INSEAD

Faculty & Research

Multinational Firms and Knowledge Diffusion: Evidence using Patent Citation Data

by

J. Singh

2004/75/SM

Working Paper Series

Multinational Firms and Knowledge Diffusion: Evidence using Patent Citation Data

Jasjit Singh

INSEAD, 1 Ayer Rajah Avenue, 138676 Singapore

+65 67995341

jasjit.singh@insead.edu

http://www.jasjitsingh.com

Original version: April 22, 2004

This version: Sept 28, 2004

Suggested running title: "MNCs and Knowledge Diffusion"

I am grateful to Bharat Anand, Richard Caves, Wilbur Chung, Ken Corts, Mihir Desai, Lee Fleming, Bronwyn Hall, Elhanan Helpman, Rebecca Henderson, Adam Jaffe, Wolfgang Keller, Tarun Khanna, Walter Kuemmerle, Josh Lerner, Megan MacGarvie, Anita McGahan, Marc Melitz, Jan Rivkin, Jordan Siegel, Olav Sorenson, Peter Thompson, Manuel Trajtenberg and seminar participants at various universities and conferences for helpful comments. I am grateful to Harvard Business School and INSEAD for funding this research. Errors remain my own.

Abstract

How actively do local subsidiaries of foreign MNCs exchange knowledge with their host countries? To what extent does this facilitate cross-border knowledge diffusion between the MNC home base and the host country? To address these issues, I use over one million patent citations to analyze micro-level knowledge flows in six most innovating countries. I find that there are significant bi-directional knowledge flows between MNCs and their host countries, but that MNCs gain more than they contribute. The pattern varies across countries and sectors, depending on the knowledge-intensity of FDI.

Keywords: Knowledge Spillovers, Technology Diffusion, Multinationals, Foreign Direct Investment, Subsidiaries

JEL Codes: F2, L2, M2, O3, O5

1. Introduction

Innovation and knowledge diffusion play a critical role in economic growth, with growth rates being highly sensitive to how easily knowledge diffuses (Romer, 1990; Grossman and Helpman, 1991; Eaton and Kortum, 1999). While economists once believed that ideas should be costless to transport, recent empirical literature has established that knowledge spillovers are geographically localized (Jaffe, Trajtenberg and Henderson, 1993; Audretsch and Feldman, 1996; Branstetter, 2001; Keller, 2002). Foreign direct investment can play an important role in overcoming this geographic constraint on the diffusion of knowledge (Caves, 1974; Aitken and Harrison, 1999; Branstetter, 2000).[1] Governments around the world continue to spend huge resources to attract multinational firms (MNCs), at least partly in the hope of knowledge gains from them. However, literature on how foreign direct investment (FDI) contributes to knowledge diffusion still remains fragmented and inconclusive.

Existing literature largely emphasizes *uni-directional* knowledge flows *from* foreign MNCs *to* host country domestic firms. However, while FDI can lead to knowledge flows for the domestic players, it can also be a channel through which domestic technology can fall into the hands of foreign competitors. Therefore, except for countries that have little unique technology of their own, it is important to consider *bi-directional* knowledge flows in studying *net* gains from FDI. The potential "leakage" of domestic knowledge through FDI is a particularly real issue for technologically advanced countries, which are the focus of this paper. For example, Dalton and Shapiro (1995) say, "Rapid growth of foreign R&D in the US has led to concerns about an erosion of US technology leadership… Some observers have questioned the quality of the research effort by foreign companies. They have argued that US research centers of foreign companies are merely

'listening posts' that focus on technology scanning." A central goal of my paper is study the extent to which this concern is valid.

It is hard to measure knowledge spillovers directly. Therefore, several studies have tried to estimate the effect of FDI on productivity of domestic firms (Caves, 1974; Aitken and Harrison, 1999). A challenge in doing so, however, has been separating knowledge spillover effects of FDI from its effect on competition (Caves, 1996; Chung, 2001; Chung, Mitchell and Yeung, 2003). An alternate empirical approach, which I follow in this paper, is to measure knowledge diffusion using patent citation data. While patent citations are an imperfect measure of knowledge diffusion and also make it hard to separate true externalities from intentional knowledge transfer (Peri, 2003), they are widely used in research as a way to directly capture micro-level knowledge flows (Jaffe and Trajtenberg, 2002). I measure *bi-directional* knowledge flows between MNC subsidiaries and domestic players, and also between MNC home base and host countries, using data on citations made by over half a million patents originating from 4,400 MNCs and domestic organizations in the US, Japan, Germany, France, UK and Canada. In its use of patent data in studying role of MNCs, the current paper builds upon Almeida (1996), Branstetter (2000) and Frost (2001), while placing much more emphasis on *bi-directional* knowledge flows, and looking at cross-country and cross-sector differences in the observed patterns.

My findings suggest that there are significant bi-directional knowledge flows between MNCs and their host countries, but that MNCs contribute less to host country knowledge than they gain from it. For intra-national knowledge flows, my specific findings are: (1a) Knowledge flows from domestic entities to local subsidiaries of foreign MNCs are as strong as those between domestic entities; (1b) Knowledge flows from MNC subsidiaries to domestic entities are weaker on an average, with the pattern differing across sectors and

[1] Multinational activity is not the only way in which global economic activity can contribute to knowledge diffusion. Trade can also play an important role (Coe and Helpman, 1995), but is not studied in this paper.

countries depending on R&D-intensity of FDI; (1c) MNC subsidiaries are particularly good at learning from each other. For knowledge flows across borders, I find that: (2a) MNCs are as good at transferring knowledge from their subsidiaries to their home base as from the home base to the subsidiaries; (2b) More intense innovative activity by MNC subsidiaries increases bi-directional knowledge flow between the host country and the MNC home base, with the gains being larger for the MNC home base than for the host country's domestic players.

This paper also makes a methodological contribution to use of patent citation data in measuring knowledge spillovers. Jaffe, Trajtenberg and Henderson (1993) pioneered a widely-used statistical technique that tries to correct for factors other than knowledge spillovers that might affect technological specialization of regions, and hence the pattern of patent citations. However, Thompson and Fox-Kean (2004) have shown that existing application of this technique often leads to over-estimation of knowledge flows. To address this, I propose a novel citation-level regression approach that estimates the probability of citation between any two patents using a choice-based sampling approach (Manski and Lerman, 1977). In addition, I use a combination of econometric techniques as well as additional robustness checks using European Patent Office (EPO) data to address concerns about using data from US Patent Office (USPTO) for international comparison.

The rest of the paper is organized as follows. Section 2 presents my formal hypotheses. Section 3 describes the patent citation data and my subsidiary-parent database. Section 4 presents preliminary analysis of knowledge flows between MNCs and domestic organizations. Section 5 describes my citation-level regression framework. Section 6 presents results on role of MNCs in both intra-national and cross-border knowledge flows. Section 7 addresses concerns regarding use of USPTO data in measuring international knowledge diffusion. Section 8 offers concluding thoughts.

2. Hypotheses

For international knowledge diffusion to be an interesting issue to study, the first fact to establish is that knowledge does not automatically transmit across countries. While previous work has found empirical support for geographic localization of knowledge spillovers (e.g., Jaffe, Trajtenberg and Henderson, 1993), recent work raises issues that could have led to over-estimation of this phenomenon (Thompson and Fox-Kean, 2004). Therefore, I revisit the following hypothesis using a new methodology that addresses the above concerns.

Hypothesis 1. *The probability of knowledge flow within a country exceeds that between different countries, even after controlling for technological specialization of countries.*

MNCs can facilitate international knowledge diffusion through their ability to transmit knowledge more effectively than would be possible through market-mediated mechanisms (Hymer, 1976; Buckley and Casson, 1976). While the transaction cost literature suggests that this happens through decreased opportunism within a firm (Williamson, 1985; Ethier, 1986; Teece, 1986), other research shows social networks and a firm's internal organization to transmit complex and tacit knowledge as the mechanisms (Hedlund, 1986; Bartlett and Ghoshal, 1989; Kogut and Zander, 1993; Nohria and Ghoshal, 1997). Distinguishing between these two is beyond the scope of this paper, but I do formally test the following hypothesis on intra-MNC knowledge flows:

Hypothesis 2. *The probability of cross-border knowledge flow within an MNC exceeds that between different firms, even after controlling for the relative technological proximity of different divisions within the same MNC.*

A central argument of this paper is that looking at *uni-directional* knowledge flows from an MNC subsidiary to its host country misses the point that knowledge could also flow

from the host country to the MNC subsidiary (Almeida, 1996; Frost, 2001), and from the subsidiary to the MNC home base (Hedlund, 1986; Bartlett and Ghoshal, 1989). My next task therefore is to empirically establish the presence of such *bi-directional* knowledge flows:

Hypothesis 3. *There are significant knowledge flows in both directions between an MNC subsidiary and its host country.*

Hypothesis 4. *There are significant knowledge flows in both directions between an MNC subsidiary and its home base.*

Existing literature also suggests that intra-national knowledge flows are particularly strong between different foreign MNC subsidiaries located in the same country (Head, Ries and Swenson, 1995; Feinberg and Majumdar, 2001; Feinberg and Gupta, 2003), which I verify next:

Hypothesis 5. *There are significant knowledge flows between local subsidiaries of different foreign MNCs.*

Next, I examine the relative strength of different knowledge flows. If local subsidiaries of foreign MNCs are involved in knowledge-intensive activities like advanced research or innovative product development, we might expect greater knowledge spillover benefits to the host country. Existing evidence suggests, however, that even MNC subsidiaries doing R&D often focus on adaptation of their parent firm's products for the local markets (Mansfield, Teece and Romeo, 1979), or on being "listening posts" to monitor local technological developments (Almeida, 1996; Florida, 1997; Frost; 2001). Surveys by Kuemmerle (1999) and Frost, Birkinshaw and Ensign (2002) reveal that, while the number of MNC subsidiaries doing advanced research has been increasing, such cases still comprise only a minority.

