Tsukuba Economics Working Papers
No. 2009-014

The Dynamics of R&D Network in the IT Industry

by

Nobuyuki Hanaki, Ryo Nakajima, Yoshiaki Ogura

November 2009

UNIVERSITY OF TSUKUBA
Department of Economics
1-1-1 Tennodai
Tsukuba, Ibaraki 305-8571
JAPAN

The Dynamics of R&D Network in the IT Industry

Nobuyuki Hanaki[a,b*], Ryo Nakajima[b], Yoshiaki Ogura[c†]

[a] *GREQAM and Université de la Méditeranée*
Centre de la Vieille Charité, 2, Rue de la Charité 13236 Marseille, cedex 02, FRANCE
[b] *Doctoral Program in Economics,*
University of Tsukuba, 1-1-1 Ten-nodai, Tsukuba, Ibaraki 305-8573, JAPAN
[c] *College of Business Administration,*
Ritsumeikan University, 1-1-1 Noji Higashi, Kusatsu, Shiga 525-8577, JAPAN

August 3, 2009

Abstract

In this paper, we provide an empirical analysis of evolving networks of successful R&D collaborations in the IT industry (consisting of firms that obtained patents in the technological category of computers and communication) in the U.S. between 1985 and 1995. We first show that the R&D network has become more extensive, more clustered, and more unequal in the sense that 'stars' have emerged in the network. We then analyze the effect of the existing network structure in the process of new R&D collaboration formation. We control for unobserved similarities among firms based on the community structures within the network that the algorithm developed by Girvan and Newman (2004) identifies and find a significant cyclic closure and preferential attachment effect.

Keywords: Dynamic networks, R&D partnerships

1 Introduction

The interfirm positive spillover of research and development (R&D) activities is widely accepted as an empirical stylized fact in the R&D literature. Existing empirical studies have found that this spillover effect is stronger among companies in a technological or geographical neighborhood (Jaffe, 1986; Jaffe et al., 1993) and among companies that invest in R&D sufficiently to maintain their absorptive capacity for external knowledge

*Corresponding Author. Tel.: +33-4-91-14-07-27. Fax: +33-4-91-14-02-27.
†E-mail addresses: nobuyuki.HANAKI@univmed.fr and hanaki@dpipe.tsukuba.ac.jp (N. Hanaki), nakajima@dpipe.tsukuba.ac.jp (R. Nakajima), and yogura@ba.ritsumei.ac.jp (Y. Ogura).

(Cohen and Levinthal, 1989). Furthermore, this spillover effect is stronger among collaborating companies than among competing companies (Branstetter and Sakakibara, 2002; Gomes-Casseres et al., 2006). These studies implicitly or explicitly recognize that knowledge exchange among innovative companies is the key determinant of R&D spillovers. In response to the recent surge of R&D collaborations in high-tech industries (Hagedoorn, 2002), many theorists have analyzed interfirm knowledge spillovers through R&D collaborations in network settings by applying the network formation theory (e.g., Goyal and Moraga-Gonzalez, 2001; Goyal and Joshi, 2003; Cowan and Jonard, 2004; Meagher and Rogers, 2004). There is ample evidence that the positioning of companies in R&D collaboration networks substantially affects their productivity in generating new knowledge, which is embodied in artifacts such as patents and new products (e.g., Powell et al., 1996; Ahuja, 2000; Schilling and Phelps, 2007).

Given the importance of network architecture in innovation performance, it is reasonable that innovative individuals and companies may want to form R&D collaborations strategically to control knowledge spillovers. In fact, the importance of the strategic formation of R&D alliances has been widely recognized in the literature.[1] In recent years, the theoretical analysis of network formation has been a lively area of research, not only in the context of R&D (e.g., Goyal and Moraga-Gonzalez, 2001) but also in other contexts; see Jackson (2006) for a review. These studies incorporate individual incentives to derive strategically stable network architecture and to analyze the efficiency characteristics of the network. However, empirical analyses of network formation are relatively rare.

The determinants of R&D collaboration between companies have been thus far investigated by industrial organization economists. They have found that the company's "absorptive capacities" (Cohen and Levinthal, 1989), such as their R&D size and intensity, significantly influence the likelihood of forming R&D alliances (e.g., Röller et al., 1998; Hernán et al., 2003). While the characteristics of firms may explain the types of companies that are likely to collaborate, they provide insufficient understanding of how companies *interact* with each other. For example, if the number of pathways for communication between companies increases, the enhanced flow of knowledge becomes attractive and, thus, any collaboration may stimulate further collaboration. This suggests that the network structure that is in place influences how new collaboration links are formed. However, existing empirical studies do not address this recursive and inductive aspect of R&D network formation. Therefore, in this paper, we address the following questions: How does collaboration between companies of one type influence the actions of other types of company? What structural characteristics are more likely to stimulate further R&D collaboration?

The goal of this paper is to study the endogenous development of R&D networks in the U.S. Information and Technology industry (defined by the technological categories of the obtained patent, as explained in the next section) by showing how companies have established new R&D collaborations with each other. Based on individual firm-level data on R&D collaboration, which was constructed from the information on granted patents, we estimate the conditional probability of new collaboration formation between any pair of companies in the industry given the network structure observed in

[1] For a recent review of the theoretical literature, see, for example, Bloch (2005)

the previous period. In particular, we estimate the impact on R&D collaboration formation of certain kinds of network topology, such as *cycles* and *stars*.

Given the complexity of the subject, the empirical analysis of the formation of an R&D network has been hampered by a scarcity of data of alliances among companies and the members of such alliances. Existing empirical studies on this subject (e.g., Gulati and Gargiulo, 1999; Hagedoorn, 2002; Schilling and Phelps, 2007) have relied on the "publicly reported alliance counting method" (Hagedoorn, 2002) for data collection; that is, information on announced interfirm R&D collaboration is collected from various publications, such as newspapers, journal articles and books. Although such literature-based data collection is extensive, the information is likely to be incomplete, as only publicly announced alliances are included in the dataset.[2]

In this paper, we adopt a different approach. We collect information on interfirm collaborations by using the NBER Patent Data File (Hall et al., 2001). Since the pioneering work of Scherer (1965) and Schmookler (1966) appeared, patent data have been used in a number of empirical studies of research collaboration among innovative companies (e.g., Singh, 2005; Cantner and Graf, 2006).

The main reason for using the NBER patent data is that these data provide information about all researchers who were involved in creating the innovation along with information on the patenting company, its geographic location, and the types of technology involved. The names of the inventors are recorded along with the name of the corporate assignee claiming each patent. We match the lists of inventors' names across different assignee companies to see if they are connected via common inventors. If the same inventors work on a particular research project across two innovating companies, we ascribe the project to an R&D partnership and identify those companies as collaborators through the inventors. Longitudinal data on an evolving R&D network are created by collecting annual snapshots of instantaneous networks.

In the process of identifying inventors, an identification error, often called the "Who is Who" problem (Trajtenberg et al., 2006) cannot be avoided. The problem refers to the possibility that the name of an individual associated with a patent may have been inadvertently spelled differently or that two people may have the same name, leading to inaccuracies in the identification process. We have deliberately used a computer matching procedure (CMP) that was recently proposed by Trajtenberg et al. (2006) to minimize errors in identifying inventors.

We start with an empirical examination of the macro-dynamic properties of R&D network structures. Several interesting features of the R&D collaboration network are identified. First, R&D networks have increased in size substantially over time. The number of nodes has increased substantially and, at the same time, the number of links per node has steadily grown. Second, networks have become increasingly connected, and the giant connected component has emerged and expanded. Third, the average distance between nodes has been stable, although network sizes have increased. Fourth, two given connected nodes tend to be linked to a common third party. The tendency to form local circles was significant in the 1990s. Given these findings, we can say that the R&D network is an emerging "small world" (Watts and Strogatz, 1998). Our

[2] Another problem is that the information on the termination of R&D collaborations is not usually published systematically, particularly for licensing and customer–supplier relationships.

final finding concerns the distribution of R&D collaborations. We find that networks have become more uneven. Our interpretation is that current R&D networks exhibit a core–periphery structure in which connected companies are becoming increasingly connected.

These findings have led us to develop a utility-based empirical model of joint R&D collaboration formation. Although agents behave nonstrategically, the model enables us to condense the issues of collaboration formation to the agents' welfare. Although our model is simple, it exhibits two mechanisms for joint collaboration: random and network-based. Hence, this model is similar to that of Jackson and Rogers (2007). Random collaboration occurs when a pair of companies meet uniformly at random and collaborate with each other by chance, given that the characteristics of the pair are controlled for. In network-based collaboration, the formation of new links depends on the exiting network structures observed by companies in advance.

We argue that two structures are important in the context of the network-based collaboration process. First, we focus on the *cyclic* structure. High search costs for collaborating partners may promote the formation of R&D alliances involving intermediate collaboration partners, which may lead to the formation of local neighborhood collaboration chains. Second, we focus on the *star* structure. Indirect benefits from spillovers may generate a positive feedback loop through which companies connect with a handful of super-connected stars that then become increasingly connected. This process is often called "preferential attachment" (Barabási and Albert, 1999). An empirical model that incorporates these features is compatible with the macroeconomic empirical findings described above.

