UNDERSTANDING THE ALLIANCE DATA

Melissa A Schilling

Stern School of Business

New York University

40 West 4th Street

New York, NY 10012

212-998-0249

mschilli@stern.nyu.edu

January, 2008

This project was made possible through funding from the National Science Foundation, Award #SES-0234075.

Key words: Alliances, Joint Ventures, Methodology, Networks, Sampling Theory

UNDERSTANDING THE ALLIANCE DATA

Abstract

A considerable body of research utilizes large alliance databases (e.g., SDC, MERIT-CATI, CORE, RECAP, and BIOSCAN) to study inter-organizational relationships. Understanding the strengths and limitations of these databases is crucial for informing database selection and research design. I conduct an analysis of five prominent alliance databases, examining their consistency of coverage and completeness, and assessing whether different databases yield the same patterns in sectoral composition, temporal trends, and geographic patterns in alliance activity. I also replicate three previously-published alliance studies to assess the impact of data limitations on research outcomes. The results suggest that the databases only report a fraction of formally announced alliances, which could have detrimental consequences for some types of research. However, the databases exhibit strong symmetries in patterns of sectoral composition, alliance activity over time, and geographic participation. Furthermore, the replications of previous studies yielded results that were highly similar to those obtained in the original studies. This study thus provides some reassurance that even though the databases only capture a sample of alliance activity, they may yield reliable results for many -- if not all -- research purposes. This information should help researchers make better-informed decisions about their choice of database and research design.

There is a considerable body of research that utilizes large alliance databases (often supplemented with data from other databases, surveys, or news retrieval searches) to explore issues related to inter-organizational relationships and collaboration (e.g., Anand & Khanna, 2000; Beckman, Haunschild & Phillips, 2004; Gulati, 1999; Hagedoorn, 2002; Kim & Higgins, 2005; Lavie & Rosenkopf 2006; Link, Paton & Siegel, 2002; Powell, Koput & Smith-Doerr, 1996; Rothermael & Deeds, 2004; Sampson, 2005; Vanhaverbeke, Duysters & Noorderhaven, 2002; Villalonga & McGahan, 2005). Some of the more prominent alliance databases include Securities Data Company (SDC), MERIT-CATI, CORE, Bioscan, or Recombinant Capital (RECAP).

Each of these databases has its own unique set of advantages and disadvantages that make it better suited to some types of research than others. It is very important for researchers to understand these advantages and disadvantages when selecting a database and creating a research design. In this paper, I conduct a comparative analysis of these databases, examining their consistency in coverage, assessing the degree to which different databases yield the same conclusions about trends in sectoral composition, alliance activity over time, and geographical participation, and explaining some of the features of the databases that differentiate them from one another. I examine alliances both in the economy at large (i.e., all-sector data), and within individual industry sectors (e.g., information technology, transportation equipment, chemicals, and biotechnology). Patterns in the alliance data are examined both graphically, and through the generation of reliability statistics across datasets. The implications of data inconsistencies or limitations are then further explored by replicating previously-published studies that use alliance data. The results of the study yield strong implications for both prior and future empirical research on alliances, and should help alliance researchers make better-informed decisions about research design.

The focus of the paper is on research and technology alliances -- those that entail some aspect of joint research or cross-technology transfer (the specific types of alliances included in each search are detailed later in the paper). This choice is made largely for pragmatic reasons: the MERIT-CATI and CORE databases are focused almost solely on this type of alliance. However, this focus has other redeeming qualities. Research and technology alliances play a particularly prominent role in both industry and research. They account for roughly half of the alliances reported in the SDC database (which seeks to capture and report a very wide scope of alliances, including simple marketing agreements or unilateral licensing deals). Research and technology alliances are also important in their capacity to enable knowledge sharing or joint knowledge creation among firms, facilitating innovation and economic growth. It is this capacity for accelerating knowledge creation and acquisition that has been responsible for a disproportionate amount of the research interest that alliances have recently garnered. The paper also focuses on the 1990 to 2005 time frame, because data was most reliably available from 1990 forwards for the set of databases utilized here.

The paper begins by giving a brief description of each of the databases utilized in the study, culminating in a summary of their key features (organized into advantages and disadvantages). It then analyzes the consistency of coverage across the databases to draw inferences about each database's completeness. Next, the consistency of patterns in the alliance data is examined (including sectoral composition patterns, temporal patterns, and geographic patterns). In the following section, a set of previously-published alliance studies are replicated to examine what impact data inconsistencies or limitations has on research outcomes. Each study is replicated using a different database than the one that was originally used. The final section summarizes the findings, and draws conclusions about how the strengths and limitations of the databases should inform research design.

THE DATABASES

This section describes the five alliance databases that are analyzed in the study. Three are multi-sector databases (SDC, MERIT-CATI, and CORE) and two are specific to the biotechnology sector (RECAP and Bioscan). A table at the end of the section summarizes some of the key features of the individual databases.

SDC

SDC is a division of Thomson Financial. The SDC database provides information on a wide range of financial transactions, including global new issues, securities trading, mergers and acquisitions, and more. The alliance data is just one slice of the data, available through the mergers and acquisitions section of the database. SDC collects data from SEC filings (and their international counterparts), trade publications, wires and news sources. SDC tracks a very wide range of agreement types, including joint ventures, strategic alliances, research and development agreements, sales and marketing agreements, manufacturing agreements, supply agreements, and licensing and distribution pacts. Furthermore, in addition to collecting agreements between industrial partners, the SDC data also includes agreements between universities and government labs, or any combination thereof. SDC manuals note that data is available from 1988 forward, though there are a small number of alliances reported in the database prior to that period, and data is quite sparse until 1990 (Anand & Khanna, 2000; also see Appendix 2). Of the databases considered here, the SDC database covers the widest range of sectors (SDC reports at least one alliance for each of 1,059 four-digit SIC codes between 1985 and 2005). Additionally, one of the SDC database's key advantages is its extensive searchability (SDC offers over 200 data elements such as the names, SIC code and nationality of participants, the terms of the deal, deal synopsis, and more for each alliance agreement) and the option to download data in a user-defined format such as an excel spreadsheet with a personalized set of columns. Though there are some occasional errors in coding (for example, a small percentage of alliances are coded as belonging to SIC codes that are obviously incorrect), in general the coding is highly accurate, and very useful in helping the researcher identify alliances of interest. SDC also provides a reference to the source(s) used in identifying the alliance, which enables the researcher to verify information provided in the database.

Since both the MERIT-CATI database and the CORE database are focused on alliances that entail some type of joint research or technology exchange, I will focus on a similar subset of the SDC data. Unless otherwise specified, the SDC alliance counts reported will refer to those alliances classified as one or more of the following types: research and development alliances, cross-licensing, cross-technology transfer, and joint ventures.[1] Furthermore, the SDC database distinguishes between announcements of alliances that are "completed" versus "pending." This distinction is not provided in most of the other databases considered here; to be as inclusive as

[1] Notably, it is often difficult to ascertain to what degree joint ventures involve joint research or technology exchange. Including joint ventures thus risks some over-inclusion (i.e., some agreements that do not entail significant joint research or technology exchange).

possible I include both completed and pending alliances for the SDC data here.[2] The database lists over 52,000 alliances between 1990 and 2005 that meet these criteria.

MERIT-CATI

The MERIT-CATI database is focused on "strategic technology agreements" which includes any alliance that entails the transfer of technology or the undertaking of joint research, such as joint research pacts, joint development agreements, R&D contracts, (mutual) second sourcing agreements and joint ventures with technology sharing or an R&D program. Simple production or marketing agreements are excluded. Furthermore, the MERIT-CATI database only includes agreements that have at least two industrial partners, thus agreements involving only universities or government labs, or one company with a university or lab, are disregarded. The MERIT-CATI database collects data from newspaper and journal articles, books, specialized journals, company annual reports, the financial times industry yearbooks, and Dun and Bradstreet's "who owns whom". The database is administered at Maastricht University in the Netherlands, and is the only non-US-based database considered here. It primarily utilizes sources written in English, though the administrators also read Dutch and German press and translated abstracts of important foreign newspapers and trade journals. Though systematic collection did not begin until 1987, one of the database's key advantages is that it utilizes retrospective data to incorporate data from earlier years (as early as 1960), and thus is very valuable for looking at long-range historical

[2] While including completed and pending alliances of all alliance types yields a much larger peak in 2000 than for the other series, this peak is significantly attenuated by restricting the alliances to those that are for purposes of joint research or technology exchanges. As shown in Appendix 1, restricting the data to joint research or technology exchanges and including both completed and pending alliances, yields a data pattern that is not very different from the pattern that is obtained by including completed alliances of any type. This suggests that joint research or technology exchange alliances drove a large portion of the variance in the SDC alliance data, and further suggests that there was no systematic bias in the portion of research and technology alliances that were completed. It does suggest, however, that a large portion of alliances announced during the 1998-2005 time period that were not research or technology alliances were not completed.

trends. It provides a numerical code for each participant that can be matched to company names in a separate file, and also provides codes for such elements as the nations of the participants, the form of cooperation, and the industrial sector of the alliance. The database lists just over 8,200 agreements over the period 1990 to 2004.[3]

CORE

The CORE database records industrial partnerships filed under the National Cooperative Research Act (NCRA) passed in 1984. The database is limited to research joint ventures, though the definition of "research joint venture" in the act is quite broad and includes cooperative efforts in research and development, production, application for patents, and the granting of licenses for the venture's results. The number of firms that choose to file under the NCRA (now the NCRPA where P standards for Production) is much smaller than the universe of organizations that engage in announced alliances, however the patterns in the database may still be informative for our purposes. The database lists 762 filings over the period 1990 to 2005.

Recombinant Capital (RECAP)

RECAP is a consulting firm specializing in the biotechnology industry. It collects data on biotechnology alliances from three primary sources: biotechnology and pharmaceutical company press releases and other literature; U.S. Securities and Exchange Commission filings; and company presentations made at investment conferences and other public meetings. Alliances can be between organizations of any type, including firms, universities, government laboratories, etc. One of the database's strengths is that it provides copies of the material contracts filed per the requirements of the U.S. Securities and Exchange Commission requirements and provides some

[3] 2004 was the last year for which data was available from MERIT-CATI at the time of this study.

analysis of the data contained therein (approximately 40% of biotechnology agreements are filed as material contracts). The database contains 24,191 biotech alliances initiated since 1973. The RECAP database lists a wide range of agreements, including acquisitions, joint ventures, licensing deals, co-development agreements, manufacturing agreements, marketing agreements, and more. The RECAP data is searchable based on a number of alliance criteria, however the output options are very limited making it a somewhat difficult database to use for large-scale analyses.

To facilitate comparison with the other databases, I chose search criteria that would yield alliances that entail some type of joint research or technology exchange. Thus the RECAP alliances reported here include all co-development agreements, cross-licensing agreements, research agreements, and joint ventures. 4,427 alliances reported in RECAP between 1990 and 2005 meet these criteria.

Bioscan

The Bioscan database is produced by American Health Consultants, and provides information pertaining to a range of activities performed by U.S. and foreign companies actively involved in biotechnology research and development. The Bioscan data is quite different in its collection and reporting methods from the other databases. Whereas SDC, MERIT, and RECAP all search a variety of news sources for reports of alliance announcements, the Bioscan database tracks the activity of a particular set of firms designated as biotechnology-related over time (the database listed 2,079 firms as of November, 2006). This can be both an advantage and a disadvantage, depending on the purposes of the researcher. On the one hand, the focus on a relatively stable population of firms can enhance the reliability of the data (at least as pertain to those firms), and facilitates the tracking of the behavior of those firms over time. On the other hand, this focus on particular firms also means that biotechnology alliances that are struck between firms that primarily operate in other industries would not be included in the Bioscan data. Furthermore, when Bioscan updates its datasets, it deletes firms that are no longer in operation (including all of their alliances), causing changes in the data reported over time and resulting in left censoring. There were 3,106 alliances reported over the 1990 to 2005 time frame in the Bioscan data as of March, 2006.

The Bioscan database cannot be systematically screened based on alliance types as it provides only limited data on the nature of the alliances included. Bioscan alliances reported here may thus be of any type.

Table 1 provides a summary of some of the key advantages and disadvantages of the databases analyzed here. Of the all-sector datasets, SDC appears to be the most inclusive in terms of types of agreements covered, types of organizations covered, and nations represented. It is by far the largest of the five alliance databases. It also provided the most searchable items of any of the databases, and it provided more information on the nature of the alliance and its participants than the other two multi-sector databases. The great strength of the MERIT-CATI dataset is that it provides retrospective data going back as far as 1960; Bioscan provides a similar retrospective perspective, but only for a limited sample of firms. The CORE dataset's key strength is that it is the only dataset examined here that captures a complete population -- all of the collaboration agreements filed under the NCRA act. Thus while it is a much smaller dataset than the others, it has some important inference advantages over the other datasets. Also, while the data provided in the actual spreadsheets is quite sparse (strictly a count organized by NAICS code and year), each entry can be matched to a notification in the *Federal Register* that provides detailed information on the collaboration. The two biotechnology databases were quite different in their relative strengths. On the one hand, RECAP reports far more alliances than Bioscan, and as will be shown later in the paper, it appears to be quite robust in terms of its coverage and reliability of temporal patterns. On the other hand, Bioscan reports a wider range of data on the set of companies that it tracks, providing a number of useful covariates that the researcher may combine with the alliance data. Furthermore, by following a defined set of companies closely, it is possible that the Bioscan data has more thorough coverage of alliances that involve those companies.

