

Collaborative Networks as Determinants of Knowledge Diffusion Patterns

Jasjit Singh

INSEAD, 1 Ayer Rajah Avenue, 138676 Singapore

+65 67995341

jasjit.singh@insead.edu

<http://www.jasjitsingh.com>

Original version: May 4, 2004

This version: November 29, 2004

Forthcoming, *Management Science*

I thank Ajay Agrawal, Juan Alcacer, Bharat Anand, Pierre Azoulay, Richard Caves, Iain Cockburn, Ken Corts, Lee Fleming, Robert Gibbons, Heather Haveman, Tarun Khanna, Steven Klepper, Josh Lerner, Jordan Siegel, Olav Sorenson, Michael Stolpe, Toby Stuart, Peter Thompson, Dennis Yao, two anonymous referees and seminar participants at CMU, Columbia, Emory, GWU, Harvard, HEC, IESE, INSEAD, Instituto de Empresa, LBS, Maryland, Minnesota, MIT, NBER, NUS, NYU, Rutgers, SMU, UNC, Vanderbilt and Wharton for helpful comments. I am grateful to Adam Jaffe and Lee Fleming for allowing me access to their data, and to Harvard Business School and INSEAD for funding this research. Errors remain my own.

Abstract

This paper examines if interpersonal networks help explain two widely documented patterns of knowledge diffusion: (1) geographic localization of knowledge flows, and (2) localization of knowledge flows within firm boundaries. I measure knowledge flows using patent citation data, and employ a novel regression framework based on choice-based sampling to estimate the probability of knowledge flow between inventors of any two patents. As expected, intra-regional and intra-firm knowledge flows are found to be stronger than those across regional or firm boundaries. I explore if these patterns can be explained by direct and indirect network ties between inventors, as inferred from past collaborations between them. The existence of a tie is found to be associated with a greater probability of knowledge flow, with the probability decreasing as the path length (geodesic) increases. Further, the effect of regional or firm boundaries on knowledge flow decreases once interpersonal ties have been accounted for. In fact, being in the same region or firm is found to have little additional effect on the probability of knowledge flow between inventors that already have close network ties. The overall evidence is consistent with a view that interpersonal networks are important in determining observed patterns of knowledge diffusion.

Keywords: Knowledge spillovers, Technology diffusion, Social networks, Interpersonal ties, Innovation

JEL Codes: F2, O3, R1, L0, M2

1. Introduction

Acquisition of knowledge can be crucial for economic success of firms, and for innovativeness and growth of geographic regions (Grossman and Helpman, 1991). However, knowledge acquisition is not easy. Even though ideas are intangible in nature, they can be extremely hard to transmit across regional or firm boundaries. In particular, two patterns of knowledge diffusion have been documented. First, knowledge flows are geographically localized (Jaffe, Trajtenberg and Henderson, 1993). Second, knowledge diffuses more easily within a firm than between firms (Kogut and Zander, 1992). In this paper, I formally examine interpersonal networks as the driver behind these patterns. While other factors such as institutions, norms, language, culture or incentives could also influence how knowledge diffuses, I estimate how much of the empirical pattern of knowledge diffusion can be explained simply by the fact that people within a region or a firm have closer interpersonal ties. While the main focus of this paper is to study the role of interpersonal networks in determining knowledge diffusion patterns, its contributions to the literature also include an improved methodology for analyzing micro-level knowledge flows, a richer dataset than is typically employed in studies on knowledge diffusion, and a separate identification of the impact of direct versus indirect ties on knowledge flow.

Extending existing research methodology for using patent citation data to measure knowledge flows (e.g., Jaffe and Trajtenberg, 2002), I employ a novel regression framework based on choice-based sampling to estimate the probability of knowledge flow between inventors of any two patents. Consistent with past research, I find that knowledge flows are stronger within regions or firms than across. In particular, even after carefully controlling for technological specialization of regions (Thompson and Fox-Kean, 2004), knowledge flow between two inventors from the same region is found to be 66% more likely than that between those from different regions. Likewise, all else being equal, knowledge flow between two inventors is three times as likely within a firm as between firms.

A rich literature in sociology has emphasized information flow through interpersonal networks (Ryan and Gross, 1943; Coleman, Katz and Menzel, 1966; Granovetter, 1973; Burt, 1992; Rogers, 1995).

This has motivated the current study for investigating whether such interpersonal networks are behind the observed patterns of knowledge diffusion mentioned above. While large-scale systematic data on interpersonal relations can be hard to obtain, an indirect way of inferring such relations is by studying secondary data on collaborations between individuals (Cockburn and Henderson, 1998; Newman, 2001; Fleming, King and Juda, 2004). Following this approach, I use collaboration information for patents registered with the U.S. Patent Office (USPTO) to construct a rich longitudinal database of interpersonal relations among *all* inventors recorded by USPTO since 1975. This forms the basis of constructing a “social proximity graph” for over one million inventors, which is then used to measure the “social distance” between any two inventors. I capture both direct and indirect interpersonal relations. For example, if an individual X has a direct tie with individual Y, and Y has a direct tie with Z, I allow the possibility that Z might learn indirectly about X’s work through his or her tie with Y. In the analysis, collaborative ties between inventors are found to be an important determinant of the probability of knowledge flow, with this probability falling with increase in social distance. For example, the probability of knowledge flow is four times as much for inventors with direct collaborative ties as for inventors which are not connected, and is 3.2 times as much for inventors having no direct tie but an indirect tie through past collaboration with the same individual.

The main analysis of the paper studies if the above interpersonal networks help explain the patterns of knowledge diffusion discussed earlier. I find that the effect of being in the same region or the same firm on the probability of knowledge flow does decrease once collaborative networks have been taken into account. While the decrease is non-trivial in magnitude and statistically highly significant, the magnitude of this effect is not as large as one might expect: once interpersonal ties have been controlled for, there is a 17% decrease in the effect of geographic co-location on probability of knowledge flow and a 12% decrease in the effect of being within the same firm on the probability of knowledge flow. However, since only a subset of all interpersonal relations are captured by a network constructed exclusively from patent collaboration data, these results should be viewed as a *lower bound* on the importance of interpersonal networks in general. This interpretation is strengthened by the additional

finding that geographic proximity and firm boundaries have little additional effect on the probability of knowledge flow between inventors that are already closely connected in the collaboration network. On the other hand, the regional and firm boundaries continue to matter a lot for knowledge flow between inventors with no ties or with only very indirect ties. As explained later in the paper, this is consistent with a view that interpersonal networks are actually quite important in causing the more intense intra-regional and intra-firm knowledge flows, and that the detected effects reported earlier are of relatively small magnitude simply because patent collaborations capture only a small portion of all relevant interpersonal ties.

The paper is organized as follows. Section 2 motivates my formal hypotheses. Section 3 describes the data on patent citations as well as on inventors. Section 4 introduces my citation-level regression framework for estimating probability of knowledge flow, and describes how I measure interpersonal ties. Section 5 reports and explains the empirical findings. Section 6 discusses some open empirical issues and possible future extensions. Section 7 offers implications and concluding thoughts.

2. Hypotheses

The analysis in this paper is comprised of three main parts, as summarized in Figure 1. The first part is to formally establish the “fact” that intra-regional and intra-firm knowledge flows are indeed more intense than that between different regions or firms. The second part is to test the extent to which existence and directness of interpersonal ties between individuals determines the probability of knowledge flow between them. The third part, which is the crux of this paper, is to combine the constructs from the first two parts in order to examine the extent to which these interpersonal networks help explain the intense intra-regional and intra-firm knowledge flows.

