Dynamics of R&D Collaboration in IT Industry*

Nobuyuki Hananki † Ryo Nakajima‡ Yoshiaki Ogura §

February 6, 2007

Abstract

This paper provides an empirical analysis of evolving networks of successful R&D collaborations in the IT industry in the United States between 1985 and 1995. We first show that the network has become more extensive, more clustered, and more unequal in the sense "stars" have emerged in the network. We then perform regression analysis in which we control for firm similarity, including unobserved similarities that we infer from the community structure of the network. The results indicate significant triadic closure as well as preferntial attachment biases.

Keywords: Dynamic networks, R&D partnerships

*We thank Gueorgi Kossinets for comments and suggestions. Financial support from the Japan Securities Scholarship Foundation is gratefully acknowledged.

†University of Tsukuba,hanaki@dpipe.tsukuba.ac.jp

‡University of Tsukuba, nakajima@dpipe.tsukuba.ac.jp

§Hitotsubashi University, ogura@ier.hit-u.ac.jp

I. Introduction

Acquisition of new knowledge is crucial for successful innovation. A large body of literature have consistently stressed the importance of knowledge spillovers between firms, and demonstrated that knowledge absorption through inter-firm partnership enhances innovative outputs (e.g., Jaffe 1986, Bernstein and Nadiri 1988). Recently, researchers have analyzed inter-firm knowledge spillovers in a network setting (e.g., Goyal and Moraga-Gonzalez 2001, Cowan and Jonard 2004, Meagher and Rogers 2004). On the empirical side there is ample evidence that positioning of firms within R&D network substantially affect their R&D performance (e.g., Powell, Koput and Smith-Doerr 1996, Ahuja 2000, Gomes-Casseres, Hagedoorn and Jaffe 2006).

These studies assume the exogenous network architecture in measuring the effects of knowledge spillover on innovation performance. R&D collaborations, however, should be considered to form endogenously and evolve over time as firms decide with whom to collaborate. In contrast with the abundance of evidence relating network position and performance, little work has been done to understand the evolution of R&D network.

The importance of strategic network formation has been widely recognized theoretically in the R&D literature.[1] Yet, the empirical analyses of network formation have been underrepresented in the literature. This paper aims to contribute to the literature by analyzing evolutionary process of R&D network in the U.S. Information and Technology (IT) industry during the period of 1985-1995. We identify an R&D collaboration relationship between inovative IT companies if at least one common researcher is observed between them at a given period of time. In order to collect the information of inter-firm collaborations, we compile patent data of all IT companies provided by NBER patent database (Hall, Jaffe and Trajtenberg 2001). The names

[1]For a recent review of theoretical literature, see, for example, Bloch (2005). Theoretical analysis of network formation is a very active area of research, not only in the context of R&D but also in other contexts as reviewed in Jackson (2006).

of inventors are recorded with the name of the assignee for each patent. We match the lists of inventors' names across different assignee companies to see if they are connected via common researchers. A longitudinal data of evolving R&D network is created by collecting instantaneous networks that are yearly snapshotted.[2]

We assume that innovative IT firms collaborate with each other such that more knowledge spillovers are expected. Following Cassiman and Veugelers (2002) who emphasize the importance of distinguishing in- and out-flows of knowledge in R&D process, we focus on firm's incentives to strategically maneuver incoming and outgoing spillovers. We assume that IT companies maximize the incoming spillovers from partners, and at the same time, minimize outgoing spillovers to non-partners in order to manage the external information effectively.

These abilities of controlling patterns of knowledge spillovers may crucially depend on the R&D network structure. A company with many collaboration links in an R&D network can benefit from inter-firm network flows of information. Such a company has better access to knowledge flow through links to other companies. Thus, maximizing incoming spillovers implies that firms try to collaborate with "more connected" companies in order to access to novel information flows. Hence the more the network links a company participating in R&D network has, the more attractive the company is. Consequently, a company with more connection attract further connection with new partners. In terms of network theory, (e.g., Barabási and Albert 1999), we can say that firms have "preferential attachment bias", that is, a firm with more

[2]The patent data has been used in the literature to identify the collaborations among innovators. For example, Cantner and Graf (2006) constructs a network of innovators in Jena, Germany, based on the patent records. Singh (2005) analyzes the relationship between collaboration among innovators and patent citation. The use of coinventors data in patents to identify R&D collaborations among firms can be justified because coinventors of a patent collaborate intensively over extended period of time as Fleming, King III and Juda (2004) reports, and such intensive collaboration will be difficult without supports from the firms these inventors are working for. In addition, the fact that patent database is publicly available gives an advantage of its use over other databases such as MERTICATI database (Hagedoorn 2002). On this other hand, using the patent has its drawback as we fail to capture those R&D collaborations that did not result in obtaining a patent. Our focus here is to analyze the evolution of successful collaboration where the success is defined in terms of obtaining a patent.

collaborations has a higher propensity to receive further.

The firm's ability to protect their proprietary knowledge may also depend on network structures that facilitate cooperation between firms in an R&D network. The theoretical literature (e.g., Shapiro and Willig 1990) demonstrates that imperfect appropriability increases the incentive of firms to free ride on the R&D efforts. But, if firms can use effective reputation or sanction mechanisms that monitor and guide external knowledge flows, innovative idea or technology will be effectively protected from free-riding by outsiders. This is reminiscent of the social capital theory proposed by Coleman (1988b). The gist of argument is that a network closure creates a reputation cost for inappropriate behavior which facilitates trustworthiness between agents involved. Thus, the firms' incentives to limit outgoing spillovers may lead to a "cyclic closure bias" (Kossinets and Watts 2006), in particular, "triadic closure bias" (Rapoport 1953, Watts 1999), that is, a firm has propensity to form a network closure among a local circle of firms.

We first measure the properties of the network over time and demonstrate that the network has become more extensive, more locally clustered, and more unequal in the sense "star" companies have emerged in the network.[3]

We then estimate the strength of various mechanisms that have been proposed in the literature as driving forces behind the network evolution. The regression analysis reveals that, after controlling for characteristics of firms as much as possible, there are significant "triadic closure" and "preferential attachment" biases.

