Collaborative Networks as Determinants of Knowledge Diffusion Patterns

Jasjit Singh

INSEAD, 1 Ayer Rajah Avenue, 138676 Singapore {jasjit.singh@insead.edu, http://www.jasjitsingh.com}

This paper examines whether interpersonal networks help explain two widely documented patterns of knowledge diffusion: (1) geographic localization of knowledge flows, and (2) concentration 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, intraregional and intrafirm knowledge flows are found to be stronger than those across regional or firm boundaries. I explore whether these patterns can be explained by direct and indirect network ties among inventors, as inferred from past collaborations among 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. Furthermore, 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 among inventors who 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.

Key words: knowledge spillovers; technology diffusion; social networks; interpersonal ties; innovation
History: Accepted by Scott Shane, technological innovation, product development, and entrepreneurship; received May 4, 2004. This paper was with the author 2 months for 1 revision.

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 et al. 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. Although 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 close interpersonal ties. Although 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 microlevel knowledge flows, a richer data set 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 than between regions or firms. In particular, even after carefully controlling for technological specialization of regions (Thompson and Fox-Kean 2005), knowledge flow between two inventors from the same region is found to be 66% more likely than that between two inventors from different regions. Likewise, all else being equal, knowledge flow between two inventors is three times as likely within than between firms.

A rich literature in sociology has emphasized information flow through interpersonal networks (Ryan and Gross 1943, Coleman et al. 1966, Granovetter 1973, Burt 1992, Rogers 1995). This has motivated the current study investigating whether such interpersonal networks are behind the observed patterns of knowledge diffusion mentioned above. Whereas it can be difficult to obtain large-scale systematic data

on interpersonal relations, we can infer such relations indirectly by studying secondary data on collaborations between individuals (Cockburn and Henderson 1998, Newman 2001, Fleming et al. 2004). Following this approach, I use collaboration information for patents registered with the United States Patent and Trademark 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 1 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 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, for inventors with direct collaborative ties the probability of knowledge flow is four times that for inventors who are not connected, and it is 3.2 times that for inventors who have no direct tie, but who have an indirect tie through past collaboration.

The main analysis of the paper studies whether 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. Although the decrease is nontrivial 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 colocation 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, because 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. 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 who are already closely connected in the collaboration network. The regional and firm boundaries continue to matter more for knowledge flow between inventors with no ties or with only very indirect ties. As explained in §5, this is consistent with a view that interpersonal networks are actually quite important in causing the more intense intraregional and intrafirm knowledge flows, and that the detected effects are

of relatively small magnitude simply because patent collaborations capture only a fraction of all relevant interpersonal ties.

This paper is organized as follows. Section 2 motivates my formal hypotheses. Section 3 describes the data on patent citations and 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.
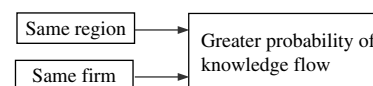
2. Hypotheses

The analysis in this paper comprises three main parts, as summarized in Figure 1. The first part formally establishes that intraregional and intrafirm knowledge flows are indeed more intense than the knowledge flows between different regions or firms. The second part tests 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, combines the constructs from the first two parts to examine the extent to which interpersonal networks help explain the intense intraregional and intrafirm knowledge flows.

Although previous work has documented geographic localization of knowledge flow (e.g., Jaffe et al. 1993), recent work raises methodological concerns that could have led to overestimation of this phenomenon (Thompson and Fox-Kean 2005). Therefore, before trying to explain the result on intraregional knowledge flows, I first test if the result does indeed continue to hold even when using a new empirical approach (as detailed in §4) that addresses some of these concerns:

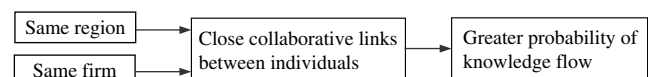**Figure 1   Summary of Hypotheses**

(a) Hypotheses 1 and 2

(b) Hypotheses 3 and 4

(c) Hypotheses 5 and 6

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 demonstrates 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 phenomenon, I formally reproduce this result 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 et al. 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 who 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 et al. 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 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 induce more trust, improving willingness of individuals to share knowledge (Tsai and Ghoshal 1998, Levin and Cross 2004). 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 et al. (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 colocation 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 region-specific factors such as 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 intraregion knowledge flows could simply be that patent collaborations reveal only a subset of actual interpersonal networks.