While previous work has documented geographic localization of knowledge flow (e.g., Jaffe, Trajtenberg and Henderson, 1993), recent work raises methodological concerns that could have led to over-estimation of this phenomenon (Thompson and Fox-Kean, 2004). Therefore, before trying to explain

the result on intra-regional knowledge flows, I first test if the result does indeed continue to hold even when using a new empirical approach (as detailed later) that addresses some of these concerns:

Hypothesis 1. *The probability of knowledge flow within a region exceeds that between different regions, even after accounting for technological specialization of regions.*

The second pattern of knowledge diffusion I study is that knowledge diffuses more effectively within a firm than would be possible through a market-mediated mechanism (Kogut and Zander, 1992). Before examining collaborative networks as a possible driver for this, I formally reproduce this result as well by testing the following:

Hypothesis 2. *The probability of knowledge flow within a firm exceeds that between different firms, even after accounting for technological specialization of firms.*

Mobility of individuals has been shown to be an important mechanism through which knowledge diffuses (Saxenian, 1994; Almeida and Kogut, 1999; Rosenkopf and Almeida, 2003). However, even in the absence of direct mobility of individuals, information can diffuse through interpersonal networks (Zander and Kogut, 1995; Zucker, Darby and Brewer, 1998; Shane and Cable, 2002; Stuart and Sorenson, 2003; Uzzi and Lancaster, 2003). In this paper, I focus specifically on direct and indirect interpersonal ties that arise from patent collaborations between inventors.¹ The next hypothesis is that such ties do enhance transmission of knowledge:

Hypothesis 3. *The probability of knowledge flow is greater between inventors with a direct or indirect collaborative tie than between inventors that are not connected in the collaborative network.*

Close network links are potentially more useful for transferring knowledge that is complex and not easily codifiable (Ghoshal, Korine and Szulanski, 1994; Uzzi, 1996; Hansen, 1999). The codified part of such knowledge (e.g., description of an innovation as recorded in a patent description) may represent just the “tip of the iceberg”, with the rest being “tacit” (Polanyi, 1966; Nelson and Winter, 1982; Kogut

¹ Stolpe (2001) uses a sample of patents on liquid crystal display technology to study knowledge diffusion through just *direct* collaborative links, but finds no evidence of it. Possible explanations could be that this technology is unusual (e.g., easier to codify), or simply that the number of collaborations in this sample was too small.

and Zander, 1992). Transmission of such knowledge might therefore be easier between individuals with close ties (Allen, 1977; Nonaka, 1994; Szulanski, 1996). In addition, direct relationships also induce more trust, improving willingness of individuals to share knowledge (Tsai and Ghoshal, 1998; Levin and Cross, 2003). Thus, transmission of complex technical knowledge should become more difficult as the “social distance”, or the number of intermediaries needed to pass knowledge from the source to the destination, increases.² This suggests the following hypothesis:

Hypothesis 4. *The probability of knowledge flow between individuals is a decreasing function of the social distance between them.*

Now I come to the main hypotheses of this paper, which involve studying the extent to which the results from Hypotheses 1 and 2 can be explained by the interpersonal networks from Hypotheses 3 and 4. Several empirical studies, such as that by Kono, Palmer, Friedland and Zafonte (1998), have established that spatial propinquity facilitates relationship formation. Sorenson and Stuart (2001) show that such localized interpersonal ties in the venture capital community lead to localized flow of information regarding investment opportunities, which in turn results in geographic localization of venture capital investments. Analogously, I test if the geographic localization of technological knowledge flows can also be explained by the fact that close direct or indirect collaborative ties are more likely to exist between inventors from the same region. This gives the following formal hypothesis:

Hypothesis 5. *Accounting for collaborative networks leads to a significant drop in the effect of geographic co-location of inventor teams on the probability of knowledge flow between them.*

An alternate hypothesis could be that geographic concentration of knowledge flows is driven not by collaborative networks but by other mechanisms such as informal interaction (“ideas in the air”) or

² Granovetter (1973) and Burt (1992) suggest that non-redundancy of resulting information flow determines the usefulness of a network tie. As a referee correctly pointed out, my “social distance” measure is defined using only the shortest network path between two teams of inventors, and hence does not capture alternate paths between two nodes and non-redundancy of information flow. Therefore, Hypothesis 4 should not be interpreted as a direct test of Granovetter’s and Burt’s theory. This issue is studied by Hansen (1999), who shows that weak and non-redundant ties are better when searching for simple information, while strong ties are better for transfer of complex knowledge.

region-specific factors like local infrastructure, institutions, regional publications, communication channels, norms, culture and government policies. Another plausible reason why patent-based collaborative networks might explain only a small portion of the intra-region knowledge flows could simply be that patent collaborations reveal only a subset of actual interpersonal networks.

Analogous to studying why intra-regional knowledge flows are strong is the question why knowledge flows are stronger within firms than between firms. Like Simon (1991) and Grant (1996), I take individuals as the unit of analysis for studying knowledge flows even within organizations. This allows me to use a unified network framework to study both inter-firm and intra-firm knowledge flows. Specifically, I explore how much of a firm's ability to transfer knowledge between its employees can be explained simply by the fact that it is a tightly knit "social community" in the specific sense of having a dense collaborative network. This gives my final hypothesis:

Hypothesis 6. *Accounting for collaborative networks leads to a significant drop in the effect of firm boundaries on the probability of knowledge flow between two teams of inventors.*

An alternate hypothesis could be that intra-firm knowledge flows are driven not by interpersonal networks but by organizational routines and processes, confidentiality requirements and different incentives for sharing knowledge with fellow employees versus outsiders. Once again, another plausible reason why networks based on patent collaborations might explain only some of the intra-firm knowledge flows is that patent collaborations surely capture only a fraction of all relevant interpersonal ties. Yet another reason why the measured "knowledge flows" are greater within firms could simply be that the role of patent citations in determining patent scope and litigation behavior makes incentives for intra-firm citations very different from those for inter-firm citations (Jaffe and Trajtenberg, 2002).

3. Patent Data

3.1. Patent Citations as Measure of Knowledge Flow

Patent citations leave behind a trail of how an innovation potentially builds upon existing knowledge. Unlike in academic papers, there is an incentive not to include superfluous patent citations as

that might reduce the scope of one's own patent. On the other hand, an inventor is legally bound to report relevant "prior art", with the patent examiner performing an objective check. Nevertheless, not all citations reflect knowledge flows. Citations might be included for strategic reasons (e.g., to avoid litigation). Also, a firm's lawyer or a patent examiner might add citations that the original inventor did not know about (Thompson, 2004; Alcacer and Gittelman, 2004). Nevertheless, recent studies comparing citation data with direct surveys of inventors show that the correlation between patent citations and actual knowledge flows is high, though not perfect (Jaffe and Trajtenberg, 2002; Duguet and MacGarvie, 2002).

I merged patent data from the US Patent Office (USPTO) with that from Jaffe and Trajtenberg (2002). The dataset had to be further enhanced to correctly identify each patent's owner, since some of a firm's patents are often listed under the name of a subsidiary instead of the parent firm. I performed parent-subsidiary match using Stopford's *Directory of Multinationals*, Dun and Bradstreet's *Who Owns Whom* directories, Compustat identifiers from Jaffe and Trajtenberg (2002), and Internet sources. About 3,300 major firms and organizations were identified, accounting for about half of USPTO patents.³ The rest of the patents were dropped, with one-third of them having only individual owners and no assignees, and the rest being spread among more than 100,000 assignees. The sample was further restricted to be from years 1986 to 1995, since the parent-subsidiary match used data sources from this period.

The geographic region of a patent was taken as one of the 337 Metropolitan Statistical Areas (MSA) or Primary Metropolitan Statistical Areas (PMSA) in the U.S., as determined by the first inventor's address. An MSA or PMSA consists of a cluster of adjacent counties with close economic and social relationships.⁴ Since I do not have systematic fine-grained geographic information for innovations arising outside the U.S., all patents with a non-U.S. address were dropped.

³ For some patents, different inventors could have different employers. However, I only have information on the first assignee, and hence had to assume that all inventors of a patent have the same employer. For brevity, I occasionally use the word "firm" to refer to firms as well as other organizations. Non-firm entities (like universities, research institutes and government bodies) own about 10% of the patents, and the results do not change even if I drop these.

⁴ I thank Lee Fleming for allowing me access to his mapping between patents and metropolitan areas, which is based on metropolitan area information available from ZipInfo (<http://www.zipinfo.com/products/z5msa/z5msa.htm>).

Even within the U.S., not all inventors live in a metropolitan area, or, even if they do, the patent-area mapping is sometimes unavailable because of data errors. These two reasons led around 20% of U.S.-based patents to be dropped as well. As a robustness check, I repeated the analysis for all U.S.-based patents using state as the unit of analysis. The main results were very similar and are therefore not reported in the paper.

3.2. Inventors

Fleming, Colfer, Marin and McPhie (2003) and Fleming, King and Juda (2004) present evidence from field interviews that collaborations recorded on patent documents meaningfully (though not perfectly) capture personal and professional interpersonal ties between inventors. Since patents are non-trivial innovations by definition, co-inventors of a patent typically collaborate intensively over an extended time period, and often maintain close interpersonal contact even years after filing the patent.

A challenge in using secondary data on collaborations is correctly identifying when two different patent records refer to the same person. I experimented with several methods to find a compromise between too many “false positives” (different individuals being incorrectly identified as the same) and too many “false negatives” (different records of the same inventor misclassified as having different inventors). Finally, I arrived at an algorithm that took two inventors as the same if and only if all of the following conditions hold:

1. The first and last names matched exactly.
2. The middle initials, if available, were the same.
3. When the middle initial field was blank in at least one of the two records, the records also overlapped on at least one of their technology subcategories.