The rest of the paper is organized as follows: The data is described in Section II. Section III discusses the evolution of structure of networks over time. Section IV shows the framework of our statistical analysis. The empirical results are discussed in Section V, and Section VI concludes.

[3]Goyal, van der Leij and Moraga-González (2006) analyzes evolution of co-authorship network among economists over time and argue that Economics is a field that has become increasingly "Small World" (Watts 1999) over past 30 years.

II. Data

We use NBER patent data (Hall et al. 2001). For each patent granted, the names of the researchers, the assignee, the assignee ID, the application year, and the technological category are listed. We focus on the patents that are classified as category two, "computers and communications." We focus on this category because our empirical strategy, which will be discussed in Section IV, is to estimate the likelihood of formation of new partnership, and according to Hagedoorn (2002), IT industry is the industry with the highest share of newly established R&D partnership among high-tech industries from the latter half of 1980s (about 40% or more).

A. *Constructing network data*

We use a three year moving window in constructing the collaboration network data, i.e., the patent applied (and consequently granted) between 1984 and 1986 are used to construct the network data for 1985, those applied between 1985 and 1987 are used to construct the network data for 1986, and so on.

We define that two companies have collaborated if the same researchers are listed in the patents owned by these two companies.[4] For example, in the example shown in table 1, company A_1 and A_2 has collaborated in the period in question because researcher R_1 and R_4 are involved in patents that are owned by the companies, namely, P_1, P_2, and P_3. In the same way, company A_1 and A_3 have collaborated because researcher R_2 is listed in the patents P_1 and P_4 that are owned by A_1 and A_3, respectively.

We follow Cantner and Graf (2006) in construct such network of collaboration. We first define a $n \times m$ matrix $\mathbf{X}$ where n is the number of companies and m is the

[4]We distinguish different researchers by their last name, first name, and the initial of the middle name. This procedure is common in constructing co-authorship network, (Newman 2004, Goyal et al. 2006). It is not perfect, of course, as Trajtenberg, Shiff and Melamed (2006) points out, two different researchers may have the same last and first name as well as middle name initial.

Table 1. An example of 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

number of researchers in the data. The i-th row of the matrix represents the company A_i, and the j-th column of the matrix corresponds to researcher R_j. The element i, j of the matrix is set equal to one if researcher R_j is listed in the patents owned by company A_i, and zero otherwise. Using the example shown in table 1, the $\mathbf{X}$ will be as follows:

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

The $n \times n$ matrix that defines the structure of the collaboration network, or *the weighted adjacency matrix*, $\mathbf{\Gamma}$, can be defined as

$$\hat{\mathbf{\Gamma}} = \mathbf{X}\mathbf{X}' = \begin{pmatrix} 3 & 2 & 1 \\ 2 & 3 & 0 \\ 1 & 0 & 2 \end{pmatrix}$$

as we do not consider a firm collaborating with itself, we replace the diagonal element of $\hat{\mathbf{\Gamma}}$ with zero and obtain

$$\mathbf{\Gamma} = \begin{pmatrix} 0 & 2 & 1 \\ 2 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

Since we consider the symmetric collaboration relations, $\mathbf{\Gamma}_{ij} = \mathbf{\Gamma}_{ji}$. The value greater

than zero represents the existence of collaboration between two companies. And the larger the value is, the more intensive the collaboration was in the sense that there were more researchers involved in the collaboration activities. For example, researchers R_1 and R_4 are both listed in the patents owned by company A_1 and A_2, which results in $\mathbf{\Gamma}_{12} = \mathbf{\Gamma}_{21} = 2$.

Most of the existing statistics to characterize the structure of network do not utilize the information regarding intensity of relationship. Also, our statistical analysis in Section IV studies only the formation the relationship and not its intensity. Therefore, what is needed is whether the relation exists or not. Thus, we can define the unweighted adjacency matrix, $\mathbf{G}$, the element of which takes the value one, if the entry to the weighted adjacency matrix, $\mathbf{\Gamma}$, is greater than zero.

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

Given the way we construct our network, it is possible that a link between two companies captures not only the genuine inter-firm collaborations but also researchers' job hopping.[5] Indeed, several surveys report that innovative companies recognize hiring away researchers from other innovative companies as an important strategic tool to acquire external knowledge. For example, Cassiman and Veugelers (2006) report that 42% of companies in their survey in Belgium replied that they actively engaged in hiring away personnel, while 37% of the companies replied that they actively engaged in R&D contracting in order to acquire external technologies. Besides this report, a study by Cantner and Graf (2006) on the local network formation in Jena, Germany, shows that a firm is more likely to form R&D collaborations with firms that employ researchers who have worked for it before. A case study by Saxenian (1994) also pro-

[5]We have chosen three year moving windows to minimize the possibility of capturing job hopping, but the possibility still remains.

vides anecdotal evidence from interviews with Silicon Valley engineers that extremely frequent job-hopping enhanced inter-firm collaborations in the Silicon Valley in the 1970s. From these circumstantial evidence, we can interepret that our R&D networks capture broader research collaborations among companies including potential ones that eventually facilitate knowledge spillovers among innovative companies.

III. Dynamics of the collaboration network

In the sample period between 1985 and 1995, we find that the total number of companies is 7773, while the number of observed collaborations between the companies is 9234. The average intensity per collaboration, i.e., the average number of researchers involved in one R&D collaboration activity, is 4.61.

Table 2 shows the basic statistics that describe the structure of collaboration network. The number of nodes is simply the number of companies in the network.[6] As there are many nodes that are not connected with others (isolated nodes), the table also report the number of linked nodes which excludes the isolated nodes. For the linked nodes, the average degree and the clustering coefficient (Watts and Strogatz 1998) are reported.

The average degree captures the average number of collaborating partners. While this has been quite stable until 1990, it has been increasing since. This shows that collaborations among companies have become increasingly more common in 1990s. At the same time, the size of the largest connected components as a fraction of the number of linked nodes in the network is increasing. This illustrate that the companies are increasingly being connected to other companies, directly or indirectly, in the web of collaboration networks.