Analogous to studying why intraregional knowledge flows are strong is the question why knowledge flows are stronger within than between firms. I follow Simon (1991) and Grant (1996) in taking individuals as the unit of analysis for studying knowledge flows

---

[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.

[2] Granovetter (1973) and Burt (1992) suggest that nonredundancy 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 or nonredundancy of information flow. Therefore, Hypothesis 4 should not be interpreted as a direct test of Granovetter and Burt's theories. This issue is studied by Hansen (1999), who shows that weak and nonredundant ties are better when searching for simple information, whereas strong ties are better for transfer of complex knowledge.

even within organizations. This allows me to use a unified network framework to study both interfirm and intrafirm 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 intrafirm 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 intrafirm 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 be that the role of patent citations in determining patent scope and litigation behavior makes incentives for intrafirm citations very different from those for interfirm 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 on existing knowledge. Unlike in academic papers, there is an incentive not to include superfluous patent citations because that might reduce the scope of an inventor's own patent. An inventor is legally bound to report relevant prior art, with the patent examiner performing an objective check, however. 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, although it is not perfect (Jaffe and Trajtenberg 2005, Duguet and MacGarvie 2005).

I merged patent data from the USPTO with that from Jaffe and Trajtenberg (2002). The data set had to be further enhanced to correctly identify each patent's owner, because some of a firm's patents are often listed under the name of a subsidiary instead of the parent firm. I performed a 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, because 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 United States, 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] Because I do not have systematic fine-grained geographic information for innovations arising outside the United States, all patents with a non-U.S. address were dropped. Even within the United States, 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 the state as the unit of analysis. The main results were very similar and are therefore not reported in the paper.

3.2. Inventors

Fleming et al. (2003, 2004) present evidence from field interviews that collaborations recorded on patent documents meaningfully (though not perfectly) capture personal and professional interpersonal ties between inventors. Because patents are nontrivial innovations by definition, coinventors 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

---

[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 to other organizations. Nonfirm entities (such as 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).

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 also by looking for an overlap of citations across patents. However, imposing either of these two conditions caused too many false negatives, because the overlap in either case was quite low even for records from the same inventor. I also considered requiring a match for street address or assignee firm, or both, as used by Fleming et al. (2003). However, I decided against it because 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. (2003) report, forcing these requirements makes the match too conservative, an issue they handle by not requiring the rule for relatively uncommon last names. Irrespective of the algorithm used, there would 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 $\Pr(K, k)$, that 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

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

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 because citations between random patents are very rare: There are only a few realized citations for every 1 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. Because 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 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. Furthermore, because the same citing patent can occur in multiple observations, the standard errors should be calculated without assuming independence across these observations.[5]

---

[5] An online appendix accompanying this paper (available at http://mansci.pubs.informs.org/ecompanion.html) 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 (2004) for an earlier application of similar methodology for predicting patent citations.

4.2. Sample Construction

The basic WESML approach samples all $y = 0$ observations with equal probability α. Because 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 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 such class pair. 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 (see footnote 5) 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 1 because 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, because 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). Furthermore, because 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. Although 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 studies.[7] I also account for other factors that 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

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 and Lissoni 2002, Agrawal et al. 2003). This assumption is also shown to hold for patent collaboration in field evidence reported by Fleming et al. (2003), and is therefore followed here. In inferring network ties that exist as of any year t (t being any year between 1986 and 1995), I include all inventors that have patented between 1975 and t (including even those not in the United States, 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 $A = \{a_{ij}\}$, where a_{ij} is 1 if the ith inventor is on the collaborating team for the jth patent, and 0 otherwise (Wasserman and Faust 1994). Figure 2 gives an example, with seven inventors A, B, C, D, E, F, and G, and seven 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

---

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

[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 of colocation of patents.

