

OLAV SORENSON

SOCIAL NETWORKS, INFORMATIONAL COMPLEXITY AND INDUSTRIAL GEOGRAPHY[□]

1. INTRODUCTION

Why do some industries reside in a limited number of highly clustered geographic locations while others spread across the landscape more broadly? Traditional explanations for variation in the spatial distribution of industrial production have focused on examining differences in the transportation costs associated with obtaining important inputs and with distributing finished goods to consumers. These transportation cost based arguments fail, however, to account for the concentration of a variety of light manufacturing and service industries, such as high technology or entertainment, where these costs make up a negligible fraction of the value of the good. Attempts to explain the clustering of these industries has led researchers to revisit the agglomeration economies proposed by Alfred Marshall (1920). Thus, recent work has elucidated the potential benefits of an extended division of labor (Romer, 1990), labor pooling (Diamond and Simon, 1990; Rotemberg and Saloner, 1990) and information spillovers (Arrow, 1962).

Although these factors undoubtedly play a role in the maintenance of many industrial districts, geographic concentration can persist even when economic efficiency (at least in production) does not support it. The explanation for this phenomenon comes from a more nuanced consideration of the process of entrepreneurship – specifically, the importance of social networks to it. Two factors must converge for a nascent entrepreneur to found a new firm. First, the potential entrepreneur must perceive an opportunity for profit in a particular segment, or market niche, of the economy. Since much of the relevant information only exists privately, awareness of potentially profitable opportunities requires connections to those with the pertinent knowledge, typically those currently engaged in business in a particular industry. Second, the individual that perceives an opportunity must build a firm – assemble the necessary capital, skilled labor and knowledge – to exploit it. Again, social relationships play a crucial role in acquiring tacit information and in convincing resource holders to join the fledgling venture, whether as employees or investors. Because the social ties that facilitate both of these antecedents rarely extend beyond the regions in which these relevant resources and knowledge reside, entrepreneurs within a given industry most frequently arise in close proximity to industry incumbents. This regularity implies that industries can remain geographically concentrated even when co-location disadvantages firms.

Though recent studies provide support for this social network based explanation (e.g., Sorenson and Audia, 2000; Klepper, 2001; Stuart and Sorenson, 2003), it suffers from a critical shortcoming: it predicts that all industries should cluster into industrial districts. Consistent with the gradual diffusion of social networks over time, some of the variation in the degree of geographic concentration stems from differences in the maturity (or age) of industries. Nevertheless, even among established industries, sectors vary in the degree to which they cluster into a small number of productive regions. Why do some industries continue to agglomerate while others spread?

This paper seeks to address this issue by answering the more focused question: when do social networks play a vital role in structuring industrial geography? This extension focuses on the interaction of social networks with the characteristics of the knowledge flowing through them – in particular, its degree of complexity. I argue that social networks become increasingly important for the transmission of knowledge as the complexity of the underlying knowledge increases. This expectation emerges from recognizing that the assimilation of information may require the receiver to engage in a process of search to fill in gaps, or correct transmission errors, in the knowledge received – a difficult task when dealing with complex knowledge. Dense social networks can diminish the need for search by facilitating high fidelity knowledge transfer: complete information with negligible noise. On the other hand, when relatively sparse social networks – such as one might find spanning the borders of salient social groups – connect the receiver to the sender, search plays an increasingly important role in transmission. Under such conditions, simple knowledge flows easily both within and across social boundaries because search can easily substitute for imperfect transmission (Rivkin, 2000). These dynamics lead us to expect that industries based on more complex knowledge will concentrate geographically to a greater degree.

This paper engages in two types of analysis to corroborate this hypothesis. First, it investigates patent data to confirm the micro-dynamics of knowledge flows. Citation patterns across patents offer something like a fossil record of the flow of knowledge. To assess our argument, we estimate the effect of knowledge complexity on the geographic dispersion of future citations using a case-control logistic analysis. Consistent with the thesis proposed, the results show that knowledge complexity plays an important role in limiting the rate at which knowledge diffuses across geographic boundaries. Next, the paper investigates the industry-level correlation between the knowledge complexity in an economic sector and its degree of geographic concentration. Again, the results show a significant relationship between the average level of informational complexity embodied in the patents associated with an industry and its spatial concentration in employment.

2. SOCIAL NETWORKS AND INDUSTRIAL GEOGRAPHY

Social networks affect the degree of geographic industrial concentration because they do not connect individuals at random. People interact most commonly with others living in the same geographic regions and with whom they share back-

grounds, interests and affiliations. These spatial patterns arise from the likelihood of having an opportunity to form a tie. To begin a relation, two individuals usually must meet one another. Because geography, interests and affiliations strongly influence daily activities, similarity on these dimensions increases the probability that two individuals will meet by chance (Festinger, Schacter and Back, 1950; Blau, 1977; Sorenson and Stuart, 2001). Even after a contact has been made, these factors continue to affect the probability of developing and maintaining a tie. Empirical research consistently reveals that individuals appear to prefer to pursue and sustain social contacts with those from similar backgrounds and with related interests (Lazarsfeld and Merton, 1954; for a review of recent research, see McPherson, Smith-Lovin and Cook, 2001). Geographic proximity also has a strong influence on the longevity of ties, as it decreases the cost of continuing a relationship. The frequent and intense interaction required to maintain a close tie can incur substantial direct costs, especially when considerable distance separates the two parties (Zipf, 1949). The opportunity costs of a tie also increase with distance as the number of equally preferred but more proximate individuals rises (Stouffer, 1940).¹ Therefore, social networks primarily connect like individuals that live in close proximity to each other.