Raising further concerns about the benefits from FDI is the adverse selection in the "knowledge intensity" of overseas operations of MNCs. Kogut and Chang (1991) find that

a disproportionately large fraction of Japanese FDI in the US is restricted to industries where the Japanese MNCs lag behind their US counterparts. Similarly, Shaver and Flyer (2000) and Chung and Alcacer (2002) find that technologically advanced MNCs are less likely to locate sophisticated facilities overseas and, when they do, are likely to locate them far from domestic players to prevent their technology from being copied. Cantwell and Janne (1999) find that foreign subsidiaries of even technologically advanced MNCs focus on the specific technologies where these MNCs lag behind. All of this raises concerns that host countries might lose more from "leakage" of domestic knowledge to MNCs than gain in the form of knowledge spillovers from MNCs, a hypothesis I directly test in this paper.

Hypothesis 6. *The probability of knowledge flow from the host country to an MNC subsidiary exceeds that from the MNC subsidiary to the host country.*

Extending the above logic, the relative extent of knowledge flows from the host country to MNCs should be most intense in settings where the domestic firms do more "knowledge-intensive" work than the MNC subsidiaries. This can be tested by seeing how the pattern of bi-directional knowledge flows varies with the relative R&D intensity (i.e., the ratio of R&D to total production) for domestic firms and MNC subsidiaries.

Hypothesis 7. *The probability of knowledge flow from the host country to MNC subsidiaries is particularly great in countries and sectors where the R&D intensity of MNC subsidiaries is significantly lower than that of the host country.*

Finally, if foreign subsidiaries of an MNC serve as listening posts for the home base, these subsidiaries should improve the absorptive capacity of the MNC *home base* for knowledge produced in the host countries. This gives the final hypothesis:

Hypothesis 8. *The relative probability of knowledge flow from a host country to a foreign MNC's home base is greatest when the MNC's local subsidiaries are most active in knowledge-related activities.*

3. Data on Patent Citations and Multinational Ownership

3.1. Patent Citations as Measure of Knowledge Flow

Patent citations leave behind a trail of how a new innovation potentially builds upon existing knowledge. An inventor is legally bound to report relevant "prior art", with the patent examiner serving as an objective check. Unlike academic papers, there is usually an incentive not to include superfluous citations, as that might reduce the scope of one's own patent. There are, however, two factors that add noise to citations as a measure of knowledge flow. First, citations might be included by the inventor for strategic reasons (e.g., to avoid litigation). Second, a patent examiner might add citations to patents that the original inventor knew nothing about. Recent studies comparing citation data with inventor surveys show that the correlation between patent citations and actual knowledge flow is indeed high, but not perfect (Jaffe and Trajtenberg, 2002; Duguet and MacGarvie, 2002). The defense given in the common research use of patent citations is that use of citations is okay in large-sample studies as long as the noise does not bias the results of interest. Note that viewing patent citations as being correlated with knowledge flows is not the same as claiming that patents themselves are the mechanism behind these knowledge flows (Singh, 2004). Consider the analogy that a PhD student may cite research papers of his advisor, even though knowledge gained by working closely with the advisor could be much more than what could be captured in the advisor's papers.

3.2. Data from US Patent Office (USPTO)

Since patents from different patent offices are not comparable to each other, it is common practice to use data from a single patent granting country like US (Jaffe and Trajtenberg, 2002) or UK (Lerner, 2002) to standardize the measure of innovation for research purposes. Following this practice, I use a data set on US patents, constructed by merging data from the US Patent Office (USPTO) with an enhanced version made available by Jaffe and Trajtenberg (2002). A major issue in using patent data is that only some of the innovations are patented (Levin, Klevorick, Nelson and Winter, 1987), with systematic differences across countries and sectors in their likelihood to file for USPTO patents. Since this makes counts of patents and patent citations misleading as raw measures, I only estimate the probability of knowledge flow between two innovations that do end up as patents, without claiming that these comprise all the innovations.

Following standard practice, the country of residence of the inventors is taken as the country where an innovation takes place. In order to ascertain whether it originated from a domestic organization or from the local subsidiary of a foreign MNC, I check whether the "home country" of the assignee organization is the same as the country of innovation. As mergers and acquisitions are a potential issue in defining the home country, I restrict my analysis to patents in a narrow time window between 1986 and 1995 as I use various data sources from around 1990 for constructing the parent-subsidiary database. I examine patents by inventors from six leading economies: US, Japan, Germany, France, UK and Canada. The number of patents from these countries for the period 1986-1995 is about 0.9 million, or about 91% of all USPTO patents (Table 1, column 1).[2] About 83% of these patents are owned

[2] Since the remaining countries account for less than 10% of the USPTO patents, I found that adding more countries did not change the aggregate results, and was not useful for extending individual country results. So I dropped these to keep the number of citing and cited country fixed effects manageable in my econometric model.

by firms or organizations (as opposed to individuals), and are the ones of interest here (Table 1, column 2).

3.3. Multinational Data

A crucial step in the data analysis was identifying whether an assignee firm has its home base in the country of innovation (e.g., IBM in the US), or if it is a local subsidiary of a foreign MNC (e.g., IBM in Germany). Unfortunately, the patent database has about 175,000 assignee names, and it is impossible to match all assignees to their parents. For example, there is no systematic rule as to whether patents originating from researchers based in a German subsidiary of IBM would be listed under the parent firm "IBM" or a separate assignee "IBM Germany" (or a name from which it is even harder to infer that this is a subsidiary of IBM).[3]

To construct my parent-subsidiary database, I inspected about 10,000 assignees as follows. First, Compustat-based parent firm identifiers (from 1989) from Jaffe and Trajtenberg (2002) were used to match around 4,600 patent assignees to 2,500 parent firms. Second, Stopford's *Directory of Multinationals* (1992) was used to match around 2,800 additional assignees with 200 parent firms. Third, using USPTO assignee information, keyword search and the Internet, about 400 government-affiliated bodies, 550 research institutes and 450 universities worldwide were identified. Finally, the ownership of another 1,000 major patent assignees was checked using a combination of *Who Owns Who* directories (1991) and data from company web sites. As Table 1 shows, the above steps account for about 556,000 patents, which is about 73% of all assigned patents. The remaining patents

[3] To avoid the situation in which a company could not be identified with a unique parent, I define an assignee to be an MNC subsidiary when a foreign firm has a *majority* stake in it. For cases where two firms had a 50-50 stake, I broke the tie in favor of the first firm. See Mowery, Oxley and Silverman (1996) or Gomes-Casseres, Jaffe and Hagedoorn (2003) for an in-depth study of alliances.

were dropped.[4] About 9% of all patents arise from foreign MNC subsidiaries, though the fraction varies a lot across countries (Table 1, column 4).[5] Although this variation is interesting in itself, exploring it is beyond the scope of this paper.

4. Preliminary Analysis

Innovations in similar technologies are likely to be located in the same region, often for reasons other than potential knowledge spillovers. Therefore, to avoid over-estimation of the localized knowledge spillover effect, it is important to control for the geographic distribution of technological activity. Jaffe, Trajtenberg and Henderson (1993) suggest a "matching" approach that takes this into account by defining the appropriate benchmark as being the citation frequency from the original patents to randomly drawn patents with similar technological and temporal characteristics as the originally cited patents.

4.1. The Matching Approach

Existing studies typically use a *3-digit* technological classification for the matching methodology suggested by Jaffe, Trajtenberg and Henderson (1993). However, Thompson and Fox-Kean (2004) show that this is not detailed enough to prevent over-estimation of localized knowledge flows (Thompson and Fox-Kean, 2004). To overcome this issue, I start by using the 9-digit subclass information available from USPTO. Since this detailed classification consists of around 150,000 sub-classes, I am able to have a much finer control for geographic distribution of technological activity. Following standard practice, all citations for which either the original or the control patent involved a self-cite from an organization to itself were excluded from the sample. Since the time lag between two patents is also an

[4] The main results reported below continue to hold if, instead of dropping any of the remaining assignees, I included them as independent entities, with the home country calculated as the country where most of its patents originate.

[5] These numbers approximately equal estimates for the fraction of national R&D coming from MNC subsidiaries in these countries, as reported by OECD (1998). This serves as an additional validation for my dataset construction.

important determinant of the probability of citation, the final sample only included those cited patents for which a control patent could be found with an application year within one year of the original. This leads to dropping about half of the citations from the original data, an issue I revisit in the next section.

To examine evidence for knowledge flows from MNC subsidiaries to domestic organizations, I examine if the fraction of MNC patents (i.e., patents originating from local subsidiaries of foreign MNCs) is higher in the set of patents cited by domestic organizations than in the set of control patents. The t-statistic used to formally test this is given by

$$t_{M\to D} = \frac{p_{M\to D} - p'_{M\to D}}{\sqrt{\frac{p_{M\to D}(1-p_{M\to D})}{N_D} + \frac{p'_{M\to D}(1-p'_{M\to D})}{N_D}}}$$

where $p_{M\to D}$ is the ratio of number of *actual* citations from domestic organizations to MNC subsidiaries to the total number of citations (N_D) made by domestic entities, and $p'_{M\to D}$ is the analogous ratio for the *control* citations. I similarly compute the t-statistics to test for domestic-to-multinational (D→M) knowledge flows.

4.2. Results from Matching

Table 2(a) gives analysis of localized knowledge diffusion from local subsidiaries of foreign MNCs to domestic organizations (M→D flows). Column (1) gives the total number of citations made by domestic organizations, and columns (2) and (3) respectively give the number and fraction of these made to patents by local subsidiaries of foreign MNCs. Columns (4) and (5) report the same analysis for patent pairs obtained by replacing each original cited patent by its control patent. Column (6) reports the difference of proportions from columns (3) and (5), and column (7) shows that a t-test rejects their equality. Column (8) gives the ratio of the two proportions (which I call the M→D index). The overall M→D index of 1.13 indicates that the probability of knowledge flow from a

patent by an MNC subsidiary to a domestic patent is 13% more likely than for two geographically random patents with similar technological and temporal characteristics.

