Appendix A: Data

US Patents. Patent data come from PatentsView, a database born out of a collaboration between American universities and the US government, and spearheaded by the USPTO. It was created in 2012 by the USPTO, the US Department of Agriculture, the American Institutes for Research, New York University, the University of California at Berkeley, and two data firms.

The PatentsView database is derived from the raw text and XML patent files held by USPTO.[1] The data contain unique inventor identifiers, geocoordinates of their location of residence at the time of applying for a patent, the technological classes of the patents, the name of the original assignees at the time of granting of the patent, their foreign status (US or non-US), and their citations to other patents.[2] We use the universe of patents granted between 1975 and 2012. This covers 6 million patents, granted to 3.6 million distinct inventors. The patent data are arranged at the patent-inventor level (one, unique patent-inventor combination per row). We keep all recorded institutional assignees in the data.[3] There are up to thirteen assignees for a single patent.[4]

Two features of the PatentsView database are essential for our purpose: (1) assignee, inventor, and location names are algorithmically disambiguated; and (2) its wide time coverage

[1] The raw patent files are publicly available at developer.uspto.gov/data or patents.reedtech.com.

[2] Additional information on inventors' gender, patents' summary text, technology classification codes such as WIPO, CPC, or IPCR, descriptions of tables and figures, patents' examiners, and the lawyers in charge of applications are made available by PatentsView. We do not use this data in the article.

[3] Most studies using patent data only use the first assignee. However, we use this information to identify collaborations between firms. Moreover, we would underestimate the total number of patents filed for foreign firms if only the first assignees were used to determine foreign R&D activities.

[4] Among the 6 million patents, only 3 have 13 assignees. Two of them—7100279 and 7780143—are assigned to 13 distinct Japanese firms. The other—US8965747—is assigned to 12 Chinese firms and Northwestern University, which is the only institutional assignee on this patent.

enables the study of technology diffusion over nearly four decades (1975–2012).[5] We provide an overview of the disambiguation procedures performed by the PatentsView team below.

Disambiguation. PatentsView uses probabilistic methods to determine whether inventors with the same name on different patents are the same person. Assignees are probabilistically disambiguated in a similar way. Addresses are disambiguated using geographic APIs.

Assignees. First, minor typos and misspellings are removed from company names using a probabilistic string-matching algorithm. Second, the main disambiguation relies on string-matching algorithms (see PatentsView website for details).

Locations. Locations are disambiguated by querying MaxMind and Google's Geocoding API, regarded as the industry standard to convert typical addresses (such as "Houghton Street, London") into geographic coordinates (such as "latitude: 51.5136, longitude: -0.1169").

Inventors. Inventor names are arguably the hardest field of the data to disambiguate. They are disambiguated via Discriminative Hierarchical Coreference, a machine learning technique that groups different spellings of a same author together via hierarchical trees, using information on the set of co-inventors of authors, the companies they patent for, their locations of residence, and the title of their patents. Details can be found on the PatentsView website. The patents are classified into 259,465 Cooperative Patent Classification (CPC) subgroups that are mapped to thirty-seven technological subcategories and six broad technological categories (Hall, Jaffe, and Trajtenberg 2001): the six broad categories are Chemicals, Computers & Communications, Drugs & Medical, Electrical & Electronic, Mechanical, and Others. The data can be accessed and downloaded freely from the PatentsView website. Because several new

[5] This feature makes it more attractive than the commonly used NBER patent citation data (Hall, Jaffe, and Trajtenberg 2001).

patent classes have been introduced since Hall, Jaffe, and Trajtenberg (2001)publication, we manually assign seventy-four of these new classes to subcategories. As a result, only 2.1 percent of patents do not have a subcategory, 92 percent of which lack patent classes in the original data. We use a patent's primary technological subcategory to determine a region-technology cell's technology.

As the disambiguation of assignees, locations, and inventors is an ongoing effort, and as the disambiguation methods are continuously refined, the same patent data sets downloaded at different times are likely to have slightly different identifiers for firms, places and people.