----------------------------Insert Table 1 About Here-------------------------------

CONSISTENCY AND COMPLETENESS OF COVERAGE

In this section I evaluate the consistency of coverage of the alliance databases to get a sense of their completeness. There is no omniscient database of alliances with which to compare the databases here. Firms are not required to report their alliances to any governing body, and while many do announce their alliances publicly, whether and how these announcements are reported in new sources or SEC filings is highly variable. Furthermore, because of the wide range of language that may be used to convey that an alliance has been formed, it is extremely difficult (perhaps impossible) to construct a series of text searches using one of the major news retrieval sources (e.g., Factiva, LexisNexis) that will yield an exhaustive, yet appropriate exclusive, list of alliance announcements.[4] For a small sample of firms, the researcher can resort to reading all of the news coverage of those firms and coding the alliances manually, but this is impractical for all but the most constrained samples. Fortunately, we can get some sense of the completeness of the datasets by comparing them to each other. If, for example, for an identically specified search of both the SDC and MERIT-CATI databases we found roughly the same number of alliances across both databases and that 95% of the alliances in one database were mirrored in the other (and thus were highly *consistent*), we would conclude that either the databases are both nearly exhaustive in their searches (highly *complete*), or both share a remarkably similar bias in their alliance collection efforts. On the other hand, if such a search yields a very different number of alliances for SDC than for MERIT-CATI and the overlap in alliances is small (i.e., low consistency), we also know that neither database is complete. For example, if identically specified searches yield 1000 alliances in SDC and 500 alliances in MERIT-CATI and only 100 of those alliances are in both databases (and assuming all retrieved alliance announcements are valid), we know that the upper bound on SDC's completeness of coverage is 71% (1000/(1000+500-100)), and the upper bound of MERIT-CATI's completeness is 36% (500/(1000+500-100)). Furthermore, the completeness is likely much lower than these upper bounds; if we assumed the firms draw their samples randomly from a population of alliances (an

[4] For example, suppose a researcher searches company press releases and newswires using LexisNexis with a search of *"alliance" w/10 ("formed" or "announced")*. This search will retrieve many announcements about the formation of alliances, but it will also retrieve many announcements about the activities of former alliances (e.g., "The Itanium Solutions Alliance today announced the availability…"), announcements of firms that have "Alliance" in their name (e.g., "Alliance Capital Management L.P. today announced that…"), and many other releases that are not actually alliance announcements, thus requiring the researcher to carefully read and screen the results. At the same time, the search will not retrieve alliance announcements that are articulated differently (e.g., "Taro Pharmaceuticals U.S.A., Inc., has entered into an agreement with Schering-Plough HealthCare Products, providing for the joint development and marketing…"), requiring the researcher to construct many different such searches in hopes of finding most publicly-announced alliances. The researcher will also will also be discouraged to discover that the inclusion of particular newswire sources and the depth of coverage from any particular source (e.g., "full coverage" versus "select coverage") changes over time in both LexisNexis and Factiva, eroding most hope of amassing an exhaustive list of alliances in this fashion. For example, while LexisNexis includes Business Wire from September 19, 1983, Factiva's first inclusion of Business Wire begins on July 28, 1988, and careful reading reveals that coverage from 1989 through 1999 was only "selected."

inappropriate assumption but useful for our purposes here), this would suggest that SDC's coverage is only 20% complete (1000/(1000 X (500/100)))[5] and MERIT-CATI's coverage is only 10% complete (500/(500 X (1000/100))).[6] We can thus get some sense of the completeness (or incompleteness) of the databases by evaluating their consistency of coverage, i.e., "to what degree do the databases report the same alliances?"

I begin by comparing the data in the multi-sector databases. The CORE database is limited to research joint ventures filed under the NCRA-JV act (which requires more effort than simply announcing an alliance, and is explicitly US based), therefore we would not expect the CORE database to list all of the technology alliances identified in the SDC and MERIT-CATI datasets. Furthermore, the vast difference in scale across the SDC and MERIT-CATI datasets already suggests that coverage is not highly consistent across the two databases. However, what remains unanswered is what percentage of alliances in the MERIT-CATI dataset can be found in the SDC dataset and vice versa?

To address this, I took three random samples of 100 alliances each from the SDC and MERIT-CATI datasets (600 total). Since consistency or completeness might change over time, the samples were taken for the years 1990, 1995, 2000. For these samples, I then manually searched for the corresponding alliance in the other database. For example, if an alliance appeared in the 1990 SDC sample of 100, I searched the 1990 MERIT-CATI data for each partner of the

[5] If a random draw of 1000 alliances only yields 20% of the MERIT alliances, and if the MERIT alliances are equally random, then this would suggest that the population of alliances is 5000

[6] If the databases have a similar bias in their collection efforts (e.g., oversampling of large, well-known firms in developed economies), their coverage statistics could be even worse (because both could be systematically neglecting large number of alliances formed by small organizations or organizations in less-developed in economies).

alliance, and if a set of matching alliance partners were found, the information on the nature of the alliance ("Deal text" in SDC; "Aim" and "Info" in MERIT-CATI) were compared to verify that the alliances were in fact the same. I tested the reliability of this search process by having MBA-level research assistants replicate each search (each sample search was replicated by at least one research assistant). A comparison of the multiple searches for the same database pair indicated almost 100% agreement -- there was only one alliance that was found in a search that was not found in its corresponding replication search. The results of the comparison are reported in table 2.

---------------------------------Insert Table 2 Here ---------------------------------

As shown in table 2, the consistency in coverage is quite low across the SDC and MERIT-CATI databases. Less than 5 percent of the alliances in the SDC samples could be found in the MERIT-CATI dataset. Furthermore, even though the SDC database is much larger than the MERIT-CATI database, only 26 percent of the alliances listed in the MERIT-CATI samples could be found in the SDC database. Part of this difference may be due to intentional differences in selection and reporting across the two datasets. For example, the MERIT-CATI database does not report alliances that do not have at least two industrial partners, whereas SDC will report alliances between any two or more organizations. Thus while SDC reports a large number of alliances between a firm and a university or between a firm and a government laboratory (as well as other combinations such as university-laboratory; university-university, laboratory-laboratory), these alliances do not appear in MERIT-CATI. The MERIT-CATI dataset also has a heavy focus on industries that would be considered high technology, whereas SDC reports a much greater number of alliances that have a research or technology focus, but are not in high technology industries. It is worth noting, however, that a great number of alliances were observed that appeared in only one dataset even though they appeared to meet the criteria of both datasets. For example, SDC reports a 1995 joint venture between Nintendo and GTE Interactive Media to jointly develop video games. This venture would seem to meet the MERIT-CATI selection criteria, and received news coverage in a number of media sources, including *Broadcasting & Cable*, *Electronic News*, *Telephony*, and the *Wall Street Journal - Eastern Edition*, yet the joint venture does not appear in the MERIT-CATI dataset. In the same year, MERIT-CATI reported a partnership signed between Asyst Technologies and Xerox to create a materials handling system to facilitate the manufacturing of flat panel displays. This deal should have met the SDC search criteria, and was written up in *Electronic Buyers' News* and *Electronic News*, and a press release issued on *Business Wire*, but the deal does not show up in the SDC database.

I next turned to comparing the consistency of coverage of biotechnology alliances by SDC, MERIT-CATI and RECAP.[7] First, I took three random samples of 100 alliances each from the RECAP dataset, the SDC biotech alliance set, and the MERIT-CATI biotech alliance set (900 total). As before, the samples were taken for years 1990, 1995, 2000, and I searched for the alliances manually in the other databases, and these searches were replicated by research assistants, again achieving nearly 100% agreement. The results are reported in table 3.

-------------------------Insert Table 3 About here -------------------------------

Table 3 indicates that despite the very similar patterns and raw numbers of alliances reported in the biotech alliance data, the consistency of coverage is still quite low. SDC identified only 7.8 percent of the biotech alliances reported in the MERIT-CATI and RECAP datasets. MERIT-

[7] The significant differences in both collection methods and reporting methods made the Bioscan data unsuitable for evaluating consistency of coverage in this way.

CATI performed somewhat better, identifying 14.2 percent of the biotech alliances reported in the SDC and RECAP datasets. RECAP performed the best, identifying an average of 22.2 percent of the biotech alliances reported in SDC and MERIT-CATI.

The implication of these analyses is that neither SDC, MERIT-CATI, nor RECAP can be considered to contain the population of research and technology alliances, and thus are at best samples. The CORE database contains a population (the population of firms that filed research joint ventures under the NCRA act), but that population is a small subset of formally-announced research and technology alliances. That said, samples can be very useful for a number of types of analysis if they are representative of the larger population. To gain some insight into this, I now turn to looking at the degree of agreement between the datasets about major patterns in alliance activity over time.

CONSISTENCY IN PATTERNS OF ALLIANCE ACTIVITY

In this section, I evaluate the degree to which the different alliance databases yield consistent patterns of alliance activity in such dimensions as sectoral composition, temporal patterns in alliance activity, and patterns in geographic representation.

Sectoral Composition

The MERIT-CATI data is coded into twenty-four different sectors, grouped into four fields of technology: Biotechnology, Information Technology, New Materials, and “Not Core Technology” (which includes transportation equipment, chemicals, and other sectors). Figure 1 shows the MERIT-CATI data decomposed into its primary constituent sectors.[8] This graph indicates that alliances in information technologies (which includes computers, industrial automation, microelectronics, software and telecommunication) and biotechnology (which includes pharmaceutical biotech, agricultural biotech, environment biotech, nutritional biotech and fine chemical biotech) account for a substantial portion of the overall worldwide alliance activity reported in the database. Over the time period 1990 to 2004, information technology accounted for an average of 41% of the worldwide alliances reported, and biotechnology accounted for an average of almost 33%. Though MERIT-CATI reports automotive and aerospace/defense alliances separately, I have collapsed these categories into a single "transportation equipment" category in order to permit comparison with the CORE data. This category appears to be the third most prominent in the MERIT-CATI dataset, accounting for an average of 12% of the alliances over the time frame.

----------------------Insert Figure 1 here --------------------

Figure 2 shows a similar decomposition for the CORE database. The CORE Database reports alliances in eleven NAICS categories corresponding to: Food Manufacturing; Petroleum & Coal Products Manufacturing; Chemicals; Fabricated Metals; Machinery; Computer & Electronic Products; Electrical Equipment, Appliances & Components; Transportation Equipment Manufacturing; Miscellaneous Manufacturing; Broadcasting & Telecommunications; Professional, Scientific & Technical Services. For ease of comparing this data to the MERIT-CATI data, I have collapsed the Computer & Electronics Products, Electrical Equipment, Appliances & Components, and Broadcasting & Telecommunications categories into a single "Information Technologies" category. What is immediately apparent is that information

[8] This sector decomposition comes from data reported by MERIT-CATI in the National Science Foundation's 2006 Science and Engineering Indicators.

technology alliances also drive a large portion of the yearly variance in the alliances reported in the CORE database over the 1990-2005 timeframe, consistent with the MERIT-CATI data. Information technology alliances account for an average of 40% of the alliances reported in the CORE database. Another similarity to the MERIT-CATI data is the prominence of the transportation equipment manufacturing series, which is the second largest series in the CORE database, accounting for 13% of the alliances on average. Chemicals alliances are the third most prominent category, accounting for 11% of the alliances on average.

There are, however, some differences in reporting between CORE and MERIT-CATI that are difficult to reconcile. First, the CORE database does not report biotechnology (or pharmaceuticals) as a separate category. Though it is likely a large portion of these alliances are in the Chemicals category, a portion of them might also be in the Professional, Scientific & Technical Services category, as the latter category includes scientific research and development services. It is clear, however, that even if the chemicals category and the Professional, Scientific & Technical services categories were combined,[9] they would not achieve the prominence that the biotechnology series exhibits in the MERIT-CATI database. Furthermore, the MERIT-CATI database reports biotech and chemicals alliances separately; if these two categories were consolidated, the series would be even more prominent.

-----------------------Insert Figure 2 About here---------------------

[9] This raises a second complexity in reconciling the data to the MERIT-CATI categories. While a significant portion of the Professional, Scientific & Technical Services category might be biotechnology related as noted, it is also possible that a significant portion of these alliances would belong in "Information Technologies" category as Professional, Scientific & Technical Services also includes computer systems design. Neither of these problems significantly influence our interpretation however, since as shown in the chart, Professional, Scientific & Technical Services do not account for a large portion of agreements reported in the CORE database.

The SDC data is available at the four-digit Standard Industrial Classification level, at the four-digit Venture Economics Industry Code level, and by custom categories that SDC creates such as "biotechnology" and "communications." For simplicity of comparison with the other datasets, I will present the data broken down into biotechnology alliances, chemicals (including pharmaceutical but excluding those coded as biotech), transportation equipment, information technology alliances, and total alliances (see Figure 3). Comparison of data based on the SDC custom category for "biotechnology" and the Venture Economics Industry Code for "biotechnology" indicated that the SDC custom category was much more congruent with the definition of biotechnology used by other databases, so the SDC custom category was used to identify biotechnology alliances.[10] The chemicals category includes everything in SIC 28--, including pharmaceuticals, but excludes those alliances coded as biotech per the SDC custom category. Transportation equipment is everything in SIC 37--, and thus includes the manufacturing of motor vehicles, aerospace equipment, and other transportation equipment. To match the MERIT-CATI data as closely as possible, I included computers (SIC 3571-3577), Microelectronics (SIC 3671-3679), software (SIC 7371-7379), telecommunications equipment (SIC 3661-3669), and communications services (4812-4899) in the information technology category.