Using only the first two conditions would have identified 1.3 million or so distinct inventors since 1975. The third condition makes the matching criteria more stringent, leading to around 1.7 million inventors. The definition of a technology “subcategory” is derived from Jaffe and Trajtenberg (2002), who group the 418 USPTO technology classes into 38 subcategories. I tried to rule out more “false positives” by requiring a finer technological classification for match across patents in the third

condition, and by also looking for an overlap of citations across patents. However, imposing either of these two conditions caused too many “false negatives” since the overlap in either case was quite low even for records from the same inventor. I also considered requiring a match for street address and/or assignee firm, as used by Fleming, Colfer, Marin and McPhie (2003). However, I decided against it since studying interaction of network ties with geography and firm boundaries is a central focus of this paper, and using these for matching might have unpredictably biased the results. Also, as Fleming et al report, forcing these requirements makes the match too conservative, an issue they handle by not requiring the rule for relatively uncommon last names. There would, irrespective of the algorithm used, be some errors in any matching process. However, unless there is a reason to believe that the matching is producing *systematic* errors, it should just lead to an attenuation bias that only makes it harder to detect an effect of collaborative networks on the probability of knowledge diffusion.

4. Empirical Methodology

4.1. Choice-Based Sampling

My empirical model estimates the probability of knowledge flow between two innovations that do end up as patents. In other words, I estimate a “citation function” $P(K, k)$, which specifies the probability that a patent K cites a patent k . Imagine a population of all such patent pairs (K, k) . In principle, we could draw a random sample from this population, and define a dependent variable y to equal 1 for pairs of patents with a citation and 0 for others. Assuming that the citation function takes a logistic functional form, y takes a value 1 for observation i with the probability

$$\Pr(y = 1 | x = x_i) = \Lambda(x_i \beta) = \frac{1}{1 + e^{-x_i \beta}}$$

where x_i is the vector of covariate values, and β is the vector of parameters to be estimated.

Unfortunately, an estimation approach based on random sampling is not practical since citations between random patents are very rare: there are only a few realized citations for every one million potential citations, making meaningful estimation impossible even with large samples. From an informational point

of view, it would be desirable to have a greater fraction of observations with $y = 1$. This can be achieved by using a “choice-based” sampling procedure: the sample is formed by taking a fraction α of the patent pairs with $y = 0$ and a fraction γ of the patent pairs with $y = 1$ from the population, with α being much smaller than γ . Since this stratification is done on the dependent variable, using the usual logistic estimates would lead to a selection bias. A technique that overcomes this problem is the *weighted exogenous sampling maximum likelihood* (WESML) estimator suggested by Manski and Lerman (1977). Intuitively, the idea is to weight each sample observation by the number of population elements it “represents” in order to make the choice-based sample “simulate” a random exogenous sample. Formally, the WESML estimator is obtained by maximizing the following weighted pseudo-likelihood function:

$$\ln L_w = \frac{1}{\gamma} \sum_{\{y_i=1\}} \ln(\Lambda_i) + \frac{1}{\alpha} \sum_{\{y_i=0\}} \ln(1 - \Lambda_i) = - \sum_{i=1}^n w_i \ln(1 + e^{(1-2y_i)x_i\beta})$$

where $w_i = (1/\gamma)y_i + (1/\alpha)(1 - y_i)$. The appropriate estimator of the asymptotic covariance matrix is White’s robust “sandwich” estimator used in pseudo-maximum likelihood estimation. Further, since the same citing patent can occur in multiple observations, the standard errors should be calculated without assuming independence across these observations.⁵

4.2. Sample Construction

The basic WESML approach samples all $y = 0$ observations with equal probability α . Since technological similarity of two patents is a strong determinant of the probability of citation, estimation efficiency can be improved by matching each patent pair having a citation (i.e., with $y = 1$) with a set of “control pairs” (i.e., with $y = 0$) such that the citing and cited patent in each control pair belong to the same respective technology class as those in the original citation.⁶ I followed this approach in matching each of the 323,820 actual citations among patents in my sample with five control pairs. In addition, to

⁵ An online appendix accompanying this paper (available at <http://www.jasjitsingh.com>) gives technical details. Please refer to Amemiya (1985, pp. 319-338) or King and Zeng (2001) for more discussion on WESML, and to Sorenson and Fleming (2001) for an earlier application of this methodology for predicting patent citations.

⁶ Sorenson and Stuart (2001) use a similar research design for estimating probability of venture capital funding.

ensure that pairs of technology classes with no citations between them were also represented in the sample, I drew a random $y = 0$ observation for each of such class pairs. These two steps led to a total of 2,217,171 control citations for the 323,820 actual citations, giving an overall sample size of 2,540,991 actual and potential citations. As the online appendix accompanying this paper shows, the WESML approach should now be generalized by noting that the sampling rate α varies by the technology class of citing and cited patents. Specifically, the weight attached to a $y = 0$ observation is now defined as the ratio of the number of $y = 0$ elements in the population to the number of $y = 0$ observations in the sample *for any given pair of technology classes*. In addition, each $y = 1$ element simply has a weight of one since I include all actual citations from the population in the sample (i.e., $\gamma = 1$).

4.3. Control Variables for Probability of Citation

To account for the fact that technologically similar patents have a greater probability of citation, existing literature typically controls for whether the 3-digit technological class of the citing and cited patents are the same. However, this can still lead to biased estimates, since there can be large heterogeneity in technology within a 3-digit class. For example, the class “Aeronautics” includes 9-digit subclasses as diverse as “Spaceship control” and “Aircraft seat belts”. To take this into account, I define dummy variables not just to control for cases with the same broad technological category (1 out of 6), the same technological subcategory (1 out of 36) and the same 3-digit primary class (1 out of 418), but also for the same 9-digit primary class (1 out of 150,000). Further, since the designation of a subclass as “primary” can sometimes be ad hoc, I also include a dummy variable that captures overlap along secondary subclasses for the citing and cited patent. While even these technology controls might not be perfect, these are the most fine-grained level possible with USPTO data, and are much more detailed than the coarse controls used in most existing studies.⁷ I also account for other factors that

⁷ Some regression-based studies use the number of citations as the dependent variable, and include a measure of “average technological distance” between citing and cited sets of patents using only a 2 or 3-digit technology classification (e.g., Jaffe and Trajtenberg, 2002). The issue of bias remains: sets with a greater fraction of patent pairs with the same 9-digit technology have a greater probability of citation, and also greater co-location of patents.

affect the probability of patent citation by including fixed effects for the time lag (in years) between the citing and cited patent, and for the application year and technological category of the citing patent.

4.4. Measuring Social Distance between Innovating Teams

In order to measure the existence and directness of collaborative ties between two teams of inventors, I define “social distance” as the minimum number of intermediaries needed to pass knowledge from the source team to the destination. This is analogous to measuring “degrees of separation” in recent studies on “small worlds” (Watts and Strogatz, 1998; Newman, 2001). In using collaboration data, it is common practice to assume that an observed collaboration marks the *beginning* of a tie between the individuals, which persists beyond the recorded collaboration date (Stolpe, 2001; Breschi & Lissoni, 2002; Agrawal, Cockburn and McHale, 2003). This assumption is supported for the case of patents by field evidence reported by Fleming, Colfer, Marin and McPhie (2003), and is therefore followed here as well. In inferring network ties that exist as of any year t (t being between 1986 and 1995), I include *all* inventors that have patented between 1975 and t (including even those not in the US, and those not associated with the 3,300 assignees used for analyzing knowledge flows).⁸

Data on inventors and inventing teams can be represented using an “affiliation matrix” $\mathbf{A} = \{a_{ij}\}$, where a_{ij} is “1” if the i th inventor is on the collaborating team for the j th patent, “0” otherwise (Wasserman and Faust, 1994). Figure 2 gives an example, with 7 inventors A, B, C, D, E, F and G, and 7 patents P1, P2, P3, P4, P5, P6 and P7. A value of “1” for element (A, P1) and “0” for element (C, P1), for example, implies that A is one of the inventors for patent P1, but C is not. The first step for studying collaborative links between inventors is to construct a “social proximity graph”. The graph for year t includes as nodes all innovations made by year t , with an edge between patenting teams X and Y if and only if the two teams have a common inventor. If there is a citation between two patents sharing a common inventor, like patents P1 and P2 in Figure 3(a), it can be seen as an inventor citing

⁸ The “Small Worlds” literature (Watts and Strogatz, 1998; Newman, 2001) uses network nodes to represent *individuals* instead of *teams*, with edges between individuals that have collaborated. For this paper, it is more natural to define the collaborating *teams* as nodes since measured knowledge flows are from one patenting *team* to another.

his or her own previous work. Since self-citations by individuals do not represent real knowledge flow, the sample used in the regression analysis later in this paper does not include such pairs of patents.