These social networks play an important role in structuring both who engages in entrepreneurship and what regions and industries they enter. By shaping both the awareness of opportunities and the ability to capitalize on those prospects, these networks tend to reify the existing distribution of industry (Sorenson and Audia, 2000). Consider the first stage in the entrepreneurial process: identifying an opportunity. Evaluating market potential often requires access to private information. Incumbents prefer to conceal, for instance, their positioning and profitability to prevent others from entering munificent market niches. Regardless, many individuals do have access to this valuable data, notably employees of incumbent firms as well as their contacts – a group that, given the local structure of social networks, likely includes others in the same industry and in the same local communities in which existing firms operate.

Once an opportunity has been identified, social networks also constrain where individuals can successfully build new firms. Even in an emerging industry, firms require capital and labor. As industries mature, firm efficiency rises as a result of investments in physical and human capital, and through the accretion of valuable knowledge. Fledgling firms need access to each of these three elements – (1) financial capital, (2) human capital, and (3) knowledge capital – to compete effectively against incumbents. Social networks facilitate access to each of these resources.

In the absence of social networks, nascent entrepreneurs may find it difficult to garner sufficient financial and human capital. All new firms face substantial and fundamental uncertainty – not only does the venture entail substantial risk, but actors even find it difficult to assess the level of risk involved. Both potential backers and recruits likely approach opportunities to join new firms with substantial suspicion. Moreover, those considering an investment in the new venture face a potential information asymmetry problem: The nascent entrepreneur probably has a better understanding of the probable success of the proposed venture than the prospective

financial supporters and employees that she attempts to recruit, an asymmetry that can lead to market failure (Akerlof, 1970). These factors create a friction in the movement of financial and human resources that social networks can help lubricate.

Social relations elevate the likelihood of mobilizing financial capital by dampening the perceived risk of the new venture. Two factors drive this effect. First, people typically consider information gathered from known parties more reliable; hence, investors more likely trust the projections of entrepreneurs with whom they share a social relation. For example, Fried and Hisrich (1994) report that venture capitalists prefer to fund companies referred to them by close contacts. Second, in the absence of a close contact, triangulating information from multiple sources might afford potential investors a greater deal of confidence regarding the reliability of their information on the prospective target (Sorensen and Stuart, 2001). The effectiveness of these close social ties in fostering the acquisition of financial capital tends to bind entrepreneurs to the regions in which they have contacts, even if other locations might seem more attractive.

Prospective entrepreneurs also need to bring human capital into the nascent firm. In industries requiring skilled labor, the largest pool of available labor frequently resides in incumbent firms of like kind (Sorensen and Audia, 2000). Considering the risks involved, entrepreneurs likely find it difficult to convince potential employees to leave their secure jobs to join an uncertain new venture. To recruit personnel at early stages, entrepreneurs frequently must use their networks of contacts within the industry to convince workers to join the fledgling firm. Particularly in the managerial and professional ranks, employees will not likely leave behind their stable positions if they do not trust the company founders and their abilities. Here also, strong social ties engender the trust necessary to recruit these scarce human resources. The need to draw on these networks, together with the workers' likely preferences to remain near family and friends, again acts to bind entrepreneurs to the regions in which they have previously lived and worked.

Finally, and most importantly for this paper's argument, social networks play a critical role in directing the flows of knowledge across firms. In a wide range of industries, knowledge of particular operating routines and technologies probably accounts for a large proportion of the heterogeneity across firms in profitability; these differences likely arise when rival firms find it difficult, for one reason or another, to replicate this scarce and valuable knowledge. Entrants with access to this information, then, potentially enjoy a large advantage over other firms (Klepper and Sleeper, 2000; Klepper, 2001). Such industry-specific knowledge resides almost entirely within the incumbents in an industry. Hence, nascent entrepreneurs require strong social ties to individuals that work in the industry, a condition that nearly requires that entrepreneurs in an industry arise from the ranks of its current employees (Sorensen and Audia, 2000). Though the odds of any tie erode with geographic and social distance, the frequency of the strong ties necessary to access this knowledge probably declines very rapidly. Developing and sustaining these strong ties entails repeated and intensive interaction, a circumstance unlikely to occur except among co-workers and close, personal friends.

All three factors, therefore, imply that entrepreneurs will find it difficult to start new ventures successfully without strong contacts to other participants in the industry they wish to enter. Moreover, due to the localized nature of social networks, only those individuals that live in close proximity to industry incumbents – and probably only those that have previously worked for one or more of these incumbents – will likely have the required connections. As a result, entry acts as a centralizing force: even in the absence of agglomeration externalities, the distribution of new entrants tends to replicate the existing distribution of production in the industry, diffusing only slowly away from the locations of the first successful entrants (Sorenson and Audia, 2000).