In Table 2(b), a similar approach shows significant knowledge flows from domestic organizations to local subsidiaries of foreign MNCs (D→M flows). The magnitude of the D→M index (1.20) is found to be even larger than the M→D case discussed above. Thus, not only does the localization of knowledge diffusion result still hold, the extent of knowledge diffusion is even stronger than the M→D case. In other words, MNC subsidiaries are better at gaining knowledge from domestic organizations than the latter are at gaining knowledge from the former. I will test this claim formally using my regression framework below.

5. Citation-Level Regression Methodology

In addition to the 3-digit vs. 9-digit technological classification issue that I have already addressed above, Thompson and Fox-Kean (2004) point out two other challenges in using the matching approach. First, dropping observations with imperfect matches can lead to a systematic bias in the measured knowledge flow patterns. Second, while the matching approach focuses on the "primary" technological classification, most patents also have several "secondary" technology classes and subclasses, with the primary versus secondary distinction not necessarily being a true reflection of a patent's fundamental characteristics. The matching approach does not capture the fact that technological relatedness of patents could show up as an overlap along any of their subclasses, and not just as their primary class or subclass being the same.

To overcome these challenges, I use a citation-level regression framework to model the probability of citation between two patents. Imagine that the probability that a patent K cites a patent k is given by a "citation function" $P(K, k)$. My interest lies in

studying how P(*K*, *k*) differs with characteristics of the cited and citing players. Among the explanatory variables, I include dummy variables for all dimensions along which I would have ideally liked to do the matching. This gives the flexibility of using multiple control variables to better control for propensity to cite even in cases where good matches do not exist.[6]

5.1. Choice-Based Sampling

Since the number of potentially citing and cited patents can be of the order of a million, the number of all possible dyads (K, k) can be of the order of a trillion. In principle, one could take a random sample of patent dyads from the population of all possible dyads. One could then define a binary variable *y* that equals 1 if the citation actually takes place, and 0 otherwise, and estimate the citation function by assuming that it can be approximated using a logistic functional form. In other words, the dichotomous dependent variable *y* would be taken as a Bernoulli outcome that takes a value 1 for observation *i* with the probability

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

where $\mathbf{x}_i$ is the vector of covariates and $\boldsymbol{\beta}$ is the vector of parameters to be estimated. However, an estimation approach based on random sampling of patent pairs is not practical because citations between random pairs of patents are very rare: there are only about seven actual citations for every one million potential citations, making estimation impossible even with very large samples.

From an informational point of view, it would be desirable to have a higher fraction of observations with $y = 1$ in the sample. This can be achieved by a "choice-based" sampling

[6] Some regression-based studies use an aggregate number of citations as the dependent variable. These models include a measure of "average technological distance" between sets of citing and cited patents using only a 2 or 3-digit technology classification. So the issue of bias remains because of within-set heterogeneity: sets with technologically closer patents have more frequent citations and also greater co-location of patents.

procedure that deliberately oversamples the patent pairs with $y = 1$.[7] In this approach, the sample is formed by taking a fraction α of the population's dyads with $y = 0$, and a fraction γ of the dyads with $y = 1$, α being much smaller than γ. Since this stratification is done on the dependent variable, however, using the usual logistic estimates would lead to a selection bias. A technique that overcomes this problem is the *weighted exogenous sampling maximum likelihood* (WESML) estimator suggested by Manski and Lerman (1977). The central idea is to explicitly recognize the difference in sampling of 0's and 1's by weighting each term in the log likelihood function by the inverse of the ex ante probability of inclusion of the corresponding observation in the sample. In other words, each sample observation is weighted by the number of elements it represents from the overall population in order to make the choice-based sample "simulate" a random exogenous sample. The WESML estimator is obtained by maximizing the following weighted "pseudo-likelihood" function:

$$\ln L_w = \frac{1}{\gamma}\sum_{\{y_i=1\}}\ln(\Lambda_i) + \frac{1}{\alpha}\sum_{\{y_i=0\}}\ln(1-\Lambda_i) = -\sum_{i=1}^{n} w_i \ln(1+e^{(1-2y_i)x_i\beta})$$

where $w_i = (1/\gamma)\,y_i + (1/\alpha)(1-y_i)$. In addition, the appropriate estimator of the asymptotic covariance matrix is White's robust "sandwich" estimator used for pseudo-maximum likelihood estimation. Further, since the same citing patent can occur in multiple observations, the standard errors should be calculated without assuming independence across these observations.

5.2. Sample Construction

Since robust standard errors can be quite large for weighted logit estimation (Green, 2003, p. 673), I use relatively large samples to ensure statistically meaningful analysis. In addition, I improve the efficiency of estimation through stratification on technological characteristics of the citing and cited patents. In other words, each actual

[7] The appendix gives technical details. Also see Amemiya (1985), Greene (2003) or King and Zeng (2001) for

citation is matched with "control citations" with the same 3-digit technology classes for the citing and cited patents. The weighted likelihood function described above has to be generalized by defining the weight attached to a $y = 0$ observation as the reciprocal of the ex ante probability of a $y = 0$ population pair *with the same respective technological cell* (i.e., the combination of technological classes for the citing and cited patents) being selected into the sample.

I define the population of possible citations as all pairs of citing and cited patents in my data (over half a million patents from 1986-1995) such that the citing year does not come before the cited year. The sample used in regression analysis was drawn from this population as follows: First, all actual citations ($y = 1$) were included in the sample, except for self-citations from a geographical division of an organization to itself. Each of these "ones" was then matched with multiple potential citations ($y = 0$) that have the same "cell" as defined by the characteristics of the actual citation. This was done while making sure that no self-citation from a geographical division of an organization was included among the control citations either. This led a sample of 5.57 million actual and potential citations.

5.3. Control Variables for Probability of Citation

As the time lag between the citing and cited patents increases, the citation probability is known to increase initially and then fall beyond a certain point (Jaffe and Trajtenberg, 2002). To control for this, I introduce dummy variables for the number of years of lag between the citing and cited patents. In addition, since the patent citation rate may change over time, additional dummy variables are used to capture the citing year fixed effects. Since patents in different industry categories have different propensities to cite others, I also include fixed effects for the broad technological category of the citing patent, as defined in the Jaffe and Trajtenberg (2002) patent database.

---

more discussion. Sorenson and Fleming (2001) have used a similar approach for predicting patent citations.

The next issue is that innovators in different countries might have a different propensity to cite patents registered with the USPTO. For example, a US patent filed by a European inventor might not necessarily cite a USPTO patent for an innovation, but might instead cite the corresponding European Patent Office (EPO) patent for that innovation. In order to avoid possible biases arising from this, all regressions include citing country fixed effects. A later section uses EPO data to carry out additional robustness checks comparing propensity to cite for MNCs and domestic firms *within the same country*.

Patents that are technologically similar have a higher probability of citation. Existing patent citation literature typically compares the 3-digit technological class information on the citing and cited patents to control for this. However, this can lead to bias estimates since there can be large heterogeneity in technology within a 3-digit class. For example, the 3-digit class "Aeronautics" includes 9-digit classes as diverse as "Spaceship control" and "Aircraft seat belts" (Thompson and Fox-Kean, 2004). To take this into account, I define dummy variables for the same broad technological category (1 out of 6), the same technological subcategory (1 out of 36), the same 3-digit primary class (1 out of 418) and the same 9-digit primary class (1 out of 150,000). Further, since the designation of a subclass as "primary" can sometimes be ad hoc, I also include a dummy variable that captures whether at least one of the secondary subclasses of a patent is the same as one of the primary *or* secondary subclasses for the other patents. While there is a chance that even these technology controls are not perfect, these are the most fine-grained level possible with USPTO data, and are much more detailed than the coarse controls used in most studies.

6. Results

6.1. Intra-Region and Intra-MNC Knowledge Flows (Hypotheses 1 and 2)

Table 3 gives a summary of relevant variables used in the regressions. The results of weighted logit regressions (WESML) appear in Table 4, where the dependent variable is 1 for patent pairs that have a citation, 0 otherwise. Column (1) reproduces the empirical "fact" that knowledge flows are particularly strong within the same country and the same MNC. These effects, however, may partly result from technological specialization of regions and firms (Jaffe, Trajtenberg and Henderson, 1993). This is found to indeed be the case in column (2), where including controls at the 3-digit classification level reduces the estimated effects for *within same country* and *within same MNC*. Column (3) addresses the concern, raised by Thompson and Fox-Kean (2004), that commonly used controls just for the 3-digit technological class are not sufficient. In particular, this specification controls for additional similarity along 9-digit primary technological classification as well as overlap of secondary technological classes between the citing and cited patents. The results show that, though absence of detailed controls was indeed leading to the biases, the estimates for *within same country* and *within same MNC* still remain significant.

While statistical significance is not a surprise given the large sample size, let us now check for economic significance. The marginal effects are reported in square brackets, after multiplying by a million for readability.[8] Since the predicted citation rate between two random patents is found to be about 5.70 in a million, the marginal effect of 2.96 for *within same country* suggests that patents from different organizations within the same country are about 52% more likely to have a citation than are otherwise similar

[8] The marginal effect of a variable j is given by $\beta_j \Lambda'(\mathbf{x}\boldsymbol{\beta})$. From the logit form, it is easy to show that this equals $\beta_j \Lambda(\mathbf{x}\boldsymbol{\beta})[1-\Lambda(\mathbf{x}\boldsymbol{\beta})]$. One can then substitute either the mean predicted probability or the population mean for

patents from different organizations in different countries. Similarly, the marginal effect of 10.4 for *within same MNC* shows that patents from different international divisions of the same MNC are around 3 times as likely to have a citation than are those from different organizations in different countries, a finding consistent with that of Gomes-Casseres, Jaffe and Hagedoorn (2003).