Notably, SDC's coverage is very wide, causing most sectors to account for a relatively small portion of the alliances over the timeframe (please see appendix 2 for a detailed breakdown of the distribution of SDC alliances by SIC code). Thus when a line is included for total alliances in the graph, it causes most of the other series to be compressed near the horizontal axis and difficult to discern. I therefore included instead a line for the total number of alliances divided by

[10] For parsimony this comparison is not provided here but is available from the author upon request.

three so that the other series were more clearly visible. Consistent with the other datasets, the chart indicates that the composite category of information technology accounts for a large portion of the total number of alliances (22% over the timeframe), and appears to account for a significant amount of the variance in alliance activity reported over 1990-2005. Biotech alliances account for 10% of the alliances, followed closely by chemicals alliances (including pharmaceutical alliances) at almost 9%. Transportation equipment alliances are also significant (just over 4 percent of the total alliances), similar to the pattern reported in both the CORE and MERIT-CATI data.

-------------------------Insert Figure 3 here-------------------------------

While these comparisons suggest that there is considerable agreement about the prominent role of information technology, transportation equipment and chemicals in the alliance databases, a chi square test comparing the distribution of alliances for these three sectors across the databases indicates that we would reject a hypothesis that the distributions are the same across the databases (Chi square statistic: 447, $p<.001$). The comparisons also raise some serious questions about the relative prominence of biotechnology alliances: biotech alliances account for almost a third of the alliances reported in MERIT-CATI over the 1990 to 2004 time frame, and for more than 50% of the alliances in some of the later years of that time period. Though biotechnology alliances are important in the SDC data, they account for only 10% of the alliances over the time frame, and it is clear they do not account for a large portion of the alliances in the CORE database. This raises the possibility that biotechnology alliances are overrepresented in the MERIT-CATI data -- a factor that could significantly impact research that uses cross-sector comparisons of alliance activity. We will return to this possibility in the “Replication of Prior Studies” section.

Temporal Patterns in Alliance Activity

To what extent do the databases agree about alliance activity over time? Casual inspection of the previous figures suggests that all of the alliance databases report a significant peak in alliance activity in 1995, but we can get more traction on this issue by standardizing the data and comparing the databases both at the all-sector level and within individual sectors.

Figure 4 provides a set of charts showing the standardized alliance counts over the 1990 to 2005 time frame from the MERIT-CATI database, the CORE database, and SDC database.[11] This chart shows strong agreement about the peak in alliance activity in 1995, and the standardization of the values further reveals that the datasets have high agreement about the relative magnitude of the peak in comparison to the overall variance in alliance announcements over the 1990-2005 time period. The chart highlights, however, that there is much more correspondence between the SDC and the CORE data (a 75% correlation), than between those datasets and the MERIT-CATI data (-.16% correlation between MERIT-CATI and SDC; -.04% correlation between MERIT-CATI and CORE). The coefficient alpha of treating these three series as multiple measures of the same phenomenon is .40, but would increase to .86 -- indicating high reliability (Nunnally, 1978) -- if the MERIT-CATI data were omitted. The differences in patterns are straightforward to identify. The SDC and CORE data both show a strong upward trend in alliances between 1989 and 1991, followed by a modest decline in 1992. MERIT-CATI, on the other hand, shows the opposite. In the MERIT-CATI data, alliances are falling from 1989 to 1991, and then turn sharply upward in 1992. All three datasets show a sharp peak in 1995, followed by declines in

[11] Data was only available through 2004 for the MERIT-CATI database, but to facilitate comparisons across graphs, the horizontal axis will show the years 1990-2005 throughout the paper.

alliance activity from 1995 to 2000 (with a fair degree of noise). However, whereas the SDC data and the CORE data show a sharp decline between 2000 and 2001 followed by further trailing downwards in alliance activity until 2004, the MERIT-CATI database shows a sharp increase in alliance activity between 2000 and 2001, followed by a dip and another sharp increase in 2003 and a more modest increase in 2004.

--------------------------Insert Figure 4 Here ---------------------------

As was noted earlier, the MERIT-CATI database is much more heavily influenced by data on biotech alliances than the other two databases. Biotech alliances account for almost 33 percent of the total alliances reported in the MERIT-CATI database on average over the time period 1990-2004. Over the same time period, biotech alliances account for only 10 percent of the joint research and technology alliances reported in the SDC database, and appear to be even less prominent in the CORE data. Since there is a strong upward trend in biotechnology alliances over the timeframe, this suggests that the prominence of the biotech data in the MERIT-CATI dataset may be disproportionately responsible for the divergence between the MERIT-CATI all-sector data and the all-sector data from the other two datasets. To explore this possibility, in figure 5 I compare standardized versions of the SDC, MERIT-CATI and CORE data with biotech alliances removed from both SDC and MERIT-CATI.[12]

-------------------Insert Figure 5 Here----------------

Removing the biotech data improves the correspondence between the graphs dramatically. The correlation between the MERIT-CATI data and the CORE data increases to 61%, and the correlation between the MERIT-CATI data and the SDC data increases to 46%, and the correlation between the SDC data and CORE data increases to 77%. If these data sets were treated as multiple items of a single measure of alliance activity, the combined measure would

[12] Biotech alliances cannot be easily removed from CORE database since it provides no coding for biotech alliances.

have a coefficient alpha of .83, and thus would be considered very reliable (Nunnally, 1978). The coefficient alpha would increase to .87 if the MERIT-CATI data were omitted from the measure.

To gain a finer-grained perspective on the reliability of temporal patterns in the alliance databases, I will next examine patterns in individual sectors that are prominent in the alliance activity across the datasets. For one such sector, biotechnology, we will also be able to consider data from RECAP and Bioscan.

Information Technology. Figure 6 shows the standardized alliance counts for the information technology sector from SDC, MERIT-CATI and CORE. As shown in figure 6, the overwhelming peak in 1995 for the information technology alliances tends to make these graphs look fairly similar. The correlation between the data is also fairly strong (32% between the SDC data and the MERIT data, 63% between the SDC data and the CORE data, and 56% between the CORE data and the MERIT data). Treated as multiple measures of the same phenomenon, they would achieve a coefficient alpha of .75 -- above typical thresholds for measure reliability (Nunnally, 1978; Peterson, 1994). Furthermore, omitting any of the databases would lower the coefficient alpha. Graphical visualization of the data does, however, suggest some interesting one- to two-year time lags between some of the peaks and valleys in the data sets, which will be discussed further in the transportation equipment section.

----------------------Insert Figure 6 Here----------------------------

Transportation Equipment. The results for the transportation equipment alliances are similar to those for the information technology alliances (see Figure 7); a peak in the mid-1990s dominates the graph. There is a particularly close correspondence between the SDC and CORE data for the transportation equipment alliances with both showing a first peak in 1991, followed by a modest decline, then another peak in the 1994-1995 time frame, followed by a long decline, and finishing up with a small peak in the 2004. The data are, again, significantly correlated (57% correlation between SDC and MERIT, 73% correlation between SDC and CORE, and 74% correlation between MERIT and CORE). Together they achieve a coefficient alpha of .87, which is again well past most thresholds for reliability testing (Nunnally, 1978; Peterson, 1994).

----------------------Insert Figure 7 Here-----------------------------

As noted in the information technology alliance data, and as apparent in the graph here, there appear to be some one-to-two year lags between some of the peaks and valleys. There are a number of potential explanations for these lags. First, there may be a time lag between when alliances are announced in the press (and thus susceptible to showing up in the SDC and MERIT-CATI datasets) versus when they are formally registered with the NCRA and thus show up in the CORE database. Second, there may be some systematic differences in coding practices. For example, an alliance might be coded with a date corresponding to its announcement -- which may be in advance of its commencement -- or its actual commencement.[13] For example, whereas SDC will report alliances that are "pending," MERIT-CATI attempts to report only those alliances that have been completed, and CORE reports only those alliances that are officially filed. It might be possible, then, for patterns in the MERIT-CATI and CORE data to slightly trail patterns in the SDC data. This could cause problems in research uses of the data if a) the researcher is not consistent in their use of a particular database, or b) the study is very sensitive to the lags specified between alliance activity and other variables of interest.

[13] It is worth noting here that SDC and RECAP report alliance dates in a month-day-year format, whereas MERIT-CATI and CORE report alliances by year only, and Bioscan reports alliance dates in a variable (either month-year or year only) format.

Chemicals. The data for the chemicals sector is more worrying: There is not any obvious correspondence across the datasets for chemicals in either the absolute number or the pattern (see figure 8). SDC reports a much larger number of chemicals technology alliances (4,982 over the time frame) and shows a strong peak in the 1993 to 1995 time range, driven in large part by pharmaceutical alliances (which account for 41% percent of the chemicals alliances on average, and more than 56% in some years). MERIT-CATI reports 581 chemistry alliances over the same time frame, with great year-to-year volatility and no noticeable trend. The CORE database reports only 85 alliances over the time range, also with significant year-to-year volatility and no noticeable trend. While there is a 45% positive correlation between the SDC data and the MERIT data (which nearly achieves significance at the $p<.05$ level), the correlation between the CORE data and SDC is only .09, and the correlation between CORE and MERIT is small and negative. The overall coefficient alpha of the three datasets would be .34, suggesting poor correspondence.

--------------------------Insert Figure 8 Here-------------------------

The fact that there is high correspondence between the datasets on information technology and transportation equipment but poor correspondence between the datasets on chemicals suggests that the data may be reliable enough to show agreement about large trends (such as the significant mid-1990s peaks in information technology and transportation technology alliances), but not reliable enough to show agreement about more minor fluctuations in alliance activity. This should not be surprising given the relatively poor results of the consistency of coverage analysis. In absence of a significant trend, the samples primarily indicate noise.

Biotechnology Alliances. For the biotechnology sector I will compare temporal patterns in alliance counts from SDC, MERIT-CATI, RECAP and Bioscan. CORE is excluded here since it does not provide any means of separating out biotechnology alliances. Because there is remarkable consensus across the databases about the absolute number of biotechnology alliances, I am able to use unstandardized alliance counts in Figure 9. The most immediate observation from Figure 9 is that the biotech data show a very different pattern than the information technology data, and that in general there is considerable agreement (with a few exceptions) about the overall pattern over time. All four data sets show a significant increase in biotech alliance activity over the time period. The RECAP, Bioscan, and MERIT-CATI show a slight downturn in 1991, followed by a long period of pronounced growth in alliance activity through 1997. SDC's pattern is somewhat different, with growth in 1991, followed by a relatively flat period through 1996, followed by a sharp increase through 1999. All four data sets show some type of wavering in the 1998 through 2000 period, perhaps reflecting the volatility in technology-news related reporting during the information technology bubble. RECAP resumes a sharp upward trend in 1999, followed by MERIT-CATI and SDC in 2000. RECAP, SDC, and MERIT-CATI all then show some flattening or decrease in alliance activity in the 2001 to 2004 time period. The Bioscan pattern is different from 1999 through 2004 time period, with a lower number of alliances reported in most of the time frame, and a relatively smooth exponential curve in alliance activity throughout. It is worth noting that there is a remarkable similarity in the absolute numbers of alliances reported from all four sources, particularly during the early 1990s, and then again toward the end of the timeframe. The correlations between the SDC, MERIT-CATI, and RECAP data are all 80% or higher. The correlation between Bioscan and SDC, MERIT-CATI and RECAP are 57%, 68% and 54% respectively. Together the measures achieve a coefficient alpha of .91 (that would increase slightly to .93 if Bioscan were omitted), indicating remarkably high reliability (Peterson, 1994).

-------------Insert Figure 9 Here ------------------

Overall then, the reliability analysis of temporal patterns in alliance activity yields more optimistic news than the consistency of coverage analysis: the five databases exhibit remarkably similar patterns in alliance activity over time (at least for those prominent sectors that have a strong trend) despite significant differences in their overall scale. This may indicate that the databases are reasonably representative samples of the true alliance activity over this time period, or that the databases exhibit highly similar biases in their sample selection.

Geographic Patterns in the Alliance Data

The last pattern that I will examine pertains to geographic scope. Do databases that are headquartered in different regions of the world exhibit different geographic biases in their data collection? Or more broadly, do the databases differ in their extent of coverage of alliance announcements from around the world? To address these questions, I will focus on SDC and MERIT-CATI, the only two databases in the study that are both explicitly international in scope and reliably report the nation of the alliance participants.

Both SDC and MERIT-CATI rely primarily on articles written in English and thus may have a bias toward North American and Western European data sources.[14] Furthermore, though both MERIT-CATI and SDC attempt to be internationally inclusive, it would be reasonable to assume that MERIT-CATI might be more Euro-centered, whereas SDC might be more US-centered based on the nationality of the organizations. The data, however, does not appear to support these

[14] The MERIT-CATI data also includes data gathered from articles in the Dutch and German press and translated abstracts of important foreign newspapers and trade journals.

speculations. While both datasets may have a western bias, the data does not suggest that the MERIT-CATI data is more Europe-centric than the SDC data (see Table 4). Table 4 shows a count of how many times each OECD country was represented in the research and technology alliances (that is, a count of each time an organization headquartered in a given OECD country appeared in an alliance announcement) for each database over the 1990 to 2005 time frame, along with percentages, and aggregates by region. As shown, both datasets report more U.S. participants than participants from any other country, and for both datasets the aggregate for North America is over 1.5 times the aggregate for the next highest region, Europe. The main difference in geographic coverage across the two datasets appears to be that SDC has far more participants reported that are from non-OECD countries (21.48% versus 3.38%). The non-OECD participants in the SDC database are overwhelmingly Asian, with the leading countries being China, Malaysia, Singapore, Hong Kong, India and Thailand, in that order, and collectively accounting for 17.28% of the country-participant counts.