3. SOCIAL NETWORKS, INFORMATIONAL COMPLEXITY AND KNOWLEDGE FLOW

Although these centripetal forces likely operate to some degree in all sectors, it seems probable that at least some of these factors act more strongly in some industries than in others. Industries vary in their degree of concentration. The factors described in the previous section, however, offer little explanation for why such heterogeneity might arise across populations of firms, except as a result of differences in their ages (slow diffusion processes can accrete over time). Although other factors relevant to the strength of centripetal and centrifugal forces may vary from one industry to the next, this section focuses on one element that might explain a portion of these industry-level differences: the characteristics of the knowledge capital necessary for success in an industry.

Certain characteristics of knowledge may reduce the ease with which it travels spatially by increasing the difficulty of transferring it. For example, even those that possess the knowledge may find it difficult to specify and communicate it precisely. As a result of causal ambiguity, the knowledge holder might lack a clear understanding of the connection between actions and outcomes (Lippman and Rumelt, 1982). Or, the use of such knowledge might call on tacit personal skills or interactions among multiple parties that the participants themselves do not fully appreciate (Polanyi, 1966; von Hippel, 1988, 1994), or that elude codification (Zander and Kogut, 1995). These factors essentially increase the incompleteness of knowledge transfer over distances (Sorenson, Rivkin and Fleming, 2002). The complexity of the information itself can also hinder knowledge transfer across space by reducing the likelihood that a recipient of the information can fill these gaps and correct transmission errors successfully.²

Informational complexity refers to the degree to which pieces of information interact to produce a desired outcome. Think of knowledge as a set of (somewhat discrete) chunks of information. Sometimes the routines represented by these chunks operate relatively independently to produce an outcome. Other times, these pieces interact in important ways. Simon (1962), thus, identifies complex knowledge as that which comprises many elements that interact richly. Similarly, scholars of information science define the complexity of knowledge according to the length of the

string required to describe it completely (Kolmogorov, 1965) – a direct function of the number of chunks and their intensity of interaction.

Complex knowledge evades easy replication. The interactions among components generate two effects that undermine copying. First, small errors in replication produce large differences in outcomes in tightly coupled systems; getting it ‘a little wrong’ does not mean having routines nearly as effective as the original – rather, such errors frequently lead to substantial degradations in performance. As a result, those attempting to replicate the original knowledge need to correct these errors to achieve anything close to the performance of the original code. The second effect, however, stymies these efforts: Interdependence also produces a rapid increase in the number of ‘local peaks’ – internally consistent, though not optimal combinations of components that thwart improvement through incremental search because altering any single component degrades the efficacy of the whole (Kauffman, 1993; Rivkin, 2000). Not only do copying errors substantially degrade the value of the knowledge, but also correcting these errors becomes increasingly difficult, as the complexity of the knowledge increases.

Social networks, therefore, become even more important to effective knowledge transfer under these conditions. These relationships can facilitate the transfer process in at least two ways. First, they may allow the actor trying to replicate the knowledge to begin with a better facsimile – either because fewer errors arose in the transmission itself or because they began working with a more complete set of information – thereby beginning their attempts to replicate and improve on the original knowledge closer to the target. Second, social networks may allow the knowledge recipient to refer back to the source of the original knowledge when problems arise with understanding and deploying it. Strong ties thus mitigate both of the factors that frustrate the flow of complex information.

3.1. Empirical corroboration

Patent data allow me to corroborate my basic argument empirically. In particular, I estimate the effect of informational complexity on the rate at which knowledge diffuses for a sample of 17,264 patents.³ Patents and their citations offer a useful setting to test these ideas. The U.S. Patent Office requires applicants to cite prior knowledge where relevant. Patent examiners review this citation data for accuracy ensuring that these data offer more objective and consistent information than one would typically find in bibliometric studies. In addition, Fleming and Sorenson (2001) have developed a means of measuring the degree of interdependence among the knowledge components that make up patents, a measure that would be difficult to replicate in other types of data. The MicroPatent database, which reports information from the U.S. Patent and Trademark Office (USPTO), provides the bulk of the data.

Fleming and Sorenson (2001) proposed an interdependence measure that entails two stages of calculation. Intuitively, interdependent pieces of knowledge should prove difficult to combine, as interactions impede the process of invention. In developing this measure, they use patent subclasses to represent knowledge

components.⁴ In the first stage, they calculate the ease with which each component recombines – the inverse of interdependence (see the first equation for subclass i used in patent j). This measure identifies every appearance of the subclass i in previous patents from 1790 to 1990.⁵ The total number of previous occurrences provides the denominator. For the numerator, Fleming and Sorenson tally the number of different subclasses appearing with subclass i on previous patents. The measure, therefore, increases when a subclass combines with a wider variety of other subclasses, controlling for the total number of applications. To generate a patent-level measure of interdependence, they take the mean of the inverted ease of recombination scores for all subclasses to which a patent belongs (the second equation below). Though some might reasonably question whether this measure successfully proxies for interdependence, Fleming and Sorenson (2004) have shown that it accords well with inventors’ own perceptions of the degree of coupling (i.e. level of interaction) among the components in their inventions. As prior research has revealed a non-monotonic relationship between this measure and the likelihood of future citations, I include both the interdependence measure and its square in the models as controls.