[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, because measured knowledge flows are from one patenting team to another.

Figure 2   An Affiliation Network

	Innovating team (patent)						
Inventor	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 4   Social Distance Between Two Nodes

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

Note. Because knowledge flows only make sense from an innovation that happens earlier to one that happens later, social distance is left undefined for P2 → P1, P3 → P1, P1 → P1, P2 → P2, etc.

two teams have a common inventor. If there is a citation between two patents sharing a common inventor, such as patents P1 and P2 in Figure 3(a), it can be seen as an inventor citing his or her own previous work. Because self-citations by individuals do not represent real knowledge flow, the actual sample used in the regression analysis in §5 does not include such pairs of patents.

The really 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 → P2 → P3 by being passed from A to C,

because A and C have 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 → P3 example above, as indicated in Figure 4. Because knowledge flows are meaningful only from an innovation that happens earlier to one that happens later, social distance need not be defined for P2 → P1, P1 → P1, etc.

Now consider Figure 3(c). The above definition suggests a social distance of 1 for P2 → P4, because there
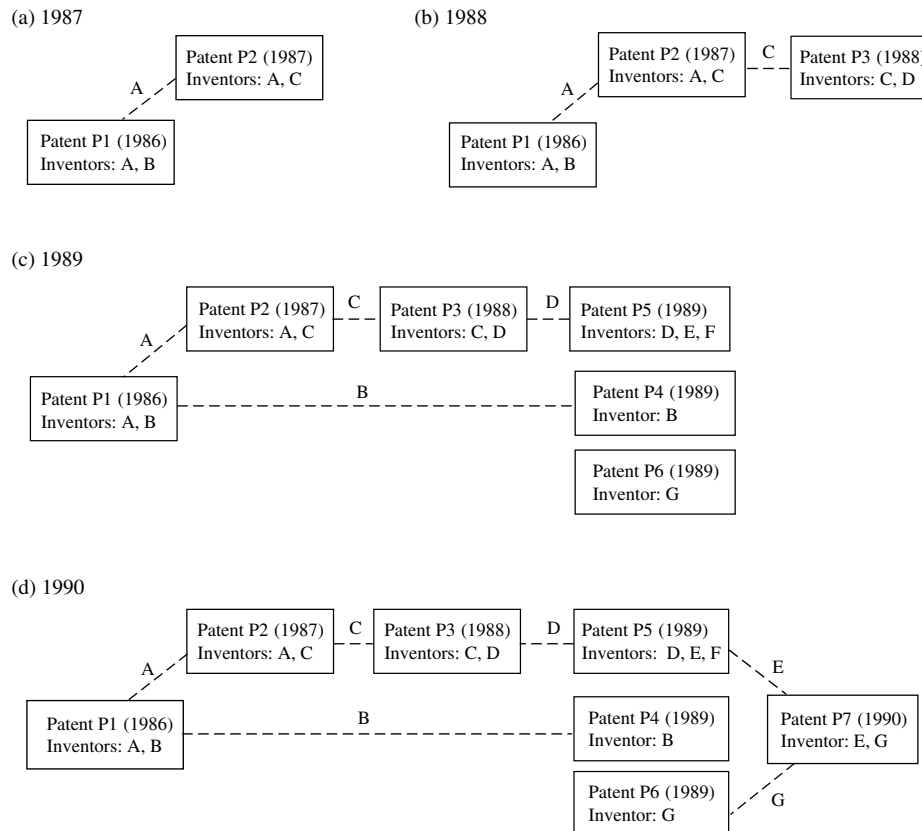
Figure 3   Social Proximity Graphs

(a) 1987

(b) 1988

(c) 1989

(d) 1990

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 United States.
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 (1 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 (1 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 USPTO 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 (1 of about 150,000) as defined in the USPTO 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 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.