The interesting case is where two patents do not have a common inventor, but have a direct or indirect collaborative tie. For example, in Figure 3(b), knowledge from P1 can flow to P3 indirectly via the path $P1 \rightarrow P2 \rightarrow P3$ by being passed from A to C, since A and C having a collaborative link as evidenced by P2. I define “social distance” as *the number of intermediate nodes on the minimum path (the geodesic) between any two nodes in the social proximity graph*. Thus, the social distance is “1” for the $P1 \rightarrow P3$ example above, as indicated in Figure 4. Since knowledge flows are meaningful only from an innovation that happens earlier to one that happens later, social distance need not be defined for $P2 \rightarrow P1$, $P1 \rightarrow P1$, etc.

Now consider Figure 3(c). The above definition suggests a social distance of “1” for $P2 \rightarrow P4$, since there is a path $P2 \rightarrow P1 \rightarrow P4$. Does this make sense even though P1 precedes P2 in time? As discussed above, a collaboration between A and B for P1 is the *beginning* of a relationship between the two. Thus, B (who is also the inventor of P4) can indeed build upon knowledge of P2 that B can gain through his or her ongoing tie with A. Thus knowledge can flow “backwards in time” along the link $P1 \rightarrow P2$, and then on to $P2 \rightarrow P4$. Likewise, knowledge from P3 can be passed by C to A, and then on to B through the chain $P3 \rightarrow P2 \rightarrow P1 \rightarrow P4$, making the social distance $P3 \rightarrow P4$ to be “2”.

Naturally, the social proximity graph evolves over time. Therefore, I actually use separate social proximity graphs for years $t=1986$ through $t=1995$ to cover all the years for which I analyze knowledge flows. To measure social distances for innovating teams from year t , I use a graph of collaborative ties already in place by t . For example, the correct value of social distance from P3 to P6

is infinity (since P6 took place in 1989, and P3 and P6 are not even in the same connected component in 1989) and not “2” (as an incorrect interpretation of the 1990 graph might suggest).⁹

Because of the large graph size, computing exact pair-wise social distances is practically impossible. Fortunately, it is still practical to classify all observations with a non-zero social distance into five mutually exclusive and exhaustive categories based on whether the social distance is 1, 2, 3, any finite value greater 3, or infinity (i.e., no social links). As the definition of variables in Table 1 shows, I capture these five cases using indicator variables *past collaboration*, *common collaborator*, *collaborators with ties*, *indirect social link* and *no social link* respectively in the analysis that follows.¹⁰

5. Results

5.1. Intra-Regional and Intra-Firm Knowledge Flows

Table 2 formally tests Hypotheses 1 and 2, i.e., that knowledge flows are particularly strong within the same region or the same firm. The weighted logit framework described above is used to estimate the probability of citation between patents, with the dependent variable being 1 when a patent pair has a citation, 0 otherwise. Column (1) finds positive and significant estimates for *within same region* and *within same firm*. However, this could result simply from technological specialization of regions and firms (Jaffe, Trajtenberg and Henderson, 1993). As column (2) shows, including controls for technological relatedness (at the level of 3-digit technological class) between patents reduces but does not eliminate the estimated coefficients for *within same region* and *within same firm*. However,

⁹ Since the social distance measure might not be comparable across years, I use year fixed effects in the regressions described later. An alternate approach could have been to use a rolling time window (e.g., use collaborations from year $t-7$ to t in defining the graph for year t) instead of using the entire history of collaborations for any year t .

¹⁰ Wasserman and Faust (1994) suggest computing pair-wise distances as follows: Define element x_{ij} of a matrix X as 1 if there is an edge between nodes i and j , 0 otherwise. The distance between i and j is then the smallest number p such that the p^{th} power matrix of X has a non-zero entry (i, j) . Unfortunately, any such approach is impractical for very large graphs (Cormen, Leiserson and Rivest, 1990). Instead, I explicitly find all pairs with a social distance of

Thompson and Fox-Kean (2004) have shown that even the 3-digit technological controls, though extensively used, are insufficient. To address this, column (3) uses additional controls based on a detailed 9-digit primary and secondary technological classification of patents. The estimates for *within same region* and *within same firm* fall further, but still remain significant. Since statistical significance could result merely from having a large sample size, I now turn to the magnitude of these effects.

The marginal effects for the weighted logit model are shown in square brackets in column (3) of Table 2, after being multiplied by a million for readability as citations are rare events.¹¹ The predicted citation rate between two random patents turns out to be about 11 in a million. Therefore, the reported marginal effect of 7.22 for *within same region* implies that patents from the same region are 66% more likely to have a citation than are otherwise similar patents from different regions. Similarly, the marginal effect of 21.6 for *within same firm* implies that patents from the same firm are around 3 times as likely to have a citation as are patents from different firms. This confirms Hypotheses 1 and 2.

5.2. The Effect of Social Distance on Probability of Knowledge Flow

As already discussed, indicator variables *past collaboration*, *common collaborator*, *collaborators with ties* and *indirect social link* capture a social distance of 1, 2, 3 and greater than 3 (but finite). Pairs of patents with a common inventor (i.e., a social distance of 0) have already been dropped from the sample since self-citations by inventors do not signify knowledge flow. As a result, if two patents belong to the same connected component in the social proximity graph, exactly one of the above four dummy variables has a value of one. Table 3a reports summary statistics for these variables in the sample. The fraction of patent pairs with no social link is only 41.7% for pairs with citations, and 50.2% for pairs with no citation. This is consistent with the Hypothesis 3 that connectedness leads to greater probability of citation. The inequality holds even for the sub-sample without self-citations by firms, where the fraction of pairs with

1, 2 or 3 by calculating the first few power matrices as they are computationally manageable. I then distinguish between having an indirect social link and no social link by finding all connected components of the graph.

¹¹ For logit, the marginal effect of a variable j can be shown to be $\beta_j \Lambda(\mathbf{x}\boldsymbol{\beta})[1-\Lambda(\mathbf{x}\boldsymbol{\beta})]$. I substitute the mean predicted probability for $\Lambda(\mathbf{x}\boldsymbol{\beta})$ into this expression in order to get an estimate of the marginal effect.

no social link is only 46.1% for pairs with citations, and 51.1% for pairs with no citation. Table 3b gives simple correlation of the social distance measures with indicator variables for having a citation, being within the same region and being within the same firm. Since these are just raw correlations from a choice-based sample, the interpretation of these numbers is limited. Nevertheless, a fact that comes out quite strikingly is that having a smaller social distance is correlated with a greater probability of patent citation as well as with a greater probability of being within the same region or the same firm.

Table 4 reports regression analysis to test Hypotheses 3 and 4, i.e., the impact of collaborative links on probability of patent citation. As a comparison of columns (1) and (2) shows, carefully controlling for technological relatedness of patents is again important since teams with collaborative links are also more likely to be technologically related. Therefore, column (2) is the preferred regression specification. Consistent with Hypothesis 3, interpersonal ties are important as the estimates for *past collaboration*, *common collaborator*, *collaborators with ties* and *indirect social link* are all positive and significant. The joint hypothesis that these social distance measures do not matter is easily rejected even at the 1% significance level, with a $\chi^2(4)$ statistic of 3372.0. The reference group for comparison in these regressions is patent pairs with no social link.

Though statistical significance could result simply from large sample sizes, the effects are also large in magnitude. If two patents are related via a past collaboration (social distance = 1), the probability of citation is about four times that for unrelated patents. If they are related via a common collaborator (social distance = 2), the probability of citation is about 3.2 times. Similarly, if they are related only because they have had collaborators that have worked with each other in the past (social distance = 3), the probability of citation is about 2.7 times. Finally, if none of these cases occur but there still exists an even more indirect collaborative link between two patents, the probability of citation is merely 4% greater than that for unrelated patents. A statistical test of equality for the estimates of different social measures is easily rejected. Thus, consistent with Hypothesis 4, the probability of citation falls as the social distance for pairs of patents increases.

5.3. Collaborative Networks and Patterns of Knowledge Flows

In this section, I test Hypotheses 5 and 6 (i.e., that knowledge flows are more intense within the same region and the same firm *because* social distances are smaller). In other words, I explore the extent to which denser collaborative networks can be seen as the *mechanism* driving more intense knowledge flows within regions and firms.