$$\text{Ease of recombination of subclass } i = E_i = \frac{\text{Count of subclasses previously combined with subclass } i}{\text{Count of previous patents in subclass } i}$$

$$\text{Interdependence of patent } j = K_j = \frac{\text{Count of subclasses on patent } j}{\sum_i E_i}$$

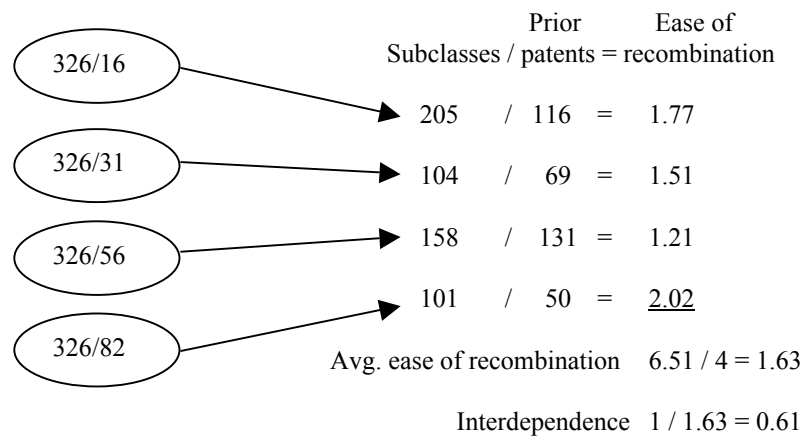


Figure 1: Calculation of interdependence measure

Figure 1 illustrates an example of this calculation for one specific patent (#5,136,185). The electronic device described by this patent belongs to four subclasses: 326/16, 326/31, 326/56, and 326/82. Each of these subclasses corresponds

to a component of the circuit. For instance, subclass 326/16 refers to the test facilitate feature, a standard feature in digital design which implements a testing mode within the semiconductor chip (McCluskey, 1986). Prior to its appearance on this patent, this subclass had been combined with 205 different subclasses across 116 patents. Hence, the ease of recombination for this subclass is $205/116 = 1.77$. Similar calculations can be made for each of the other components: 326/31 has appeared on 69 patents with 104 different subclasses ($104/69 = 1.51$); 326/56 combined with 158 different subclasses on 131 prior inventions ($158/131 = 1.21$); and 326/82 has been coupled with 101 different components in 50 patents ($101/50 = 2.02$). The average of these ease of recombination scores is 1.63; since interdependence relates inversely to the ease of recombination, inverting this score provides us with a measure of the degree of interdependence, or coupling.

Analyzing the dispersion of future citations also necessitates a measure of distance. Since patents list the inventor's address, we can locate patents according to where their inventors live, matching 3-digit zip codes⁶ to the latitudes and longitudes of the centroids of these postal codes. The distance separating two points, i and j , on a sphere is:

$$d_{ij} = C \arccos(\sin(lat_i) \sin(lat_j) + \cos(lat_i) \cos(lat_j) \cos(long_i - long_j))$$

with latitude (lat) and longitude ($long$) measured in radians. C translates the result into linear units; $C = 3437$ corresponds to miles. Taking the natural log of this measure generates a functional form consistent with theoretical expectations (Stouffer, 1940; Zipf, 1959; cf. Sorenson and Stuart, 2001, for empirical results). Interacting this term with the interdependence measure allows a test of our thesis; if social networks play an especially strong role in structuring the diffusion of complex knowledge, patents with a high degree of interdependence should diffuse even more slowly.

To explore the diffusion of knowledge, we estimate the likelihood that a future patent cites a focal patent conditional on its distance from the focal patent, the complexity of the information embodied in the focal patent, and several control variables. This procedure corresponds to an analysis of tie formation. Although many researchers approach such data by generating a matrix of all potential dyads (i.e. a case for every future patent that could have cited a focal patent), using logistic regression to estimate the effects of covariates, this approach has two weaknesses. First, it fails to account correctly for non-independence across cases. This weakness could prove particularly problematic here, as it would tend to underestimate the standard errors for properties of the patent (e.g., complexity). Second, generating and manipulating such a large matrix can prove unwieldy. For example, analyzing all potential dyads in our sample would require the generation of a data matrix with more than eleven billion cells. Sampling from this set offers one potential solution; however, this approach fails to account for the fact that the citations (as opposed to the non-citations) provide most of the information in the estimation of the likelihood function (Coslett, 1981; Imbens, 1992).