is a path P2 → P1 → 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 on knowledge of P2 that B can gain through his or her ongoing tie with A. Thus, knowledge can flow "backward in time" along the link P1 → P2, and then on to P2 → P4. Likewise, knowledge from P3 can be passed by C to A, and then on to B through the chain P3 → P2 → P1 → P4, making the social distance P3 → P4 equal to 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 (because 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 pairwise social distances is practically impossible. Fortunately, it is still practical to classify all observations with a nonzero 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. Intraregional and Intrafirm Knowledge Flows
Table 2 formally tests Hypotheses 1 and 2, i.e., that knowledge flows are particularly strong within

---

[9] Because the social distance measure might not be comparable across years, I use year fixed effects in the regressions described in §5. 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 pairwise 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 pth power matrix of X has a nonzero entry (i, j). Unfortunately, such an approach is impractical for very large graphs (Cormen et al. 1990). Instead, I explicitly find all pairs with a social distance of 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.

Table 2   Intraregion and Intrafirm Knowledge Flows

	(1)	(2)	(3)
Within same region	1.428**	0.886**	0.656**
	(0.049)	(0.019)	(0.030)
	[15.71]	[9.75]	[7.22]
Within same firm	3.277**	2.432**	1.964**
	(0.043)	(0.021)	(0.029)
	[36.05]	[26.75]	[21.60]
Technological relatedness			
Same tech category		0.797**	1.261**
		(0.020)	(0.020)
Same tech subcategory		1.911**	2.403**
		(0.018)	(0.019)
Same primary tech class		4.320**	4.822**
		(0.014)	(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

Notes. 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.

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 are in parentheses, with clustering on citing patent. Marginal effects are in square brackets after multiplication with 1,000,000. Fixed effects were included for technological category, application year, and time lag, but are not reported here.

*Significant at 5%; **significant at 1%.

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 et al. 1993). As Column (2) shows, including controls for technological relatedness (at the level of a 3-digit technological class) between patents reduces but does not eliminate the estimated coefficients for *within same region* and *within same firm.* However, Thompson and Fox-Kean (2005) 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. Because 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, because 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 patents from different regions that are otherwise similar. 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 is consistent with 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 because 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 1. Table 3(a) 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 Hypothesis 3 that connectedness leads to greater probability of citation. The inequality holds even for the subsample without self-citations by firms, where the fraction of pairs with no social link is only 46.1% for pairs with citations, and 51.1% for pairs with no citation. Table 3(b) 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. Because these are just raw correlations from a choice-based sample, the interpretation of these numbers is limited. Nevertheless, a fact that emerges 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 because teams with collaborative links are also more likely to be technologically related. Therefore,

---

[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 to get an estimate of the marginal effect.

Table 3(a)   Summary Statistics for Social Distance Measures

	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

Note. 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.

Table 3(b)   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

Column (2) is the preferred regression specification. Consistent with Hypothesis 3, interpersonal ties are found to be important, as indicated by the estimates for *past collaboration*, *common collaborator*, *collaborators with ties*, and *indirect social link* being 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 3,372.0. The reference group for comparison in these regressions is patent pairs with no social link.

Although 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 intraregion and intrafirm 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 colocation on probability of citation falls from 65.6% to 54.4%.[12] Likewise, the coefficient estimate for *within same firm*

---

[12] Normally, in nonlinear models, one should only compare marginal effects and not coefficient estimates across models.

Table 4   Effect of Social Distance on Probability of Citation Between Patents

	(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

Notes. 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.

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 are in parentheses, with clustering on citing patent. Marginal effects are in square brackets after multiplication with 1,000,000. Fixed effects were included for technological category, application year, and time lag but are not reported here.

*Significant at 5%; **significant at 1%.

Table 5   Does Social Distance Help Explain Intraregional and Intrafirm Knowledge Flows?

	(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)
Number of observations	2,540,991	2,540,991	2,540,991

Notes. 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.