The clustering coefficient measures the degree of local connectedness of the net-

[6] As discussed above, since we take three year moving windows, the network for 1985 is based on the granted patents that have been filed between 1984 and 1986.

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

year	Number of nodes*	Number of linked nodes	Average degree	Clustering coefficient	Variance of Bonacich centrality	The size of the largest connected component+	The number of connected components
1985	1919	197	1.45	0.10	0.31	0.18	65
1986	2088	220	1.41	0.04	0.36	0.13	75
1987	2282	266	1.49	0.12	0.37	0.24	87
1988	2500	350	1.51	0.05	0.44	0.26	97
1989	2705	383	1.43	0.03	0.47	0.21	123
1990	2874	426	1.45	0.05	0.58	0.25	136
1991	2981	446	1.64	0.08	0.69	0.42	118
1992	3126	516	1.69	0.08	0.72	0.41	131
1993	3499	613	1.76	0.08	0.78	0.41	153
1994	4011	774	2.03	0.13	0.85	0.55	144
1995	4607	980	2.35	0.15	0.85	0.60	175

All measures exclude isolated nodes unless otherwise noted.
$*$ Including isolated nodes.
$+$ As a fraction of the number of lined nodes.

work.[7] The clustering coefficient is much higher than expected if the connections are made randomly among the existing nodes in the network.[8] The high clustering have been found not only in various social networks, such as collaboration among economists (Goyal et al. 2006) or researchers in other fields (Newman 2004), but also in the physical or neural networks (Watts and Strogatz 1998). The clustering coefficient shows the increasing trend since 1990, which suggests that the search for the collaborating partners have become increasingly more local (Jackson and Rogers 2006), and resulted to give a "triadic closure bias" (Rapoport 1953, Watts 1999) in the way the collaboration network has evolved.

Table 2 also reports the variance of Bonacich centrality (Bonacich 1987) which measures the importance of central players, or "star" companies, in the network.[9] The low variance of the Bonacich centrality across the nodes in the network implies that relative positions of nodes are similar to each other, while high variance represents "core-periphery" structure in which there are a small number of "star" companies that position themselves in the center of collaboration network. Table 2 shows that the variance of the Bonacich centrality has been increasing throughout the sample

[7]It is defined as follows: let k_i be the number nodes connected with node i (degree of node i). The possible number of connections among those k_i nodes is $k_i(k_i-1)/2$. The clustering coefficient for node i is defined as the ratio of the actual number of connection exists among these k_i nodes to the possible number, $k_i(k_i-1)/2$. Thus, the more collaborations there are among the collaborating partner for company i, the larger the clustering coefficient for company i becomes. The extreme case is where the clustering coefficient is one. In such a case, there are many isolated groups of companies that collaborate with all the other companies within the group, but do not collaborate with those outside of the group.

[8]The clustering coefficient for random graphs with the same number of linked nodes are less than 0.002 for all the years.

[9]The Bonacich centrality for node i, bc_i, is defined as follows:

$$bc_i(\alpha,\beta) = \sum_j (\alpha + \beta bc_j)\mathbf{\Gamma}_{ij}$$

where α is set so that $\sum_i bc_i(\alpha,\beta)^2 = 1$, and $\beta = 0.1$. The value of β can be any value between 0 and 1.0. But it is common to set it to 0.1 (Haynie 2001). Setting it to other values does not change the qualitative results much. Ballester, Calvó-Armengol and Zenou (2005) has recently demonstrated a relationship between Bonacich centrality measures and Nash equilibrium action of a player in a certain class of network games. It is of an interesting future research to exploit such relationship in R&D collaboration network.

period. Thus, as the collaborations have become more common, a few star companies seem to have emerged in the network.

IV. Statistical Model

In this section we turn to statistical analysis of the R&D collaboration formation by the innovating IT firms.

We use the standard latent utility framework to model the collaboration network formation between companies. We assume that the latent utility of company i from collaboration with company j at time $t+1$ is given by the following linear function:

$$u_{ij} = \alpha + \beta_{own} X_i(t) + \beta_{other} X_j(t) + \gamma Z_{ij}(t) + \sum_k \rho_k d_{ij}^{k-1}(t) + \varepsilon_{ij}(t). \tag{1}$$

The first component incorporates the systematic utility from collaboration where X_i and X_j is the background characteristics (e.g., firm size) of company i, and Z_{ij} is the *common* background characteristics of company i and j (e.g., similarity of production process). The second component involves the kth cyclic closure bias. We define d_{ij}^{k-1} as the dummy variable that takes one if the shortest distance between i and j is equal to k and zero otherwise. If $\rho_k > 0$, then the company obtains a positive utility by forming the cycle of length k. So the parameter ρ_k measures the degree of the tendency to form kth cyclic closure. It is the generalized notion of the triadic closure bias (Kossinets and Watts 2006). The third component is an idiosyncratic random error shock ε_{ij}. As in the standard random utility model, we assume that the random errors $\varepsilon_{ij}(t)$s are independent across companies and over time, and are identically distributed from the logistic distribution.

Without the loss of generality, the utility from no collaboration can be normalized to be zero. Thus, if $u_{ij} > 0$, the company i is willing to collaborate with company j

at time $t+1$. The probability is given by

$$\text{Prob}(u_{ij} > 0) = F\left[\alpha + \beta_{own} X_i(t) + \beta_{other} X_j(t) + \gamma Z_{ij}(t) + \sum_{k \geq 3} \rho_k d_{ij}^{k-1}(t) + \varepsilon_{ij}(t)\right], \quad (2)$$

where F is the cumulative distribution function of the logistic distribution.