6.2. Details of Intra-National Knowledge Flows (Hypotheses 3, 4, 5 and 6)

Table 5 breaks up the *within same country* knowledge flows into 4 types: between domestic entities (D→D), from domestic entities to local subsidiaries of foreign MNCs (D→M), from MNC subsidiaries to domestic entities (M→D) and between MNC subsidiaries (M→M). Figure 1 illustrates these definitions for clarity. The reference category is the cross-border inter-organizational knowledge flows, compared with which D→D knowledge flow probability is found to be greater by 3.0 in a million, D→M probability is greater by 3.0 in a million, M→D probability is greater by 2.1 in a million and M→M probability is greater by 4.4 in a million. Given that the average citation rate between two random patents is 5.7 in a million, all four kinds of intra-national knowledge flow effects are quite large in relative magnitude. The fact that M→D and D→M flows are both positive and significant, with the latter exceeding the former, is consistent with the earlier findings using matching (Table 2).

Table 5 also breaks down the *within same MNC* category into two sub-categories: knowledge flows from a foreign subsidiary of an MNC to its home base (S→H), and from its home base to the foreign subsidiary (H→S). The comparable (and statistically indistinguishable) estimates suggest that the probability with which a patent from a foreign subsidiary cites one from the MNC's home base is about the same as that with which a

---

$\Lambda(\boldsymbol{x\beta})$ for getting an estimate of the marginal effect. I report the former. The latter estimate is typically slightly higher in value.

patent from the home base cites one from the subsidiary. This is consistent with a view of MNCs as a "learning organization", where subsidiaries not only build upon the knowledge of the home base but also contribute to further learning (Kogut and Zander, 1993; Dunning, 1993).

The bottom of the table reports the relative magnitude and statistical comparison of different estimates. The coefficient for M→D flows is 30% smaller than for D→M flows, as indicated by the ratio $\beta_{M\to D}$ / $\beta_{D\to M}$ of 0.7. A test of equality of $\beta_{M\to D}$ and $\beta_{D\to M}$ is rejected at the 1% significance level. Similarly, M→D flows are statistically smaller than the D→D flows (by 30%). D→M flows, on the other hand, are not any weaker in strength than D→D flows. Thus, the intensity of knowledge flows from domestic organizations to MNC subsidiaries is statistically no different from that between domestic organizations themselves. There is little evidence that MNC subsidiaries face a "liability of foreignness" (i.e., are unable to tap into the localized knowledge exchange in a country). To summarize, while MNCs are as good at learning from domestic organizations as domestic organizations are at learning from each other, MNCs contribute somewhat less to local learning.[9]

It is interesting to note that multinational subsidiaries are also really good at learning from each other, with the M→M estimate being much greater than that for even D→D or D→M knowledge flow. This is consistent with previous findings on knowledge spillovers between MNC subsidiaries (Head, Ries and Swenson, 1995; Feinberg and Majumdar, 2001; Feinberg and Gupta, 2003). In analysis not reported here, I found the

[9] In order to rule out the possibility the result is due to knowledge flows from domestic universities/research labs to MNC subsidiaries, I included separate dummy variables for whether the D→M flows were originating from domestic firms or domestic universities/research labs. I found that the D→M flows originating from domestic firms are actually slightly *higher* rather than lower than the D→M flows from domestic universities/research labs.

M→M effect to be driven largely by the probability of knowledge flow being very high between foreign subsidiaries of MNCs from the same home country.

6.3. Cross-Country Differences in Bi-directional Knowledge Flows (Hypothesis 7)

What is the underlying mechanism for the result that knowledge flows from the host countries to the MNCs exceed those back from the MNCs to the host countries? To dig deeper into this issue, I repeat the above analysis for the six individual countries. In Table 6, I interact each of the six indicator variables discussed earlier with dummy variables for countries. I find evidence of strong intra-national knowledge flows in all countries.

The aggregate finding that D→M knowledge flows are stronger than M→D knowledge flows holds true for the US, Japan and Germany.[10] The equality of the two-way flows cannot be rejected for France and Canada, while the trend actually reverses for the UK. One explanation for this pattern is that the domestic firms and organizations in the US, Japan and Germany are, on an average, technologically more advanced than the average subsidiary of a foreign multinational based there, and therefore have much less to learn from the latter. R&D data from OECD (1998) supports this explanation: the R&D intensity (i.e., R&D/production) of domestic firms and foreign MNCs differs most in Germany and Japan, with the domestic R&D intensity being almost twice of that for MNC subsidiaries. It is therefore no surprise that the disparity between D→M and M→D flows is also highest for these two countries. Likewise, the fact that UK is the only country where D→M knowledge flows are significantly *weaker* than M→D knowledge flows is consistent with the fact that UK is the only country where the R&D intensity of MNCs *exceeds* that of domestic players.

[10] Thus, though Japanese firms gain by investing in the US, US firms also gain by investing in Japan, giving no evidence of Japanese firms being worse overall at sharing knowledge, a finding consistent with Spencer (2000).

6.4. Cross-Sector Differences in Bi-directional Knowledge Flows (Hypothesis 7)

To investigate the heterogeneity in knowledge flows further, I now look at cross-sector differences since learning-related incentives for location choice are greater for technologies where new knowledge plays an important role (Audretsch and Feldman, 1996). In particular, when locating abroad can lead to learning, both industry laggards and leaders have an incentive to open overseas subsidiaries. On the other hand, when the learning opportunities are small compared with potential leakage of their own technology, the leaders have less incentive to locate abroad. To explore this, I now break down analysis of innovations originating in the US into six broad technology categories.[11]

The sample used in Table 7 includes only the citing patents from the US. I interact each of the six indicator variables discussed earlier with dummy variables for technological categories. Although this coarse technological classification surely hides heterogeneity within technological categories, some interesting patterns still emerge. First, "Drugs & Medical" and "Chemical", two of the most R&D intensive sectors, show high levels of knowledge exchange among all players. This is consistent with Chung and Alcacer (2002), who suggest that these are sectors where not just the foreign industry laggards but also industry leaders actively locate advanced facilities in the US. For example, all foreign pharmaceutical firms invest heavily in R&D in the US in order to keep abreast with the latest developments in a sector that involves discrete product innovation and a long uncertain product innovation process: R&D intensity for Pharmaceuticals is 10.5% for MNC subsidiaries, which is even higher than the 6.5% figure for domestic firms (OECD, 1998). Since MNC subsidiaries in these sectors are quite advanced, it is natural that the issue of weak M→D flows resulting from adverse selection in the technological competence of subsidiaries would not exist in these sectors.

Two individual sectors where D→M knowledge flows are indeed significantly stronger than M→D knowledge flows are "Computers & Communication" and "Electrical & Electronics". This is consistent with Chung and Alcacer's (2002) finding that FDI in these sectors is dominated by industry laggards. For example, R&D intensity for Computers is 4.5% for MNC subsidiaries and 13.5% for domestic firms in the US (OECD, 1998). This is also consistent with Florida's (1997) finding that 37% of the MNC subsidiaries in the US for these sectors have a "listening post" role, as opposed to only 17% in "Chemicals" and 25% in "Drugs & Medical." For the "Mechanical" category, all three kinds of localized knowledge flows involving MNC subsidiaries are weaker than D→D flows, possibly because it is not a particularly knowledge-intensive sector.

6.5. Cross-Border Citations between Different Firms (Hypothesis 8)

The above analyses study intra-national, inter-firm knowledge flows (D→D, D→M, M→D and M→M) and cross-border, within-firm knowledge flows (S→H and H→S). Taken together, the two show that MNC subsidiaries are an intermediary for *cross-border, inter-firm* knowledge flow. I now look for possible direct effect of an MNC's subsidiary activity on the probability of *cross-border citation between different firms* (i.e., between host country domestic players and the MNC home base). Two caveats should be made: First, this is a very strong test. While an increased probability of cross-border citation between different firms suggests intense knowledge flow, a zero effect *does not* indicate an absence of such knowledge flow since knowledge flowing indirectly through a subsidiary need not result in *cross-border* citation between different firms. Second, the findings are based on a cross-sectional comparison, without modeling the endogeneity of the decision to locate overseas.

[11] I would have liked to repeat the sector-level analysis for other individual countries, and for a finer sector classification, but the smaller resulting sample sizes for patents by MNC subsidiaries made that impractical.

I define the "presence" of the citing assignee in the cited country as the logarithm of the number of patents originating from its subsidiary in the cited country. This can be seen as a measure of its local absorptive capacity (Cohen and Levinthal, 1989). Similarly, I define the "presence" of the cited assignee in the citing country as the logarithm of the number of patents originating from its subsidiary in the citing country. In addition to the control variables already discussed above, additional controls used are the logarithm of worldwide patenting by the citing assignee and by the cited assignee. This ensures that the foreign presence variables do not simply pick up overall scale effects, which would arise if larger assignees systematically differ in the propensity to cite or be cited.

Since I am now interested only in cross-border patent citations between different players, all patent pairs from the same firm or the same country are now dropped. The regression results are reported in Table 8. The negative estimate for the global scale of the citing assignee suggests that larger firms rely much less on external sources of knowledge, perhaps because they build more upon their own internal knowledge. Similarly, the positive estimate for the global scale of the cited assignee suggests that patents from larger firms have a greater likelihood of being cited by other firms.

As discussed above, the variables of most interest to us are the presence of the citing assignee in the cited country, and that of the cited assignee in the citing country. The marginal effects of these variables can be interpreted follows. A 1% increase in inventive activity by a foreign MNC's local subsidiary increases the citation probability by the foreign MNC's *home base* to the host country's domestic players by 3% (recall that regressions use log of presence, hence the percentage interpretations). In contrast, there is only a 1.1% increase in citation probability by the host country's domestic players to the foreign MNC's home base when the MNC's local innovative activity goes up by 1%.