This analysis approaches the problem by using a matched sample design. All 70,271 citations that actually occur enter the data, along with a comparison set of four⁷ randomly selected patents that did not cite each focal patent. Although this sampling yielded 139,487 dyads, the analyses only consider the 76,807 cases where both inventors reside in the U.S. Since focal patents still enter the data multiple times, the tables report Huber-White robust standard errors. As logistic regression can yield biased estimates when the sample design correlates with the dependent variable (as it does here), the models correct for this effect using the method proposed by King and Zeng (2001).⁸

Table 1: Rare events logit models of future citations*

	Model 1	Model 2	Model 3	Model 4 No self cites
Distance (logged miles)	-.453** (.012)	-.464** (.051)	-.492** (.057)	-.431** (.056)
Distance X interdependence	-.118** (.011)	-.109* (.052)	-.108* (.051)	-.098* (.048)
Interdependence	1.27** (.143)	1.28* (.646)	1.23 (.648)	1.28** (.714)
Interdependence ²	-.185** (.045)	-.176 (.185)	-.177 (.184)	-.232 (.196)
Same assignee		.335 (.242)	.381 (.248)	
Same class		3.63** (.469)	3.62** (.459)	3.64** (.486)
Subclass overlap		4.85** (.761)	4.82** (.734)	4.75** (.764)
Activity control		.736* (.348)	.727* (.348)	.723* (.356)
Distance control			-.192 (.163)	
Distance control X distance			.026 (.025)	
Constant	-4.50** (.083)	-6.39** (.357)	-6.16** (.409)	-6.57** (.394)
Log-likelihood	-49064.5	-22848.5	-22804.7	-20549.5
Pseudo R ²	.076	.562	.562	.552

* 76,807 cases, 51.7% represent ties (vs. .00059% in population) • $p \leq .05$ ** $p \leq .01$; Standard errors shown in parentheses

Table 1 reports the results of these models. The first model, which includes only the covariates of interest, reveals that the likelihood of a future patent citing a focal patent declines with distance; moreover, as expected, the strength of this effect in-

creases with the interdependence of the technology embodied in the focal patent. Interdependence also exhibited its expected non-monotonic relationship to the likelihood of future citation.

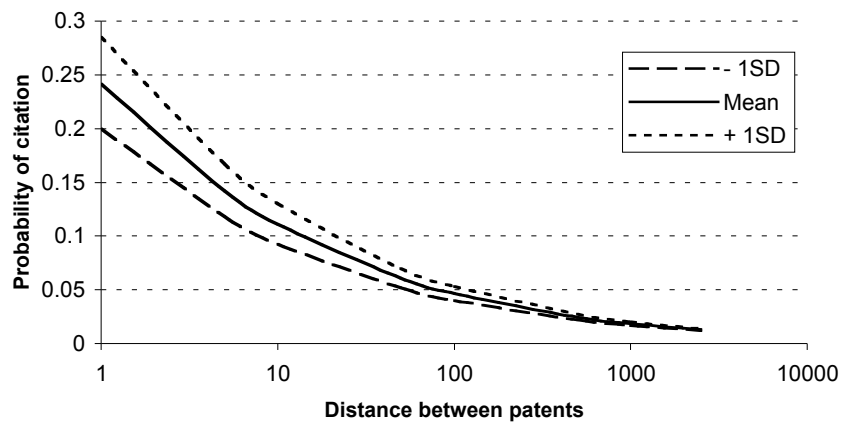
Model 2 introduces a variety of control variables: *Same assignee* takes a value of one if both the focal patent and the potential citing patent have the same assignee listed (i.e. a self-citation); though studies of patent citations typically find higher rates of self-citation, these effects do not appear robust to the fine-grained controls for technological similarity used here. *Same class* is a dummy variable with a value of one if both patents belong to the same primary technological class; patents in the same class cite one another with much greater frequency than those in different classes. *Subclass overlap*, the proportion of subclasses on the potentially citing patent that appears on the focal patent, provides a much finer-grained measure of the technological similarity between two patents; this measure also reveals a strong relationship between the technological similarity of two patents and the likelihood of a citation. Finally, the *activity control* accounts for the total degree of activity in the technology classes to which the focal patent belongs – essentially, it amounts to a weighted average of the number of citations received by a typical patent in each of the classes in which the focal patent falls. As one would expect, patents in more active technological areas receive future citations at a higher rate. Regardless, the effects of distance and the interaction between interdependence and distance still hold: interdependence retards the geographic dispersion of future citations.

Despite these strong results, the distribution of research activity in a technological area might explain these results. In other words, industries characterized by highly complex knowledge might concentrate more intensely. Though the argument being made here suggests precisely such a relationship, the direction of causality could run in the opposite direction of the one proposed here: co-location might lead to the development of more complex technologies, rather than informational complexity limiting the geographic flow of the diffusion of knowledge. To account for such an effect, the models incorporate a *distance control*. The distance control averages the natural log of the distance between all pairs of patents, granted in 1990, in the same primary class as the focal patent; in essence, it captures the average distance that one would expect to see between citing patents in a technological field if geography did not influence who cited whom. Model 3 includes both this variable and its interaction with the distance term as controls; however, these controls neither have significant effects on the likelihood of a citation, nor do they affect the estimates for the interaction between distance and interdependence.

The final column, model 4, reports an estimate without self-citations. Since the process described in section 2 would require the movement of knowledge across firms, this model restricts the analysis to these cases. Once again, the results appear robust to this alternative specification of the model: interdependence depresses the likelihood that a distant patent cites the focal patent. Figure 2 displays this effect graphically. A patent of average interdependence is more than four times more likely to receive a citation from another patent in the same zip code as from one 100 miles away; by comparison, a patent with an interdependence score one standard deviation above the mean is six times more likely to be built on by a local patent

than one 100 miles distant. Hence, the figure illustrates that informational complexity intensifies the localization of spillovers.