To gain more insight into the circumstances under which R&D collaboration occurs, we estimate an empirical model that controls for a number of company background characteristics. The estimation results show that similarities in the size of R&D input have a non-monotonic impact on the probability of collaboration, which suggests that companies whose scale of R&D activity is similar but not too similar tend to collaborate. We find that there is a significant relationship between network structure and R&D collaboration. Companies with many collaborators are likely to attract further collaborations in the next period. This suggests significant preferential attachment in collaboration formation. At the same time, we find evidence of significant cyclic closure effects, which suggests that companies are willing to collaborate with other companies that form part of a chain of common third-party collaborators within the network.

There are at least two explanations for these empirical findings regarding the role of network structure in R&D collaboration formation: the importance of referral in the search for collaborating partners, and the effort made by firms to increase the appropriability of the fruits of joint R&D projects. Referrals from current collaborators reduce the cost of finding new and reliable collaborators. Because referrals are made between firms that have collaborated before, this results in cyclic closure. In addition, if a firm has collaborated with many others in the past, referrals can lead the firm to attract more collaboration in the future. Finding a reliable partner is important for a firm entering into joint R&D projects in order to increase the appropriability of the fruits of the project by maximizing incoming knowledge spillovers while minimizing outgoing knowledge spillovers (Cassiman and Veugelers, 2002). Outgoing spillovers

can also be reduced by forming a dense web of collaborations and strengthening the threat of joint punishments against deviators. Such considerations also enable firms to find new partners in local circles. Unfortunately, we are unable to discriminate the effect of these two mechanisms from our data.

Existing empirical analyses of network formation among firms (see, for example, Gulati, 1995; Gulati and Gargiulo, 1999; Powell et al., 2005) have also found that network-based mechanisms such as preferential attachment and cyclic closure are important. These studies, however, tend to overestimate the importance of such mechanism because of the lack of control for unobserved similarities among firms. In our analysis, we control for unobserved similarities among firms by identifying the community structures in the existing network using the algorithm developed by Girvan and Newman (2004). The result shows that the effect of the existing network structure remains statistically significant even after unobserved similarities among firms are controlled for.

The rest of the paper is organized as follows. The data are described in Section 2. Section 3 is a discussion of the evolution of the structure of networks over time. Section 4 is a description of the framework of our statistical analysis. The empirical results are discussed in Section 5. Section 6 is the conclusion.

2 Network Data

The data used for this study are drawn from NBER patent data. We restrict our attention to the subsample of companies that have obtained patents classified as "Computers and Communications" by Hall et al. (2001).[3] This technological category is closely related to the R&D activities of the Information and Technology (IT) industry, and we refer to the set of firms in our dataset as the IT industry for expositional purposes.

The industry provides an interesting context for our study because companies in this industry actively obtained patents for their intellectual property (Levin et al., 1987). It is also noted by Hagedoorn (2002) and others that the IT industry has been one of the most active industries in forming R&D alliances. In fact, the IT industry has had the highest share of newly established R&D partnerships among high-tech industries since the late 1980s. Thus, there are a sufficient number of R&D collaborations in this industry.

We use data on patents at the U.S. Patent and Trademark Office that were registered between 1985 and 1995. We choose 1985 as the initial year of our sample period because the number of R&D collaborations established before 1985 is too small for a meaningful analysis of R&D networks.

2.1 Network Construction

We use a schematic graph to represent an R&D network, which is a collection of innovating companies (nodes) and a collection of joint collaborations (links) between

[3] According to the classification of Hall et al. (2001), this technological class includes the following subcategories: Communications; Computer Hardware & Software; Computer Peripherals; and Information Storage.

Table 1: An example of the raw data.

Patent	Assignee	Researchers
P_1	A_1	R_1, R_2
P_2	A_2	R_1, R_3, R_4
P_3	A_1	R_1, R_4
P_4	A_3	R_2, R_5

them. In what follows, network nodes consist of *independent* companies. Subsidiary companies are not included as network nodes because the contribution of intrafirm R&D collaboration is considered to differ from that of interfirm R&D collaboration.

To identify the parent–subsidiary relationship, we supplement the NBER patent database with corporate and noncorporate name-matching results available from Bronwyn Hall's Web page of *The Patent Name-Matching Project*.[4] In addition, we have supplemented our data with *SDC Platinum, the Worldwide Mergers and Acquisitions Database*, issued by Thomson-Reuters. Among all the M&As since 1979 that are reported in *SDC Platinum*, we select the cases in which the acquiring company obtains all of the stock of the target company. We then consider those two companies to be in a parent–subsidiary relationship and treat them as one company after the merger.

We follow Cantner and Graf (2006) in constructing the adjacency matrix corresponding to the network graph illustrating R&D collaborations. Assuming that group-based inventors are involved in joint R&D projects, a collaboration link is identified between two companies if there is at least one common inventor listed in the patents owned by the companies.

As noted in the Introduction, we have utilized the computer-matching procedure (CMP) recently proposed by Trajtenberg et al. (2006) to identify inventors. The essence of CMP is to adjust for possible spelling errors of inventors' names on patents to avoid identifying one inventor as two individuals and minimize the possibility of identifying two inventors as the same person by utilizing other information such as addresses, assignees, and patent classes. This is a much better identification method than using exclusively last names, first names, and middle initials, which have so far been used in constructing co-authorship networks.[5] Identification methods that are similar to CMP have been employed in several recent studies.[6]

The following example illustrates the methodology. Let us assume that four patents are owned by three corporate assignees with five inventors, as shown in Table 1. The table shows, for example, that patent P_1, which is owned by assignee company A_1, was invented by two researchers, R_1 and R_2.

Define an $n \times m$ matrix X, where n is the number of companies and m is the

[4] `http://www.econ.berkeley.edu/ bhhall/pat/namematch.html`

[5] See, for example, Newman (2004); Goyal et al. (2006)

[6] See, for example, McHale (2006); Schankerman et al. (2006); Marx et al. (2007); Hoisl (2007)

number of researchers recorded in the dataset. By using the above example, we have:

$$X = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{pmatrix}.$$

The adjacency matrix, Γ, which summarizes all the collaboration relationships between the assignee companies, is thus given by:

$$\Gamma = XX' = \begin{pmatrix} 3 & 2 & 1 \\ 2 & 3 & 0 \\ 1 & 0 & 2 \end{pmatrix}.$$

The ith diagonal element of Γ represents the total number of inventors involved in the collaborative activities between company i and the other companies. The off-diagonal element of Γ represents the number of inventors involved in the collaborative activities between the two companies. Thus, the larger the value of the element, the more intense the R&D collaboration between the two companies.

Like most existing studies of network formation, the following statistical analysis does not utilize information about the intensity of the collaborative relationships. Thus, we focus on the following *unweighted* adjacency matrix, G, the off-diagonal elements of which are either zero or unity, as follows:

$$G = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

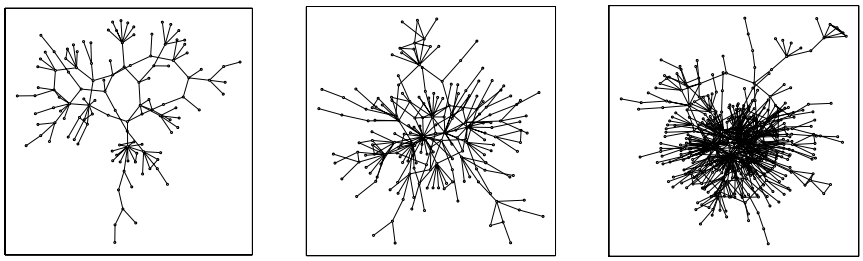
An entry is unity if the corresponding entry of the weighted adjacency matrix, Γ, is positive. The unweighted adjacency matrix indicates whether there is a collaborative relationship between a pair of companies. Note that the diagonal elements of G are zero because we do not consider intrafirm collaborations in our analysis.

To conduct a dynamic analysis of R&D alliance networks, we 'slice' the collaboration data into several snapshots. One issue is that although we know that a joint research project existed in the year in which the patents were applied for, we have no information about the start and end dates of the project. Given that R&D collaborations typically last for more than a year, following previous studies (e.g., Schilling and Phelps, 2007), we make the conservative assumption that R&D partnerships last for *three years*.[7] In other words, we assume that the network for year t includes all R&D collaborations represented by patents applied for during years $t-1$ and $t+1$. By using such a three-year window, we obtain information on 10 waves of instantaneous networks between 1985 and 1995.[8]

Given the way we identify networks, there is a risk of creating spurious collaboration links between two companies. This happens if researchers switch their jobs

[7] The three-year window may be justified by the data. For example, based on a survey of the top managers of 52 companies in the biotechnology industry, supplemented by data from *Bioscan*, Deeds and Hill (1998) found that R&D collaborations lasted for an average of 3.47 years.

[8] As explained above, because we use three-year moving windows, the network for 1985 is based on the granted patents filed between 1984 and 1986.