The analysis appears in Table 5. For easy comparison, column (1) reproduces the intra-region and intra-firm results from column (3) of Table 1. Column (2) adds the social distance measures to the econometric model. Upon doing so, the coefficient estimate for *within same region* drops from 0.656 to 0.544, with its marginal effect falling from 7.22 to 5.98. In other words, once social distance has been accounted for, the incremental effect of geographic co-location on probability of citation falls from 65.6% to 54.4%.¹² Likewise, the coefficient estimate for *within same firm* drops from 1.964 to 1.726, with the marginal effect falling from 21.6 to 19.0. Put differently, once social distance has been controlled for, the incremental effect of being in the same firm on the probability of citation falls from 196% to 173%. To summarize, accounting for collaborative ties diminishes the result of localized knowledge flows as well as intra-firm knowledge flows. Not only is the decrease non-trivial in magnitude for both cases, it is also found to be statistically significant.¹³ However, the decrease turns out to be much smaller than one would expect if social networks were the *main* driver of knowledge diffusion. As discussed earlier in the theoretical discussion of hypotheses 5 and 6, a culprit for not having a stronger effect is probably that a network constructed using only patent collaborations captures just a fraction of all relevant network ties.

To investigate this further, column (3) of Table 5 considers a richer regression model that allows the possibility that direct and indirect ties do not operate similarly for transferring knowledge. In other

¹² Normally, in non-linear models, one should only compare marginal effects and not coefficient estimates across models. However, for rare events, the marginal effect $\beta_j \Lambda(\mathbf{x}\boldsymbol{\beta})[1-\Lambda(\mathbf{x}\boldsymbol{\beta})]$ can be approximated as $\beta_j \Lambda(\mathbf{x}\boldsymbol{\beta})$, making β_j directly interpretable as the fractional change in probability of citation when binary variable j goes from 0 to 1.

¹³ To test statistical significance, the coefficients of *within same region* in columns (1) and (2) were interpreted as means of samples drawn from normally distributed populations. A t-test was then used to test the hypothesis that the two means could arise from the same population. An analogous test was done for *within same firm*.

words, there might be interaction effects between social distance and geographic co-location, and between social distance and firm boundaries, in determining probability of knowledge flow. Since column (3) includes both these sets of interaction variables, the direct coefficients for *within same region* and *within same firm* now should be interpreted only as the effects for the reference case where the citing and cited patents have no social link. Interestingly, the interaction effects for *within same region* with *past collaboration*, *common collaborator* and *collaborators with ties* are similar in magnitude but opposite in sign as compared to the direct effect for *within same region*. In other words, conditional on the social distance being small (i.e., 1, 2 or 3), geographical co-location has a relatively small unexplained net effect on citation probability. For example, conditional on having a social distance of 1, the difference in net coefficient for patent pairs within the same region versus those that are not is given by $(0.868 - 0.697) = 0.171$. This translates into just a 17% increase in citation probability from being in the same region conditional on already having a social distance = 1. Statistically, this is in fact indistinguishable from having no unexplained effect of co-location once the standard errors have been taken into account. On the other hand, for patents that are connected only through indirect social links or are not connected at all, geographic co-location continues to affect citation probability significantly. For example, for patent pairs with no social link, the difference in net coefficient for pairs that arise within the same region versus those that do not is 0.868, which translates into an 87% difference in estimated citation probability between inventor teams that are geographically co-located versus those that are not. An explanation might be that, for teams with no close ties apparent from collaboration data on patents, there might still exist missing ties that are both geographically concentrated and beneficial for knowledge flow. These could, for example, be collaborations that did not lead to patents, and hence did not get captured in patent data. These could also be fundamentally different kinds of professional and social interaction, such as meeting at conferences and professional get-togethers, or even at golf clubs and coffee shops.

Analogously, the interaction effects for *within same firm* with *past collaboration*, *common collaborator* and *collaborators with ties* are all quite large in magnitude and opposite in sign to the main effect for *within same firm*. In other words, for patent pairs with a small social distance of 1, 2 or 3, being

in the same firm matters much less for the citation probability. In fact, a formal hypothesis that the effect is zero for the case of social distance of 1 cannot be rejected. Likewise, the net incremental effect of being within the same firm is much smaller for other cases of relatively short social distance (29% for social distance of 2, and 81% for social distance of 3) than it is for cases with indirect social link (158%) or no social link (203%). In other words, being in the same firm affects knowledge flows more in cases where close interpersonal links do not exist. Once again, this could perhaps be a result of interpersonal ties not captured in patent collaboration data.

6. Limitations and Future Research

This paper takes network ties as given. However, interpersonal ties can arise endogenously as a result of deliberate steps taken by rational actors (Coleman, 1988; Glaeser, Laibson and Sacerdote, 2002). If tie formation is more likely in settings where more knowledge flows are expected, regression estimates could overstate the true causal influence of collaborative links on knowledge flows. Similarly, if an inventor *X* who has cited inventor *Y* is also more likely to cite *Y*'s work in the future and also more likely to try to develop a collaborative relationship with *Y*, we might observe a correlation between collaborations and citations that need not signify a causal relationship. Addressing such causality issues would require explicitly modeling the tie formation process in an empirical framework.

While adopting a network perspective allows a study of within-firm and cross-firm knowledge flows in a single framework, it does not do full justice to a broader view of "organizational knowledge" (Levitt and March, 1988; Huber, 1991; Kogut and Zander, 1992; Nonaka, 1994). Another issue in studying intra-firm knowledge flows is that patent citations could be more common within firms simply because a firm does not lose anything by making citations to itself. The most conservative interpretation of my results is therefore to view the *within same firm* dummy only as a control variable, and to interpret the results as being only about geographic localization of knowledge flow.

Regarding methodology for using patent citations to measure knowledge flows, the measure could in principle be improved by omitting citations added by patent examiners and not by inventors

(Thompson, 2004; Alcacer and Gittelman, 2004). However, doing so was not practical for me since USPTO has started making the distinction between citations by patent examiners versus inventors only since 2001, and even that data is not easily available in a machine-readable form.

Another issue is that patent collaborations capture only a subset of relevant interpersonal relations. A natural extension of could therefore be to supplement patent collaboration data with additional data sources regarding interpersonal ties (e.g., collaboration on research papers or projects), and to see if patterns of knowledge flow can be explained more completely as a result. On the theoretical side, one could try to go beyond the “social distance” measure based on minimum path length, and to try to capture more nuanced network structural effects like the role of non-redundancy of information flow (Granovetter, 1973; Burt, 1992; Ahuja, 2000). Another interesting direction of research would be exploring how the role of different kinds of network ties differs across technologies with differences in complexity and codifiability of knowledge (Hansen, 1999; Sorenson, Rivkin and Fleming, 2004).

7. Implications and Conclusion

An extensive literature has emphasized that knowledge diffusion tends to be restricted by regional and firm boundaries. This paper rigorously examines *why* this happens, suggesting distribution of interpersonal networks as an explanation. I find evidence that interpersonal networks are quite important in determining patterns of intra-regional and intra-firm knowledge flow, even though their full impact might be hard to measure since collaborations on patents represent only a small portion of the overall set of social relations. The importance of studying the interplay of social networks and geography is echoed in other ongoing research efforts. For example, Agrawal, Cockburn and McHale (2003) show that patents from inventors who move from one region to another continue to be cited by their former collaborators, reflecting that direct ties from past collaborations facilitate knowledge flow across regions. Likewise, Breschi and Lissoni (2002) find the association between patent citations and geographic co-location in Italy to be greater for socially connected patent teams than others, suggesting important interaction effects between geographic co-location and collaborative links. The focus of my paper has been enriching this

stream of research through a rigorous measurement of the extent to which collaborative networks help explain observed patterns of intra-regional and intra-firm knowledge flow. In the process, I also introduce an improved methodology for studying micro-level knowledge flows, a much more exhaustive dataset than is typically used by any study of this kind, and an analysis that allows separate estimation of the impact of direct versus indirect interpersonal ties in collaborative networks.

This paper has important implications for management. The results emphasize that interpersonal networks are crucial for management of complex knowledge, despite growing emphasis on formal knowledge management systems. Further, while geography matters for knowledge diffusion, and it is so at least in part because interpersonal networks tend to be regional in nature. This suggests that an important component of a firm's human resource management should be not only to track the knowledge base of its employees but also to understand their participation in key interpersonal networks that span regional and firm boundaries. Further, a firm could learn more from its environment by encouraging its employees to build external collaborative links rather than merely opening divisions close to "hi-tech clusters" with a hope that knowledge gains would follow on their own. The analysis on interaction between social distance and geographic co-location shows that, for inventors connected via short path lengths, geographic co-location has a smaller residual effect on the probability of knowledge flow. This suggests that geographic constraints can be overcome by fostering interpersonal links across regions.