Consider the collaboration formation between two companies, i and j, which are *not connected* at time t. We assume that the two companies i and j collaborate only if they are willing to collaborate *at the same time*. Let $G_{ij}(t)$ denote the R&D collaboration between company i and j at time t. The conditional probability that company i and j will initiate collaboration at time $t+1$ given that they do not collaborate at time t is presented by

$$\text{Prob}(G_{ij}(t+1) = 1 | G_{ij}(t) = 0) = \text{Prob}(u_{ij} > 0) \cdot \text{Prob}(u_{ji} > 0) \quad (3)$$

The equality follows from the assumption that $\varepsilon_{ij}(t)$ and $\varepsilon_{ji}(t)$ are independent. The possibility that the random shocks are *correlated* will be discussed later.

The conditional likelihood that currently unconnected companies initiate the R&D collaboration in the next period is given by

$$L(\theta) = \prod_{i<j} \text{Prob}(G_{ij}(t+1) = 1 | G_{ij}(t) = 0), \quad (4)$$

where θ is the vector of parameters such that $\theta = (\alpha, \beta, \gamma, \rho_k)$. The product $\prod$ is taken over all possible pairs of (i, j) such that $i < j$. Because of Equation (4), the conditional likelihood can be rewritten as

$$L(\theta) = \prod_{i} \prod_{j \neq i} \text{Prob}(u_{ij} > 0). \quad (5)$$

Thus the total log-likelihood function over the sample periods is presented by

$$\ell(\theta) = \sum_t \sum_i \sum_{j \neq i} \ln F \left[\alpha + \beta_{own} X_i(t) + \beta_{other} X_j(t) + \gamma Z_{ij}(t) + \sum_{k \geq 3} \rho_k d_{ij}^{k-1}(t) + \varepsilon_{ij}(t) \right]. \quad (6)$$

The structural parameter θ is estimated from the log likelihood function. Since the F is the cumulative distribution function of the logistic distribution, the standard logistic regression method is used for estimation.

One final note concerns the identification of parameters. It is found that two parameters β_{own} and β_{other} cannot be separately identified from the data. The reason is simple: Since these parameters are *symmetric* in the log likelihood function, any two sets of symmetric values of $(\beta_{own}, \beta_{other})$, such as (b_1, b_2) and (b_2, b_1), yields the same log likelihood value for equation (6). Thus in the empirical analysis below, we simply report $\beta = (\beta_{own} + \beta_{other})/2$. The parameter β can be interpreted as the average effect between own and other's background characteristics on link formation.

V. Empirical Results

We included a number of independent variables that are thought to influence the firm's decision to form R&D collaborations. Table 3 presents the definitions of the variables that we used in the estimation.

The first set of such variables contains information on firms' characteristics. These include firms' sales, R&D expenditures, and R&D intensity defined by the share of total R&D expenditures in total sales. These variables concerning firm size were shown to affect cooperation in R&D in previous literature, e.g., Colombo and Garrrone (1996), and Hernán, Marín and Siotis (2003), which argue that they will increase firms' "absorptive capacity" (Cohen and Levinthal 1989) of new innovation. We also included the number of R&D collaborations in the current period as a predictor of new collaboration in the next period. After controlling for firms' "absorptive capac-

Table 3. Description of Variables

Variable	Definition
Sale size	Firms' sales in 10^6 US dollars.
R&D size	Firms' R&D expenditure in 10^6 US dollars.
R&D intensity	R&D expenditure share in total sales.
Number of current collaborations	Number of R&D collaborations.
Similarity of production process	A dummy variable that takes one if the 5-digit NAICS codes match between firms, and zero otherwise.
Similarity of research activity	Sample correlation of subcategories of applied patents between firms. All patents applied in the sample period are used to compute the similarity measure.
Similarity of sale size	$= 1/(1 +$ difference in sale size).
Similarity of R&D size	$= 1/(1 +$ difference in R&D size).
Same state dummy	A dummy variable that takes one if the firms are located in the same state, and zero otherwise.
Same county dummy	A dummy variable that takes one if the firms are located in the same county, and zero otherwise.
Time trend	Time trend starting from 0 if the observation is in 1985, up to 9 if the observation is in 1995
d_{ij}^k (dummy variables)	$= 1$ if the shortest distance btween inovating firm i and j is equal to k where $k = 2, 3, 4, 5$.

ity", this captures "preferential attachment bias" (Barabási and Albert 1999), i.e., tendency for those firms with many partners to form more collaborations in the future. This bias, if it exists, explains the emergence of a few "star" companies in the network of R&D collaborations we discussed in Section III.

The second set of control variables reflects similarities between firms. We expect that the firms share many of the characteristics are more likely to cooperate with each other than those are not. We computed similarity measure of firms' production processes using sub-industry grouping system by NAICS (North American Industry Classification System). The similarity index takes one if two firms fall into the same five-digit NAICS sub-industry class, and zero otherwise. As an alternative similarity measure, we also computed non-centerred correlation coefficients of the technological portfolios of applied patents between firms. We used a citation-probability-adjusted

version of the correlation coefficient defined by Jaffe (1986).[10] It can be considered to measure a similarity or closeness among technological categories of firms' research activity. We also considered the similarities of firms' potential capacity. The similarity indices are computed from firms' sale size and R&D sizes, respectively.

In addition to these similarity indices, we included measures of geographical proximity between firms. We expect that the firms that are located in the same state or county are more likely to form R&D collaboration than those that are not.

Finally, we included the cyclic closure biases to measure tendencies for local clustering. To capture the k-th cyclic closure bias, we computed the shortest (geodesic) path between innovating firms in R&D collaboration network, and construct a dummy variable d_{ij}^{k-1}, which takes one if the shortest-distance between a pair of firm i and j is $k-1$, and zero otherwise. Since the pairs of firms that are connecting with more than 5 distances are very rare, we included up to 6-th cyclic closure biases in the estimation below.