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 are in parentheses, with clustering on citing patent. Marginal effects are in square brackets after multiplication with 1,000,000. Fixed effects were included for technological category, application year, technological relatedness, and time lag between patents but are not reported here.

*Significant at 5%; **significant at 1%.

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 intrafirm knowledge flows. Not only is the decrease nontrivial 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

---

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*.

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 words, there might be interaction effects between social

distance and geographic colocation, and between social distance and firm boundaries, in determining probability of knowledge flow. Because 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 with the direct effect for *within same region.* In other words, conditional on the social distance being small (i.e., 1, 2, or 3), geographical colocation 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 of 1. Statistically, this is in fact indistinguishable from having no unexplained effect of colocation once the standard errors have been taken into account. For patents that are connected only through indirect social links or are not connected at all, however, geographic colocation 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 colocated versus those that are not. One possible explanation is 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 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 0 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 possibly be a result of interpersonal ties not captured in patent collaboration data.

6. Limitations and Future Research

This paper takes network ties as exogenous. However, interpersonal ties could arise endogenously as a result of deliberate steps taken by rational actors (Coleman 1988, Glaeser et al. 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 inventor X who has cited inventor Y is also more likely to cite Y's work in the future while also trying 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.

Whereas 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 intrafirm 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* variable only as a control, and to interpret the results as being 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 because USPTO has started making the distinction between citations by patent examiners versus citations by inventors only since 2001, and even those data are not easily available in a machine-readable form.

Another issue is that patent collaborations capture only a subset of relevant interpersonal relations. An extension could 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 capture more nuanced network structural effects such as 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 et al. 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 intraregional and intrafirm knowledge flow, even though their full impact might be hard to measure because 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 et al. (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 colocation in Italy to be greater for socially connected patent teams than others, suggesting important interaction effects between geographic colocation and collaborative links. The focus of this paper has been to enrich this stream of research through a rigorous measurement of the extent to which collaborative networks help explain observed patterns of intraregional and intrafirm knowledge flow. In the process, I also introduce an improved methodology for studying microlevel knowledge flows, a much more exhaustive data set 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. Furthermore, geography matters for knowledge diffusion, 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. Furthermore, 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 colocation shows that, for inventors connected via short path lengths, geographic colocation 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, an issue explored further by Singh (2005).

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, because 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 to both these puzzles 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 intrafirm 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. 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 et al. (1996), Rosenkopf and Almeida (2003), and Gomes-Casseres et al. (2005).

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 to ensure 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 also ultimately the capacity of regions to innovate (Fleming et al. 2004). Although 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 noncompete agreements, and through subsidies for joint research and development (R&D) projects and joint regional conferences.

The finding that collaborative networks can help overcome geographic distance is particularly important for developing 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 toward external learning by directly tapping into foreign collaborative networks. 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 been seen as erosion of local technical base. Instead, governments might even consider setting up incentives and mechanisms for their well-trained emigrants to continue to maintain close interpersonal ties with the country's citizens, and encourage local firms to use foreign subsidiaries to access foreign knowledge by tapping into foreign interpersonal networks.

An online appendix to this paper is available at http://mansci.pubs.informs.org/ecompanion.html.

Acknowledgments

References

Agrawal, A., I. Cockburn, J. McHale. 2003. Gone but not forgotten: Labor flows, knowledge spillovers, and enduring social capital. National Bureau of Economic Research working paper 9950, Cambridge, MA.

Ahuja, G. 2000. Collaboration networks, structural holes, and innovation: A longitudinal study. *Admin. Sci. Quart.* **45** 425–455.

Alcacer, J., M. Gittelman. 2004. How do I know what you know? Patent examiners and the generation of patent citations. Mimeo, New York University, New York.

Allen, T. J. 1977. *Managing the Flow of Technology.* MIT Press, Cambridge, MA.

Almeida, P., B. Kogut. 1999. The localization of knowledge and the mobility of engineers in regional networks. *Management Sci.* **45**(7) 905–917.