(a) 1991 (b) 1993 (c) 1995

Figure 1: Evolution of IT R&D Networks

and work for different companies. For example, if the inventor of a patent owned by a company move to another company and patents a new invention there, our method indicates a spurious link between companies that, in fact, did not collaborate.

Because there are no comprehensive databases that track the affiliations of researchers, we are unable to avoid the errors arising because of inventor mobility. However, to some extent, this potential for error could be mitigated by using *short* spans of years to construct the window for snapshooting networks. Given empirical evidence that researchers typically stay at one company for several years,[9] the three-year window seems sufficiently short to control for the noise caused by job switching. Furthermore, even if job switching does occur within the three-year period, it seems unlikely that the job switcher, who is new to the research team, would be listed as one of the inventors of a patent generated by an existing research project. Thus, our network data may suggest negligible spurious links between before-job-switching and after-job-switching companies.

Based on the method of network construction described above, we find that 6,746 companies applied for at least one patent, of which 3,315 companies (49.1 percent) had at least one R&D collaboration link with other companies over the 10-year period from 1985 to 1995.

Figure 1 illustrates a small selection of these networks to highlight both the evolving topology of the R&D networks and the processes by which new links are added. Each presented network shows the largest cluster of companies that are linked directly or indirectly to each other through the chain of R&D collaborations. The figures reveal that the network has grown over time. Detailed network statistics are analyzed in the next section.

[9]This can be inferred from the survey conducted by Almeida and Kogut (1999), which shows that 428 inventors switched jobs 335 times between 1974 and 1994. This suggests that each inventor averaged 1.82 jobs during this period and, thus, stayed with one company for an average of $11.05 (= 20/1.82)$ years. Additional support for our assumption comes from the employment tenure statistics available from the U.S. Bureau of Labor Statistics. These data show that, for 1991, the median tenure of engineers was 6.7 years, and that of mathematical and computer scientists was 4.2 years.

3 Dynamics of the R&D Network

We study the following network statistics. Let N be the set of independent companies in the R&D network for a given year. The number of companies, n, is defined as $n = |N|$. Recall that the collection of all R&D collaborations is represented by the adjacency matrix G; that is, for two companies i and j, $G_{ij} = 1$ represents the existence of an R&D partnership.

Let $N(i)$ be the set of companies collaborating with company i. The total number of collaborators working with company i is referred to as the *degree* of company i and is defined as $\eta(i) = |N(i)|$. We refer to companies that have no collaborators as *isolated nodes*. We term companies that have at least one collaborator as *linked nodes*. Thus, $\eta(i) = 0$ if i is an isolated node, and $\eta(i) \geq 1$ if i is a linked node. In what follows, we use the subscript l to denote a linked node. Thus, $G_l \subset G$ denotes its network. We use n_l to denote the number of linked nodes. The proportion of all nodes that are linked nodes is measured by n_l/n.

The average degree of a network G is thus defined by $\eta(G) = \sum_{i \in N} \eta(i)/n$. There is a path between companies i and j if there is a direct connection, $G_{ij} = 1$, or if there is a set of distinct intermediate collaborators $k_1, k_2, \cdots, k_m$ such that $G_{ik_1} = G_{k_1k_2} = \cdots = G_{k_mj} = 1$. Two companies are *connected* if there is a path between them. A connected component is a set of companies that are connected to each other. In this section, we focus on the *giant component* of a network, which is the largest connected component in the network, among connected components. The giant component is denoted by $G_g \subset G$, and the number of member nodes is denoted by n_g. We measure the size of a giant component by the proportion of member companies that are linked to the network, which is given by n_g/n_l.

We compute the *clustering coefficient* (Watts and Strogatz, 1998) of company i, which is defined as:

$$C_i(G) = \frac{\sum_{j \in N(i)} \sum_{k \in N(i)} G_{jk}}{\eta_i(\eta_i - 1)}.$$

We set $C_i(G) = 0$ if $\eta(i) \leq 1$. This can be interpreted as the percentage of a company's collaborators who are collaborating with each other. For a network G, we can compute the *average clustering coefficient*, which is denoted by $C(G)$. Let d_{ij} denote the distance between two connected companies i and j in a network G, which is defined by the length of the shortest path between them. For the largest connected component of a network, G_g, we compute the *average distance*, which is defined by:[10]

$$d(G_g) = \frac{\sum_{i \in N_g} \sum_{j \in N_g} d_{ij}}{n_g(n_g - 1)}.$$

We also compute a nodal *centrality* (Bonacich, 1987) measure which is defined as

[10]The distance between two unconnected companies is not defined. Thus, when we compute the average distance, we measure the average distance of the companies in the largest connected component of the network G.

Table 2: The evolution of the R&D collaboration network structure

year	network size	linked node fraction	average degree	giant component fraction	average distance	clustering coefficient	variance nodal centrality
t	n	n_l/n	$\eta(G_l)$	n_g/n_l	$d(G_g)$	$C(G_l)$	$\mathrm{Var}(\phi)$
1985	1503	0.12	1.42	0.07	2.79	0.11	0.26
1986	1640	0.13	1.31	0.07	2.79	0.05	0.27
1987	1791	0.14	1.38	0.12	4.03	0.10	0.33
1988	1919	0.15	1.42	0.12	4.41	0.03	0.38
1989	2059	0.15	1.33	0.13	4.41	0.02	0.39
1990	2183	0.16	1.38	0.08	3.54	0.06	0.37
1991	2282	0.16	1.50	0.33	6.02	0.06	0.47
1992	2413	0.18	1.54	0.34	4.79	0.05	0.60
1993	2672	0.19	1.65	0.36	4.63	0.08	0.72
1994	3015	0.21	1.88	0.47	4.90	0.10	0.80
1995	3318	0.25	2.09	0.52	4.58	0.12	0.86

G_l represents the network of linked nodes.
G_g represents the network of the giant connected component.

follows:[11]

$$\phi_i = \sum_j (a + b\phi_j) G_{ij},$$

where a is a normalization scale factor such that $\sum_i \phi_i^2 = 1$ and b is a decay factor that scales down the relative weight of longer paths.[12] As the above definition suggests, the centrality of a node is recursively related to the sum of the centralities of the other nodes to which it is connected. Thus, a node that is connected to many well-connected nodes is assigned a high degree of centrality, whereas a node that is connected with only a few poorly connected nodes is assigned a low degree of centrality. To capture the heterogeneity of connectedness among the nodes, or the "core–periphery" structure of the network, we compute the variance of nodal centrality $\mathrm{Var}(\phi_i)$ to characterize the network.

Table 2 reports the basic statistics that describe the structure of R&D collaboration networks. It shows that the number of nodes has steadily increased, which suggests that an increased number of companies have applied for patents (and later successfully been granted those patents). This number more than doubled in the 10-year period under study.

[11] Ballester et al. (2006) have recently demonstrated a relationship between Bonacich centrality measures and the Nash equilibrium action of a player in a particular class of network games.

[12] Although the value of the decaying factor b can take any value between zero and unity, we follow the common practice of setting b to 0.1 (e.g., Haynie, 2001). Using different values of b does not greatly affect our results.

As for the patterns of connection between nodes, the reported fraction of linked nodes, (n_l/n), increased substantially from about 10 percent to about 25 percent in 10 years. The average degree, $\eta(G_l)$, which is computed for the linked nodes, increased significantly in the 1990s. This suggests that companies tend to collaborate with more companies over time. The growth of the connected component, measured by the fraction of the giant component, (n_g/n_l), expanded substantially in the 1990s as well. In fact, in the 1980s, the giant component was only less than 10 percent of the linked nodes but, by 1995, more than 50 percent of linked nodes were included in the giant component. These observations indicate that companies are increasingly being connected with more than half of all linked companies through collaboration chains.

The average distance between the nodes in the giant component is given by $d(G_g)$. The average distance shows an increasing trend in the 1980s and wide fluctuation in 1990 and 1991. However, it became rather stable between 4.5 and 5.0 after that. The size of the giant component grew *exponentially* at a rate of above 20 percent in 1990. Thus, in the 1990s, the average distance between nodes in the giant component grew more slowly than did the logarithm of the number of nodes. Growth in the average distance scales logarithmically with the network size represents evidence of a "small-world" effect (Newman, 2003). The implication of a small-world network is that because many pairs of companies are connected by a few collaboration links, the spread of information, or knowledge spillovers, is fast in the R&D network.

The degree of interconnectedness of collaboration is measured by the clustering coefficient. We report the clustering coefficient of the linked node, $C(G_l)$. The clustering coefficients reported in Table 2 are computed for linked nodes. The clustering coefficients are much higher than expected when connections are made randomly among the existing nodes in the network.[13] Generally, the clustering coefficient is high and has exhibited an increasing trend since 1992. This suggests that companies tend to collaborate with other companies that are located in their local circle.