Because the background characteristics are not available for all firms, we restrict our attention to the subsample to the innovating IT companies that are listed on the NYSE, NASDAQ, and AMEX in the sampling period. The data for these companies are obtained from S&P's COMPUSTAT. We find that there are 478,654 possible collaborations among those stock-market-listed firms, while only 288 collaborations (0.06 percent) are newly formed in the sample period. It should be noted that, although the sample are restricted by the stock-market-listed firms, the network related

[10]The correlation coefficient is difined by

$$\text{Similarity of research activity}_{ij} \equiv \frac{f_i W f_j'}{[(f_i W f_i')(f_j W f_j')]^{1/2}}$$

where f_i is a row vector of the number of patent applications in each technological subcategory taken out by firm i, and W is the citation probability matrix among each technological subcategories that are computed from the entire U.S. patent citation data from 1981 to 1999. The citation probability matrix instead of the identity matrix is used as a weight so that the similrarity measure can pick up the similarity or closeness among technological categories. By doing so, we coped with the problem of a potential discrepancy between the artificial classification and the true linkage among technologies.

Table 4. Descriptive Statistics of Variables

	min	max	mean	standard deviation
Sale size	0.000060	152.172000	4.719287	12.648110
R&D size	0.000051	7.035800	0.209263	0.622785
R&D intensity	0.002682	24.650000	0.108906	0.596994
Number of current collaborations	1	55	1.431433	2.171766
Similarity of production process	0	1	0.050838	0.219668
Similarity of research activity	0.000000	1.000000	0.237604	0.243423
Similarity of sale size	0.006529	1.000000	0.427505	0.322739
Similarity of R&D size	0.124444	1.000000	0.849150	0.200142
Same state dummy	0	1	0.093057	0.290512
Same county dummy	0	1	0.028806	0.167261
d^2	0	1	0.002737	0.052243
d^3	0	1	0.005495	0.073922
d^4	0	1	0.005465	0.073726
d^5	0	1	0.004103	0.063925

measures, such as the shortest distance, are calculated based on the full sample.

The descriptive statistics on the independent variables are reported in Table 4.

A. *Baseline Estimation Result*

Table 5 reports the marginal effect of the independent variables on the probability of forming new collaboration and the estimates of the k-th cyclic closure bias parameter ρ_k for $k = 3, 4, 5, 6$. The specification (1) and (2) provide the baseline estimates of the structural parameters. The difference between these two specifications is whether the liner time trend variable is included or not. The estimation results present that the time trend coefficient is positive and statistically significant, and thus suggests that the growth of R&D collaborations might be explained by a secular technological progress in the sampling period.

Not surprisingly, we found that the similarities of firms' production process and research activity have highly significant and positive effect on the likelihood of R&D collaboration. This suggests that the similar firms in line of industry are more likely to cooperate in R&D activity, as is consistent with our prior expectation.

Table 5. Baseline Estimation Results of Logit Regression for R&D Collaboration

Variable or Parameter	(1)	(2)	(3)
Constant	−7.3459***	-13.8774	-13.8516
	(0.2724)	(2.4210)	(2.3732)
Time trend		0.0722***	0.0699***
		(0.0265)	(0.0262)
Similarity of production process	0.8693***	0.8567***	0.8605***
	(0.1623)	(0.1620)	(0.1625)
Similarity of research activity	3.7343***	3.7139***	3.7868***
	(0.2391)	(0.2377)	(0.2374)
Similarity of sale size	−1.1538***	−1.1774***	−1.1292***
	(0.3669)	(0.3677)	(0.3575)
Similarity of R&D size	−2.4467***	−2.3904***	−2.2932***
	(0.3921)	(0.3926)	(0.3466)
Same state	0.696***	0.6717***	0.6744***
	(0.2121)	(0.2124)	(0.2126)
Same county	0.4726*	0.4713*	0.5206**
	(0.2486)	(0.2487)	(0.2479)
Sale size	−0.0246***	−0.0255***	
	(0.0091)	(0.0090)	
R&D size	0.3123**	0.3479**	
	(0.1502)	(0.1509)	
R&D intensity			0.0489
			(0.1378)
Number of current collaborations	0.0470***	0.0418***	0.0391***
	(0.0068)	(0.0070)	(0.0058)
ρ_3	2.5436***	2.4268***	2.4946***
	(0.1811)	(0.1843)	(0.1829)
ρ_4	1.6534***	1.5224***	1.5533***
	(0.2234)	(0.2271)	(0.2267)
ρ_5	1.5355***	1.4009***	1.4059***
	(0.2693)	(0.2724)	(0.2725)
ρ_6	1.1188***	0.9964**	1.0236**
	(0.4206)	(0.4222)	(0.4219)
χ^2	1335.71	1343.34	1333.19
$log-likelihood$	−1692.4	−1688.58	−1693.65
Pseudo R^2	0.283	0.285	0.282
N	425084	425084	425084

Notes: Standard errors are in parentheses.
$*$ Significant at the 10-percent level.
$**$ Significant at the 5-percent level.
$***$ Significant at the 1-percent level.

Different results, however, are reported for firm size similarities. The estimated coefficients of similarities in sale size and R&D size are both *negative* at the 1 percent significance level. These results indicate that the firms that are asymmetric in size have strong incentive to participate in R&D cooperation.

We also found that geographical proximities matter for forming new R&D collaborations between innovating IT firms. The result suggests that the firms that are located in the same county and same state are more likely to collaborate with each other than those that are not.

An interesting difference emerges between the effects of firm's sizes, measured in terms of sales and R&D expenditures, on R&D collaboration formation. The estimation result indicates that R&D expenditure has significantly positive impact on R&D cooperation.[11] On the contrary, the estimated effect of sales on R&D cooperation is negative and statistically significant. Yet, as indicated by the estimates, the R&D size has far larger impact on joint R&D participation than the sale size does. All these results seem to suggest that the R&D expenditures is critical in building absorptive capacity.

It should be noted that the alternative specification (3) presents that R&D intensity, which is widely used in the previous literature (e.g., Colombo and Garrrone 1996) as a proxy for absorptive capacity, is insignificant. This suggests that the absolute level of R&D expenditure level, rather than relative level of R&D expenditure, might be relevant to a firm's absorptive capacity.