Thus, though increased MNC activity is associated with increased cross-border patent citations in both directions, the asymmetry found in intra-national citations exists even for cross-national patent citations: the MNC *home base* gains more in terms of inter-organizational knowledge spillovers from its overseas investments than the domestic players in the host country do. These findings are consistent with similar results found in more specialized settings by Branstetter (2000) for Japanese FDI in the US, and Globerman, Kokko and Sjöholm (2000) for inward and outward FDI for Sweden. Further, when I analyzed the data separately for the six countries, increased presence of citing MNC in cited country had a positive and significant effect on citation probability in five of the six countries: US, Japan, France (at 10% significance level), UK and Canada. On the other hand, increased presence of potentially cited MNC in the citing country had a positive and significant effect in only 2 countries: Japan and Canada (at 10% level). Once more, this suggests that the latter result is weaker than the former.

7. Further Issues in Using USPTO Patent Citations

All regressions in this paper include country fixed effects to control for systematic cross-country differences in propensity to cite USPTO patents. However, this does not resolve a related concern that MNC subsidiaries and domestic organizations *even within the same country* might differ in their propensity to cite USPTO patents. In particular, patents from MNC subsidiaries might have a systematically different tendency to cite USPTO patents and instead cite a patent representing the same innovation but registered with another country's patent office.[12] To look into this possibility, I examined citations made to both USPTO and European Patent Office (EPO) patents by a random sample of 1,612 USPTO patents from 1995, about half of them originating in domestic organizations

and the other half in MNC subsidiaries. For each patent in the sample, I identified if one or more cited EPO patents did have equivalent USPTO patents that could equivalently have been cited, and therefore represent "missing citations" in USPTO data. The mapping from EPO to USPTO patents was done using the "OECD Triadic Patent Families" database, which has information on patents filed for the same innovation at both USPTO and EPO.

The results are summarized in Table 9. The mean number of citations to USPTO patents by a patent from the above sample was 5.85, while the mean number of citations to EPO patents was 1.13. A large fraction of the EPO citations did not have an equivalent USPTO patent, hence do not reflect any bias in the estimate of probability of citation between just the innovations captured by USPTO patents. The mean number of citations to EPO patents that do have equivalent USPTO patents, which really gives the number of "missing citations" described above, was only 0.32 per patent. The missing citations are thus quite small in number compared with citations that do get made to USPTO patents. Further, as Table 9 shows, the average number of missing citations per patent from MNC subsidiaries (0.43) is a little higher than those for domestic organizations (0.22). This holds both in the sub-sample of patents originating in the US (0.39 for MNC subsidiaries and 0.24 for domestic organizations), and for those that originate elsewhere (0.46 for MNC subsidiaries and 0.21 for domestic organizations). In either sub-sample, the missing citation bias therefore is in the direction of underestimating the extent of localized knowledge diffusion to MNCs more than to domestic organizations. In other words, if we could correct for this bias in the previous analysis, it would slightly strengthen the main result that probability of D→M knowledge flow exceeds that of M→D knowledge flow.

[12] Since USPTO patents provide patent protection only in the US, a patent needs to be separately applied for in Europe for protection there.

8. Discussion and Concluding Remarks

Much of the recent debate on globalization has centered on whether MNCs contribute as much as they gain from their host countries. To address one aspect of this broad issue, I study how the extent of knowledge flows from MNCs to a host country compares with knowledge acquisition by MNCs from the host country. Analysis of patent citation data reveals that, while local subsidiaries of foreign MNCs help a country gain access to knowledge originating in foreign firms, they also cause domestic technology to fall into the hands of foreign competitors. Thus, knowledge spillovers from inward FDI, particularly in countries that possess valuable technology of their own, are not free – they come at the cost of significant "leakage" of domestic knowledge. Knowledge flows from domestic organizations to MNCs are found to significantly exceed those from MNCs to domestic organizations for three of the six largest economies (US, Japan and Germany), and two of the six broad technological categories ("Computers & Communications" and "Electrical & Electronics").

The above patterns are consistent with a hypothesis that *net* knowledge flows from foreign MNC subsidiaries to domestic players are strongest in countries and industries where MNC subsidiaries are involved in knowledge-intensive activities. For the policy maker, it implies that not just the magnitude of FDI but also its level of sophistication should be considered in pursuing knowledge spillovers. Policies should focus on attracting FDI that is technologically sophisticated, and on sectors where the host country is a technological laggard. Further, the findings suggest that *outward* FDI might sometimes be more effective than *inward* FDI for acquiring knowledge originating abroad. Thus, instead of only promoting inward FDI and discouraging outward FDI, a country might gain from encouraging its domestic firms to also seek out foreign sources of knowledge.

There are three caveats to any policy interpretation of my results. First, knowledge diffusion effects are only a part of the overall welfare effects of MNCs. Second, patent citation data does not allow separate measurement of knowledge *transfers* (which are planned, priced and paid for) and knowledge *spillovers* (which are unintended externalities). Third, endogeneity of the MNC's decision of whether and where to locate overseas is not incorporated in my model.

The focus of this paper has been developed countries, partly because patent data is not as meaningful a source of information for developing countries. In particular, knowledge spillovers in developing countries lead less often to radical innovation and more often result in adoption of existing technologies. Also, since domestic organizations are rarely as advanced as foreign MNCs, the learning effect in developing countries might be weaker for MNCs and stronger for domestic organizations. But the general point made in the paper should still apply: not only the magnitude but also the knowledge content of investments by foreign MNCs affects the possibility of knowledge spillovers. Different kinds of MNC activity, like state-of-the-art R&D or production facilities versus simple assembly operations, might have different implications for knowledge flows. Future research on FDI should therefore focus less on just measurement of knowledge spillovers, and more on the conditions needed for and the mechanisms driving such spillovers.

References

Aitken, B., and A. Harrison (1999), "Do Domestic Firms Benefit from Foreign Investment? Evidence from Venezuela" *American Economic Review* 89: 605-618

Almeida P. (1996), "Knowledge sourcing by foreign multinationals: Patent citation analysis in the U.S. semiconductor industry" *Strategic Management Journal* 7: 155-165.

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

Audretsch, D.B. and M.P. Feldman (1996), "R&D Spillovers and the Geography of Innovation and Production" *American Economic Review*, 86(3): 630-640.

Bartlett, C.A. and S. Ghoshal (1989). *Managing Across Borders: The Transnational Solution*. Harvard Business School Press: Boston, MA.

Branstetter, L. (2000), "Is foreign direct investment a channel of knowledge spillovers? Evidence from Japan's FDI in the United States" National Bureau of Economic Research Working Paper # 8015.

Branstetter, L. (2001), "Are Knowledge Spillovers International or Intranational in Scope? Microeconometric Evidence from the U.S. and Japan" *Journal of International Economics* 53: 53-79.

Buckley, P.J. and M.C. Casson (1976). *The Future of the Multinational Enterprise.* London: Holmes & Meier.

Caves, R.E. (1974), "Multinational firms, competition and productivity in host-country markets" *Economica*, 41: 176-193.

Caves, R.E. (1996) *Multinational Enterprise and Economic Analysis.* Cambridge University Press (Second Edition).

Cantwell, J.A. and O. Janne (1999), "Technological globalisation and innovative centres: the role of corporate technological leadership and locational hierarchy," *Research Policy*. 28:119-144.

Chung, W., W. Mitchell and B Yeung (2003), "Foreign direct investment and host country productivity: The American automotive component industry in the 1980s," *Journal of International Business Studies*. 34(2): 199-218.

Chung, W. (2001), "Identifying Technology Transfer in Foreign Direct Investment: Influence of Industry Conditions and Investing Firm Motives," *Journal of International Business Studies*, 32 (3): 211-229.

Chung, W., and J. Alcacer (2002), "Knowledge Seeking and Location Choice of Foreign Direct Investment in the United States" *Management Science* 48(12):1534-1554.

Coe, D.T., and E. Helpman (1995), "International R&D Spillovers" *European Economic Review* 39: 859-887.

Cohen, W., and D. Levinthal (1989), "Innovation and learning: The two faces of R&D", *Economic Journal* 99: 569-596.

Dalton, D.H. and M.G. Shapiro (1995). *Globalizing Industrial Research & Development*. Office of Technology Policy, US Dept of Commerce.

Duget, E. and M. MacGarvie (2002), "How Well Do Patent Citations Measure Knowledge Spillovers?" Working paper.

Dunning, J.H. (1993). *Multinational Enterprises and the Global Economy.* Addison-Wesley.

Eaton, J., and S. Kortum (1999), "International Patenting and Technology Diffusion: Theory and Measurement", *International Economic Review* 40: 537-570.

Ethier, W. (1986), "The Multinational Firm", *Quarterly Journal of Economics* 101:805-834.

Feinberg, S. and S.K. Majumdar (2001), "Technology Spillovers From Foreign Direct Investment In The Indian Pharmaceutical Industry. Journal of International Business Studies," 32(3): 421-438.

Feinberg, S., and A.K. Gupta (2003), "Knowledge Spillovers and the Assignment of R&D Responsibilities to Foreign Subsidiaries" *Strategic Management Journal*.

Florida, R. (1997), "The globalization of R & D: results of a survey of foreign-affiliated R&D laboratories in the USA" *Research Policy* 26(1): 85-103.

Frost, T.S. (2001), "The geographical sources of foreign subsidiaries' innovations" *Strategic Management Journal* 22: 101-123.

Frost, T.S., J.M. Birkinshaw and P.C. Ensign (2003), "Centers of Excellence in Multinational Firms." *Strategic Management Journal*, 23: 997-1018.

Globerman, S., A. Kokko, and F. Sjöholm (2000), "International Technology Diffusion: Evidence from Swedish Patent Data", *Kyklos* 53: 17-38. 55

Gomes-Casseres, B., A.B. Jaffe and J. Hagedoorn (2003), "Do Alliances Promote Knowledge Flows?" Mimeo.