The R&D network has two features that characterize the *small-world* network (Watts and Strogatz, 1998): (i) relatively short distances between nodes in the giant component, and (ii) a large clustering coefficient. This has been observed in various social networks, such as collaborations among economists (Goyal et al., 2006) and researchers in other fields (Newman, 2004). Deroian et al. (2007) shows similar properties for the R&D partnership networks in the pharmaceutical industry as well. [14]

The variance of nodal centrality, $\mathrm{Var}(\phi)$, increased substantially over the sample period (last column of Table 2). A low variance of centrality implies that the relative positions of nodes are similar to each other, whereas a high variance of centrality implies that there are small numbers of super-connected 'center' companies and, at the same time, large numbers of 'peripheral' companies with fewer connections. Thus, as collaborations have become more common, a few star companies have emerged in the network.

In summary, the key empirical regularities of the R&D network are as follows. First, the R&D network has become more extensive. The numbers of both nodes and

[13]The clustering coefficients of random graphs with the same number of linked nodes are less than 0.006 for the whole sampling period from 1985 to 1990, one order of magnitude lower than what is reported.

[14]Deroian et al. (2007) constructs their partnership network for 1985-2005 on the basis of the SDC Platinum database. They also report the decline of inter-firm alliances after 2000.

links have increased substantially. Second, innovative companies are becoming increasingly connected through R&D collaborations, and distances between them are shortening. Third, R&D alliances have become locally clustered, and companies tend to connect in local dense neighborhood circles. Fourth, the R&D network has developed unevenly, and a core–periphery structure has emerged.

4 Empirical Strategy

In this section, we turn to a statistical analysis of R&D collaboration. Although one could allow for new link formation, dissolution, and reformation of dissolved links at the same time, a dynamic model incorporating such simultaneous decision making about multiple choices is too complicated, as each decision may reflects varying consideration of firms. Thus, we focus on *newly established* collaborations and formulate the conditional probability of link formation between companies that have not collaborated before. In other words, we ignore company decisions about the maintenance and abolition of existing links as well as the recreation of links that have existed in the past. Given our interest in analyzing expanding R&D networks, conditioning on the formation of new links suits our purpose and simplifies the estimation of the structural parameters.

The mechanism underlying the formation of new R&D collaboration is considered to be a random matching process similar to that proposed by Jackson and Rogers (2007). Let us assume that companies that have never collaborated meet with each other and decide simultaneously whether to collaborate. Two types of matching processes are considered. First, matching occurs randomly at each date. Under the random matching process, the willingness of companies to collaborate may be affected by their exogenous attributes, such as the congruence of their research interests. However, when these characteristics have been controlled for, the chance of joint collaboration can be considered to be purely random and independent of the network architecture. We refer to such a process as collaboration through *random matching*. Second, matching occurs on a particular topology of the existing network. For example, companies may benefit from searching locally for potential collaborators and may collaborate with their neighbors in the current network. The network-matching process illustrates how the existing network structure influences the formation of R&D links. Such a process is referred to as collaboration through *network-based* matching.

In terms of the standard random-utility framework, the model of collaboration formation through random or network-based matching can be written as follows. Let us assume that companies i and j have not collaborated with each other before period t. Let $u_{ij}(t)$ be the latent utility of company i derived from initiating collaboration with company j at time t. We assume that the potential utility derived from the collaboration, u_{ij}, is represented by the following linear function of observed and unobserved terms:

$$u_{ij}(t) = \alpha + \beta W_{ij}(t-1) + \sum_k \rho_k d_{ij}^{k-1}(t-1) + \varepsilon_{ij}(t-1), \tag{1}$$

where $W_{ij}(t-1)$ represent the lagged background characteristics of individual com-

panies and pairs of companies. To be precise, this can be expressed as $W_{ij}(t-1) = [X_i(t-1)\ X_j(t-1)\ Z_{ij}(t-1)]$. Thus, $\beta W_{ij}(t-1) = \beta_1 X_i(t-1) + \beta_2 X_j(t-1) + \beta_3 Z_{ij}(t-1)$. The individual firm-specific variables, $X_i(t-1)$ and $X_j(t-1)$, are the lagged background characteristics of companies i and j respectively. The term $Z_{ij}(t-1)$ represents lagged pair-specific common attributes. These exogenous variables determine the systematic part of the random-matching collaboration mechanism. The idiosyncratic part is represented by an unobserved random error term, $\varepsilon_{ij}(t-1)$, which is assumed to be independent across pairs of companies and over time. The possibility that the random error terms are correlated is discussed later. For simplicity, we assume that the error has a logistic distribution.

In the latent-utility model described above, the network-based collaboration mechanism is represented by the term $\sum_k \rho_k d_{ij}^{k-1}(t-1)$. We define $d_{ij}^{k-1}(t-1)$ as a dummy variable that takes a value of unity if the shortest distance between i and j is equal to $k-1$ at period $t-1$ and is zero otherwise. If $\rho_k > 0$, the company obtains a positive benefit by forming a cycle of length k. Thus, the parameter ρ_k measures the degree to which there is a tendency to form the kth cyclic closure; this is henceforth termed the cyclic closure preference. If the cyclic closure preference is significant, the scope for new collaborations is restricted to a local chain of existing collaborations.

We allow another mechanism for network-based collaboration. Given the existence of knowledge spillovers through interfirm collaborations, companies may be more willing to make alliances with companies that have more collaborators and, thus, provide more knowledge spillovers. Firms' preference to be connected with other firms that have many connections is referred to as "preferential attachment" following Barabási and Albert (1999). Let us assume that the utility derived by company i from collaborating with company j is proportional to the potential partner's number of existing collaborators. Noting that the number of collaborators of company j at time $t-1$ is given by the degree, $\eta_j(t-1)$, a simple variant of Equation (1) is given by:

$$u_{ij}(t) = \alpha + \beta W_{ij}(t-1) + \lambda \eta_j(t-1) + \sum_k \rho_k d_{ij}^{k-1}(t-1) + \varepsilon_{ij}(t-1). \qquad (2)$$

Although the structural model of R&D collaboration is simple, it exhibits the macrodynamic features described in the previous section. Intuitive explanations are as follows. Network-based collaboration, arising either because of a preference for cyclic closure or a preferential attachment, may generate a feedback loop that facilitates further collaborations. For example, an increase in R&D collaboration shortens the link between any pair of companies and, thus, promotes further links between other companies through the cyclic closure preference. Similarly, the preferential attachment mechanism implies that an increase in the number of collaborations provides a conduit for information, such as knowledge and skills, from collaborators and, thus, stimulates further collaboration. These positive feedback mechanisms may generate a self-organizing expansion of R&D links over time. The small-world effect may be a product of preferential attachment because the prevalence of hubs can serve as a 'shortcut' that brings many nodes into close proximity. It is apparent that local clustering in the R&D network is a product of the cyclic closure preference. The emerging core–periphery structure may be driven by preferential attachment. This increases collaboration with

center nodes, which become increasingly connected, and leaves peripheral nodes relatively detached.

Without loss of generality, the utility derived from not collaborating can be normalized to zero. Thus, if $u_{ij}(t) > 0$, company i is willing to collaborate with company j at time t. Using Equation(2) to represent utility, the corresponding probability of collaboration is

$$\text{Prob}(u_{ij}(t) > 0) = F\left[\alpha + \beta W_{ij}(t-1) + \lambda \eta_j(t-1) + \sum_k \rho_k d_{ij}^{k-1}(t-1)\right], \quad (3)$$

where F is the logistic cumulative distribution function. Given this specification, the model assumes that a company's current collaboration decisions are affected by individual and joint background characteristics and existing network structures from the previous period.

We assume that both collaborators must *mutually* decide to collaborate for R&D collaboration to take place. That is, R&D collaboration only occurs if two companies are willing and agree to collaborate with each other *at the same time*. Let $G_{ij}(t)$ denote the R&D collaboration between company i and company j at time t. The conditional probability that companies i and j initiate collaboration at time t, given that they have not collaborated before, is given by

$$\text{Prob}(G_{ij}(t) = 1 | G_{ij}(s) = 0; s < t) = \text{Prob}(u_{ij}(t) > 0) \cdot \text{Prob}(u_{ji}(t) > 0). \quad (4)$$

Equality follows from the assumption that $\varepsilon_{ij}(t-1)$ and $\varepsilon_{ji}(t-1)$ are independent.

Given the above specification, the likelihood of collaboration for all possible pairs of companies, which is defined by

$$L(\theta; t) = \prod_{i<j} \text{Prob}(G_{ij}(t) = 1 | G_{ij}(s) = 0; s < t), \quad (5)$$

can be written as

$$L(\theta; t) = \prod_{i \neq j} \text{Prob}(u_{ij} > 0). \quad (6)$$

This equation follows because u_{ij} and u_{ji} are symmetric. Equation (3) implies that the overall sample log-likelihood function is given by

$$\ell(\theta) = \sum_t \sum_{i \neq j} \ln F\left[\alpha + \beta W_{ij}(t-1) + \lambda \eta_j(t-1) + \sum_k \rho_k d_{ij}^{k-1}(t-1)\right]. \quad (7)$$

The structural parameter $\theta = (\alpha, \beta, \lambda, \rho_k)$ is estimated from the log-likelihood function. Because F is the logistic cumulative distribution function, the standard logistic regression method is used for the estimation. Note that the product relates to all possible pairs of (i, j) such that $i \neq j$. Hence, the observation unit for the log-likelihood function is each pair of companies, (i, j), not the individual companies, i or j. For maximum likelihood estimation, the sample size, N, refers to all possible pairs of companies.