The estimated coefficient of the number of current collaborations is positive and statistically significant. It should be noted that this is after controlling for firm's absorptive capacity by R&D size or R&D intensity. Therefore, we can attribute this result to the existence of "preferential attachment bias" (Barabási and Albert 1999), that is, firms try to collaborate with others that currently have many collaborating

[11] As discussed in Section IV, the estimated coefficients are average effect of own and other's characteristics.

partners in order to gain access, although indirectly, to the novel information and technology pooled therein to increase incoming spillovers as much as possible. This bias explains why we observe a few "star" companies emerging in the R&D network.

The estimates of the closure biases are reported in the bottom rows of Table 5. The cyclic closure biases are all positive and statistically significant. This implies that innovating IT firms are willing to form new R&D collaboration with firms that are inside the circle of their current partners, i.e., within a chain of few intermediaries. It should be noted that the strength of the cyclic closure bias decrease monotonically with distance between firms. Hence, two innovating firms with shorter chain of intermediates are more likely to cooperate in R&D research activities. This provides a reason why the R&D network clusters locally.

The results on cyclic closure bias shed light on how firms are managing the pattern of spillovers "to maximize the incoming spillovers *from* partners and nonpartners, while at the same time minimizing spillovers *to* nonpartners" (Cassiman and Veugelers 2002) in addition to the preferential attachment bias discussed above. Coleman (1988a) argues that the closing the local cycles promotes the cooperative behavior, in our context, lowers the possibility of outgoing spillovers, because such closure raises reputation costs of inappropriate behavior and creates a possibility of collective sanctioning. Ohta and Sekiguchi (2006) analyzes such mechanisms in sustaining cooperative behavior in the context of repeated game where behaviors of a player in one relation is not directly observable in other relationships with the player. The finding that estimated coefficients become smaller as the distance between two firms increases shows that firms try to collaborate locally and reduce the possibility of outgoing spillovers by relying on the possible reputation or sanction mechanisms that dense local interactions create.

B. *Estimation Result Controlling Unobserved Common Factors*

While there are evidences of strong cyclic closure biases in firms' R&D collaborations, there is a possible source of omitted variables problem. If there is unobserved common factors that affect the collaboration decisions of a group of firms in the R&D network, the effect of those omitted factors might not be separately identified from the cyclic closure bias. As illustration, suppose that research managers of companies were former collegues at a company or former classmates at a business school and were in close promixity. Then their companies may be more likely to cooperate in R&D activities due to their parsonal association (see Saxenian 1994). If we cannot observe such common predisposition of research managers that leads their companies to collaborate, we may mistakenly attribute the effect to a cyclic closure bias.

To examine the possibility of omitted variables problem described above, we added to the model dummy variables that explicitly accounts for unobserved factors that are common to all companies in proximity in the R&D network. Two types of dummy variables are considered. First we include a dummy variable that takes value one for a pair of innovating firms that belong to the same *connected subnetwork*. Second we include a dummy variable that takes value one for a pair of innovating firms that belong to the same *community subnetwork*.

The primary assumption of the empirical strategy is that a common factor, which is unobservable to researchers, affects all firms in the same connected or community subnetwork. Thus the latent utility model of R&D collaboration formation is modified as follows:

$$u_{ij} = \alpha + \beta_{own} X_i(t) + \beta_{other} X_j(t) + \gamma Z_{ij}(t) + \sum_k \rho_k d_{ij}^{k-1}(t) + \delta g_{ij} + \varepsilon_{ij}(t). \qquad (7)$$

where g_{ij} is either a connected subnetwork dummy variable or community subnetwork dummy variable, which takes value one if i and j belong to the same connected or

community subnetwork. We expect that $\delta \geq 0$ and thus the unobserved common factor facilitates R&D collaborations between firms that belong to the *same* connected or community subnetwork. The factor g_{ij} yields correlation in the neighborhood of the network if it is not taken into account. In that case, the unobserved error term, $(\delta g_{ij} + \varepsilon_{ij}(t))$, is correlated between inovating firms that belong to the same connected or community subnetwork due to the common factor g_{ij}.

To find community structure, we use a network partition method proposed by Girvan and Newman (2004). Roughly speaking, a community is a subset of nodes within the network (i.e., subnetwork) such that connections among them are denser than connections with the rest of the nodes in the network. Their community detection algorithm is based on the idea of "betweenness" of links in the network, where betweenness of a link is a measure that favors links that lie between communities.[12] Thus, if the links with high betweenness scores are removed, the community subnetworks are left behind out of the entire network.[13]

Given that the utility is specified by Equation (7), we can employ the same estimation method as we did for the baseline empirical model. We assumed the new R&D collaboration is formed between firm i and j only when both firms agree to do so. Thus the conditional likelihood is represented by Equation (5) with different u_{ij} that is specified above.

Table 6 reports the estimation results. We used the best-fit-specification (the specification (2) in the previous table) for selecting the independent variables. We

[12]The betweenness of a link, or "edge betweenness" is measured by counting, among the shortest-path between all the pairs of nodes that are connected, the number of shortest-path going through the edge under consideration. Since a path between two nodes that belong to different communities must go through edges that lies between these communities, the edge betweenness of such edges will be higher.

[13]How many communities should we expect in a network? This is a difficult question to answer without some prior knowledge about how the network is formed. Girvan and Newman (2004) proposes the use of modularity measure, and defines the community when the measure is the highest. The modularity measures the difference between the fraction of the edges in the network that connect nodes within the same community and the expected value of the same quantity with the same community division but connections between the nodes are random. They demonstrate that this procedure works very well for the network with a priori known community structure.