--------------------------Insert Table 4 About here --------------------------------

On the one hand, this analysis suggests that it might be naïve to presume that a database's geographic location leads to a bias in favor of alliance announcements in its home region. On the other hand, the databases do appear to have significant differences in the geographic scope of their coverage of alliance announcements. Because we do not know which, if either, of the databases has a more representative sample, this analysis suggests that we should be hesitant to use either of the databases to draw conclusions about geographically-based differences in alliance activity.

REPLICATION OF PRIOR STUDIES

The previous analyses raised a number of important issues for how differences and similarities in the alliance databases might impact research results. To further explore this, in this section I perform replications of three previously published studies that represent highly different uses of the data. These studies include 1) a study that describes temporal and sectoral patterns in alliance activity using the MERIT-CATI data, 2) a study that combines SDC alliance data with other variables of interest to address industry differences in organizational form, and 3) a study that conducts a network analysis using Bioscan alliance data. I replicate each of the studies using a different database than the one that was used in the original study, as described below.

1. Patterns in alliance data: Replication of Hagedoorn's (2002) study of relative internationalization indexes

In 2002, Hagedoorn published a study in *Research Policy* that examined major trends and patterns in inter-firm R&D partnerships since 1960. One of the key aspects of this study was an examination of the internationalization of R&D partnerships, that is the percent of R&D partnerships that include participants from different nations, based on data from the MERIT-CATI database. If we think that internationalization may vary by geography (which is a reasonable assumption), then the significantly different geographic scope of the MERIT-CATI database versus the SDC database might lead to significantly different patterns when the study is replicated using the SDC database.

Hagedoorn reports his data in four different time periods (1960-1969; 1970-1979; 1980-1989, and 1990-1998) permitting me to replicate his study for the 1990-1998 time period using SDC data. This examination considers both the percent of international R&D partnerships overall, and a breakdown by sector. Hagedoorn notes that in the MERIT-CATI data, international alliances (i.e., those with participants from two or more nations) account for "about 60%" in the early 1990s, and decline to below 50% of all newly made R&D partnerships by the late 1990s. Using the SDC data the percent of international R&D alliances averages 67% for the period 1990 to 1994, and 66% from 1995 to 1998 (in conformance with the Hagedoorn study, I use data only through 1998). Thus the percentages are roughly comparable to those found in the Hagedoorn study, but without the significant decline reported in the Hagedoorn study. Notably, if one includes 1999 and 2000 data in the SDC replication, the average does decline to 64%.

Hagedoorn (2002) also calculates a "relative international partnering" index (RII) per sector whereby the relative distribution of the sectoral number of international partnership (IPi) and sectoral domestic partnerships (DPi) are set against the distribution of all international partnerships (TIP) and all domestic partnerships (TDP):

$$RII_i = \frac{IP_i / DP_i}{TIP/TDP}$$

Hagedoorn's key findings here are that a) contrary to what one might expect, the high tech sectors do not exhibit higher than average internationalization in their partnering, and b) the propensity for international partnering varies considerably across industries. Using the SDC data yields similar conclusions, though the absolute magnitude of the relative internationalization indexes are often quite different (see Figure 10). Within the high tech sector, the results using SDC data are very similar to those found in Hagedoorn (2002): pharmaceuticals and information technology both have relatively low internationalization indexes in comparison to aerospace and defense. The results for instruments and medical equipment, chemicals, and oil and gas are also very comparable across the two studies. In the SDC data, however, the automotive, consumer electronics, electrical equipment, food and beverage, and metals industries all exhibit significantly higher relative internationalization indexes than those reported in the Hagedoorn study.

--------------------------Insert Figure 10 About Here----------------------

What could drive the substantial differences in relative internationalization indexes reported for these industries? Casual inspection indicates that cross-border alliances in these industries have disproportionately higher representation by Asian countries in the SDC database (e.g., 34% of the crossborder alliances across these three industries list a partner from Japan, compared to only 20% for non-crossborder alliances, and 21% of the crossborder alliances list a partner from China compared to only 5% for non-crossborder alliances), suggesting that the difference in the results may be driven in part by the greater coverage of Asian alliances in the SDC database. This reaffirms our earlier concerns about geographic coverage and raises a new concern: If geographic areas of the world are differentially represented across sectors of industry (Japan's prominence in automobiles and consumer electronics, and Europe's prominence in pharmaceuticals, for example), then differences in the scope of geographic coverage could lead to biases in studies that examine sectoral differences in alliance activity. Though the earlier analysis of sectoral composition suggested considerable agreement about the sectors that make heaviest use of alliances, there could still be significant differences when more sectors are considered or when sectors are defined more narrowly. The next replication, which looks at inter-industry differences in alliance activity, addresses this concern more directly.

2. Combining alliance data with other variables of interest: Replication of Schilling & Steensma's (2001) study of modular organizational forms

In 2001, Schilling and Steensma published a study in the *Academy of Management Journal* that examined whether modular systems theory could predict the use of loosely-coupled organizational forms such as contract manufacturing, alternative work arrangements, and alliances. Their study combined SDC alliance data with other archival industry-level measures to (among other things) predict inter-industry differences in use of alliances. Differences in sectoral coverage across the databases could, therefore, yield highly different outcomes in a replication of this study.

The sample in the Schilling and Steensma (2001) study included 330 four-digit (SIC) manufacturing industries. They created a measure of alliance formation based on a count of alliances per industry by using the SIC code of each partner in each alliance, as reported by SDC. This count was then normalized by dividing it by the number of firms in the industry as reported by the Census Bureau.[15] Both counts used 1997 data. They then regressed this measure against labor intensity (total industry employment divided by industry sales, both for 1992), heterogeneity of inputs and demands (a compound measure based on the 1992 Benchmark Input Output data released by the Bureau of Economic Analysis), technological change (average total factor productivity growth rate from 1980 to 1994, as reported in the Bartelsman-Gray database

[15] Notably, this count of firms is only a rough proxy of industry size since Census data is restricted to US firms.

at the National Bureau of Economic Research), competitive intensity (a composite measure of number of firms, the inverse of the four-firm concentration ratio, and the inverse of the Herfindahl Hirschmann index), availability of industry standards (a dummy variable based on the existence of a standards-setting organization in the industry), and several interaction terms.

Schilling and Steensma's key findings were that heterogeneity of inputs and demands, technological change, and industry standards were significantly and positively related to alliance formation, consistent with the modular systems theory. I attempted to replicate this study as closely as possible using the MERIT-CATI dataset. I was able to use the original data for the independent variables collected by Schilling and Steensma. I then coded the MERIT-CATI 1997 alliance data into four digit SIC codes using a combination of the categorization scheme used in the MERIT-CATI database, and the deal text descriptions of each alliance. The MERIT-CATI database reports 581 alliances formed in 1997. Of these, 339 could be classified into a four-digit manufacturing industry code, 23 were too broadly defined to be classified, and the remainder was in non-manufacturing industries (primarily software and business services). This count was then normalized using the same firm numbers used in Schilling and Steensma (2001), and the regressions were replicated (see Table 5).

------------------------Insert Table 5 About Here--------------------------------

As shown, the results are very similar to, though weaker than, those obtained in the original study. Heterogeneity of inputs and demands, technological change, and industry standards are still positively related to alliance formation though some of the other results disappear (e.g., a negative effect for competitive intensity found in the original study disappears), and the R^2 of both the restricted and full model are lower. This result is reassuring in that it suggests that inter-industry patterns in alliance activity appear to be relatively consistent across the databases. If differences in geographic scope impact sectoral coverage, this impact does not appear to be severe enough to obscure significant relationships between alliance activity and other variables of interest.

3. Alliance network dynamics: Replicating Powell, Koput and Smith-Doerr's (1996) study of biotechnology collaborations

In 1996, Powell, Koput and Smith-Doerr published a study in *Administrative Science Quarterly* that has become a seminal article on networks as the locus of innovation. Replicating this study with another database should prove particularly illuminating. Network analyses can be highly sensitive to the omission of important nodes, thus having a sample rather than a population can raise serious concerns for studies that attempt to draw conclusions about network structure. If, for example, an important hub is inadvertently missed, the connectivity of the network may be greatly underestimated.

Powell et al. gathered data on 225 dedicated biotechnology firms over the time period 1990 to 1994, relying primarily on Bioscan and supplementing this data when necessary with data from other industry sources, annual reports, SEC filings, and when necessary, calling the companies directly. Their primary measures were counts of R&D alliances, counts of non-R&D alliances, each firm's degree and closeness centrality in each year (calculated using UCINET), and a measure of network portfolio diversity calculated as a Blau index of heterogeneity,

$y_{it} = 1 - \sum p^2_{it,j}$

where $p_{it,j}$ is the proportion of firm *i*'s ties of type *j* in year *t*, out of *i*'s total number of ties in year *t*.

Some of Powell, Koput and Smith-Doerr's key hypotheses were that 1) the greater the number of R&D alliances a firm has at a given time, and experience at managing alliances the firm has at a given time, the greater the number of non-R&D collaborations it subsequently pursues, and the more diverse its future portfolio of ties will become, controlling for prior levels; 2) the greater the firm's number of R&D alliances, its diversity of ties, and its experience at managing alliances, the more centrally connected the firm subsequently becomes, controlling for prior connectedness; and 3) the greater a firm's centrality in a network of relationship at a given time, the greater its number of subsequent R&D collaborations, controlling for prior R&D activity. Collectively, these hypotheses suggest considerable endogeneity in network dynamics. I attempted to replicate this study as closely as possible using the SDC database. Since I was unable to obtain the original list of companies used by Powell et al.,[16] I began with the companies listed in Bioscan as of March 2006 that had at least one alliance between 1990 and 1994. This resulted in 217 companies. I removed all suffixes from these companies (Corp., Ltd, Co., Inc., LLC, etc.) and then searched for these companies within the set of all alliances in SDC coded as biotech. Using SDC's alliance classifications of R&D, marketing/licensing, manufacturing, and joint ventures, I created counts for each company of the number of alliances

[16] I contacted the authors to attempt to obtain this list but it was unavailable, as it had been subjected to periodic updating for follow-on research projects.

of each type it had in each year. [17] Any given alliance could be classified as more than one type, so an alliance could, for example, be both an R&D alliance and a joint venture, and be counted for both categories. I also used these counts to recreate Powell et al.'s diversity measure, and used UCINET to calculate normalized degree centrality and closeness centrality for each company. To calculate R&D network experience and non-R&D network experience, I calculated the number of years since the firm's first R&D alliance or non-R&D alliance in this set. Since I only used data going back to 1990 for SDC for reasons noted previously, the maximum value this measure can take in the replication is four years; however, 95% of the alliances reported in Bioscan are also post 1990 so this limitation should have limited effect on the replication.[18]

As shown in Table 6, I obtained results that were similar to those found in the original Powell et al. study. First, like Powell et al., I found that the number of R&D ties at a given time appear to be related to its subsequent marketing/licensing ties, but not significantly related to its subsequent manufacturing ties. Second, I obtained the same coefficient between R&D ties and network diversity as that obtained in the Powell et al. study, though in the replication this coefficient fails to achieve statistical significance. Third, consistent with Powell et al., I found that the more R&D ties a firm has and the greater its network diversity, the more centrally connected the firm subsequently becomes. On the other hand, my results for the effect of network experience on network portfolios and on degree centrality, were different from Powell et al.'s, suggesting that perhaps my inability to utilize data prior to 1990 had a greater effect on the experience measure than I anticipated. Furthermore, whereas Powell et al hypothesize and obtain

[17] Joint venture ties are used in the "total ties" measure and in the calculation of network diversity, however there was insufficient variation in joint ventures to use it as a dependent variable in the panel models.
[18] I did not attempt to gather financial data on these firms (e.g., size, whether they were publicly-held) as this would have been excessively burdensome (particularly for the privately-held firms) and these variables tend to have little temporal variance, and thus would largely be captured by the firm fixed effects.

a significant positive relationship for closeness centrality on subsequent R&D ties, I obtained a significant and negative relationship. Finally, there were also a number of curious differences in the coefficients obtained for control variables which might suggest that the replication suffers from omitted variable bias due to my inability to include all of the controls used in the original study. In sum, like Powell et al., I found significant indication of network endogeneity with the primary differences in results being in those for network experience, and in the effect of centrality on future R&D ties.

-------------------------Insert Table 6 About Here------------------------------

Given the potentially significant differences in our starting sample, the differences in our analytic techniques (I used negative binomial panel models for the models for which the dependent variable is a count of ties), and my omission of some control variables, I interpret the similarities in these results as generally reassuring: despite being incomplete, the databases yielded similar conclusions about network dynamics. Perhaps this should not, however, be so surprising. Networks of organizations have strongly skewed degree distributions, meaning that some organizations have many more connections than others and are disproportionately responsible for the connectivity of the network and other structural statistics of interest. As mentioned previously, sampling the nodes in such networks can raise serious problem. However, unlike most network sampling methods, the alliance databases here (with the exception of Bioscan) are sampling on the "links" (the alliances) rather than the "nodes" (the organizations). This means that the likelihood of an organization making it into the sample is directly related to the number of alliances it publicly announces, reducing the likelihood of an important hub being overlooked. Furthermore, an organization's size and prominence is directly related to both the number of alliances it is likely to have, and the amount of press attention it is likely to receive, further reducing the likelihood of a major hub being overlooked, at least in the datasets that consider all forms of organizations.[19] Finally, because there is considerable stability in the hubs (that is, firms that are hubs in a given period are highly likely to be hubs in preceding or subsequent periods), these hubs create resilient scaffolding for the rest of the network. Collectively, these features mean that if one captures the activity of the hubs (and as noted above, the alliance databases are likely to capture the hubs), one will capture much of the structure and dynamics of the network.