As a final note regarding the statistical model, we discuss the identification of some structural parameters. It transpires that β_1 and β_2, which relate to individual-specific characteristics, cannot be separately identified from the data. The simple reason is that because these parameters are *symmetric* in the log-likelihood function, any two sets of symmetric values of (β_1, β_2), such as (b_1, b_2) and (b_2, b_1), yield the same log-likelihood value for Equation (7). Thus, in the following empirical analysis, we simply report $\beta = (\beta_1 + \beta_2)/2$. The parameter β can be interpreted as the average effect on link formation of a company's and its collaborator's background characteristics.

5 Empirical Results

5.1 Estimation Sample

As previously reported, the main dataset used for estimation is constructed from the NBER patent data. In addition, we use the S&P's COMPUSTAT database to supplement detailed information about individual companies.

Two datasets are considered in the following estimation. First, we collect the data of all IT companies, defined by the technological categories of patents, that applied for at least one patent in each year of the observation period. Second, we use the subsamples of the IT companies that are listed on the stock markets of NYSE, NASDAQ or AMEX. The first dataset covers substantially more companies than the second one, but it does not contain detailed information on the firm characteristics because COMPUSTAT information is not available for companies that are not listed on the stock markets. In what follows, we call the former *the full set of companies'* data, and call the latter as the *market-listed companies'* data.

Both datasets contain the information about a pair of companies $(i\ j)$ across time t. The dependent variable $G_{ij}(t)$ is a dummy variable that takes one if a new link is observed between company i and j in year t. As explained in the model section, since we focus on *newly established* collaborations, we only collect pairs of companies that had *never* collaborated before for both datasets. [15] Hence, once companies collaborate, they are excluded from the estimation samples for subsequent periods. The data also contain a set of independent variables $W_{ij}(t-1)$, which are the lagged background characteristics of individual companies and pairs of companies. In what follows, a time index is omitted for notational convenience.

We categorize the control variables W_{ij} into three groups. The first group is a set of variables that measures the "absorptive capacity" (Cohen and Levinthal, 1989) of innovating companies, which has been shown to affect companies' R&D cooperation decisions in previous studies. We consider both the *research size* and *research quality* to measure the absorptive capacity. As for the "size" of research, we use the ***R&D expenditure***, which is available from the COMPUSTAT database for the market-listed companies. On the other hand, the variable is not available for non-marketed-listed companies, which account for the majority of the full set of companies' data. Therefore, we instead use the *number of patents* for which the company applied in a given year as a proxy for the company's R&D size. This variable is available for the full set of

[15] To check for collaborations prior to the sample period 1985–1995, we tracked back to 1975.

companies. As for the "quality" of research, we take the sum of the *cumulative number of patents* applied by the company and the *cumulative number of patent citation* referred to by other companies. For these two variables, we aggregate the information over the past 10 years. This variable, constructed from the NBER patent database, is available for both market-listed and non-market-listed companies; thus, it is included in both the market-listed companies' data and the full set of companies' data. It is noteworthy that these research size and quality variables are included in the form of the natural logarithm in the regressions.

The second group is related to similarity between two companies, which is defined in terms of technology, R&D size, and location. First, regarding *technological similarity*, we essentially treat the research areas of two companies as similar if the companies applied for patents in similar technological fields. Following Jaffe (1986), we measure technological similarity, denoted by TS_{ij}, by using an uncentered correlation of patent application subcategories between companies i and j.[16] The second similarity has to do with research size. The similarly, denoted by RS_{ij}, is given by the inverse of the absolute difference in a measure of absorptive capacity.[17] Finally, locational similarity is considered as many previous studies have stressed that geographical origin is an important factor in collaboration decisions. We define the locational similarity measure, denoted by LS_{ij}, to be equal to one if the companies' headquarters locate in the same state and zero otherwise. For the market-listed companies' data, the headquarters information is available from COMPUSTAT data. For the full set of companies' data that include both market-listed and non-market-listed companies, we impute the state of the headquarters by the state that appears most frequently as inventors' addresses which are available from the NBER patent data.

The third and the most important group of variables is related to network structure. As shown in the previous section, this includes the *cyclic closure structure* variables and the *star structure* variable, which represent a firm's preference to form a cyclic closure and a preferential attachment tendency, respectively. To capture the kth cyclic closure preference effect, we compute the shortest (geodesic) path between innovating companies in the R&D collaboration network and construct a dummy variable, d_{ij}^{k-1}, which is equal to unity if the shortest distance between companies i and j is $k-1$ and zero otherwise. For convenience, we include up to the sixth ($k = 6$) cyclic closure preference because of the rarity of network closures to more than the fifth degree. To

[16] To be precise, the correlation coefficient, TS_{ij}, is defined as:

$$TS_{ij} \equiv \frac{f_i C f_j'}{[(f_i C f_i')(f_j C f_j')]^{1/2}},$$

where f_i is a row vector of the number of patent applications in each technological subcategory taken out by company i and C is the citation probability matrix for each technological subcategory computed from all U.S. patent citation data from 1981 to 1999. We use the citation probability matrix, rather than the identity matrix, as the weight so that our similarity measure reflects the similarities between technological categories.

[17] The researsh size similarity is defined by

$$RS_{ij} = \frac{1}{1 + |R_i - R_j|},$$

where R_i and R_j are the research size measures of companies i and j, respectively. This similarity measure is inspired by the "social distance" measure that was proposed by Akerlof (1997).

Table 3: Definitions of Controlled Variables

Variable Name	Description
Absorptive Capacity Variables:	
RD	Total R&D expenditure in millions of U.S. dollars in the last period.
NPATENT	The number of patent applications in the last period.
CUMPATENT	The cumulative number of patent applications by the last period.
CITED	The cumulative number of cited patents by the last period.
Similarity Variables:	
TS	Technological similarity, defined by the uncentered correlation of patent application subcategories.
RS	Research size similarity, defined by the inverse of the difference in R&D size.
LS	Locational similarity, defined to be one if the states of the companies 'headquarters coincide.
Network Variables:	
d^{k-1}	A dummy variable that takes one if the shortest distance between a pair of companies is $k-1$ in the last period.
η	The number of collaborations in the last period.

capture the preferential attachment effect, we include the natural log of the nodal degree of R&D collaborations, which is denoted by $\log(\eta_i)$ for company i. To account for the likelihood contribution of network formation between companies i and j, the degree variables of both companies, $(\log(\eta_i)\ \log(\eta_j))$, are included. Yet, as already noted, because the effects of own-company and partner characteristics cannot be separately identified simultaneously, the average of the preferential attachment effect of its own and that of the partner company is reported.

Table 3 presents the definitions of the control variables for the two datasets. In addition to the variables explained above, we include trend variables with a base year of 1985. Table 4 presents the descriptive statistics of the dependent and independent variables used in our analysis.

5.2 Baseline Estimation Results

Table 5 reports the estimation results. Column (1) presents the estimates for the market-listed companies' data while column (2) presents the estimates for the full set of companies' data. The robust standard error of each estimated coefficient is reported in parentheses.

We find that the estimated coefficients of most variables have the expected signs. The estimates of the absorptive-capacity-related variables, $\log(RD)$, $\log(NPATENT)$ and $\log(CUMPATENT + CITED)$, show that the variable representing research size and research quality has positive and statistically significant effects on collaboration formation. This evidence is found consistently in both the market-listed companies' data (column (1)) and the full set of companies data (column (2)). Given these facts, we conclude that the absorptive capacity is an important factor that facilitates R&D collaboration between companies.

As for similarity-related variables, we find that technological similarity, TS, is negative and statistically significant in column (1), although the effect is not statistically

Table 4: Summary Statistics

	Market-listed Companies	Full set of Companies
Dependent Variable:		
G	0.0010	0.0002
	(0.0310)	(0.0125)
Research Size:		
$\log(RD)$	3.6364	
	(2.1399)	
$\log(NPATENT)$		3.2323
		(1.6556)
Research Quality:		
$\log(CUMPATENT + CITED)$	5.3496	4.6677
	(2.2283)	(1.9123)
Technological Similarity:		
TS	0.1738	0.3291
	(0.1845)	(0.2233)
Research Size Similarity:		
RS	0.3805	0.4512
	(0.2110)	(0.2155)
Locational Similarity:		
LS	0.1068	0.0035
	(0.3089)	(0.0591)
Cyclic Closure:		
third degree ($k = 3$)	0.0044	0.0010
	(0.0658)	(0.0316)
fourth degree ($k = 4$)	0.0068	0.0022
	(0.0821)	(0.0465)
fifth degree ($k = 5$)	0.0055	0.0027
	(0.0740)	(0.0523)
sixth degree ($k = 6$)	0.0038	0.0022
	(0.0615)	(0.0473)
Nodal Degree :		
$\log(\eta)$	0.3925	0.3003
	(0.6603)	(0.5039)
Sample Size	185619	5494790

significant in column (2). This suggests that, at least for market-listed companies, innovative synergy would not arise very easily if two companies had similar research areas, and such companies would collaborate less often than those that do not share research fields. However, it is consistently shown in both columns (1) and (2) that the research size similarity, RS, has a significant positive impact on collaboration. Interestingly, the relationship between similarity and collaboration is non-monotonic. Seeing from the negative estimates of the quadratic term of the research size similarity, it can be interpreted that companies may interact with other companies that are similar, but not too similar, in research capacity.