One might notice that the expected main effects of interdependence on the likelihood of a citation falls below traditional levels of significance. Though this fact appears inconsistent with the baseline relationship between interdependence and future citations proposed by Fleming and Sorenson (2001), a careful look at the models reveals that the lack of significance stems from an increase in the size of the standard errors, rather than a decline in the point estimates of the effect size. Hence, multi-collinearity may well account for this result, by reducing the model's statistical power.



The lines illustrate effects at the mean level of interdependence, as well as at one standard deviation above and below this mean.

Figure 2: The likelihood of a within class citation by complexity and distance

4. COMPLEXITY AND INDUSTRIAL GEOGRAPHY

Pushing this mechanism up to the industry level of analysis generates a fairly straightforward set of predictions. Consider the following premises: a) Knowledge capital represents an important resource necessary for successfully entering an industry; b) Incumbents form the primary repository for this information; c) Particularly when this knowledge is complex – containing many elements that depend sensitively on each other – entrepreneurs require strong social networks to access this information effectively; and d) Social networks tend to localize in geographic and social space. Taken together, these factors imply that social networks should matter most in industries with relatively complex technologies at their core; as a result, industries based on these technologies should diffuse more slowly in space.

Ideally, one would like longitudinal information on the spatial dispersion of a variety of industries. The data requirements, however, make such information infeasible. The evidence presented here relies on a cross-section instead. In particular, I compare the dispersion of industry to the average level of interdependence found in the patents related to that industry.

The measure of industrial dispersion comes from the figures reported in the appendix of Krugman (1991). These statistics purport to describe the level of geographic dispersion in employment across US states using a Herfindahl measure for 106 3-digit SIC industries. The highest possible score on this measure (one) would correspond to an industry in which all employees worked in a single state. A score of zero (or near zero) suggests that production spreads across all states in a manner roughly proportional to the populations in those states.

Relating the characteristics of patents to industries involves some assumptions. In the US, patents themselves do not include SIC codes. The aggregation to industries here makes use of a concordance between USPTO classes and SIC codes developed by Brian Silverman (1999). In its first step, this concordance uses data from Canadian patents to assign USPTO classes to a probability distribution of International Patent Codes (IPC; i.e. if all patents from a USPTO class fell in the same IPC class it would translate one-to-one, but if only 50% fell in a particular IPC class, it would receive a weight of .5 in translating from one system to the other). The second step involved relating the IPC classes to industries; since Canadian patents include information on both, this entailed the relatively straightforward process of generating a probability distribution of likely industry applications associated with each IPC. To return back to the US system, the concordance takes the third step of relating Canadian industry codes to US SIC codes (for details on this measure, see Silverman, 1999). The concordance provides a noisy, though unbiased, means of associating the characteristics of patents with industries. To generate the industry level measures of interdependence, I averaged the level of interdependence across all patents within a particular technological class. To arrive at industry level measures, I averaged these scores across the technological class weights indicated by the concordance.

Figure 3 depicts the correlation between Krugman's geographic concentration measure and the average level of informational complexity for these 106 3-digit SIC industries. As the figure illustrates, industries based on more complex knowledge appear more likely to remain tightly clustered in a small number of regions. These results fit quite nicely with the expected pattern. In total, informational complexity can account for roughly 15% of the degree of dispersion across industries ($p < .001$). Though that number might seem low, one must remember that a large number of other factors likely influence the degree of geographic concentration within an industry, including the location of key inputs, transportation costs, and the age of the industry itself (since diffusion processes take time).

Even a cursory perusal of the chart, however, suggests that two outliers may play a heavy role in reducing this correlation. To verify that these cases do not account for the results, I reran the analysis excluding them. Although the strength of the relationship declines somewhat – without the two outliers, informational complexity

only explains 10% of the variation in the geographic concentration of industries – the association remains significant ($p < .005$). Thus, the results appear broadly consistent with the thesis that social networks matter most in industries drawing on relatively complex information.

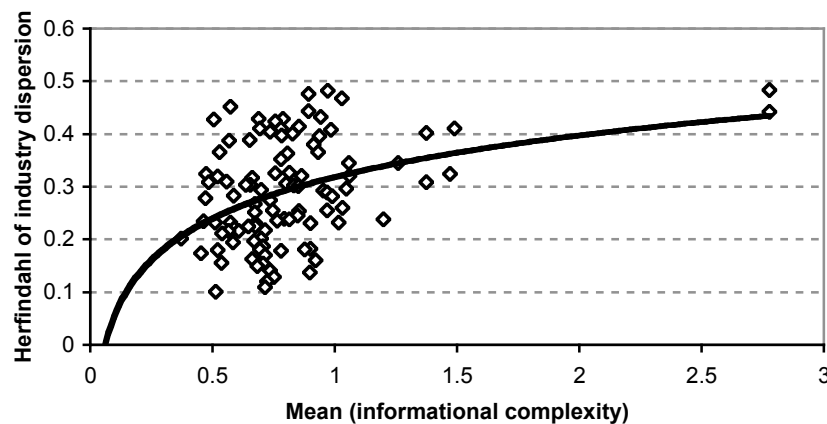


Figure 3: Industry dispersion by informational complexity

Though my account emphasizes the importance of access to information in the entrepreneurial process, clustering as a result of informational complexity might also reflect a type of agglomeration economy. To the extent that the knowledge concerned improves firm performance, firms that can access this information without incurring the full cost of acquiring it will benefit relative to those that cannot. Firms might therefore cluster so that they can ‘share’ the costs of acquiring valuable information (Arrow, 1962), or alternatively the labor efficiency embedded in employees with access to this difficult-to-transfer knowledge.