Bioscan essentially samples on nodes since it identifies a group of relevant companies first and then searches for alliance data on those companies. However, Powell et al. went to considerable effort to ensure that their initial sample was complete (as detailed in their paper), supplementing the Bioscan data with data from many other news sources and interviews with biotech specialists. Thus perhaps Powell et al.'s diligence overcame any sampling problems inherent in Bioscan, and SDC's practice of sampling on links helped to ensure that the main scaffolding of the network was faithfully replicated.

CONCLUSIONS

In this paper I have examined five popular alliance databases to provide some insight into their advantages and disadvantages for use in management research. I first provided a descriptive account of some of the databases' key features. I then analyzed their consistency and completeness of coverage, the reliability of sectoral and temporal patterns, and consistency in geographic scope. Finally, I replicated several well-known studies that utilize alliance data to examine whether the differences observed between the databases would significantly impact

[19] Notably, a number of important "hubs" are universities or government labs, thus networks based on datasets that restrict alliances to those with two or more industrial partners will tend to understate the connectivity of the network

research conclusions. These analyses reveal some of the key strengths and weaknesses of the individual databases, as well as providing useful insight into how alliance databases can (and cannot) be effectively used.

One of the major findings of this examination is that the consistency of coverage between the alliance databases is rather poor, suggesting that none of the databases includes the population of formally-announced alliances -- they are all samples.[20] Only a small portion of the alliances reported in a given database are mirrored in the other databases. Furthermore, there are a number of reasons to assume that the sample effectively obtained by each database is unlikely to be random (all of the databases analyzed here are likely to have a bias towards alliances reported in English-language news sources, and are also likely to over-represent alliances formed between large firms due to news reporting biases), nor are the same biases likely to be shared across datasets given differences in collection purposes (e.g., commercial versus academic), and populations of interest (e.g., firms only versus organizations more generally).

Despite these limitations, however, the datasets exhibited similarity in sectoral composition, and marked symmetries in patterns of alliance activity over time. The three multi-sector databases exhibited agreement about the prominence of the information technology, transportation equipment, and chemical sectors in alliance activity, though they exhibit some discrepancy about the importance of biotechnology alliances.[21] The three multi-sector databases also exhibited high

[20] The CORE database contains the population of alliances filed under the NCRA-JV act, but this population is a very small portion of alliance activity.

[21] Some of the differences in the reported biotech alliances may be due to the fact that biotechnology is still a somewhat nebulous category. It is not assigned a code under the Standard Industrial Classification system or the North American Industrial Classification System, nor is it possible to identify a group of industry codes that make up biotechnology the way that the information technology category was created, because many of the industry codes that would be invoked in biotechnology alliances are also invoked for alliances that do not constitute biotechnology.

reliability in the temporal pattern of alliance activity in the multi-sector data, and within the information technology and transportation equipment sectors. Additionally, all four databases that track biotechnology alliances exhibited remarkable agreement about both the absolute number of biotechnology alliances and their pattern over time. The only real weak spot in the assessment of reliability of temporal patterns was for the chemicals sector, which exhibited low reliability across the three multi-sector databases. This may indicate that the databases only reliably report symmetric patterns when there is a strong trend (the information technology, transportation equipment, and biotechnology sectors all exhibited strong trends, whereas the chemicals sector indicated no such trend).

These results provide some reassurance that even though each databases only captures a sample of alliance activity, they may yield reliable results for many -- if not all -- research purposes. To get at this question more directly, I replicated three published alliance studies using a different database from that used in the original study. The results of these tests yielded more positive than negative news: most of the replication results yielded conclusions that were highly similar to those yielded by the original studies. It is important to note that I would not have expected to obtain *identical* results in the replications even if the databases had been entirely consistent. It is very difficult to precisely replicate a previously published study unless one works closely with the original authors and data. Many minor details of analysis are not well documented (how

---

This makes it difficult to verify that all of the databases are coding the same type of alliances as biotech alliances. While all of the biotechnology databases appear to attempt to capture the medical biotechnology alliances that would be considered archetypal for this field (e.g., an alliance to jointly develop a drug target based on a genetically engineered agents), the databases might vary in the extent to which they include categories such as agricultural products (e.g., the development of genetically-engineered seeds or bioinsecticides), machinery or other supplies used in the research or production processes of biotechnology (e.g., the co- development of software used to screen drug targets, or cross-licensing of technologies used in a molecular distillation dryer), or other industrial categories not typically associated with biological products (e.g., an alliance between petroleum companies to develop microbially-produced polymers). These definition issues may drive some of the disparity between the biotechnology data.

outliers are handled, how alliances by firms that merge during the time frame of the analysis are handled, etc.) Furthermore, in the case of the Powell study, I was neither working from the same original sample, nor did I duplicate all of the control variables in the original study. Given these sources of error, it is encouraging that roughly the same pattern of results emerged.

It is important to note, however, that the fact that the databases contain only a portion of the actual alliance activity could have detrimental consequences for some types of research. If, for example, a researcher wishes to utilize alliance data for a group of firms that overall have relatively low numbers of alliances (e.g., a study of small or new firms, a study of firms in sectors such as textiles or paper that make relatively little use of alliances, a network study that wishes to obtain accurate information about periphery firms), it would be risky to rely on any single database. Similarly, if the researcher wishes to obtain complete information about a particular firm, the researcher would be well-advised to supplement the data with their own targeted search efforts. It is also important to note that informal agreements are another significant form of collaboration between firms, and such agreements are not represented in any of these databases. Thus networks based on these databases likely understate the true degree of connectivity among firms. On the other hand, to the degree that informal collaborations between firms tend to lead to more formal collaborations over time, as has been suggested by other research (Rosenkopf, Metiu & George 2001), the structure of the formal collaboration network may provide some insight into the structure of the informal collaboration network.

Overall, the results here suggest that the alliance databases are a valuable, and generally reliable (if not exhaustive) resource for the study of inter-organizational relationships, so long as the research design takes the sample limitations into account. When the databases are being used in a research context that is characterized by high numbers of alliances (for example, when data is being aggregated up to the multi-sector level, or when the firms of interest tend to be large and operate in technology-intensive industries), the datasets are likely to yield reliable results. The results here also provide some insight into the relative strengths and weaknesses of the individual databases. Armed with this knowledge, researchers should be able to both make more informed decisions about which databases to use for their research questions, and to design their studies in a manner that makes best use of a database's strengths while attenuating its weaknesses.

REFERENCES

Anand BN, Khanna T. 2000. The structure of licensing contracts. *Journal of Industrial Economics*, **48**:103-135.

Beckman CM, Haunschild PR, Phillips DJ. 2004. Friends or strangers? Firm-specific uncertainty, market uncertainty, and network partner selection. *Organization Science*, **15**:259-275.

Grannis R. 2004. Sampling the Structure of Large-Scale Social Networks. *Proceedings of the Sixth International Conference on Social Science Methodology: Recent Developments and Applications in Social Research Methodology*. Amsterdam, the Netherlands.

Gulati R. 1999. Network location and learning: The influence of network resources and firm capabilities on alliance formation. *Strategic Management Journal*, **20**:397-421.

Hagedoorn J. 2002. Inter-firm R&D partnerships: An overview of major trends and patterns since 1960. *Research Policy*, **31**:477-492.

Kim JW, Higgins MC. 2005. Where do alliances come from? The effects of upper echelons on alliance formation. *Academy of Management Proceedings*, J1-J6.

Lavie D, Rosenkopf L. 2006. Balancing exploration and exploitation in alliance formation. *Academy of Management Journal*, **49**:797-818.

Link AN, Paton D, Siegel DS. 2002. An analysis of policy initiatives to promote strategic research partnerships. *Research Policy*, **31**:1459-1466.

Morris, M. 2004. *Network Epidemiology: A Handbook for Survey Design and Data Collection.* Oxford University Press: London.

Nunnally JC. 1978. *Psychometric Theory*, 2d ed., McGraw Hill: New York.

Peterson RA. 1994. A meta-analysis of Cronbach's coefficient alpha. *Journal of Consumer Research*, **21**:381-391.

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

Rosenkopf, L., A. Metiu & V.P. George. 2001. From the bottom up? Technical committee activity and alliance formation. ***Administrative Science Quarterly***, **46**: 748-772.

Rothaermel FT, Deeds DL. 2004. Exploration and exploitation alliances in biotechnology: A system of new product development. *Strategic Management Journal*, **25**:201-221.

Sampson RC. 2005. Experience effects and collaborative returns in R&D alliances. *Strategic Management Journal*, **26**: 1009-1031.

Schilling, M.A. & Steensma, K. 2001. The use of modular organizational forms: An industry level analysis. *Academy of Management Journal*, **44**: 1149-1169.

Vanhaverbeke W, Duysters G, Noorderhaven N. 2002. External technology sourcing through alliances or acquisitions: An analysis of the Application-Specific Integrated Circuits industry. *Organization Science*, **13**:714-733.

Villalonga B, McGahan AM. 2005. The choice among acquisitions, alliances, and divestitures. *Strategic Management Journal*, **26**:1183-1208.

Table 1: Comparison of Alliance Databases

Database	Sectors	Advantages	Disadvantages
CORE Database	Multi-sector	• Reports the population of agreements filed under the NCRA Act. Data is highly reliable data for this population. • It is possible to obtain the original documents filed with each agreement reported in this dataset through the *Federal Register*. • Data goes back to 1985. • 762 filings listed between 1990-2005.	• Explicitly U.S. focused. • NCRA collaborations are only a very small subset of the collaboration activities struck between U.S. firms. • Only count data provided in spreadsheet form.
Securities Data Company (SDC)	Multi-sector	• Very broad coverage of sectors and agreement types. • Highly searchable, with over 200 data elements, including coded fields and keyword searches. • User-defined output formats available, facilitating use of data in large-scale analysis. • Over 52,000 research and technology agreements listed between 1990-2005. • Includes many non-OECD alliances.	• Data is highly sparse prior to 1990 • Some data elements rely on subjective coding; subject to reliability issues. • Bias toward English-language sources.
MERIT-CATI	Multi-sector	• Fairly broad coverage of sectors. • Utilizes retrospective data to incorporate data as early as 1960. • Over 8,200 agreements listed between 1990-2004.	• Very limited deal text provided prior to 1997. • Bias toward English-language sources.
Recombinant Capital (RECAP)	Biotech	• Provides great depth of information on individual alliance agreements. • Broad coverage of agreement types. • Searchability is moderately high with both keywords and coded fields. • Reports alliance data as far back as 1973. • Over 20,000 alliances reported between 1990-2005.	• Output options are limited, making it difficult to use for large-scale analyses. • Heavy focus on U.S. SEC filings may cause U.S. firms to be over-represented. • Likely bias toward English-language news sources.
Bioscan	Biotech	• Tracks activities of a fairly stable-set of firms, permitting better longitudinal assessments of firm behavior. • Provides a detailed profile of each firm, including key employees, major products, business strategy, and stock history. Patenting data also available. • Reports alliance data as early as 1985. • 3,106 alliances listed between 1990-2005.	• Searchable only with key words (not codes) -- makes searching less reliable. • Limited output options. • Updating procedure drops data from firms that no longer exist, causing shifts in alliance counts over time. • Likely bias toward English-language news sources.