Griliches, Z. (1990), "Patent statistics as economic indicators: A survey" *Journal of Economic Literature* 28: 1661-1797.

Greene, W. (2003) *Econometric Analysis*. Prentice Hall, 5th Edition.

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

Head K, J. Ries and D. Swenson (1995), "Agglomeration benefits and location choice: Evidence from Japanese manufacturing investments in the United States." *Journal of International Economics*, 38: 223-247.

Hedlund, G. (1986), "The hypermodern MNC: A heterarchy?" *Human Resource Management* **25**(1): 9-35.

Hymer, S.H. (1976). *The International Operations of National Firms: A Study of Direct Investment.* MIT Press, Boston, MA.

Jaffe, A.B., M. Trajtenberg and R. Henderson (1993), "Geographic localization of knowledge spillovers as evidenced by patent citations" *Quarterly Journal of Economics* 434: 578-598.

Jaffe, A.B. and M. Trajtenberg (2002). *Patents, Citations & Innovations: A window on the knowledge economy*. MIT Press, Cambridge, MA.

Keller, W. (2002), "Geographic localization of international technology diffusion", *American Economic Review* 92(1).

King, G. and L. Zeng (2001), "Logistic Regression in Rare Events Data", *Political Analysis* 9(2): 137-163

Kogut, B. and S.J. Chang (1991), "Technological Capabilities and Japanese Foreign Direct Investment in the United States" *The Review of Economics and Statistics* 73 (3): 401-413.

Kogut, B. and U. Zander (1993), "Knowledge of the Firm and the Evolutionary Theory of the Multinational Corporation." *Journal of International Business Studies*. 24(4), pp. 625-645.

Kuemmerle, W. (1999), "Foreign direct investment in industrial research in the pharmaceutical and electronics industries – results from a survey of multinational firms" *Research Policy* 28: 179-193.

Lerner, J. (2002), "150 Years of Patent Protection" *American Economic Review* 92 (2).

Levin, R., A. Klevorick, R. Nelson and S. Winter (1987), "Appropriating the Returns from Industrial Research and Development." *Brookings Papers on Economic Activity* 3:783-820.

Mansfield, E., D. Teece and A. Romeo (1979), "Overseas research and development by US-based firms." *Economica* 46(182) 187-196.

Manski, C.F. and S.R. Lerman (1977), "The Estimation of Choice Probabilities from Choice Based Samples," *Econometrica* 45(8): 1977-88.

Mowery, D.C., J.E. Oxley and B.S. Silverman (1996), "Strategic Alliances and Inter-firm Knowledge Transfer," *Strategic Management Journal*, 17: 77-91.

Nohria, N. and S. Ghoshal (1997). *The Differentiated network: Organizing Multinational Corporations for Value Creation.* Jossey-Bass Publishers, San Francisco.

OECD (1998). *Internationalisation of Industrial R&D: Patterns and Trends.*

Peri, G. (2003), "Knowledge Flows, R&D Spillovers and Innovation," Mimeo.

Romer, P.M. (1990), "Endogenous Technological Change." *Journal of Political Economy* 98 (5): S71-S102.

Shaver, J.M. and F. Flyer (2000), "Agglomeration Economies, Firm Heterogeneity, and Foreign Direct Investment in the United States," *Strategic Management Journal* 21: 1175-1193.

Singh, J. (2004), "Social Networks as Determinants of Knowledge Diffusion Patterns." Mimeo. Available at http://www.jasjitsingh.com/academic/papers.html

Sorenson, O. and L. Fleming (2001) "Science and the Diffusion of Knowledge." Working paper, Harvard Business School.

Spencer, J.W. (2000), "Knowledge Flows in the Global Innovation System: Do U.S. Firms Share More Scientific Knowledge than their Japanese Rivals?" *Journal of International Business Studies*. 31(3): 521-530.

Teece, D.J. (1986), "Transaction cost economics and multinational enterprise" *Journal of Economic Behavior and Organization*, 7: 21-45.

Thompson, P. and M. Fox-Kean (2004), "Patent Citations and the Geography of Knowledge Spillovers: A Reassessment." *American Economic Review*, forthcoming.

Williamson, O.E. (1985). *The Economic Institutions of Capitalism*. The Free Press.

Table 1: Overview of patent data

Country	Total patents 1986-95 in NBER database	Total number of assigned patents	Assigned patents with clean parent information	Fraction of patents from multinational subsidiaries
	(1)	(2)	(3)	(4)
United States	546,824	418,045	287,787	8.5%
Japan	217,313	212,427	183,870	2.1%
Germany	74,041	67,154	45,869	19.5%
France	29,791	27,120	17,289	20.4%
United Kingdom	26,631	23,968	15,131	40.3%
Canada	20,700	13,015	5,697	50.0%
Subtotal 6 countries	915,300	761,729	555,643	9.0%
Other countries	94,924	73,115	38,402	27.3%
Total worldwide	1,010,224	834,844	594,045	10.2%

Table 2(a): Knowledge diffusion from MNC subsidiaries to domestic organizations (M→D)

		Actual Citations		Control Citations		Comparison		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Country	Total citations by domestic	Citations by domestic to mult sub	%Citations by domestic to mult sub	Citations by domestic to mult sub	%Citations by domestic to mult sub	(3) - (5)	t-ratio	(3)/(5)
United States	430,262	17,010	3.95%	15,136	3.52%	0.44%	10.7	1.12
Japan	245,441	2,082	0.85%	1,879	0.77%	0.08%	3.2	1.11
Germany	27,326	658	2.41%	542	1.98%	0.42%	3.4	1.21
France	12,727	124	0.97%	101	0.79%	0.18%	1.5	1.23
United Kingdom	7,895	197	2.50%	149	1.89%	0.61%	2.6	1.32
Canada	3,536	32	0.90%	15	0.42%	0.48%	2.5	2.13
Total	**727,187**	**20,103**	**2.76%**	**17,822**	**2.45%**	**0.31%**	**11.9**	**1.13**

Table 2(b): Knowledge diffusion from domestic organizations to MNC subsidiaries (D→M)

		Actual Citations		Control Citations		Comparison		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Country	Total citations by mult sub	Citations by mult sub to domestic	%Citations by mult sub to domestic	Citations by mult sub to domestic	%Citations by mult sub to domestic	(3) - (5)	t-ratio	(3)/(5)
United States	41,272	22,590	54.73%	18,799	45.55%	9.19%	26.5	1.20
Japan	5,156	2,464	47.79%	2,083	40.40%	7.39%	7.6	1.18
Germany	10,841	1,302	12.01%	985	9.09%	2.92%	7.0	1.32
France	3,856	166	4.30%	114	2.96%	1.35%	3.2	1.46
United Kingdom	9,689	220	2.27%	274	2.83%	-0.56%	-2.5	0.80
Canada	3,457	38	1.10%	25	0.72%	0.38%	1.6	1.52
Total	**74,271**	**26,780**	**36.06%**	**22,280**	**30.00%**	**6.06%**	**24.9**	**1.20**

Table 3: Summary of variables used for regressions analysis

Same tech category	Indicator variable that is 1 if both the citing and the potentially cited patent belong to the same broad industry category (one of 6) as defined in the Jaffe and Trajtenberg (2002) database
Same tech subcategory	Indicator variable that is 1 if both the citing and the potentially cited patent belong to the same broad technical subcategory (one of 36) as defined in the Jaffe and Trajtenberg (2002) database
Same primary tech class	Indicator variable that is 1 if both the citing and the potentially cited patent belong to the same 3-digit primary technology class (one of about 450) as defined in the US Patent classification system
Same primary subclass	Indicator variable that is 1 if both the citing and the potentially cited patent belong to the same 9-digit primary technology subclass (one of about 150,000) as defined in the US Patent classification system
Secondary subclass overlap	Indicator variable that is 1 if at least one of the secondary 9-digit subclasses of one patent is the same as a primary or secondary subclass of the other patent in the dyad
Within same country	Indicator variable that is 1 if the citing and cited patents originate from inventors located in the same country
Within same MNC	Indicator variable that is 1 if the citing and cited patents are from two divisions (located in different countries) of the same MNC
D→D	Indicator variable that is 1 if both the citing and potentially cited patent belong to the same country, with assignees for both being domestic players in the country
D→M	Indicator variable that is 1 if both the citing and potentially cited patent belong to the same country, with assignee for the former being a local subsidiary of a foreign multinational and for the latter being a domestic player
M→D	Indicator variable that is 1 if both the citing and potentially cited patent belong to the same country, with assignee for the former being a domestic player and for the latter being a local subsidiary of a foreign multinational
M→M	Indicator variable that is 1 if both the citing and potentially cited patent belong to the same country, with assignees for both local subsidiaries of foreign multinationals
S→H	Indicator variable that is 1 if citing patent is from the home base of an MNC and the cited patent is from a foreign subsidiary (located abroad) of the same MNC
H→S	Indicator variable that is 1 if citing patent is from the local subsidiary of a foreign MNC and the cited patent is from the home base (located abroad) of the same MNC
Presence of citing assignee in cited country	Log(1 + number of patents that originate in the same country as the potentially cited patent and are assigned to the citing entity)
Presence of cited assignee in citing country	Log(1 + number of patents that originate in the same country as the citing patent and are assigned to the potentially cited entity)
Scale of citing assignee	Log(number of worldwide patents for 1980-99 that are assigned to the citing entity)
Scale of cited assignee	Log(number of worldwide patents for 1980-99 that are assigned to the cited entity)

Table 4: Intra-national and intra-MNC knowledge flows

	(1)	(2)	(3)
Within same country	0.672**	0.578**	0.520**
	(0.009)	(0.005)	(0.009)
	[3.83]	[3.29]	[2.96]
Within same MNC	3.291**	2.110**	1.825**
	(0.110)	(0.026)	(0.050)
	[18.76]	[12.03]	[10.40]
Technological relatedness:			
Same tech category		1.148**	1.108**
		(0.011)	(0.012)
Same tech subcategory		1.246**	1.218**
		(0.014)	(0.015)
Same primary tech class		3.243**	1.930**
		(0.011)	(0.015)
Same primary subclass			2.282**
			(0.028)
Secondary subclass overlap			4.111**
			(0.012)
Number of observations	5,577,206	5,577,206	5,577,206