Note, however, that this account still assumes that informational complexity curtails the flow of knowledge through social networks. It simply does not imply any social inefficiency to this outcome. Though beyond the scope of this chapter, one means of differentiating between these two accounts might entail examining the distribution of *de novo* and *de alio* entrants separately. Whereas all firms would benefit from these externalities, the constraints described in sections 2 and 3 should primarily affect entrepreneurs (and hence the localization of *de novo* firms).

5. CONCLUSION

Social networks matter to industrial geography. In particular, these networks play an important role in determining who can successfully start a new venture in an industry, and where they can start it. Potential entrepreneurs need access to a variety of resources – including knowledge of the key technologies in the industry – to begin

operations and to compete successfully; social networks facilitate access to these resources, a notion supported by several recent studies (e.g., Sorenson and Audia, 2000; Klepper, 2001; Stuart and Sorenson, 2003).

Industries do not always agglomerate, however. If social networks strongly structure this process, all industries should exhibit clustering. Nevertheless, industries vary substantially in the degree to which they concentrate geographically, begging the question: Why do some industries concentrate while others spread? Put differently, one might ask: When do social networks play a vital role in structuring industrial geography?

The paper maintains that the characteristics of the knowledge underlying the key technologies in the industry determine the ease with which distant entrepreneurs can attempt entry. Social networks become increasingly important for the knowledge transmission as the complexity of the information increases. Entrepreneurs trying to replicate the success of incumbents must successfully mimic their understanding of key technologies, a process that involves trial-and-error learning if the entrepreneur does not begin with a perfect understanding of this knowledge. Such trial-and-error learning proves difficult, however, with complex knowledge. Dense social networks can diminish the need for search though, by facilitating high fidelity information transfer: complete information with little noise. Hence, networks prove most important to entrepreneurship, and most limiting to the diffusion of industry, when complex knowledge contributes importantly to success in the industry.

To corroborate this thesis, the paper offers two types of evidence. At the micro-level, analysis of patent citations reveals that these citations bridge more extensive geographic expanses when the underlying knowledge contains a low level of interdependence, or coupling, between its components. Moving to the industry-level, a cross-sectional analysis of the geographic dispersion of production across industries shows that the degree of informational complexity (measured using the average level of interdependence for patents related to an industry) explains a significant portion of this cross-industry variation. Hence, the evidence appears to support the notion that the complexity of the information flowing through a network has some bearing on its relevance for the geographic distribution of economic activity.

Anderson Graduate School of Management, University of California – Los Angeles

6. NOTES

□ This paper reports research that builds substantially on work that the author has done in conjunction with Lee Fleming and Jan Rivkin. Though faults in this piece remain the responsibility of the author, Lee and Jan deserve credit for many of its merits. I also thank Brian Silverman for generously providing access to his patent class-SIC code concordance. Guido Bünstorf and the participants of the Max Planck Institute workshop on “The role of labor mobility and informal networks for knowledge transfer” provided many comments that proved useful in the development of this paper.

¹ Stouffer (1940) and Zipf (1949) both derive formal models to predict the functional relationship between distance and the probability of interaction – opportunity costs drive Stouffer’s model while direct costs play a larger role in Zipf’s approach. Both expect that the probability of a tie should vary in proportion to the reciprocal of the distance between two actors.

² Sorenson, Rivkin and Fleming (2002) provide an extended discussion and analysis of this process.

³ The focal data include information on May and June of 1990 (see Fleming and Sorenson, 2001). We selected the year to make maximal use of the data, since we needed several years of data following the sample to follow citations, while still retaining recency. We chose the start month for 1990 at random and limited the set to two months of data to maintain a manageable computational burden – even so, the computation of the interdependence independent variable required roughly 50 billion calculations.

⁴ The USPTO classifies each patent into one or more fine-grained subclasses (nearly 100,000).

⁵ Although some might question the stability of this measure over time, all of the results remain robust to the use of a second interdependence measure based only on the data from 1980 to 1990.

⁶ Although the USPTO reports the 5-digit zip, the cleaned data, available from CHI, which called every patent holder to verify inventor location, only provides these data at the 3-digit level.

⁷ The inclusion of four patents ensures that the ‘control’ group has roughly the same size as the set of realized ties.

⁸ The traditional logistic regression model treats the dichotomous outcome variable as a Bernoulli probability function that takes a value 1 with the probability π :

$$\pi_i = \frac{1}{1 + e^{-X_i\beta}}$$

where X represents a vector of covariates and β denotes a vector of parameters. King and Zeng (2001) show that the following weighted least squares expression estimates the bias in β from over-sampling rare events:

$$\text{bias}(\hat{\beta}) = (X'WX)^{-1} X'W\beta$$

where $\beta = 0.5Q_n[(1+w_1)\beta - w_1]$, the Q are the diagonal elements of $Q = X(X'WX)^{-1}X'$, $W = \text{diag}[\beta(1-\beta)w_i]$, and w_i represents the fraction of ones (events) in the sample relative to the fraction in the population. Essentially, one regresses the independent variables on the residuals using W as the weighting factor. Tomz (1999) has developed a Stata command, `relomit`, which corrects for this bias.