It is also shown that the locational similarity, LS, has a statistically significant effect on collaboration. This finding is robust for the results of both columns (1) and (2). These findings suggest that companies with headquarters in the same state are more likely to collaborate with each other. It is noteworthy, however, that the locational similarity variable in our analysis only measures the proximity of company headquarters on the state level. Therefore, it may be more appropriate to focus attention on the proximity of company research units or business establishments. Thus, further investigation with a focus on location will be necessary.

The estimates of the network variables indicate that the cyclic closure effects, ρ_k,

Table 5: Baseline Regression Results

	Market-listed Companies (1)	Full set of Companies (2)
CONSTANT	-11.7681^{***}	-12.7136^{***}
	(0.7349)	(0.4257)
Research Size:		
$\log(RD)$	0.2203^{***}	
	(0.0704)	
$\log(NPATENT)$		0.2848^{***}
		(0.0414)
Research Quality:		
$\log(CUMPATENT + CITED)$	0.2078^{***}	0.1831^{***}
	(0.0754)	(0.0304)
Technological Similarity:		
TS	-2.5537^{**}	-0.0949
	(1.0037)	(0.5486)
TS^2	1.9821	0.0611
	(1.4185)	(0.6526)
Research Size Similarity:		
RS	6.2846^{***}	4.9020^{***}
	(1.7398)	(1.1125)
RS^2	-4.6214^{***}	-4.8361^{***}
	(1.6377)	(1.0481)
Locational Similarity:		
LS	0.8560^{***}	0.8439^{***}
	(0.1817)	(0.2576)
TREND	0.0222	0.0567^{***}
	(0.0359)	(0.0165)
Cyclic Closure Effect ρ:		
third degree ρ_3	2.5991^{***}	3.4289^{***}
	(0.2489)	(0.1440)
fourth degree ρ_4	2.1358^{***}	2.4342^{***}
	(0.2439)	(0.1475)
fifth degree ρ_5	2.2072^{***}	1.7809^{***}
	(0.2869)	(0.1892)
sixth degree ρ_6	1.2148^{**}	1.8651^{***}
	(0.5245)	(0.2156)
Preferential Attachment Effect λ:		
$\log(\eta)$	0.4946^{***}	0.3790^{***}
	(0.1533)	(0.0721)
Log-likelihood	-1090.4146	-7150.1950
Sample Size	185619	5494790

are positive and statistically significant for all k at the one percent significance level. This evidence is consistently found in columns (1) and (2). Furthermore, the triadic closure effect ρ_3 has the largest effect among other cyclic closure effects. This suggests that companies that have mutual third-party collaborators in their network neighborhood tend to establish a new R&D collaboration. The mechanism seem to work consistently, irrespective of whether or not they are listed on the stock market.

We consider at least two reasons that R&D collaborations are more likely to be undertaken between companies that have mutual third-party collaborators. First, previous collaborators may play a referral role for new collaborators. Companies would probably gain by reducing any uncertainty about whether a potential partner would behave opportunistically. [18] Given such uncertainty, companies may want to use preexisting

[18]For example, a potential collaborator may free-ride by limiting contributions to a collaboration and/or may take advantage of a close relationship to use resources or information in ways that may damage their interests.

collaboration as conduits of information about the reliability of potential partners, and thus, they would prefer to establish "secure" relationships with known local partners in their close circle rather than begin relationships with new collaborators with unknown behavior.[19] Second, it may be motivated by the appropriability of the fruits of joint R&D, about which companies might be seriously concerned when starting joint R&D projects (Cassiman and Veugelers, 2002). To prevent collaborators from appropriating the technology generated by joint R&D projects for their own interests against the interests of others, collaborating firms must be able to punish deviators.[20] The existence of an indirect link through mutual collaborators may enhance the effectiveness of penalties and improve the appropriability of the fruits of joint R&D projects.[21]

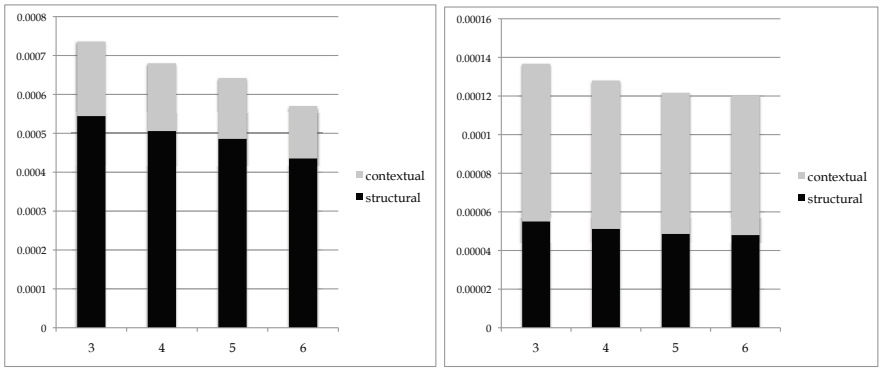
We also find in Table 5 that the preferential attachment effect, λ, is positive and statistically significant at the one percent level in columns (1) and (2). This finding reinforces our argument that the attractiveness of collaboration is determined by the resources and capabilities that a potential partner possesses. It has been argued that collaboration ties are key vehicles through which companies obtain access to external knowledge, including technical breakthroughs and new insights to problems and failures (e.g., Powell et al., 1996). Our estimation results confirm that access to the resources of potential partners may promote R&D collaboration. In addition, companies that have knowledge repositories from previous collaborations are likely to attract more collaborators and, thus, more opportunities for collaboration. However, it is also possible that companies' unobserved capabilities for R&D collaboration are captured by the preferential-attachment variable.

The results presented above clearly indicate that two types of mechanisms may explain R&D collaboration formation between IT companies. First, companies decide to establish a new R&D collaboration with each other because of their cyclic closure preference or preferential attachment predisposition. Such a collaboration mechanism is considered as *structural* or *endogenous*. Secondly, companies collaborate "contextually" with each other even without taking into account the generated network structure. For example, as shown in estimation results, companies with similar research size, or companies that are located in the same state, will collaborate with each other. Such collaboration due to characteristic similarities is considered *contextual* or *exogenous*. The distinction between the structural and contextual mechanisms is important. Whereas the structural effect generates an additional mechanism that facilitates further collaborations with others, the contextual effect does not.

[19] The same idea is found in Gulati and Gargiulo (1999).

[20] For example, if two companies are collaborating on an R&D project and this is the only connection between them, then the only way to penalize malicious technology diverters is to expel them from current collaborations and exclude them from future ones. However, if these two companies are connected not only directly but also indirectly through mutual third-party collaborators, it is possible for deviators to be punished by all other collaborators who are aware of such behavior; deviators could be punished by being ostracized from the alliance comprising third-party collaborators.

[21] A similar sanctioning logic has been widely proposed by several sociologists (e.g., Coleman, 1990; Putnam, 1992; Walker et al., 1997), who hypothesize that such punishments are more easily enforced among agents belonging to a closely knit network. In the literature of economics, Spagnolo (1999) uses a repeated-game framework to analyze transfers of trust from social to production relationships that facilitate cooperation in production and shows that the amount of credible punishment, as embodied in social capital, strengthens cooperation. Lippert and Spagnolo (2004) provides rigorous support for the sanctioning role of network closure by using a game-theoretic model of networks of relational contracts.



(a) market-listed companies' data (b) all companies' data

Figure 2: Decomposition of the Collaboration Probability into Structural and Contextual Effects

To disentangle these effects, we decompose the collaboration probability into two components: (i) the structural effect, which arises because of cyclic closure preferences or preferential attachment, and (ii) the contextual effect, which arises because of non-structural factors. Figure 2 presents the results of the decomposition for the market-listed companies' data (a) and the full set of companies' data (b). The height of each bar graph represents the average predicted probability that a new R&D collaboration is established between two companies that never collaborated before and is computed with the parameter estimates presented above. For figures (a) and (b), we present the predicted probabilities with up to $d = 5$ network distances. The structural part of the collaboration probability between companies with d network distance is computed by the probability that is attributable to the cyclic closure preference with $(d + 1)$ degree after controlling for the effect of similarity-related variables so that the effect is zero. The contextual part is given by the residual part of the probability that cannot be explained by the structural part.