However, two puzzles still remain regarding firm strategy. First, if the knowledge gains from locating in a geographic region depend on the extent to which a firm's employees are connected in the broader network, how does a firm capture at least a part of the rents from "knowledge spillovers" rather than these accumulating completely to employees in the form of higher wages? Second, since collaborative links with outsiders can lead to not just knowledge *inflows* but also knowledge *outflows* (Singh, 2004), how does a firm prevent loss of its competitive position resulting from "leakage" of its own knowledge to competitors? The answer probably lies in the firm's ability to employ unique complementary assets that make some of the knowledge more valuable inside the firm than when used by its competitors. However, this is an issue worth future exploration.

The findings on intra-firm knowledge flow have important implications as well. For example, the analysis on interaction between social distance and firm boundaries shows that firm boundaries *per se* need not constrain knowledge flow if strong collaborative links can be established with outsiders. On the other hand, even mergers or acquisitions might not be sufficient for knowledge flow if the employee networks of the two former firms fail to be integrated. Likewise, the success of alliances and joint ventures as a means for knowledge transfer also depends on fostering close interpersonal ties between employees from the two sides, an argument consistent with findings of Mowery, Oxley and Silverman (1996), Rosenkopf and Almeida (2003), and Gomes-Casseres, Jaffe and Hagedoorn (2003).

The results should also be of interest to a policy-maker interested in a region's economic development. For example, incentives given to encourage outside firms only to open a local division may not be enough in themselves for ensuring knowledge spillovers to local firms. Such knowledge flows can be enhanced through deliberate cultivation of interpersonal networks, for example, by encouraging mobility and interaction of people across firm and regional boundaries. By influencing the structure of networks, a policy-maker might be able to influence not just the knowledge flows but ultimately the capacity of regions to innovate (Fleming, King and Juda, 2004). While firms might not be open to direct interference in such matters, regional leaders can have indirect influence over regional interpersonal networks through policy instruments, e.g., through lax implementation of non-compete agreements, and through subsidies for joint R&D projects and joint regional conferences.

The finding that collaborative networks can help overcome geographic distance is particularly important for underdeveloped regions and countries. This suggests that, besides trying to entice advanced firms from elsewhere to open local subsidiaries, regions can also take an active approach towards external learning by tapping into foreign collaborative networks directly. For example, overseas movement of people ("brain drain") from developing countries need not be welfare-reducing, and location of R&D laboratories overseas by local firms should not be seen as erosion of local technical base. Instead, governments might consider setting up incentives and mechanisms for their well-trained emigrants to

continue to maintain close interpersonal ties with the locals, and encourage local firms to use foreign subsidiaries to access foreign knowledge by tapping into foreign interpersonal networks.

References

- Agrawal, A., I. Cockburn and J. McHale (2003), "Gone But Not Forgotten: Labor Flows, Knowledge Spillovers, and Enduring Social Capital." NBER Working Paper No. 9950
- Ahuja, G. (2000), "Collaboration Networks, Structural Holes, and Innovation: A Longitudinal Study," *Administrative Science Quarterly*, 45: 425-455.
- Alcacer, J. and M. Gittelman (2004), "How Do I Know What You Know? Patent Examiners and the Generation of Patent Citations," Mimeo.
- Allen, T.J. (1977). *Managing the Flow of Technology*. MIT Press, Cambridge.
- Almeida, P. and B. Kogut (1999), "The Localization of Knowledge and the Mobility of Engineers in Regional Networks," *Management Science*, Vol. 45(7), 905-917.
- Amemiya, T. (1985) *Advanced Econometrics*. Harvard University Press, Cambridge.
- Breschi, S. and F. Lissoni (2002), "Mobility and Social Networks: Localised Knowledge Spillovers Revisited" Paper presented in workshop on "The Role of Labour Mobility and Informal networks for Knowledge Transfer" at the Max Plank Institute.
- Burt, R.S. (1992) *Structural Holes: The Social Structure of Competition*. Harvard University Press, Cambridge.
- Cockburn, I.M. and R.M. Henderson (1998), "Absorptive Capacity, Coauthoring Behavior, and the organization of Research in Drug Discovery," *Journal of Industrial Economics*, 46(2): 157-182.
- Coleman, J.S., E. Katz and H. Menzel (1966). *Medical Innovation*. Bobbs-Merrill, New York.
- Coleman, J.S. (1988), "Social Capital in the Creation of Human Capital" *The American Journal of Sociology*, Vol. 94 Supplement: S95-S120.
- Cormen, T.H., C.E. Leiserson and R. L. Rivest (1990). *Introduction to Algorithms*. MIT Press, Cambridge.
- Duguet, E. and M. MacGarvie (2002), "How Well Do Patent Citations Measure Knowledge Spillovers?" Mimeo.
- Fleming, L., L. Colfer, A. Marin and J. McPhie (2003), "Why the Valley Went First: Agglomeration and Emergence in Regional Inventor Networks," Mimeo.
- Fleming, L., C. King and A. Juda (2004), "Small Worlds and Regional Innovative Advantage," Mimeo.
- Ghoshal, S., H. Korine and G. Szulanski (1994), "Interunit communication in multinational corporations," *Management Science* 40: 96-110.
- Glaeser, E.L, D. Laibson, and B. Sacerdote (2002), "The Economic Approach to Social Capital," *Economic Journal*.
- Gomes-Casseres, B., A.B. Jaffe and J. Hagedoorn (2003), "Do Alliances Promote Knowledge Flows?" Mimeo.

- Granovetter, M.S. (1973), "The Strength of Weak Ties," *American Journal of Sociology* 78: 1360-1380.
- Grant, R.M. (1996), "Toward a Knowledge-Based Theory of the Firm," *Strategic Management Journal* 17: 109-122.
- Grossman, G., and E. Helpman (1991), *Innovation and Growth in the World Economy*. MIT Press, Cambridge.
- Jaffe, A.B., M. Trajtenberg and R. Henderson (1993), "Geographic localization of knowledge spillovers as evidenced by patent citations" *Quarterly Journal of Economics* 434: 578-598.
- Jaffe, A.B. and M. Trajtenberg (2002). *Patents, Citations & Innovations: A window on the knowledge economy*. MIT Press, Cambridge.
- Hansen, M.T. (1999), "The Search-Transfer Problem: The Role of Weak Ties in Sharing Knowledge across Organization Subunits," *Administrative Science Quarterly* 44: 82-111.
- Huber, G.P. (1991), "Organizational Learning: The Contributing Processes and the Literatures," *Organization Science* 2(1): 88-115.
- King, G. and L. Zeng (2001), "Logistic Regression in Rare Events Data", *Political Analysis* 9(2): 137-163
- Kogut, B. and U. Zander (1992), "Knowledge of the Firm, Combinative Capabilities, and the Replication of Technology," *Organization Science* 3 (3): 383-397.
- Kono, C., D. Palmer, R. Friedland and M. Zafonte (1998), "Lost in Space: The Geography of Corporate Interlocking Directorates," *American Journal of Sociology* 103(4): 863-911.
- Levin, D. and R. Cross (2003), "The Strength of Weak Ties You can Trust: The Mediating Role of Trust in Effective Knowledge Transfer," *Management Science*, Forthcoming.
- Levitt, V. and J.G. March (1988), "Organizational Learning," *Annual Review of Sociology* 14: 319-340.
- Manski, C.F. and S.R. Lerman (1977), "The Estimation of Choice Probabilities from Choice Based Samples." *Econometrica* 45(8): 1977-88.
- Mowery, D.C., J.E. Oxley and B.S. Silverman (1996), "Strategic Alliances and Inter-firm Knowledge Transfer," *Strategic Management Journal* 17: 77-91.
- Nelson, R. and S. Winter (1982). *An Evolutionary Theory of Economic Change*. Harvard University Press, Cambridge.
- Newman, M.E.J. (2001), "The Structure of Scientific Collaboration Networks." *Proceedings of U.S. National Academy of Science* 98: 404-409.
- Nonaka, I. (1994), "A Dynamic Theory of Organizational Knowledge Creation," *Organization Science* 5(1): 14-37.
- Polanyi, M. (1966). *The Tacit Dimension*. Routledge & Kegan Paul, London.
- Rogers, E.M. (1995). *Diffusion of Innovations* (fourth edition). Free Press, New York.
- Rosenkopf, L. and P. Almeida (2003), "Overcoming Local Search through Alliances and Mobility." *Management Science* 49(6). 0751-0766.
- Ryan, B. and N. Gross (1943), "The diffusion of hybrid seed corn in two Iowa communities." *Rural Sociology* 8(1): 15-24.