Table 6. Results of Logit Regression for R&D Collaboration; Including Subnetwork Dummies

Variable or Parameter	(4)	(5)	(6)
Constant	−13.4373***	−13.3891***	−13.6332***
	(2.4382)	(2.4469)	(2.4379)
Time trend	0.0667**	0.0662**	0.0688**
	(0.0267)	(0.0268)	(0.0267)
Similarity of production process	0.8285***	0.8278***	0.8368***
	(0.1625)	(0.1626)	(0.1626)
Similarity of research activity	3.6939***	3.6947***	3.6887***
	(0.2386)	(0.2387)	(0.2384)
Similarity of Sale size	−1.1560***	−1.1569***	−1.1656***
	(0.3692)	(0.3692)	(0.3701)
Similarity of R&D size	−2.3758***	−2.3732***	−2.3637***
	(0.3941)	(0.3942)	(0.3955)
Same state	0.6830***	0.6836***	0.7076***
	(0.2126)	(0.2126)	(0.2126)
Same county	0.4524*	0.4514*	0.4001
	(0.2488)	(0.2488)	(0.2501)
Sale size	−0.0249***	−0.0249***	−0.0246***
	(0.0091)	(0.0091)	(0.0091)
R&D size	0.3358**	0.3356**	0.3304***
	(0.1513)	(0.1514)	(0.1517)
Number of current collaborations	0.0425***	0.0423***	0.043***
	(0.0070)	(0.0070)	(0.0070)
Connected subnetwork dummy	1.7151***	1.7159***	1.7084***
	(0.3064)	(0.3064)	(0.3065)
Community subnetwork dummy		−0.0768	0.4797**
		(0.3114)	(0.2372)
ρ_3	0.78600**	0.8036**	0.6114*
	(0.3359)	(0.3432)	(0.3490)
ρ_4	−0.1174	−0.1092	−0.2041
	(0.3607)	(0.3622)	(0.3643)
ρ_5	−0.2392	−0.2343	−0.2739
	(0.3909)	(0.3914)	(0.3915)
ρ_6	−0.6467	−0.6452	−0.6748
	(0.5073)	(0.5073)	(0.5078)
χ^2	1363.82	1363.89	1367.75
$log - likelihood$	−1678.3363	−1678.3055	−1676.3755
Pseudo R^2	0.2889	0.2889	0.2897
N	425084	425084	425084

Notes: Standard errors are in parentheses.
$*$ Significant at the 10-percent level.
$**$ Significant at the 5-percent level.
$***$ Significant at the 1-percent level.

include the connected subnetwork dummy variable to control for the unobserved group heterogeneity, which is presented by the specification (4).

In addition, we also included the community subnetwork dummy variable, which is presented by the specifications (5) and (6). To determine the community structure of the R&D network, we employed both the *unweighted* adjacency matrix $\mathbf{G}$, which represents collaboration links as binary variables, and the *weighted* adjacency matrix $\mathbf{\Gamma}$, where its collaboration links are weighted by the number of common researchers between firms. In other words, we took into account of the "strength" of the collaboration for identifying distinct communities in the R&D network.[14] The specification (5) and (6) present the estimation results for unweighted and weighted community structure respectively.

We found that the connected subnetwork dummy variable is positive and statistically significant for all the specifications. On the other hand, the community subnetwork dummy variable is not significant for the specification using the unweighted adjacency matrix (specification (5)), while it is positive and statistically significant for the specification using the network structure weighted by the number of common researchers (specification (6)). This result may reflect the fact that the more accurate community structure can be identified by using the information of the strength of R&D collaborations. We expect that the stronger the R&D collaboration between firms, or, equivalently saying, the more researcher are involved in the R&D projects between firms, the closer the connection becomes between them. Thus, those firms having closer connections with each other are more likely to belong to the same community subnetwork.

As far as the firms' individual and common background characteristics are concerned, the point estimates in Table 6 are very similar to those in Table 5. All signs

[14]For the weighted network, the "edge betweenness" is calculated reflecting the weight on edges. Namely, the distance between two directly connected nodes are defined to be $1/\Gamma_{ij}$, and shortest-path are calculated based on this weighted distance.

of the estimates are as before, and the variables that are significant in Table 5 are also significant in Table 6.

Interestingly, all the estimates of cyclic closure biases become significantly smaller if the unobserved subnetwork factors are controlled for. For example, the estimate of the triadic closure bias ρ_3 decreases from 2.4268 for the specification (2) in Table 5 to 0.6114 for the specification (6) in Table 6. Yet, it is important to keep in mind that the estimates the triadic closure bias is still positive and statistically significant at least at the 10 percent significance level. On the other hand, the cyclic closure biases with more than the third degrees become negative but statistically insignificant once the subnetwork dummy variables are included. These results suggest that at least some closure bias is not mainly driven by unobserved group factors. The evidence can be interpreted in favor of positive triadic closure bias in forming new R&D collaboration as its effect on limiting the outgoing spillover through reputation and sanction mechanisms is the greatest.

VI. Conclusion

In this paper we have studied the evolution of successful R&D collaboration in the U.S. IT industry between 1985 and 1995 using the information contained in the granted patent in the U.S. The descriptive statistics of collaboration network suggest that the collaboration patterns have become more extensive, more locally clustered, and more unequal in the sense stars have emerged in the network. The regression analysis reveals that there is significant *triadic closure bias* and *preferential attachment bias* in the choice of collaboration partners even after controlling for characteristics of firms as much as possible, including some unobserved similarities between firms that we infer from the community structure of the network. The triadic closure and preferential attachment biases can be seen as firms trying to maximize incoming spillovers from partners and non-partners while minimizing out-going spillovers to non-partners in

their search of collaborating partners.

While focusing on the evolution of the structure of collaboration networks, we have not addressed other interesting questions such as: How are such dynamics of collaboration network related with the dynamic patterns of knowledge flows among firms? What is the relationship between a firm's position in collaboration network with its R&D productivity? We await for future research to answer these questions.

References

Ahuja, Gautam, "Collaboration Networks, Structural Holes, and Innovation: A Longitudinal Study.," *Administrative Science Quarterly*, 2000, *45*, 425–455.

Ballester, Coralio, Calvó-Armengol, Antoni, and Zenou, Yves, "Who's Who in Networks. Wanted: The Key Player," *Econometrica*, 2005, *forthcoming.*

Barabási, Albert-László and Albert, Réka, "Emergence of Scaling in Random Networks," *Science*, 1999, *286*, 509–512.

Bernstein, Jeffrey I. and Nadiri, Ishaq M., "Interindustry R&D Spillovers, Rates of Return, and Production in High-Tech Industries," *American Economic Review*, 1988, *78*.