The figures show that, for the market-listed companies' data (a), the structural effect on collaboration formation exceeds the contextual effect, while the magnitude relationship is reversed for the full set of companies' data (b). It thus can be said the R&D collaboration formation of market-listed companies is driven, for the most part, by the structural mechanism that facilitates further collaboration formation in a self-fulfilling manner. In contrast, the structural mechanism is less important than the contextual mechanism for non-market-listed companies.

In summary, we have established three main findings. First, a company's absorptive capacities are important determinants of R&D collaboration. Second, companies with similar, but not too similar, research capacity are highly likely to collaborate with each other. Third, network-based collaboration is significant. Companies with more collaborators are more likely to collaborate with each other. Furthermore, companies that are closer in the existing R&D network are more likely to collaborate with each other.

5.3 Controlling for Unobserved Common Factors

Although evidence of a strong cyclic closure preference is found, there may be an omitted-variables problem. If there are unobserved common shocks that affect the decisions of a *group* of companies, the effect of those omitted factors might be incorporated into the cyclic closure preference effect. For example, let us assume that a company's research managers formerly worked for a nearby company. The two companies may be more likely to collaborate in R&D activities because of this personal association.[22] If the predisposition of research managers is not observed by researchers, the effect may be mistakenly attributed to a cyclic closure preference.

To address this problem, we modify the model to incorporate a group-specific unobserved factor. The primary assumption underlying this specification is that such common factors, which are unobservable to researchers, affect all companies in the same group. Let us assume that company i and company j both belong to group g. Then, the utility function is

$$u_{ij}(t) = \alpha + \beta W_{ij}(t-1) + \lambda \eta_j(t-1) + \sum_k \rho_k d_{ij}^{k-1}(t-1) + \delta_g(t-1) + \varepsilon_{ij}(t-1), \quad (8)$$

where $\delta_g(t-1)$ is a fixed effect that is common to companies i and j belonging to group g; other companies probably also belong to the group. The fixed effect, δ_g, if not taken into account, causes the overall error term, $\delta_g + \varepsilon_{ij}$, to be correlated between the companies in group g.

The omitted group-specific factor can arise from the endogenous group formation of companies based on unobserved characteristics. Companies are likely to sort themselves into groups in which members share common unobserved characteristics. Hence, the group structure should capture hidden sorting across companies and their shared attributes. We employ a simple strategy to identify hidden group structures. In other words, each group is identified by a "community structure", as referred to by Girvan and Newman (2004). Essentially, a community is a subset of nodes within a network such that connections between them are deeper than connections to other nodes in the network.[23] Our group identification strategy relies on the assumption that companies are densely connected with each other because they share common unobserved group-level attributes. Using the algorithm proposed by Girvan and Newman (2004), we identify 161 and 48 communities for the all companies' dataset and market-listed companies' dataset, respectively, in the sample period between 1985 and 1995. The fixed effect term δ_g is assigned to each pair of companies i and j if both companies belong to group g.

[22] See Saxenian (1994) for anecdotal evidence in this regard.

[23] Girvan and Newman's community detection algorithm is based on the idea of the "betweenness" of links in the network, in which this betweenness is a measure that emphasizes links between communities. They propose using a modularity index to determine the number of communities. The modularity index measures the difference between the proportion of edges in the network that connect nodes within the same community and the expected value of the proportion of edges that have the same community division but in which the connections between nodes are random. We use the community structure that provides the *highest* modularity.

Table 6: Community Fixed Effects Regression Estimation Results

	Market-listed Companies (3)	Full set of Companies (4)
Research Size:		
$\log(RD)$	0.2245^{***}	
	(0.0755)	
$\log(NPATENT)$		0.2731^{***}
		(0.0381)
Research Quality:		
$\log(CUMPATENT + CITED)$	0.2215^{***}	0.1882^{***}
	(0.0679)	(0.0255)
Technological Similarity:		
TS	-2.6361^{**}	-0.1885
	(1.1783)	(0.4959)
TS^2	1.7260	0.0470
	(2.0805)	(0.5643)
Research Size Similarity:		
RS	5.9228^{***}	4.9421^{***}
	(1.6811)	(1.4123)
RS^2	-4.2508^{***}	-5.2924^{***}
	(1.6148)	(1.2646)
Locational Similarity:		
LS	0.8541^{***}	0.2633
	(0.1848)	(1.0602)
TREND	0.0372	0.0750^{***}
	(0.0402)	(0.0200)
Cyclic Closure Effect ρ:		
third degree ρ_3	2.5944^{***}	3.2356^{***}
	(0.2717)	(0.1622)
fourth degree ρ_4	2.1589^{***}	2.3347^{***}
	(0.2599)	(0.1552)
fifth degree ρ_5	2.2452^{***}	1.7691^{***}
	(0.2816)	(0.2100)
sixth degree ρ_6	1.2261^{**}	1.8549^{***}
	(0.5216)	(0.1851)
Preferential Attachment Effect λ:		
$\log(\eta)$	0.4463^{***}	0.3863^{***}
	(0.1548)	(0.0577)
Log-likelihood	-1049.9447	—
Sample Size	185228	10% bootstrap

Given the utility specification by Equation (8), the log-likelihood function for new collaboration links is derived analogously to Equation (7). Because the cumulative distribution function of the error term is assumed to be logistic, the fixed effects logit estimator is used to estimate the model.

Table 6 reports the estimation results for the fixed effects model. Column (4) presents the estimates for the market-listed companies' dataset. We use the same variable specification as that used for the baseline estimation. For the data of market-listed companies, 6 community-specific fixed effects are included in the estimation.[24]

We also tried the same fixed effects estimation to the full set of companies' data. However, because of the immense sample size ($N = 5,494,790$) and the lack of machine power, the estimation could not be implemented. We, thus, instead used 10% bootstrapped samples of the original sample and estimated the fixed effects model for the subsample. We repeated this subsample estimation procedure for 50 times with the subsample being bootstrapped each round and computed the average of the parameter estimate. Column (4) in Table 6 reports the bootstrap estimates, and the bootstrap estimates of the standard errors are shown in parentheses. [25]

It is shown that the coefficients of most of the variables are similar to those from the baseline specification but less precisely estimated. The qualitative results are the same as before with one exception: the variables measuring locational similarity are not significant for the full set of companies' data.

The estimated effects of the network variables are statistically significant and have the predicted signs. The effect of preferential attachment remains positive and significant. The point estimates are similar to those from the baseline specification. Hence, the inclusion of group-specific effects does not negate these network effects. This suggests that closure preference is not primarily driven by unobserved group characteristics. Furthermore, triadic closure preference continues to have the strongest effect. Thus, provided we are allowed to assume that most of the omitted variables only vary at the community level, our results suggest evidence of positive triadic closure preference.

6 Conclusion

In this paper, we have studied the evolution of successful R&D collaboration in the U.S. IT industry between 1985 and 1995 by using information on patents granted in the U.S. The descriptive statistics on R&D networks provide insights into how networks have evolved. Put simply, collaboration has become more extensive, more locally clustered, connected over shorter distances, and increasingly unequal. Considering the laws that guide these macrodynamic features, through regression analysis, we found, having controlled for as many company characteristics as possible, that the choice of collaboration partners is significantly affected by cyclic closure preference and preferential

[24] The fixed effects of communities in which no collaboration links are observed are dropped from the estimation sample. To estimate the fixed effects model, within-group variation in the response variable is required.

[25] We employ the subsampling bootstrap method proposed by Politis and Romano (1994). In the subsampling bootstrap procedure, resampling should be implemented *without replacement*. For details, see Theorem 2.1 in Politis and Romano (1994) and Politis et al. (1999).

attachment preference. Our results support the following two hypotheses relating to interfirm collaboration formation: (1) the local partner search hypothesis, according to which *the connected get more interconnected*; and (2) the accumulation advantage hypothesis, according to which *the connected get more connected*.

Our paper makes a unique contribution to the literature on R&D alliances by measuring the impact of existing collaboration structures on the formation of new collaborations and by explaining how the collaboration network emerges. Although many previous studies have analyzed interfirm collaboration structures that can enhance innovation, none has explained how the collaboration network can be systematically structured among companies that are uncoordinated *ex ante*. Our findings suggest that firms search for reliable collaboration partners through referrals from previous collaborators. The results also suggest that companies attempt to form dense webs of local collaborators to increase the threat of joint punishment to minimize the risk of malicious behavior by their collaboration partners. However, our results do not identify the separate roles of these factors in explaining the behavior of firms. More theoretical work is needed to identify the different dynamic structural effects produced by firms' networking strategies.

By focusing on the formation of new collaborations, we did not analyze the patterns and determinants of the dissolution of existing collaborations. The question of how companies decide to terminate existing partnerships and start new ones was not examined. We have based our empirical framework on rather myopic and adaptive agents making decision about with whom to make a new link. Forward-looking agents can be considered in a more dynamic context. In such a set-up, there is possible endogeneity between a firm's current position in the network and its future decision about the termination of existing links or the creation of new links. Future research on both the creation and dissolution of ties must also address this endogeneity issue as well to gain a full understanding of the dynamics of networks.