- Saxenian, A.L. (1994). *Regional Advantage: Culture and Competition in Silicon Valley and Route 128*. Harvard University Press, Cambridge.
- Shane, S. and D. Cable (2002), "Network Ties, Reputation, and the Financing of New Ventures," *Management Science* 48 (3): 364-381.
- Simon, H.A. (1991), "Bounded Rationality and Organizational Learning," *Organization Science* 2: 125-134.
- Singh, J. (2004), "Multinational Firms and Knowledge Diffusion: Evidence using Patent Citation Data." In D.H. Nagao (Ed.), Best Paper Proceedings of the 2004 Meeting of the *Academy of Management*.
- Sorenson, O. and L. Fleming (2001), "Science and the Diffusion of Knowledge." Working paper 02-095, Harvard Business School.
- Sorenson, O., J.W. Rivkin and L. Fleming (2004), "Complexity, Networks and Knowledge Flow" Mimeo.
- Sorenson, O. and T.E. Stuart (2001), "Syndication Networks and the Spatial Distribution of Venture Capital Investments," *American Journal of Sociology* 106(6): 1546-88.
- Stolpe, M. (2001), "Mobility of Research Workers and Knowledge Diffusion as Evidenced in Patent Data The Case of Liquid Crystal Display Technology" Kiel Working Paper No. 1038.
- Stuart, T. and O. Sorenson (2003), "The Geography of Opportunity: Spatial Heterogeneity in Founding Rates and the Performance of Biotechnology firms." *Research Policy* 32: 229-253.
- Szulanski, G. (1996), "Exploring internal stickiness: Impediments to the transfer of best practice within the firm," *Strategic Management Journal* 17: 27-43.
- Thompson, P. and M. Fox-Kean (2004), "Patent Citations and the Geography of Knowledge Spillovers: A Reassessment." *American Economic Review*, forthcoming.
- Thompson, P. (2004), "Patent Citations and the Geography of Knowledge Spillovers: What do Patent Examiners Know?" Mimeo.
- Tsai, W. and S. Ghoshal (1998), "Social capital and value creation: The role of intrafirm networks," *Academy of Management Journal* 41: 464-476.
- Uzzi, B. (1996), "The sources and consequences of embeddedness for the economic performance of organizations: The network effect," *American Sociological Review* 61: 674-698.
- Uzzi, B., and R. Lancaster (2003), "Relational embeddedness and learning: The case of bank loan managers and their clients," *Management Science* 49: 383-399.
- Wasserman, S. and K. Faust (1994). *Social Network Analysis: Methods and Applications*. Cambridge University Press, UK.
- Watts, D.J. and S. Strogatz (1998), "Collective Dynamics of Small World Networks." *Nature* 393: 440-442.
- Zander, U. and B. Kogut (1995), "Knowledge and the speed of the transfer and imitation of organizational capabilities: An empirical test," *Organization Science* 6: 76-91.
- Zucker, L.G., M.R. Darby and M.B. Brewer (1998), "Intellectual Human Capital and the Birth of U.S. Biotechnology Enterprises" *American Economic Review* 88(1): 290-306.

Table 1: Definition of variables

Citation	The binary dependent variable, which equals 1 if there is actually a citation between the potentially citing and cited patents, 0 otherwise
Within same region	Indicator variable that is 1 if the citing and cited patents originate from inventors located in the same region, i.e., the same metropolitan area within the U.S.
Within same firm	Indicator variable that is 1 if the citing and cited patents are owned by the same parent firm
Same tech category	Indicator variable that is 1 if both the citing and the potentially cited patent belong to the same broad industry category (one of 6) as defined in the Jaffe and Trajtenberg (2002) database
Same tech subcategory	Indicator variable that is 1 if both the citing and the potentially cited patent belong to the same broad technical subcategory (one of 36) as defined in the Jaffe and Trajtenberg (2002) database
Same primary tech class	Indicator variable that is 1 if both the citing and the potentially cited patent belong to the same 3-digit primary technology class (one of about 450) as defined in the US Patent classification system
Same primary subclass	Indicator variable that is 1 if both the citing and the potentially cited patent belong to the same 9-digit primary technology subclass (one of about 150,000) as defined in the US Patent classification system
Secondary subclass overlap	Indicator variable that is 1 if at least one of the secondary 9-digit subclasses of one patent is the same as a primary or secondary subclass of the other patent in the dyad
Past collaboration	Indicator variable that is 1 if there is no common inventor between the two patents, but at least one inventor of the citing patent has collaborated with an inventor of the cited patent in the past. This corresponds to a social distance of 1.
Common collaborator	Indicator variable that is 1 if there is no past collaboration, but there is a common collaborator who has worked with an inventor of the citing patent and an inventor of the cited patent in the past. This corresponds to a social distance of 2.
Collaborators with ties	Indicator variable that is 1 if there is neither of the last two cases hold, but at least one former collaborator of someone from the citing team has in the past collaborated with a former collaborator for someone from the citing team. This corresponds to a social distance of 3.
Indirect social link	Indicator variable that is 1 if none of the last three cases hold, but the two patents still belong to the same connected component of the social proximity graph. This corresponds to social distance of >3 but finite.
No social link	Indicator variable that is 1 if there is no network path between the citing and the cited teams, i.e., the two are in different connected components of the social proximity graph. This corresponds to social distance of infinity.

Table 2: Intra-region and intra-firm knowledge flows

This table shows that the probability of knowledge flow is greater between two patenting teams from the same region or firm, even after accounting for technological relatedness of the citing and cited patents. It also shows that inadequate controls for technology, as used in existing literature, can bias the knowledge spillover results.

	(1)	(2)	(3)
Within same region	1.428** (0.049) [15.71]	0.886** (0.019) [9.75]	0.656** (0.030) [7.22]
Within same firm	3.277** (0.043) [36.05]	2.432** (0.021) [26.75]	1.964** (0.029) [21.60]
Technological relatedness:			
Same tech category		0.797** (0.020)	1.261** (0.020)
Same tech subcategory		1.911** (0.018)	2.403** (0.019)
Same primary tech class		4.320** (0.014)	4.822** (0.020)
Same primary subclass			6.415** (0.127)
Secondary subclass overlap			0.942** (0.059)
Number of observations	2,540,991	2,540,991	2,540,991

A weighted logit regression is used, with the dependent variable being 1 if there is a citation between two patents and 0 otherwise
 Robust standard errors in parentheses, with clustering on citing patent
 Marginal effects in square brackets after multiplication with 1,000,000
 Fixed effects for technological category, application year and time lag
 ** significant at 1%; * significant at 5%

Table 3a: Summary statistics for social distance measures

This table gives the mean value, expressed as a percentage, of each social distance variable in the sample indicated by the column corresponding to each entry.

	Entire sample		No self-citations by firms	
	Citations (N=323,820)	Controls (N=2,217,171)	Citations (N=240,724)	Controls (N=2,110,421)
Past collaboration (Social distance = 1)	7.18%	0.35%	1.05%	0.04%
Common collaborator (Social distance = 2)	4.41%	0.58%	1.04%	0.13%
Collaborators with ties (Social distance = 3)	2.82%	0.78%	1.11%	0.34%
Indirect social link (Social distance > 3 but finite)	43.88%	48.11%	50.72%	48.42%
No social link (Social distance = infinity)	41.71%	50.18%	46.08%	51.08%

Table 3b: Sample correlations of social distance measures with other variables

	Citation	Within same region	Within same firm
Past collaboration (Social distance = 1)	0.207	0.225	0.345
Common collaborator (Social distance = 2)	0.124	0.186	0.290
Collaborators with ties (Social distance = 3)	0.067	0.140	0.216
Indirect social link (Social distance > 3 but finite)	-0.028	-0.042	-0.076
No social link (Social distance = infinity)	-0.057	-0.074	-0.103

Table 4: Effect of social distance on probability of citation between patents

This table shows that the probability of knowledge flow increases as the social distance between two teams of inventors decreases, even after technological similarity of the citing and cited patents has been accounted for.

	(1)	(2)
Past collaboration (Social distance = 1)	6.801** (0.051) [74.81]	3.018** (0.078) [33.20]
Common collaborator (Social distance = 2)	4.984** (0.082) [54.82]	2.170** (0.075) [23.87]
Collaborators with ties (Social distance = 3)	3.527** (0.114) [38.80]	1.674** (0.047) [18.41]
Indirect social link (Social distance > 3 but finite)	0.115** (0.016) [1.27]	0.043** (0.015) [0.47]
Technological relatedness:		
Same tech category		1.322** (0.018)
Same tech subcategory		2.478** (0.020)
Same primary tech class		5.028** (0.020)
Same primary subclass		6.608** (0.138)
Secondary subclass overlap		0.938** (0.063)
Number of observations	2,540,991	2,540,991

A weighted logit regression is used, with the dependent variable being 1 if there is a citation between two patents and 0 otherwise
 Robust standard errors in parentheses, with clustering on citing patent
 Marginal effects in square brackets after multiplication with 1,000,000
 Fixed effects for technological category, application year and time lag
 ** significant at 1%; * significant at 5%

Table 5: Does social distance help explain intra-regional and intra-firm knowledge flows?