Amemiya, T. 1985. *Advanced Econometrics*. Harvard University Press, Cambridge, MA.

Breschi, S., F. Lissoni. 2002. Mobility and social networks: Localised knowledge spillovers revisited. Paper presented at workshop on *The Role of Labour Mobility and Informal Networks for Knowledge Transfer*. Max Plank Institute, Germany.

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

Cockburn, I. M., R. M. Henderson. 1998. Absorptive capacity, coauthoring behavior, and the organization of research in drug discovery. *J. Indust. Econom.* **46**(2) 157–182.

Coleman, J. S. 1988. Social capital in the creation of human capital. *Amer. J. Sociology* **94**(Suppl.) S95–S120.

Coleman, J. S., E. Katz, H. Menzel. 1966. *Medical Innovation*. Bobbs-Merrill, New York.

Cormen, T. H., C. E. Leiserson, R. L. Rivest. 1990. *Introduction to Algorithms.* MIT Press, Cambridge, MA.

Duguet, E., M. MacGarvie. 2005. How well do patent citations measure knowledge spillovers? Evidence from French innovative surveys. *Econom. Innovation New Tech.* **14**(5).

Fleming, L., C. King, A. Juda. 2004. Small worlds and regional innovative advantage. Working paper 04-008, Harvard Business School, Boston, MA.

Fleming, L., L. Colfer, A. Marin, J. McPhie. 2003. Why the valley went first: Agglomeration and emergence in regional inventor networks. Mimeo, Harvard Business School, Boston, MA.

Ghoshal, S., H. Korine, G. Szulanski. 1994. Interunit communication in multinational corporations. *Management Sci.* **40** 96–110.

Glaeser, E. L., D. Laibson, B. Sacerdote. 2002. The economic approach to social capital. *Econom. J.* **112** F437–F458.

Gomes-Casseres, B., A. B. Jaffe, J. Hagedoorn. 2003. Do alliances promote knowledge flows? *J. Financial Econom.* Forthcoming.

Granovetter, M. S. 1973. The strength of weak ties. *Amer. J. Sociology* **78** 1360–1380.

Grant, R. M. 1996. Toward a knowledge-based theory of the firm. *Strategic Management J.* **17** 109–122.

Grossman, G., E. Helpman. 1991. *Innovation and Growth in the World Economy*. MIT Press, Cambridge, MA.

Jaffe, A. B., M. Trajtenberg. 2002. *Patents, Citations and Innovations: A Window on the Knowledge Economy*. MIT Press, Cambridge, MA.

Jaffe, A. B., M. Trajtenberg, R. Henderson. 1993. Geographic localization of knowledge spillovers as evidenced by patent citations. *Quart. J. Econom.* **108** 578–598.

Hansen, M. T. 1999. The search-transfer problem: The role of weak ties in sharing knowledge across organization subunits. *Admin. Sci. Quart.* **44** 82–111.

Huber, G. P. 1991. Organizational learning: The contributing processes and the literatures. *Organ. Sci.* **2**(1) 88–115.

King, G., L. Zeng. 2001. Logistic regression in rare events data. *Political Anal.* **9**(2) 137–163.

Kogut, B., U. Zander. 1992. Knowledge of the firm, combinative capabilities, and the replication of technology. *Organ. Sci.* **3**(3) 383–397.

Kono, C., D. Palmer, R. Friedland, M. Zafonte. 1998. Lost in space: The geography of corporate interlocking directorates. *Amer. J. Sociology* **103**(4) 863–911.

Levin, D., R. Cross. 2004. The strength of weak ties you can trust: The mediating role of trust in effective knowledge transfer. *Management Sci.* **50** 1477–1490.

Levitt, V., J. G. March. 1988. Organizational learning. *Ann. Rev. Sociology* **14** 319–340.

Manski, C. F., S. R. Lerman. 1977. The estimation of choice probabilities from choice based samples. *Econometrica* **45**(8) 1977–1988.