Table 2: Consistency of Coverage for SDC and MERIT-CATI All-Sector Data

	Percent found in...	
Samples from...	SDC	MERIT-CATI
SDC 1990	*	11%
SDC 1995	*	1%
SDC 2000	*	2%
MERIT 1990	32%	*
MERIT 1995	31%	*
MERIT 2000	15%	*
Averages	**26.0%**	**4.7%**

Table 3: Consistency of Coverage for SDC, MERIT-CATI, and RECAP Biotech Data

	Percent found in...		
Samples from...	SDC	MERIT-CATI	RECAP
SDC 1990	*	6	7
SDC 1995	*	27	28
SDC 2000	*	9	16
MERIT 1990	6	*	18
MERIT 1995	8	*	28
MERIT 2000	11	*	36
RECAP 1990	2	9	*
RECAP 1995	12	20	*
RECAP 2000	8	14	*
Averages	**7.8**	**14.2**	**22.2**

Table 4: OECD Nations, Times Represented in Alliances, SDC versus MERIT-CATI

	SDC				MERIT-CATI			
OECD Nations	**Count**	**Percentage**	**Region Totals (OECD only)**		**Count**	**Percentage**	**Region Totals(OECD Only)**	
Canada	4094	3.69%	North America	38.83%	583	1.48%	North America	49.81%
Mexico	430	0.39%			34	0.09%		
United States	38577	34.76%			19029	48.25%		
Austria	406	0.37%	Europe	21.88%	72	0.18%	Europe	32.64%
Belgium	583	0.53%			394	1.00%		
Czech Republic	420	0.38%			6	0.02%		
Denmark	347	0.31%			169	0.43%		
Finland	584	0.53%			165	0.42%		
France	2977	2.68%			1906	4.83%		
Germany	4607	4.15%			2684	6.81%		
Hungary	403	0.36%			38	0.10%		
Iceland	14	0.01%			16	0.04%		
Ireland	310	0.28%			73	0.19%		
Italy	1671	1.51%			962	2.44%		
Luxembourg	85	0.08%			20	0.05%		
Netherlands	1573	1.42%			1778	4.51%		
Norway	415	0.37%			140	0.35%		
Poland	359	0.32%			5	0.01%		
Portugal	158	0.14%			13	0.03%		
Slovak Republic	72	0.06%			2	0.01%		
Spain	790	0.71%			160	0.41%		
Sweden	851	0.77%			701	1.78%		
Switzerland	994	0.90%			611	1.55%		
Turkey	256	0.23%			27	0.07%		
United Kingdom	6406	5.77%			2930	7.43%		
Japan	13939	12.56%	Asia	14.21%	4996	12.67%	Asia	13.62%
Korea	1834	1.65%			377	0.96%		
Australia	3614	3.26%	Australia	3.60%	202	0.51%	Australia	0.55%
New Zealand	380	0.34%			16	0.04%		
Non-OECD	23841	21.48%	Non-OECD	21.48%	1332	3.38%	Non-OECD	3.38%
Total	110990				39441			

Table 5: Replication of Schilling & Steensma (2001)

	Schilling & Steensma (2001) Study using SDC Data		Replication using MERIT-CATI Data	
Variables	**Restricted**	**Full**	**Restricted**	**Full**
Constant	.08***	.08***	.001	.001
Control				
Labor intensity	-2.64	-4.12†	-.01	-.01
Direct Effects				
Heterogeneity	.03***	.01	.002**	0
Technological change	3.10***	2.63***	.21*	.18
Competitive intensity	-.01**	-.04***	-.00	-.01
Industry standards	.04†	.03†	.01**	.01**
Indirect Effects				
Technological change X Heterogeneity		.39		.00
Competitive intensity X Heterogeneity		.01**		.00
Industry standards X Heterogeneity		.03*		.01*
$R^{2\ b}$	.12***	.16***	.07***	.09***
ΔR^2		.04**		.02†

[a]N = 330 for Schilling & Steensma, 2001; 292 for replication.
b Significance levels refer to the F statistics associated with the variance explained.
†p < .1, *p < .05, **p < .01, ***p < .001

Table 6: Results of Panel Regressions: Powell, Koput & Smith-Doerr (1996) versus Replication using SDC Data

a) Determinants of Network Portfolios

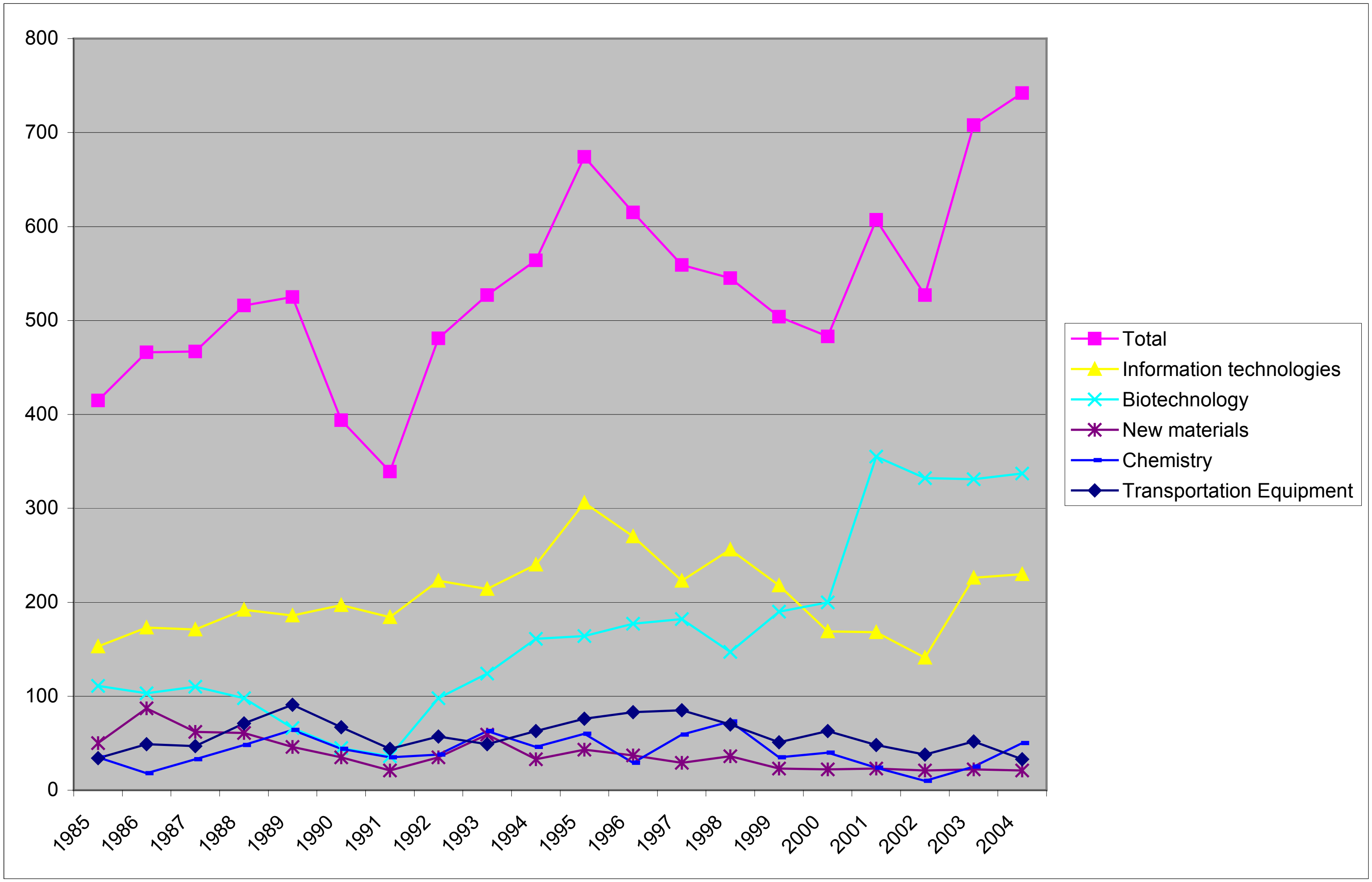
	Powell, Koput & Smith-Doerr (1996) using Bioscan data[a]					Replication using SDC Data[b]				
	R&D ties $_{t+1}$	Manuf. $_{t+1}$	Mktg $_{t+1}$	Total ties $_{t+1}$	Diversity $_{t+1}$	R&D ties $_{t+1}$	Manuf $_{+1}$	Mktg $_{t+1}$	Total ties $_{t+1}$	Diversity $_{t+1}$
R&D ties			.16*	.47*	.01*		.27	.28†	.35**	.01
Non-R&D network experience			.11*			-7.25	-7.15	-6.93	-7.01	-.05**
Closeness centrality	.22*					-1.08*	-1.20	-.83*	-1.38**	
Lagged dependent variable	.49*	.30*	.32*	.42*	.18*	.27†	-.34	-.22	-.20**	.76**
Total ties	.03*					-.18*	-18.97	-.11		.03*
R-squared	.84	.92	.95	.93	.82	[c]	[c]	[c]	[c]	.78

b) Determinants of Degree Centrality

	Powell, Koput & Smith-Doerr (1996) using Bioscan data		Replication using SDC Data	
	Degree Centrality $_{t+1}$	Closeness Centrality $_{t+1}$	Degree Centrality $_{t+1}$	Closeness Centrality $_{t+1}$
R&D ties	1.50*		.08**	.08**
R&D network experience		.06*	-.05**	-.05**
Non-R&D network experience	.83*	.04*	-.03**	-.03**
Network diversity	6.34*	.22*	.20**	.20*
Lagged dependent variable	.24*	.21*	.74**	.72**
Total ties	.38*	.01*	-.03*	-.01
R-squared	.90	.75	.87	.69

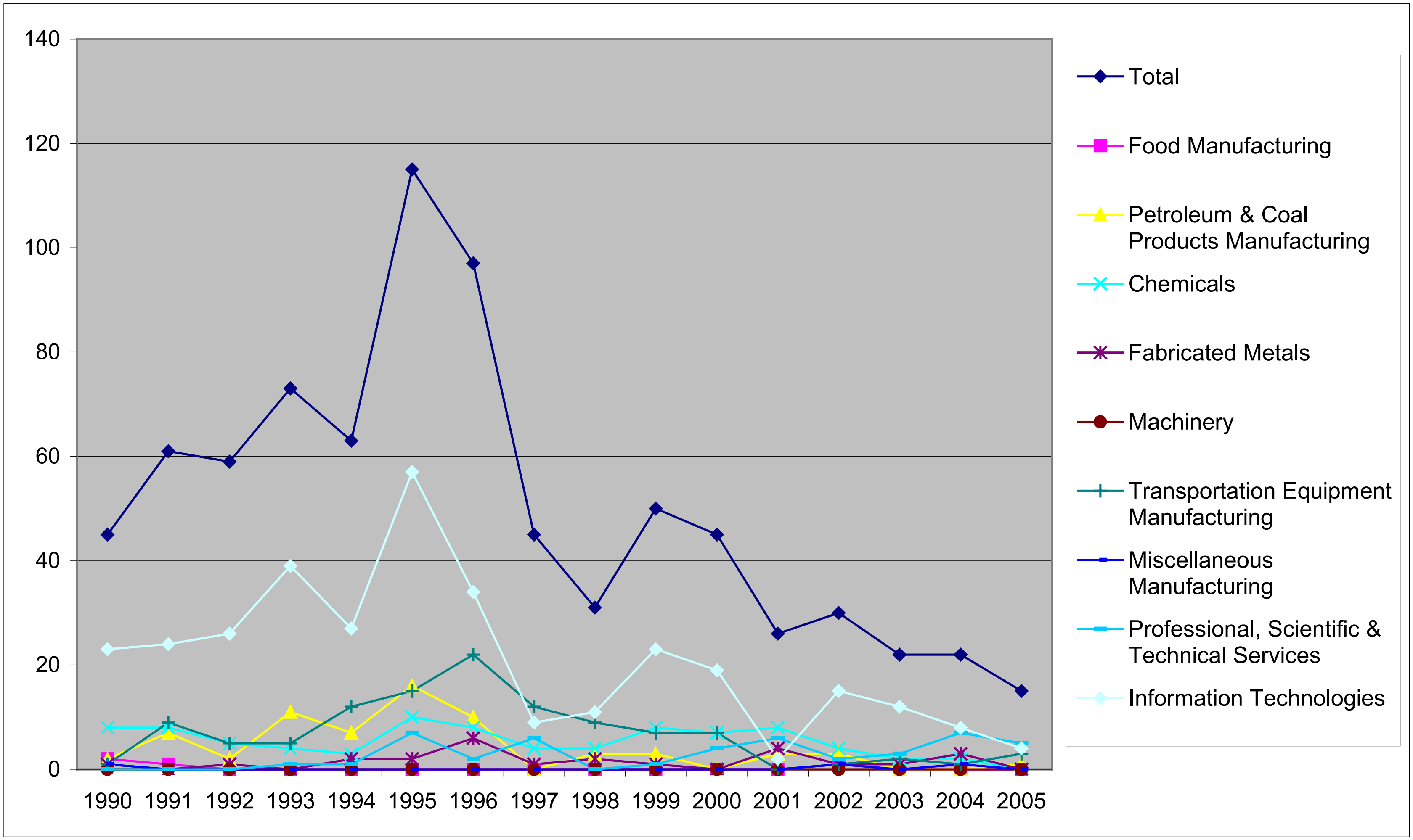
[a]For Powell et al. results, all reported coefficients (except the effect of total ties on R&D ties) are significant at or beyond the .05 level.
[b]For replication results, † p<.10, * p<.05 (one-tailed test); ** p<.01 (one tailed test); Fixed firm effects are included in all models. Fixed year effects (dummies) are included in all models except when the dependent is Diversity $_{t+1}$, Degree Centrality $_{t+1}$, or Closeness Centrality $_{t+1}$; for these latter models, inclusion of fixed year effects caused collinearity problems.
[c] Because the dependent can take on only non-negative integer values for these variables (and exhibited significant overdispersion), I used a fixed effects negative binomial panel model for these analyses; thus R squared is not reported.

