shown. * p<0.1, ** p<0.05, *** p<0.01

Table C-4. Robustness checks: alternative fixed effects.

Depvar = co-invents dummy	**(1)**	**(2)**	**(3)**
inverse linear distance between i and j	0.00159*** (0.000)	0.00159*** (0.000)	0.00159*** (0.000)
inventor pair same applicant	0.0276*** (0.002)	0.0276*** (0.002)	0.0276*** (0.002)
inventor pair same type of applicant	-0.0000610*** (0.000)	-0.0000605*** (0.000)	-0.0000605*** (0.000)
inventor pair same CEL subgroup	0.0000186 (0.000)	0.0000158 (0.000)	0.0000136 (0.000)
Base specification	Y	N	N
Inventor pair fixed effect	N	Y	N
Compound fixed effects x2	N	N	Y
Observations	843894	843894	843894
F	28.537	.	.
r2	0.767	0.767	0.767

Source: KITES-PATSTAT.
Notes: All models use time dummies + individual + inst + area + hist + techfield fixed effects. Robust standard errors in parentheses. Constant not shown. * p<0.1, ** p<0.05, *** p<0.01

Table C-5. Robustness checks: clustering standard errors.

Depvar = co-invents dummy	(1)	(2)
inverse linear distance between i and j	0.00159*** (0.000)	0.00159*** (0.000)
inventor pair same applicant	0.0276*** (0.002)	0.0276*** (0.002)
inventor pair same type of applicant	-0.0000610*** (0.000)	-0.0000610*** (0.000)
inventor pair same CEL subgroup	0.0000186 (0.000)	0.0000186 (0.000)
Robust standard errors	Y	N
Standard errors clustered on inventor pair	N	Y
Observations	843894	843894
F	28.537	28.396
r2	0.767	0.767

Source: KITES-PATSTAT.
Notes: All models use time dummies + individual + inst + area + hist + techfield fixed effects.
Constant not shown. * p<0.1, ** p<0.05, *** p<0.01

Table C-6. Robustness checks: non-linear estimator, reduced form equation.

Depvar = co-invents dummy	OLS	NBREG	NBREG / MFX
inverse linear distance between i and j	0.0377*** (0.004)	2.933*** (0.363)	0.000340*** (0.000)
inventor pair same applicant (d)	0.651*** (0.032)	7.512*** (0.227)	0.200*** (0.047)
inventor pair same type of applicant (d)	-0.000252 (0.000)	0.317 (0.502)	0.0000373 (0.000)
inventor pair same CEL subgroup (d)	0.000920* (0.000)	0.469 (0.308)	0.0000554 (0.000)
Observations	856007	856007	856007
F-statistic	27.137		
R-squared	0.059		
Log-likelihood		-5404.545	-5404.545
Chi-squared		4596.533	4596.533

Source: KITES-PATSTAT.
Notes: All models use time dummies + individual + inst + area + hist + techfield fixed effects. Robust standard errors in parentheses. Constant not shown. * p<0.1, ** p<0.05, *** p<0.01

Table C-7. Robustness checks: sample construction.

Depvar = co-invents dummy	original	base_1	base_2	base_3	base_4	base_5
inverse linear distance between i and j	0.00159***	0.000891***	0.00158***	0.00113***	0.00162***	0.000801***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
inventor pair same applicant	0.0276***	0.0317***	0.0303***	0.0390***	0.0355***	0.0302***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
inventor pair same type of applicant	-0.0000610***	-0.0000448*	-0.0000651***	-0.00000392	-0.0000426*	-0.0000761***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
inventor pair same CEL subgroup	0.0000186	0.0000551*	0.0000277	0.0000545*	0.0000449	0.0000757***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Observations	843894	865177	879183	865617	849983	859085
F-statistic	28.537	29.462	31.939	25.144	27.434	24.911
R-squared	0.767	0.786	0.783	0.743	0.762	0.781

Source: KITES-PATSTAT.
Notes: All models use time dummies + individual + inst + area + hist + techfield fixed effects. Robust standard errors in parentheses. Constant not shown.
* p<0.1, ** p<0.05, *** p<0.01