Mowery, D. C., J. E. Oxley, B. S. Silverman. 1996. Strategic alliances and inter-firm knowledge transfer. *Strategic Management J.* **17** 77–91.

Nelson, R., S. Winter. 1982. *An Evolutionary Theory of Economic Change*. Harvard University Press, Cambridge, MA.

Newman, M. E. J. 2001. The structure of scientific collaboration networks. *Proc. National Acad. Sci.* **98** 404–409.

Nonaka, I. 1994. A dynamic theory of organizational knowledge creation. *Organ. Sci.* **5**(1) 14–37.

Polanyi, M. 1966. *The Tacit Dimension*. Routledge & Kegan Paul, London, UK.

Rogers, E. M. 1995. *Diffusion of Innovations*, 4th ed. Free Press, New York.

Rosenkopf, L., P. Almeida. 2003. Overcoming local search through alliances and mobility. *Management Sci.* **49**(6) 751–766.

Ryan, B., 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, MA.

Shane, S., D. Cable. 2002. Network ties, reputation, and the financing of new ventures. *Management Sci.* **48**(3) 364–381.

Simon, H. A. 1991. Bounded rationality and organizational learning. *Organ. Sci.* **2** 125–134.

Singh, J. 2004. Multinational firms and knowledge diffusion: Evidence using patent citation data. D. H. Nagao, ed. *Proc. Sixty-Third Annual Meeting* (*New Orleans*) *Acad. Management* (*CD*). Academy of Management, 1543–8643.

Singh, J. 2005. Distributed R&D, cross-regional ties and quality of innovative output. Mimeo, INSEAD, Singapore.

Sorenson, O., L. Fleming. 2004. Science and the diffusion of knowledge. *Res. Policy* **33** 1615–1634.

Sorenson, O., T. E. Stuart. 2001. Syndication networks and the spatial distribution of venture capital investments. *Amer. J. Sociology* **106**(6) 1546–1588.

Sorenson, O., J. W. Rivkin, L. Fleming. 2004. Complexity, networks and knowledge flow. Mimeo, University of California, Los Angeles, CA.

Stolpe, M. 2001. Mobility of research workers and knowledge diffusion as evidenced in patent data: The case of liquid crystal display technology. Working paper 1038, Kiel Institute for World Economics, Germany.

Stuart, T., O. Sorenson. 2003. The geography of opportunity: Spatial heterogeneity in founding rates and the performance of biotechnology firms. *Res. Policy* **32** 229–253.

Szulanski, G. 1996. Exploring internal stickiness: Impediments to the transfer of best practice within the firm. *Strategic Management J.* **17** 27–43.

Thompson, P. 2004. Patent citations and the geography of knowledge spillovers: What do patent examiners know? Mimeo, Florida International University, Miami, FL.

Thompson, P., M. Fox-Kean. 2005. Patent citations and the geography of knowledge spillovers: A reassessment. *Amer. Econom. Rev.* Forthcoming.

Tsai, W., S. Ghoshal. 1998. Social capital and value creation: The role of intrafirm networks. *Acad. Management J.* **41** 464–476.

Uzzi, B. 1996. The sources and consequences of embeddedness for the economic performance of organizations: The network effect. *Amer. Sociological Rev.* **61** 674–698.

Uzzi, B., R. Lancaster. 2003. Relational embeddedness and learning: The case of bank loan managers and their clients. *Management Sci.* **49** 383–399.

Wasserman, S., K. Faust. 1994. *Social Network Analysis: Methods and Applications.* Cambridge University Press, Cambridge, UK.

Watts, D. J., S. Strogatz. 1998. Collective dynamics of small world networks. *Nature* **393** 440–442.

Zander, U., B. Kogut. 1995. Knowledge and the speed of the transfer and imitation of organizational capabilities: An empirical test. *Organ. Sci.* **6** 76–91.

Zucker, L. G., M. R. Darby, M. B. Brewer. 1998. Intellectual human capital and the birth of U.S. biotechnology enterprises. *Amer. Econom. Rev.* **88**(1) 290–306.