Bloch, Francis, "Group and Network Formation in Industrial Organization: A Survey," in Gabrielle Demange and Myrna Wooders, eds., *Group Formation in Economics: Networks, Clubs and Coalitions*, New York,NY: Cambridge University Press, 2005, pp. 335–353.

Bonacich, Phillip, "Power and Centrality: A Family of Measures," *The American Journal of Sociology*, 1987, *92*, 1170–1182.

Cantner, Uwe and Graf, Holger, "The Network of Innovators in Jena: An Application of Social Network Analysis," *Research Policy*, 2006, *35*, 463–480.

Cassiman, Bruno and Veugelers, Reinhilde, "R & D Cooperation and Spillovers: Some Empirical Evidence from Belgium," *American Economic Review*, 2002, *92*, 1169–1184.

___ **and** ___ , "In Search of Complementarity in Innovation Strategy: Internal R & D External R&D Acquisition," *Management Science*, 2006, *52*, 68–82.

Cohen, Wesley M. and Levinthal, Daniel A., "Innovation and Learning: The Two Faces of R&D," *The Economic Journal*, 1989, *99*, 569–596.

Coleman, James S., "Free Riders and Zealots: The Role of Social Networks.," *Sociological Theory*, 1988a, *6*, 52–57.

___ , "Social Capital in the Creation of Human Capital," *American Journal of Sociology*, 1988b, *94*, S95–S120.

Colombo, Massimo G. and Garrrone, Paola, "Technological Cooperative Agreements and Firms R&D Intensity. A Note on Causality Relations," *Research Policy*, 1996, *25*, 923–932.

Cowan, Robin and Jonard, Nicolas, "Network Structure and the Diffusion of Knowledge," *Journal of Economic Dynamics & Control*, 2004, *28*, 1557–1575.

Fleming, Lee, III, Charles King, and Juda, Adam, "Small Worlds and Innovation," Working paper 04-008, Harvard Business School 2004.

Girvan, Michelle and Newman, Mark. E. J., "Finding and Evaluating Community Structure in Networks," *Physical Review E*, 2004, *69*, 026113.

Gomes-Casseres, Benjamin, Hagedoorn, John, and Jaffe, Adam B., "Do Alliance Promote Knowledge Flows?," *Journal of Financial Economics*, 2006, *80*, 5–33.

Goyal, Sanjeev and Moraga-Gonzalez, Jose Luis, "R & D Network," *The RAND Journal of Economics*, 2001, *32*, 686–707.

___ **, van der Leij, Marco, and Moraga-González, José Luis**, "Economics: Emerging Small World?," *Journal of Political Economy*, 2006, *114*, 403–412.

Hagedoorn, John, "Inter-Firm R & D Partnerships: an Overview of Major Trends and Patterns since 1960," *Research Policy*, 2002, *31*, 477–492.

Hall, Bronwyn, Jaffe, Adam, and Trajtenberg, Manuel, "The NBER Patent Citations Data File: Lessons, Insights and Methodological Tools," Working Paper 8498, National Bureau of Economics Research 2001.

Haynie, Dana L., "Delinquent Peer Revisited: Does Network Structure Matter?," *American Journal of Sociology*, 2001, *106*, 1013–1057.

Hernán, Roberto, Marín, Pedro L., and Siotis, George, "An Empirical Evaluation of the Determinants of Research Joint Venture Formation," *Journal of Industrial Economics*, 2003, *51*, 75–89.

Jackson, Matthew O., "The Economics of Social Networks," in Richard Blundell, Whitney Newey, and Torsten Persson, eds., *Advances in Economics and Econometrics, Theory and Applications: Ninth World Congress of the Econometric Society*, Vol. 1, Cambridge University Press, 2006.

—— **and Rogers, Brian W.**, "Meeting Strangers and Friends of Friends: How Random are Social Networks?," *mimeo*, California Institute of Technology 2006.

Jaffe, Adam B., "Technological Oppoetunity and Spillovers of R&D: Evidence from Firms' Patents, Profits and Market Value," *American Economic Review*, 1986, *76*, 984–1001.

Kossinets, Gueorgi and Watts, Duncan J., "Empirical Analysis of an Evolving Social Network," *Science*, 2006, *311*, 88–90.

Meagher, Kieron and Rogers, Mark, "Network Density and R & D Spillovers," *Journal of Economic Behavior & Organization*, 2004, *53*, 237–260.

Newman, Mark. E. J., "Coauthorship Networks and Patterns of Scientific Collaboration," *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 2004, *101*, 5200–5205.

Ohta, Katsunori and Sekiguchi, Tadashi, "Multilateral Repeated Games: Possibility of Cooperation under Limited Observability," Working Paper, Institute of Economic Research, Kyoto University 2006.

Powell, Walter W, Koput, Kenneth W, and Smith-Doerr, Laurel, "Interorganizational Collaboration and the Locus of Innovation: Networks of Learning in Biotechnology," *Administrative Science Quarterly*, 1996, *41*, 116–145.

Rapoport, Anatol, "Spread of Information through a population with sociostructural bias: I. Assumption of Transitivity," *Bullein of Mathematical Biophysics*, 1953, *15*, 523–533.

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

Shapiro, Carl and Willig, Robert D., "On the Antitrust Treatment of Production Joint Ventures," *Journal of Economic Perspectives,*, 1990, *4*, 113–130.

Singh, Jasjit, "Collaborative Networks as Determinants of Knowledge Diffusion Patterns," *Management Science*, 2005, *51*, 756–770.

Trajtenberg, Manuel, Shiff, Gil, and Melamed, Ran, "The "Names Game": Harnessing Inventors' Patent Data for Economic Research," Working Paper 12479, National Bureau of Economics Research 2006.

Watts, Duncan J., *Small Worlds: The Dynamics of Networks Between Order and Randomness*, Princeton University Press, Princeton, New Jersey, 1999.

—— **and Strogatz, Steven H.**, "Collective Dynamics of 'Small-World' Networks," *Nature*, 1998, *393*, 440–442.