Although we have information on the intensity of collaboration measured by the number of researchers common to both companies, we did not utilize it in our analysis. We have treated all links as homogeneous and abstracted from the strength of ties and their types. The strength and types of ties, however, are known to play important roles in information flows (See Granovetter (1973) for an example from the labor market). The manner in which the strength and type of existing links will affect the creation and termination of ties needs to be examined. The dynamics of the strength of the ties will need to be examined as well. The first step to answer such questions is to establish a framework for analyzing the dynamics of weighted, as opposed to unweighted, networks. We leave this important and interesting task to future research.

References

Ahuja, G., 2000. Collaboration networks, structural holes, and innovation: A longitudinal study. Administrative Science Quarterly 45, 425–455.

Akerlof, G. A., 1997. Social distance and social decisions. Econometrica 65, 1005–1027.

Almeida, P., Kogut, B., 1999. Localization of knowledge and the mobility of engineers in regional networks. Management Science 45, 905–917.

Ballester, C., Calvó-Armengol, A., Zenou, Y., 2006. Who's who in networks. wanted: The key player. Econometrica 74, 1403–1417.

Barabási, A.-L., Albert, R., 1999. Emergence of scaling in random networks. Science 286, 509–512.

Bloch, F., 2005. Group and network formation in industrial organization: A survey. In: Demange, G., Wooders, M. (Eds.), Group Formation in Economics: Networks, Clubs and Coalitions. Cambridge University Press, New York, pp. 335–353.

Bonacich, P., 1987. Power and centrality - a family of measures. American Journal Of Sociology 92, 1170–1182.

Branstetter, L. G., Sakakibara, M., 2002. When do research consortia work well and why? evidence from japanese panel data. American Economic Review 92, 143–159.

Cantner, U., Graf, H., 2006. The network of innovators in jena: An application of social network analysis. Research Policy 35, 463–480.

Cassiman, B., Veugelers, R., 2002. R&d cooperation and spillovers: Some empirical evidence from belgium. American Economic Review 92, 1169–1184.

Cohen, W. M., Levinthal, D. A., 1989. Innovation and learning: The two faces of r&d. The Economic Journal 99, 569–596.

Coleman, J. S., 1990. Foundations of Social Theory. Harvard University Press, Cambridge.

Cowan, R., Jonard, N., 2004. Network structure and the diffusion of knowledge. Journal of Economic Dynamics & Control 28, 1557–1575.

Deeds, D. L., Hill, C. W. L., 1998. An examination of opportunistic action within research alliances - the analysis of discrete structural alternatives. Journal of Business Venturing 14, 141–163.

Deroian, F., M'Chirgui, Z., Milelli, C., 2007. Evidences on inter-firm r&d partnerships in three high-tech industries. *GREQAM Working Paper* no. 2007-09, GREQAM.

Girvan, M., Newman, M. E. J., 2004. Finding and evaluating community structure in networks. Physical Review E 69, 026113.

Gomes-Casseres, B., Hagedoorn, J., Jaffe, A. B., 2006. Do alliance promote knowledge flows? Journal of Financial Economics 80, 5–33.

Goyal, S., Joshi, S., 2003. Networks of collaboration in oligopoly. Games and Economic Behavior 43, 57–85.

Goyal, S., Moraga-Gonzalez, J. L., 2001. R&d network. The RAND Journal of Economics 32, 686–707.

Goyal, S., van der Leij, M. J., Moraga-González, J., 2006. Economics: Emerging small world. Journal of Political Economy 114, 403–412.

Granovetter, M., 1973. The strength of weak ties. American Journal of Sociology 78, 1360–1380.

Gulati, R., 1995. Social structure and alliance formation patterns: A longitudinal analysis. Administrative Science Quarterly 40, 619–652.

Gulati, R., Gargiulo, M., 1999. Where do interorganizational networks come from? American Journal of Sociology 104, 1439–1493.

Hagedoorn, J., 2002. Inter-firm r&d partnerships: an overview of major trends and patterns since 1960. Research Policy 31, 477–492.

Hall, B. H., Jaffe, A. B., Trajtenberg, M., 2001. The nber patent citations data file: Lessons, insights and methodological tools. National Bureau of Economic Research, Inc, NBER Working Papers: 8498.

Haynie, D. L., 2001. Delinquent peer revisited: Does network structure matter? American Journal of Sociology 106, 1013–1057.

Hernán, R., Marín, P. L., Siotis, G., 2003. An empirical evaluation of the determinants of research joint venture formation. Journal of Industrial Economics 51, 75–89.

Hoisl, K., 2007. Tracing mobile inventors – the causality between inventor mobility and inventor productivity. Research Policy 36, 619–636.

Jackson, M. O., 2006. The economics of social networks. In: Blundell, R., Newey, W., Persson, T. (Eds.), Advances in Economics and Econometrics, Theory and Applications: Ninth World Congress of the Econometric Society. Vol. 1. Cambridge University Press.

Jackson, M. O., Rogers, B. W., 2007. Meeting strangers and friends of friends: How random are social networks? forthcoming in American Economic Review.

Jaffe, A. B., 1986. Technological opportunity and spillovers of r&d: Evidence from firms' patent, profits, and market value. American Economic Review 76, 984–1001.

Jaffe, A. B., Trajtenberg, M., Henderson, R., 1993. Geographic localization of knowledge spillovers as evidenced by patent citations. Quarterly Journal of Economics 108, 577–598.

Levin, R. C., Klevorick, A. K., Nelson, R. R., Winter, S. G., 1987. Appropriating the returns from industrial research and development. Brookings Papers on Economic Activity 3, 783–831.

Lippert, S., Spagnolo, G., 2004. Networks of relations. Mimeo, University of Mannheim Economic Department.

Marx, M., Strumsky, D., Fleming, L., 2007. Noncompetes and inventor mobility: Specialists, stars, and the michigan experiment. *HBS Working Paper* no. 07-042, Harvard Business School.

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

Meagher, K., Rogers, M., 2004. Network density and r& d spillovers. Journal of Economic Behavior & Organization 53, 237–260.

Newman, M. E. J., 2003. The structure and function of complex networks. SIAM Review 45, 167–256.

Newman, M. E. J., 2004. Coauthorship networks and patterns of scientific collaboration. Proceedings of the National Academy of Sciences of the United States of America 101, 5200–5205.

Politis, D. N., Romano, J. P., 1994. Large sample confidence regions based on subsamples under minimal assumptions. Annals of Statistics 22, 2031–2050.

Politis, D. N., Romano, J. P., Wolf, M., 1999. Subsampling. Springer Series in Statistics. Springer-Verlag, New York.

Powell, W. W., Koput, K. W., Smith-Doerr, L., 1996. Interorganizational collaboration and the locus of innovation: Networks of learning in biotechnology. Administrative Science Quarterly 41, 116–145.

Powell, W. W., White, D. R., Koput, K. W., Owen-Smith, J., 2005. Network dynamics and field evolution: The growth of interorganizational collaboration in the life sciences. American Journal of Sociology 110, 1132–1205.

Putnam, R. D., 1992. Making Democracy Work. Princeton University Press, Princeton.

Röller, L.-H., Tombak, M. M., Siebert, R., 1998. The incentives to form research joint ventures: Theory and evidence. Wissenschaftszentrum Berlin (WZB), Working Papers.

Saxenian, A., 1994. Regional Advantage: Culture and Competition in Silicon Valley and Route 128. Harvard University Press.

Schankerman, M., Shalem, R., Trajtenberg, M., 2006. Software patents, inventors and mobility. mimeo.

Scherer, F. M., 1965. Firm size, market structure, opportunity, and the output of patented inventions. American Economic Review 55, 1097–1125.

Schilling, M. A., Phelps, C. C., 2007. Interfirm collaboration networks: the impact of small world connectivity on firm innovation. forthcoming in Management Science.

Schmookler, J., 1966. Invention and Economic Growth. Harvard University Press, Cambridge.

Singh, J., 2005. Collaborative networks as determinants of knowledge diffusion patterns. Management Science 51, 756–770.

Spagnolo, G., 1999. Social relations and cooperation in organizations. Journal of Economic Behavior & Organization 38, 1–25.

Trajtenberg, M., Shiff, G., Melamed, R., 2006. The "names game": Harnessing inventors' patent data for economic research. NBER Working Papers 12479.

Walker, G., Kogut, B., Shan, W., 1997. Social capital, structural holes and the formation of an industry network. Organization Science, 8, 109–125.

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

Acknowledgment

We would like to thank for vaulable comments from Gueorgi Kossinets, Hidehiko Ichimura, Daiji Kawaguchi, Tomoya Mori, Ryo Kambayashi, Reiko Aoki, Hiroyuki Chuma, Sadao Nagaoka, and anonymous referees. Of course, the responsibility of all remaining errors is ours. Financial support from the Japan Securities Scholarship Foundation, Inamori Foundation and the Kikawada Foundation (21-seiki bunka gakujyutsu zaidann) are gratefully acknowledged. This research was also supported by Grant-in-Aid for Young Scientists (B)awarded from The Ministry of Education, Culture, Sports, Science, and Technology of Japan (Grant. No. 19730165)