Figure 1: MERIT-CATI Data, with Sector Decomposition[22], 1990-2004



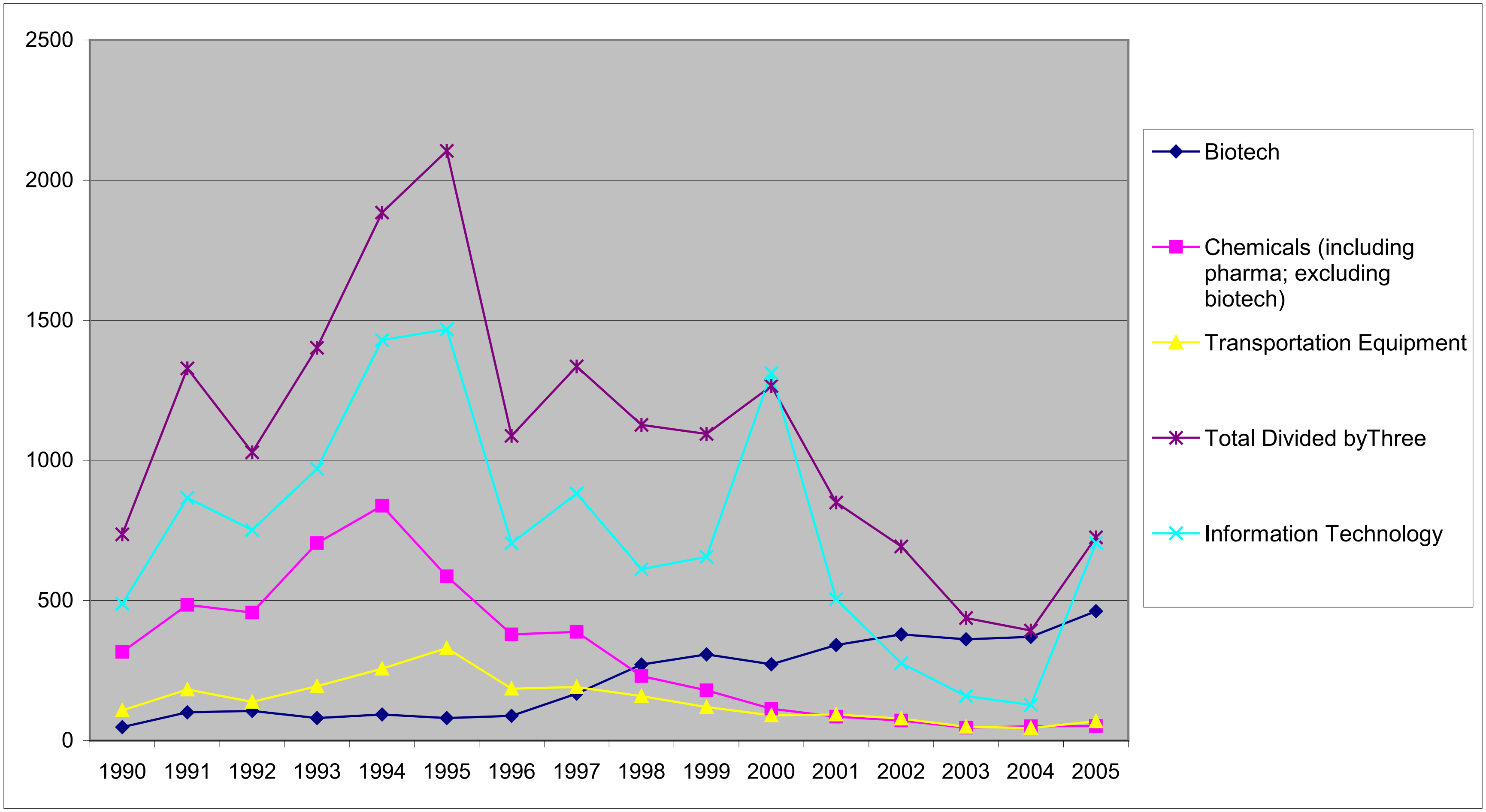
[22] MERIT-CATI reports automotive and aerospace/defense alliances separately, however, I have collapsed these categories into a single "transportation equipment" category in order to permit comparison with the CORE data

Figure 2: CORE Database, with Sector Decomposition[23], 1990-2005



[23] For comparison purposes, the Computer & Electronics Products, Electrical Equipment, Appliances & Components, and Broadcasting & Telecommunications categories were collapsed into a single "Information Technologies" category.

Figure 3: SDC Database, Sectoral Decomposition, 1990 - 2005[24]



[24] Note, because the total number of technology alliances reported in SDC is very large compared to those reported in individual sectors, including a line for the total causes all of the other series to be close to the horizontal axis and difficult to discern. Thus I show here the total divided by three so that the trends in the individual series can be seen more clearly.

Figure 4: All Sector Alliances from SDC, MERIT-CATI and CORE, Standardized, 1990-2005

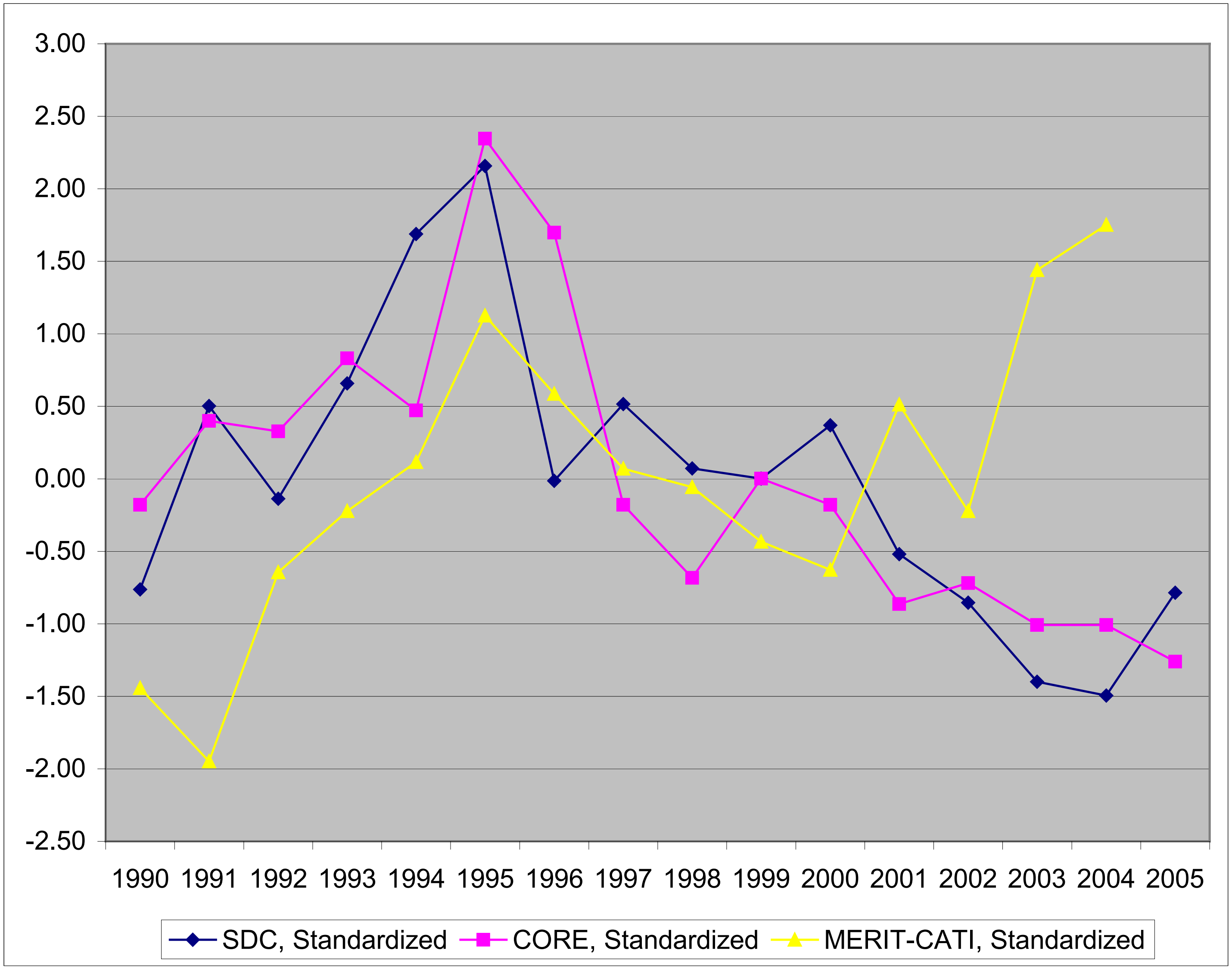


Figure 5: Standardized All-Sector Alliance Data with Biotech Alliances Omitted, 1990-2005

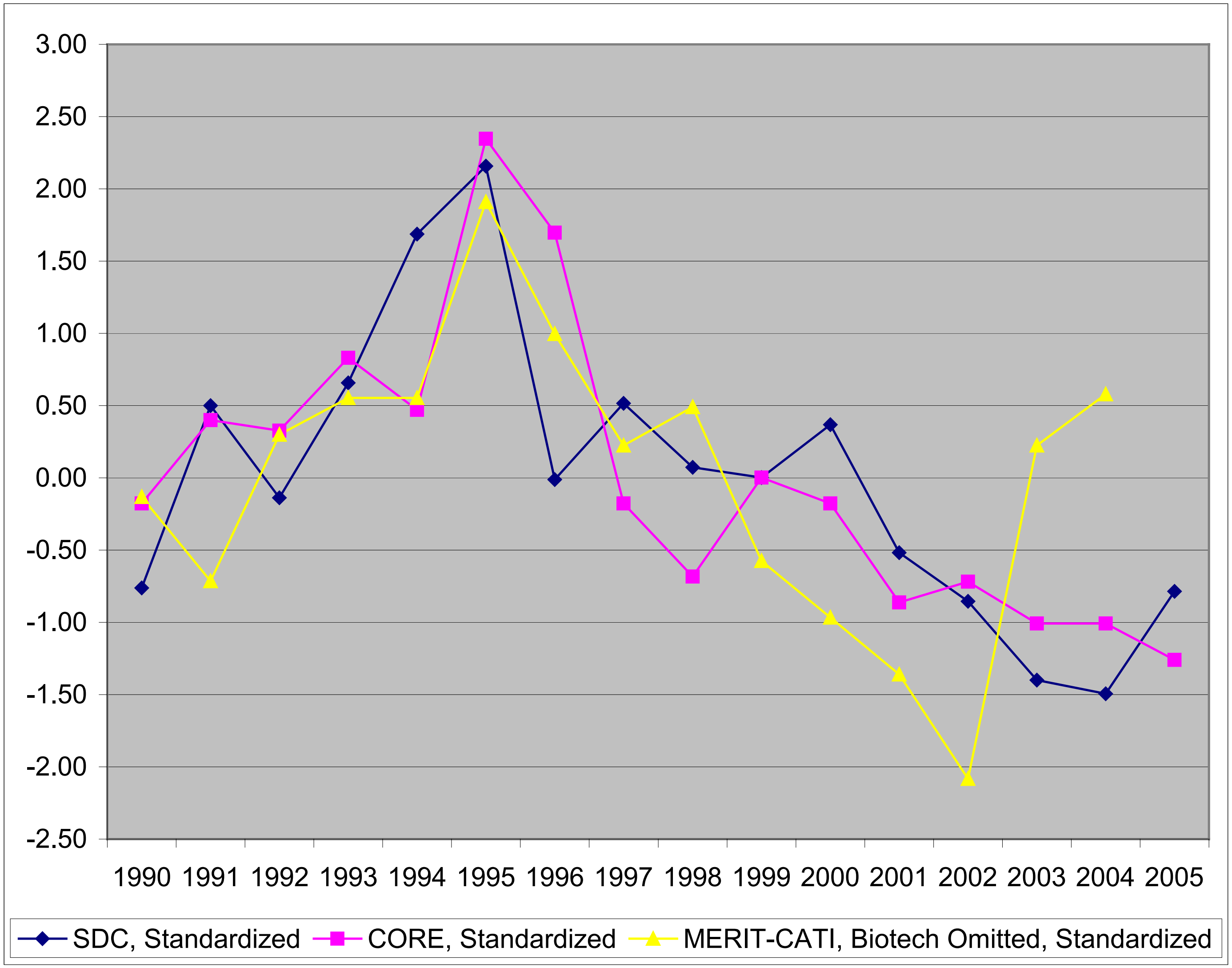


Figure 6: Information Technology Alliances from SDC, MERIT-CATI and CORE, Standardized, 1990-2005

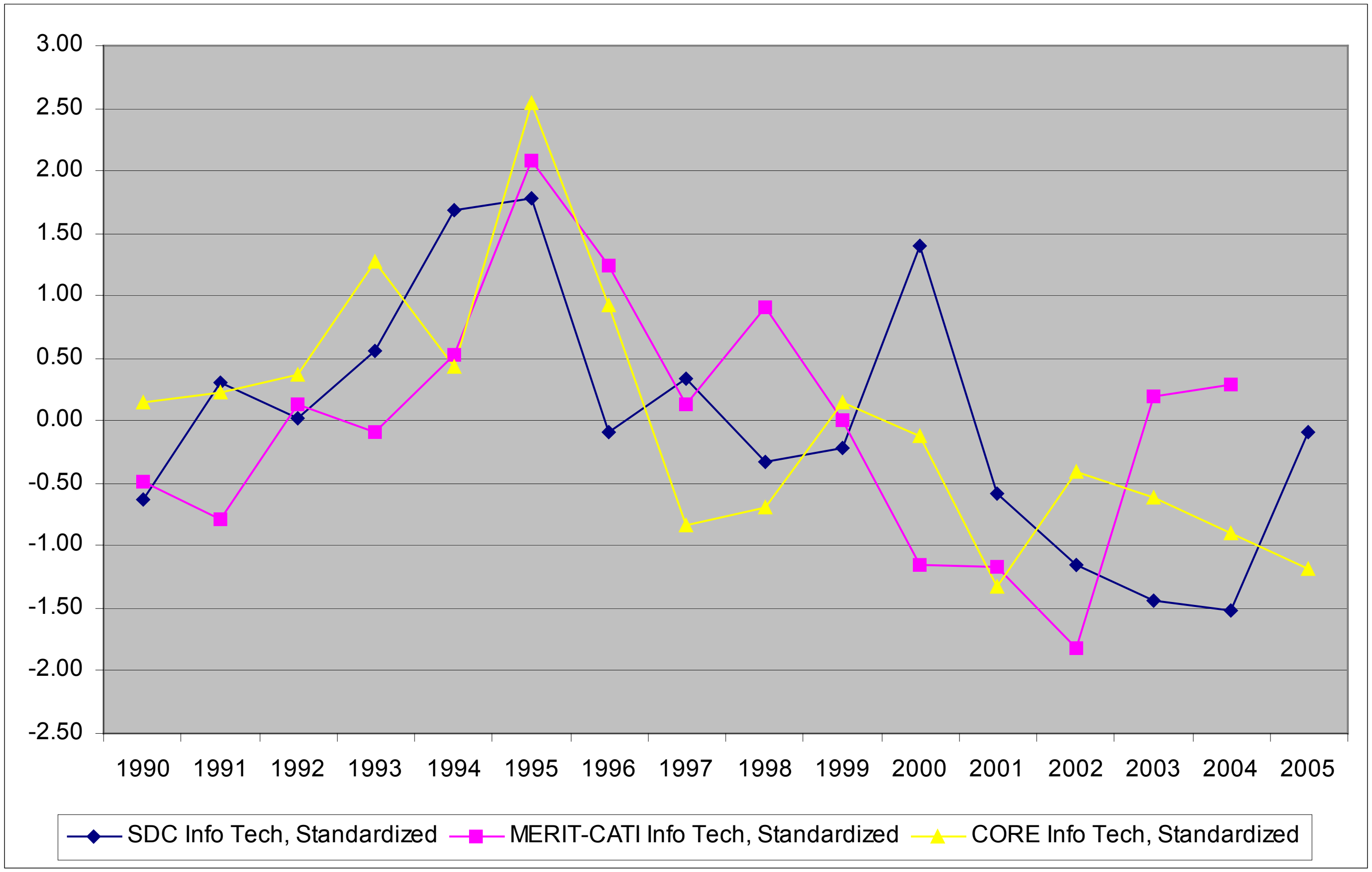


Figure 7: Transportation Equipment Alliances from SDC, MERIT-CATI and CORE, Standardized, 1990-2005