A weighted logit regression is used, with the dependent variable being 1 if there is a citation between two patents and 0 otherwise
Robust standard errors in parentheses, with clustering on citing patent
Marginal effects in square brackets after multiplication with 1,000,000
Fixed effects used for technological category of citing patent, country of citing patent, citing patent year and time lag between patents
** significant at 1%; * significant at 5%

Table 5: Detailed break-up of intra-national and intra-MNC knowledge flows

Within same country	
D→D	0.525** (0.010) [2.99]
D→M	0.521** (0.032) [2.97]
M→D	0.366** (0.030) [2.09]
M→M	0.768** (0.096) [4.38]
Within same MNC	
S→H	1.796** (0.080) [10.24]
H→S	1.848** (0.061) [10.53]
Observations	5,577,206
$\delta_{D\to M}$ / $\delta_{D\to D}$	0.99
$\delta_{M\to D}$ / $\delta_{D\to D}$	0.70**
$\delta_{M\to M}$ / $\delta_{D\to D}$	1.46**
$\delta_{M\to D}$ / $\delta_{D\to M}$	0.70**
$\delta_{H\to S}$ / $\delta_{S\to H}$	1.03

A weighted logit regression is used, with the dependent variable being 1 if there is a citation between two patents and 0 otherwise
Robust standard errors in parentheses, with clustering on citing patent
Marginal effects in square brackets after multiplication with 1,000,000
Controls for technological similarity of citing and cited patent included in regression, but not shown here to enhance readability
Fixed effects used for technological category of citing patent, country of citing patent, citing patent year and time lag between patents
** significant at 1%; * significant at 5% (In case of ratios, whether statistically different from 1 is tested)

Table 6: Intra-national and intra-MNC knowledge flows in different countries

	Country of origin of citing patent					
	US	**Japan**	**Germany**	**France**	**UK**	**Canada**
Within same country						
D→D	0.517**	0.535**	0.503**	0.526**	0.688**	1.406**
	(0.013)	(0.016)	(0.042)	(0.089)	(0.141)	(0.173)
D→M	0.491**	0.579**	0.941**	0.700**	0.281*	0.865**
	(0.037)	(0.081)	(0.114)	(0.148)	(0.109)	(0.213)
M→D	0.371**	0.255*	0.461**	0.719**	0.670**	1.015**
	(0.032)	(0.103)	(0.082)	(0.149)	(0.143)	(0.245)
M→M	0.695**	1.357**	0.633**	1.738**	0.934**	1.061**
	(0.120)	(0.354)	(0.235)	(0.338)	(0.167)	(0.309)
Within same MNC						
S→H	1.925**	1.771**	1.153**	1.357**	1.920**	2.383**
	(0.107)	(0.212)	(0.204)	(0.192)	(0.211)	(0.292)
H→S	1.607**	2.097**	2.203**	1.964**	1.644**	2.177**
	(0.115)	(0.251)	(0.145)	(0.120)	(0.095)	(0.100)
Country fixed effect	-	-0.384**	-0.319**	-0.248**	-0.064	-0.022
		(0.014)	(0.021)	(0.018)	(0.038)	(0.028)
$\delta_{D\to M} / \delta_{D\to D}$	0.95	1.08	1.87**	1.33	0.41*	0.62*
$\delta_{M\to D} / \delta_{D\to D}$	0.72**	0.48**	0.92	1.37	0.97	0.72
$\delta_{M\to M} / \delta_{D\to D}$	1.34	2.54*	1.26	3.30**	1.36	0.75
$\delta_{M\to D} / \delta_{D\to M}$	0.76**	0.44**	0.49**	1.03	2.38*	1.17
$\delta_{H\to S} / \delta_{S\to H}$	0.83*	1.18	1.91**	1.45**	0.86	0.91

A weighted logit regression is used, with the dependent variable being 1 if there is a citation between two patents and 0 otherwise
Robust standard errors in parentheses, with clustering on citing patent
Controls for technological similarity of citing and cited patent included in regression, but not shown here to enhance readability
Fixed effects used for technological category of citing patent, country of citing patent, citing patent year and time lag between patents
** significant at 1%; * significant at 5% (In case of ratios, whether statistically different from 1 is tested)

Table 7: Intra-national and intra-MNC knowledge flows for different sectors in the U.S.

	Technological category of citing patent					
	Chemical	**Computers & Communications**	**Drugs & Medical**	**Electrical & Electronic**	**Mechanical**	**Other**
Within same country						
D→D	0.390**	0.650**	0.671**	0.438**	0.251**	0.826**
	(0.029)	(0.021)	(0.068)	(0.025)	(0.028)	(0.055)
D→M	0.401**	0.687**	0.645**	0.420**	0.151	0.587**
	(0.065)	(0.056)	(0.185)	(0.082)	(0.102)	(0.112)
M→D	0.400**	0.390**	0.650**	0.100	0.169*	0.760**
	(0.063)	(0.064)	(0.103)	(0.079)	(0.073)	(0.121)
M→M	0.492*	0.745**	1.633**	0.401	-0.124	1.749**
	(0.208)	(0.184)	(0.228)	(0.358)	(0.285)	(0.239)
Within same MNC						
S→H	1.861**	1.780**	2.270**	1.747**	2.504**	1.895**
	(0.231)	(0.147)	(0.406)	(0.249)	(0.252)	(0.488)
H→S	1.875**	1.024**	2.351**	1.638**	2.052**	1.461*
	(0.212)	(0.190)	(0.336)	(0.275)	(0.290)	(0.656)
Category fixed effect	-	0.900**	-0.725**	0.511**	0.612**	-0.372**
		(0.027)	(0.059)	(0.029)	(0.030)	(0.048)
$\delta_{D\to M} / \delta_{D\to D}$	1.03	1.06	0.96	0.96	0.60	0.71**
$\delta_{M\to D} / \delta_{D\to D}$	1.03	0.60**	0.97	0.23**	0.67	0.92
$\delta_{M\to M} / \delta_{D\to D}$	1.26	1.15	2.43**	0.92	-0.49	2.12**
$\delta_{M\to D} / \delta_{D\to M}$	1.00	0.57**	1.01	0.24**	1.12	1.29
$\delta_{H\to S} / \delta_{S\to H}$	1.01	0.58**	1.04	0.94	0.82	0.77

A weighted logit regression is used, with the dependent variable being 1 if there is a citation between two patents and 0 otherwise
Robust standard errors in parentheses, with clustering on citing patent
Controls for technological similarity of citing and cited patent included in regression, but not shown here to enhance readability
Fixed effects used for technological category of citing patent, country of citing patent, citing patent year and time lag between patents
** significant at 1%; * significant at 5% (In case of ratios, whether statistically different from 1 is tested)

Table 8: Effect of MNC subsidiary activity on cross-border citations between different firms

Presence of citing assignee in cited country	0.030** (0.004) [0.16]
Presence of cited assignee in citing country	0.011** (0.004) [0.06]
Scale of citing assignee	-0.012* (0.006) [-0.06]
Scale of cited assignee	0.031** (0.005) [0.17]
Observations	3,027,928

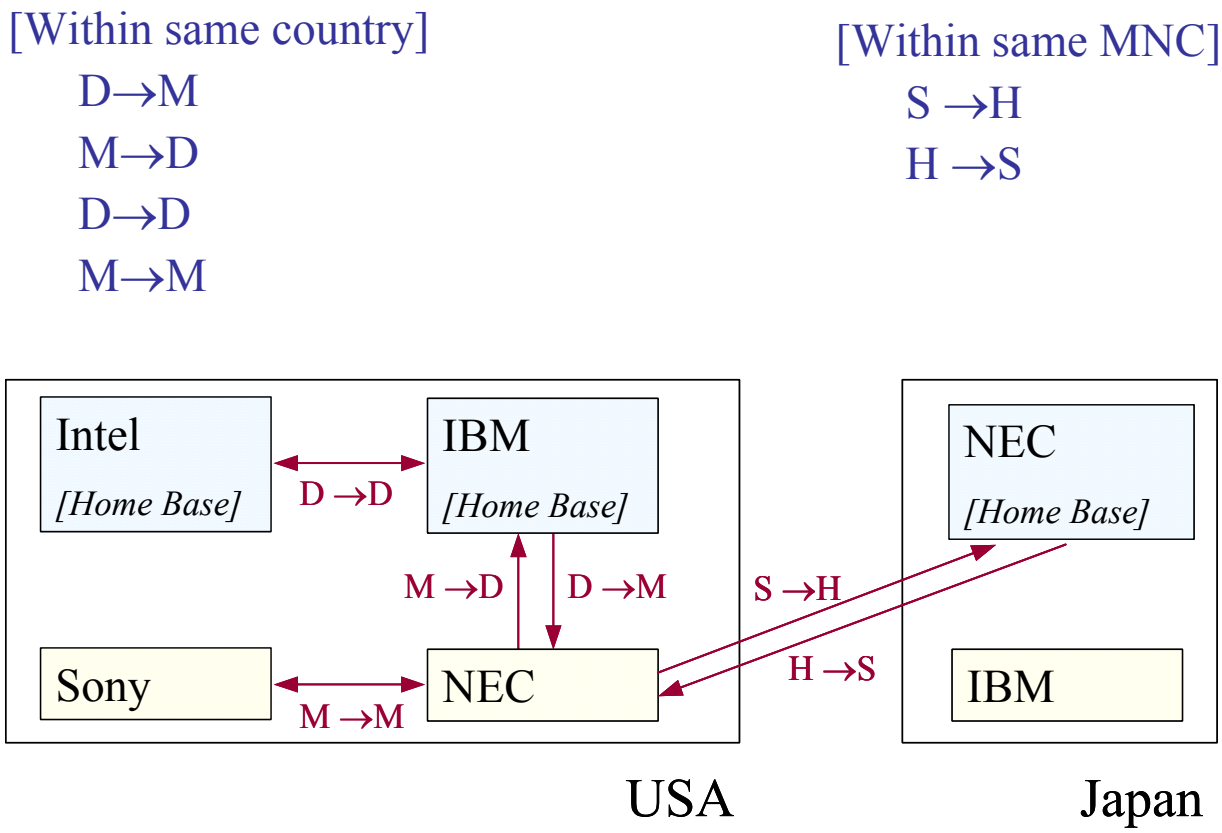
A weighted logit regression is used, with the dependent variable being 1 if there is a citation between two patents and 0 otherwise
Robust standard errors in parentheses, with clustering on citing patent
Marginal effects in square brackets after multiplication with 1,000,000
Controls for technological similarity of citing and cited patent included in regression, but not shown here
Fixed effects used for technological category of citing patent, country of citing patent, citing patent year and time lag
** significant at 1%; * significant at 5%