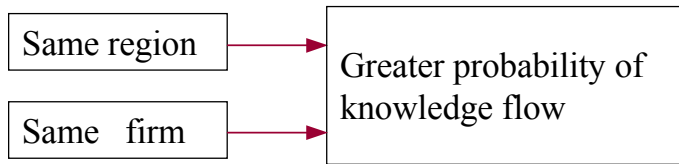
This table studies if interpersonal ties help explain the greater probability of knowledge flow between two patenting teams from the same region or firm. Column (1) reproduces the results from column (3) of Table 2. Column (2) shows that accounting for social distance reduces the *within same region* and *within same firm* estimates for probability of patent citation. Column (3) shows that there are important interaction effects.

	(1)	(2)	(3)
Within same region	0.656** (0.030) [7.22]	0.544** (0.035) [5.98]	0.868** (0.037) [9.55]
Within same firm	1.964** (0.029) [21.60]	1.726** (0.027) [18.99]	2.034** (0.038) [22.37]
Past collaboration (Social distance = 1)		1.340** (0.079)	3.364** (0.121)
Common collaborator (Social distance = 2)		0.561** (0.078)	2.447** (0.069)
Collaborators with ties (Social distance = 3)		0.376** (0.048)	1.495** (0.084)
Indirect social link (Social distance > 3 but finite)		0.010 (0.015)	0.095** (0.016)
Within same region * Past collaboration			-0.697** (0.171)
Within same region * Common collaborator			-1.220** (0.152)
Within same region * Collaborators with ties			-0.930** (0.092)
Within same region * Indirect social link			-0.183** (0.046)
Within same firm * Past collaboration			-2.141** (0.166)
Within same firm * Common collaborator			-1.743** (0.097)
Within same firm * Collaborators with ties			-1.207** (0.102)
Within same firm * Indirect social link			-0.452** (0.045)
Technological relatedness	Y	Y	Y
Number of observations	2,540,991	2,540,991	2,540,991

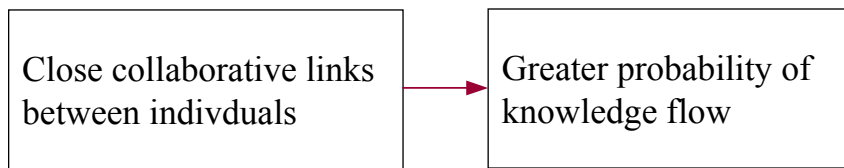
A weighted logit regression is used, with the dependent variable being 1 if there is a citation between two patents and 0 otherwise
 Robust standard errors in parentheses, with clustering on citing patent
 Marginal effects in square brackets after multiplication with 1,000,000
 Fixed effects for technological category, application year and time lag between patents
 ** significant at 1%; * significant at 5%

Figure 1: Summary of hypotheses

(a) Hypotheses 1 and 2:



(b) Hypotheses 3 and 4:



(c) Hypotheses 5 and 6:

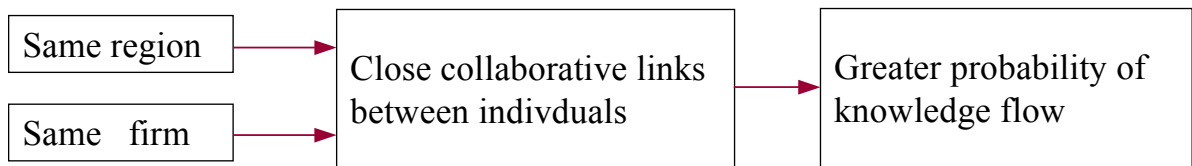


Figure 2: An affiliation network

Inventor	Innovating Team (Patent)						
	P1	P2	P3	P4	P5	P6	P7
A	1	1	0	0	0	0	0
B	1	0	0	1	0	0	0
C	0	1	1	0	0	0	0
D	0	0	1	0	1	0	0
E	0	0	0	0	1	0	1
F	0	0	0	0	1	0	0
G	0	0	0	0	0	1	1

Year 1986 1987 1988 1989 1989 1989 1990

Figure 3: Social proximity graphs

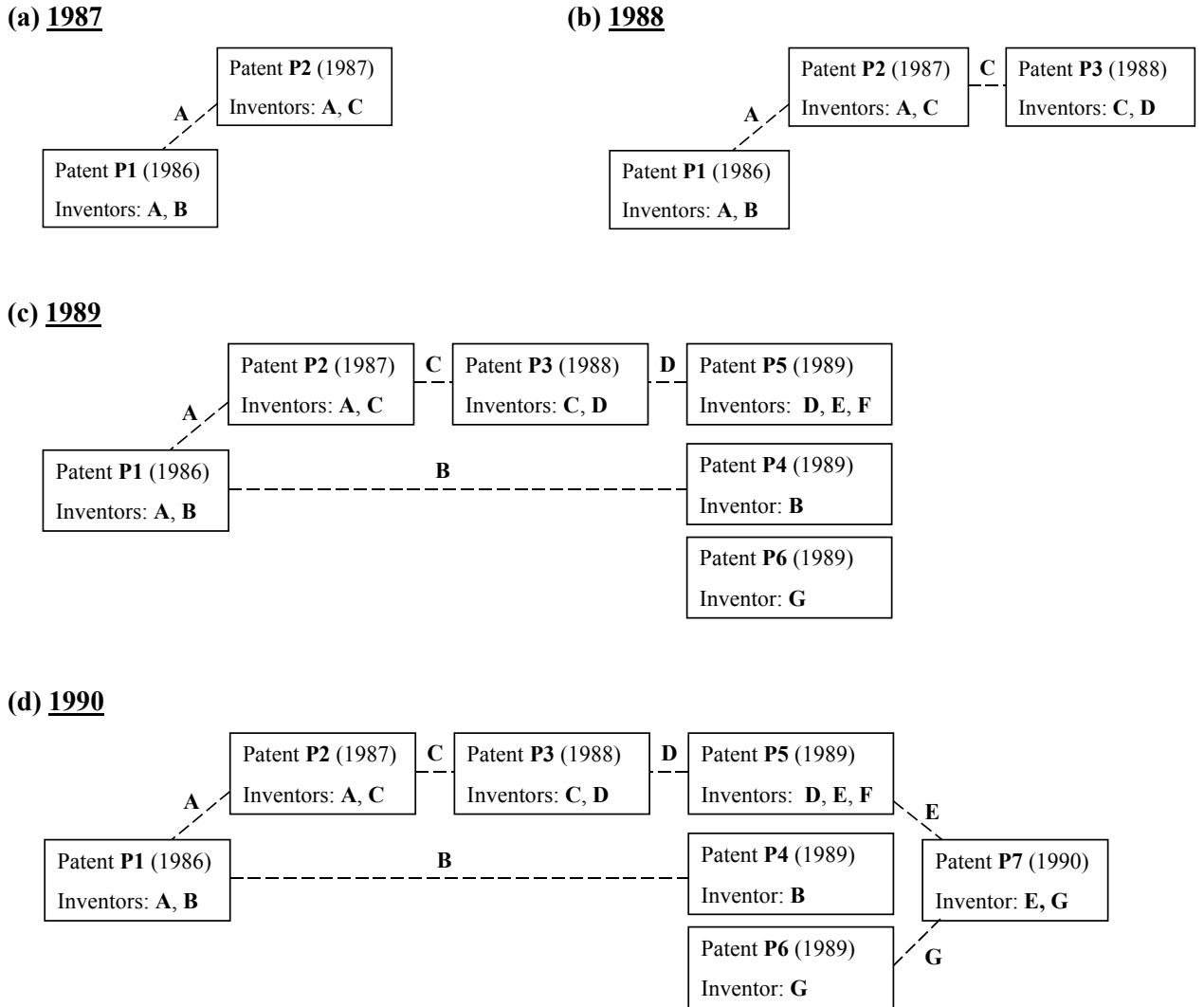


Figure 4: Social distance between two nodes

Since knowledge flows only make sense from an innovation that happens earlier to one that happens later, *social distance* is left undefined for $P2 \rightarrow P1$, $P3 \rightarrow P1$, $P1 \rightarrow P1$, $P2 \rightarrow P2$, etc.

		Destination						
		P1	P2	P3	P4	P5	P6	P7
Source	P1	-	0	1	0	2	∞	3
	P2	-	-	0	1	1	∞	2
	P3	-	-	-	2	0	∞	1
	P4	-	-	-	-	3	∞	4
	P5	-	-	-	3	-	∞	0
	P6	-	-	-	∞	∞	-	0
	P7	-	-	-	-	-	-	-