7. REFERENCES

- Akerlof, G.A. (1970). The market for ‘lemons’: qualitative uncertainty and the market mechanism. *Quarterly Journal of Economics*, 84, 488-500.
- Arrow, K.J. (1962). The economic implications of learning by doing. *Review of Economic Studies*, 29, 155-173.
- Blau, P.M. (1977). *Inequality and heterogeneity*. New York, Free Press.
- Coslett, S.R. (1981). Maximum likelihood estimator for choice-based samples. *Econometrica*, 49, 1289-1316.
- Diamond, C., & Simon, C. (1990). Industrial specialization and increasing returns to labor. *Journal of Labor Economics*, 8, 175-201.
- Festinger, L., Schacter, S., & Back, K.W. (1950). *Social pressures in informal groups*. New York, Harper.
- Fleming, L., & Sorenson, O. (2001). Technology as a complex adaptive system: evidence from patent data. *Research Policy*, 30, 1019-1039.
- Fleming, L., & Sorenson, O. (2004) Science as a map in technological search. *Strategic Management Journal*, 25, forthcoming.
- Fried, V.H., & Hisrich, R.D. (1994). Toward a model of venture capital investment decision-making. *Financial Management*, 23, 28-37.
- Imbens, G. (1992). An efficient method of moments estimator for discrete choice models with choice-based sampling. *Econometrica*, 60, 1187-1214.
- King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, 9, 137-163.

- Klepper, S. (2001). Employee startups in high-tech industries. *Industrial and Corporate Change*, 10, 639-674.
- Klepper, S., & Sleeper, S. (2000). Entry by spinoffs. Mimeo
- Kolmogorov, A.N. (1965). Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1, 1-17.
- Krugman, P. (1991). *Geography and Trade*. Cambridge, MA, MIT Press.
- Lazarsfeld, P.F., & Merton, R.K. (1954). Friendship as a social process: a substantive and methodological analysis. In: Berger, M., Abel, T., & Page, C.H. (Eds.). *Freedom and control in modern society*. New York, Van Nostrand.
- Lippman, S., & Rumelt, R. (1982). Uncertain imitability: an analysis of interfirm differences in efficiency under competition. *Bell Journal of Economics*, 13, 418-438.
- Marshall, A. (1890). *Principles of economics*. London, MacMillan.
- McCluskey, E. (1986). *Logic design principles with emphasis on testable semi-custom circuits*. Englewood Cliffs, NJ, Prentice Hall.
- McPherson, M., Smith-Lovin, L., & Cook, J.M. (2001). Birds of a feather: homophily in social networks. *Annual Review of Sociology*, 27, 415-444.
- Polanyi, M. (1966). *The tacit dimension*. New York: Anchor Day.
- Rivkin, J.W. (2000). Imitation of complex strategies. *Management Science*, 46, 824-844.
- Romer, P. (1987). Growth based on increasing returns to specialization. *American Economic Review*, 77, 56-62.
- Rotemberg, J.J., & Saloner, G. (1990). Competition and human capital accumulation: a theory of interregional specialization and trade. Mimeo
- Simon, H.A. (1962). The architecture of complexity. *Proceedings of the American Philosophical Association*, 106, 467-482.
- Silverman, B.S. (1999). Technological resources and the direction of corporate diversification: toward an integration of the resource-based view and transaction cost economics. *Management Science*, 45, 1109-1124.
- Sorenson, O., & Audia, P.G. (2000). The social structure of entrepreneurial opportunity: geographic concentration of footwear production in the United States, 1940-1989. *American Journal of Sociology*, 106, 424-462.
- Sorenson, O., Rivkin, J.W., & Fleming, L. (2002). Complexity, networks and knowledge flow. Mimeo.
- Sorenson, O., Stuart, T.E. (2001). Syndication networks and the spatial distribution of venture capital financing. *American Journal of Sociology*, 106, 1546-1588.
- Stouffer, S.A. (1940). Intervening opportunities: a theory relating mobility and distance. *American Sociological Review*, 5, 845-867.
- Stuart, T.E., & Sorenson, O. (2003). The geography of opportunity: spatial heterogeneity in founding rates and the performance of biotechnology firms. *Research Policy*, 32, 229-253.
- Tomz, M. (1999). relogit (Stata ado file). Available at <http://gking.harvard.edu/stats.shtml>
- von Hippel, E. (1988). *The sources of innovation*. New York, Oxford University.
- von Hippel, E. (1994). Sticky information and the locus of problem solving: Implications for innovation. *Management Science*, 40, 429-439.
- Zander, U., & Kogut, B. (1995). Knowledge and the speed of transfer and imitation of organizational capabilities: an empirical test. *Organization Science*, 6, 76-92.
- Zipf, G.K. (1949). *Human behavior and the principle of least effort*. Reading, MA, Addison-Wesley.