Table 9: Frequency of USPTO and EPO citations by a USPTO patent

	Citing patents from all countries			Citing patents from US		Citing patents not from US	
	All assignees (N=1,612)	Domestic (N=810)	MNC (N=802)	Domestic (N=436)	MNC (N=369)	Domestic (N=374)	MNC (N=433)
Mean number of citations to USPTO patents	5.84	5.68	6.00	6.75	6.95	4.42	5.19
Mean number of citations to EPO patents	1.12	0.83	1.41	0.77	1.42	0.89	1.41
Mean number of citations to EPO patents with "equivalent" US patents in the OECD triadic database	0.32	0.22	0.43	0.24	0.39	0.21	0.46

Figure 1: Six kinds of knowledge flows

This figure illustrates the six kinds of knowledge flows studied in this paper. Four of these are knowledge flows within the same country but across different firms and organizations: D→M, M→D, D→D and M→M (where "D" refers to "Domestic firm or organization" like Intel and IBM in the U.S., and "M" refers to "MNC subsidiary" like subsidiaries of foreign MNCs Sony and NEC in the US). The remaining two knowledge flows are those within the same MNC but across different countries: S→H and H→S (where "S" refers to "Subsidiary" and "H" refers to the "Home Base" of an MNC). The reference category (i.e., knowledge flows not included in any of these six groups) is the cross-border inter-organizational knowledge flows.



Appendix

A Note on Choice-Based Sampling and WESML

In samples where the fraction of y=1 observations (the "ones") is very small, the information content is much greater in the ones rather than the zeroes. To see this, recall that the asymptotic covariance matrix for the MLE for logit is given by (see Greene, 2003, p. 672)

$$\left[\sum_{i=1}^{n} \Lambda_i (1-\Lambda_i) x_i x_i^{'}\right]^{-1}$$

If the logit model has some explanatory power, Λ_i is larger (i.e. closer to 0.5 for rare events) when y_i =1. Thus $\Lambda_i(1-\Lambda_i)$ is larger, implying that having a higher fraction of 1's in the sample would reduce variance. Choice-based sampling tries to achieve this by over-sampling on the "ones" from the population. The sample is formed by taking a fraction α of the population's dyads with $y = 0$, and a fraction γ of the dyads with $y = 1$, where α is much smaller than γ. The probability of a citation *conditional on the dyad being in the sample* flows from Bayes' rule:

$$\Lambda_i^{'} = \frac{\gamma \Lambda_i}{\gamma \Lambda_i + \alpha(1-\Lambda_i)} = \frac{\gamma}{\gamma + \alpha e^{-\beta X_i}} = \frac{1}{1 + e^{-(\ln\left(\frac{\gamma}{\alpha}\right) + \beta X_i)}}$$

The extra term $ln(\gamma/\alpha)$ in the exponent leads to a bias. However, since the functional form is still logistic, a simple estimation strategy is to simply subtract $ln(\gamma/\alpha)$ from the estimate for the constant term of the usual logit. The efficiency of the correction, however, depends crucially on the logit functional form not being misspecified (Manski and Lerman, 1977; Cosslet, 1981). An alternate method, which is not as sensitive to model misspecification, is the *weighted exogenous sampling maximum likelihood* (WESML) estimator suggested by Manski and Lerman (1977). The WESML estimator is obtained by maximizing the following weighted "pseudo-likelihood" function:

$$\ln L_w = \frac{1}{\gamma} \sum_{\{y_i=1\}} \ln(\Lambda_i) + \frac{1}{\alpha} \sum_{\{y_i=0\}} \ln(1-\Lambda_i) = -\sum_{i=1}^{n} w_i \ln(1 + e^{(1-2y_i)x_i\beta})$$

where $w_i = (1/\gamma)\, y_i + (1/\alpha)(1-y_i)$.

In other words, each sample observation is weighted by the number of elements it represents from the overall population in order to make the choice-based sample "simulate" a random exogenous sample. Here is some intuition on why WESML works: Let the joint probability density be g(*x*,*y*) for the sample, and

g*(*x*,*y*) for the population. Let the fraction of elements with $y = j$ be f(*j*) in the sample, and f*(*j*) in the population (j = 0,1). Let n and N be sample size and population size respectively, and n_j and N_j be the number with $y = j$. Using conditional probability rules,

$$g(x,j) = \Pr(x \mid y = j) f(j) = \frac{g^*(x,j) f(j)}{f^*(j)} = \frac{g^*(x,j)(n_j / n)}{N_j / N} = \frac{N/n}{w(j)} g^*(x,j)$$

where $w(j) = N_j/n_j$ is the reciprocal of the sampling rate for observations with $y = j$. Let P(y_i) be the probability of $y = y_i$ conditional on $x = x_i$ in the population. Then, the expected value of the weighted likelihood function is Thus, ignoring the constant scaling factor N/n, the expected value of the weighted

$$E \ln L_w = \int \left(\sum_{i=1}^{n} w(y_i)[\ln P(y_i)] \right) g(x, y_i) dx = \sum_{i=1}^{n} \left(\int w(y_i)[\ln P(y_i)] \frac{N/n}{w(y_i)} g^*(x, y_i) dx \right) = \frac{N}{n} \int \left(\sum_{i=1}^{n} [\ln P(y_i)] \right) g^*(x, y_i) dx$$

log likelihood equals the expected log likelihood for the same sample resulting through random exogenous sampling from the population. As shown formally in Amemiya (1985, section 9.5.2), this ensures consistency of WESML estimation.

The choice-based WESML procedure described above can be extended to allow "matched samples". This involves taking all actual citations (y=1) and matching each of these with k "control citations" (y=0) along a dimension z (e.g., the "cells" indexed by the vector combination of the citing technological class and cited technological class). Without loss generality, denote the values z can take as 1, 2, …, T. For a matching-based sampling design, it is easier to think of not just y but (z, y) as the dependent variable. In forming the likelihood function, I will use the result that

$$\Pr(z = z_i \text{ and } y = j \mid x = x_i) = \Pr(z = z_i \mid x_i) \Pr(y = j \mid z = z_i \text{ and } x = x_i)$$

$$= \Pr(z = z_i \mid x_i) \Pr(y = j \mid x = x_i)$$

The second equality assumes that the vector **x** includes all information about z that affects citation outcome y, i.e., **x** is a sufficient statistic for z. The log-likelihood function for estimation using an exogenous random sample of size n would therefore be

$$\ln L = \sum_{i=1}^{n} \ln\left[\Pr(z = z_i \text{ and } y = y_i \mid x_i)\right]$$

$$= \sum_{i=1}^{n} \left\{ y_i \ln\left[\Pr(z = z_i \mid x_i)\Lambda(x_i\beta)\right] + (1 - y_i) \ln\left[\Pr(z = z_i \mid x_i)\left(1 - \Lambda(x_i\beta)\right)\right] \right\}$$

This forms the basis for deriving the pseudo-likelihood function for choice-based sampling. Each log likelihood function term has to be weighted by the inverse of the probability that the corresponding population

element will be included in the sample. To derive these weights, denote the number of elements with $z = t$ and y=j as n_{tj} for the sample and N_{tj} for the population. Matching ensures that, from each cell, I pick all elements with y=1 and k times as many elements with y=0. In other words, $n_{t1} = N_{t1}$ and $n_{t0} = kN_{t1}$. Also, since N_{tj} is known, the probability p_{tj} of a population element with z = t and y = j getting selected in our sample is easily calculated as $p_{t1}= n_{t1}/N_{t1}=1$ and $p_{t0}= n_{t0}/N_{t0} = kN_{t1}/N_{t0}$ for all values of t. Denoting $w_{tj} = 1/p_{tj}$, the weighted likelihood function for choice-based sampling is the given by

$$\ln L_w = \sum_{i=1}^{n} \left\{ y_i w_{z_i 1} \ln\left[\Pr(z = z_i \mid x_i)\Lambda(x_i\beta)\right] + (1 - y_i) w_{z_i 0} \ln\left[\Pr(z = z_i \mid x_i)\left(1 - \Lambda(x_i\beta)\right)\right] \right\}$$

$$= C - \sum_{i=1}^{n} w_i \ln\left(1 + e^{(1-2y_i)X\beta}\right)$$

$$\text{where} \quad w_i = y_i w_{z_i 1} + (1 - y_i) w_{z_i 0} \quad \text{and} \quad C = \sum_{i=1}^{n} w_i \ln\left[\Pr(z = z_i \mid x_i)\right]$$

Since C is independent of β, it can be ignored in the maximum likelihood procedure. Thus, a weighted logit estimation can be used, where the weights of the observations are now given by w_i. Unlike the simple WESML with random sampling from the y=0 observations, the weights now depend not just on the value of y but also on the cell that the observations falls into.