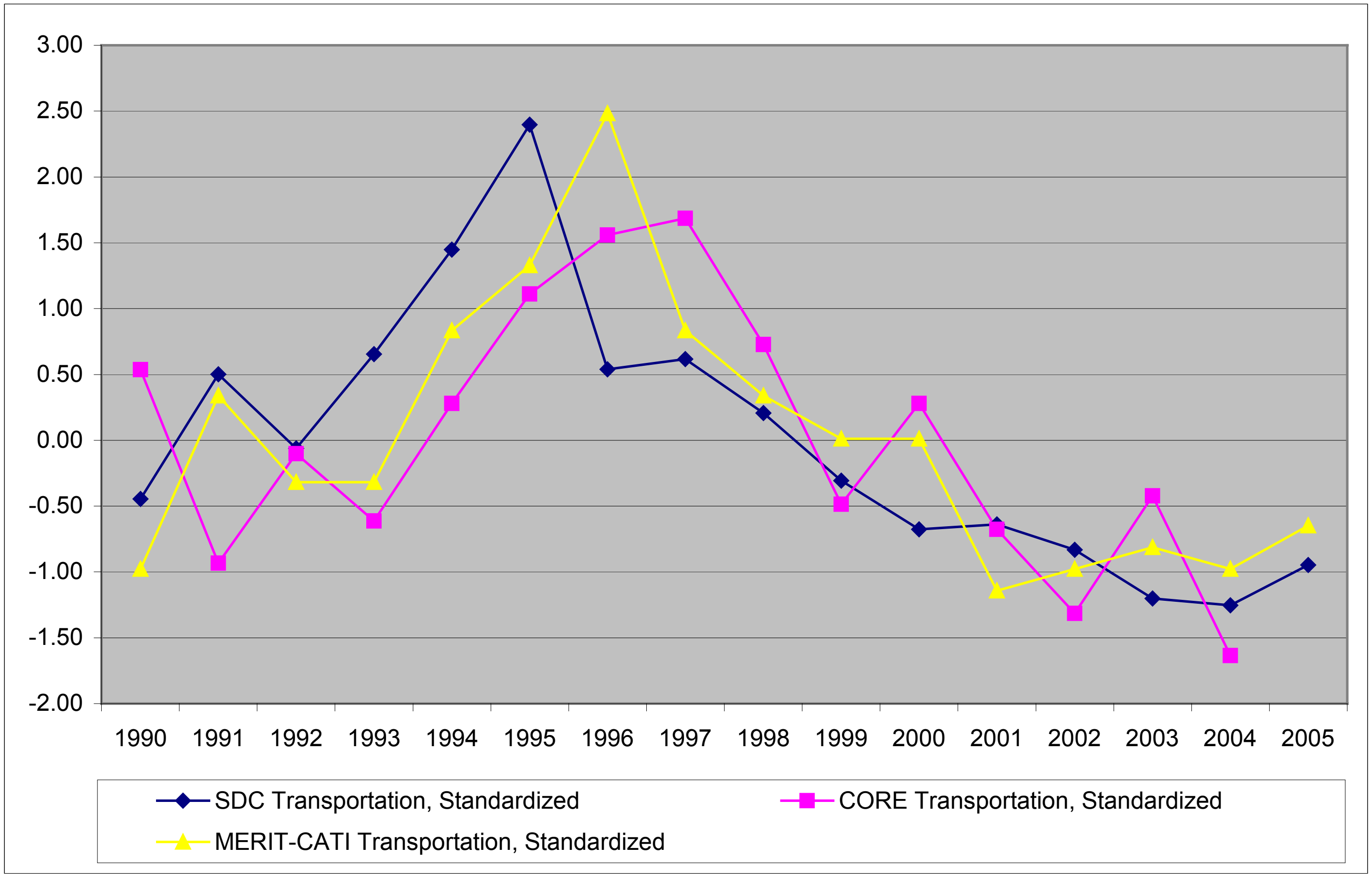


Figure 8: Chemicals Alliances from SDC, MERIT-CATI and CORE, Standardized, 1990-2005

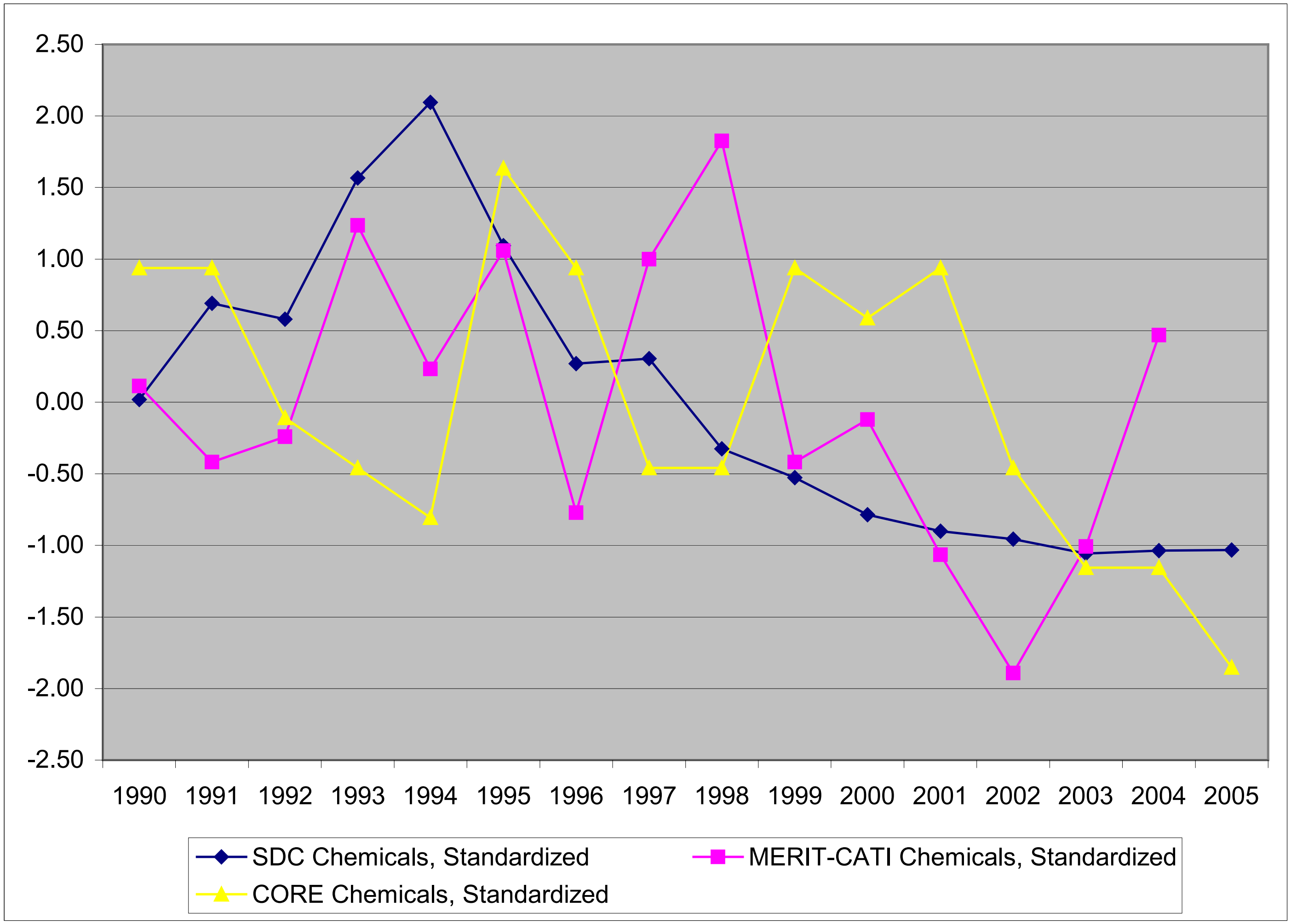


Figure 9: Biotechnology Alliances from SDC, MERIT-CATI, RECAP and Bioscan, 1990-2005

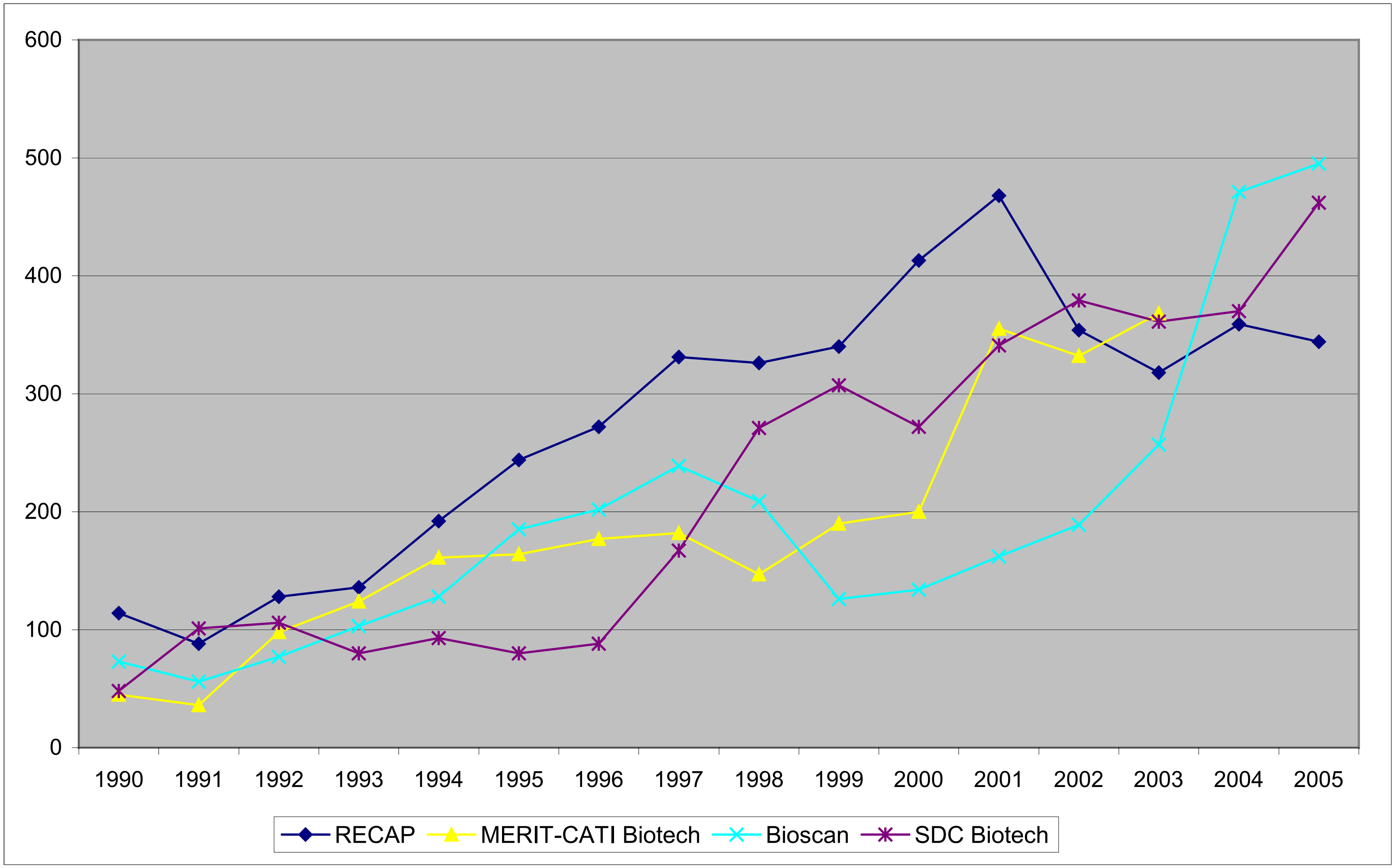


Figure 10: Comparison of Relative Internationalization Indexes, Hagedoorn 2002 versus Replication with SDC