Subnational data. Next, we assign the patent-inventor couples to 1,549 regions. We have data on GDP per capita (at the national and the regional level), average years of education and population for 1,456 of these regions. The data are described and used in Gennaioli et al. (2014), and they were available discontinuously from 1960 to 2010 for most regions. Typically, two data points in a region would be five years apart. We linearly interpolate missing values between any two available values, so as to have data for each year. The regions for which we have observables cover 97.2 percent of the USPTO patents. Based on the disambiguated locations of inventors, we assign each latitude-longitude pair associated with the location of residence of inventors to a regional polygon via a spatial join algorithm.[6]

The map below (Figure A.1), taken from Gennaioli et al. (2014), shows the geographic coverage of the regional data. We are indebted to Nicola Gennaioli, Rafael La Porta, Florencio Lopez De Silanes, and Andrei Shleifer for sharing their data and shapefile.

[6] geoinpoly (Picard 2015).

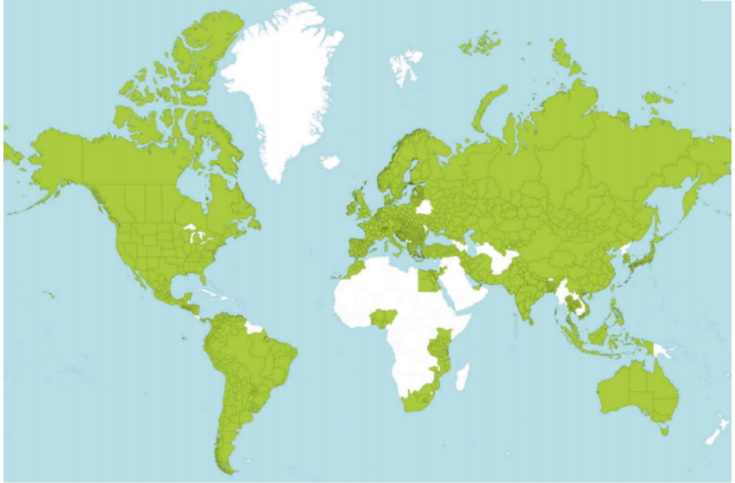

Figure A.1. Coverage of regional data.

Match rate of treatment patents to ORBIS. In the main text, we argue that we achieve a matching rate between our treatment patents and ORBIS records of the original assignees that should be considered to be (very) high. Here we further substantiate this claim.

1. *Attrition in treatment branch plants*

   ORBIS only contains records for firms that existed at the time that the data were retrieved, that is, in 2017. We therefore expect that our match rate will be higher for more recent treatments than for less recent ones. Moreover, we expect that branch plants of technology leaders have a higher survival rate. Both effects are visible in the figure below (Figure A:2), which shows the ORBIS match rate for top 5 percent and bottom 95 percent treatment patents.

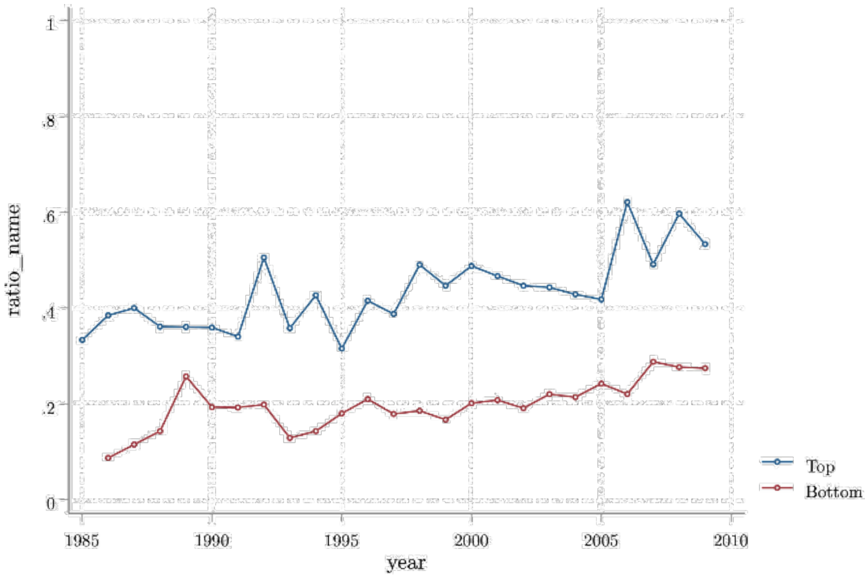

Figure A.2. Attrition in ORBIS match.

The attrition rate in early years is over 60 percent for treatments of technology leaders against about 90 percent for lower ranking MNEs.[7] However, attrition rates more or less halve towards the end of the period.

2. *Overall coverage of branch plants in ORBIS*

The ORBIS data set itself has far from perfect coverage. To get a sense of the problem, we compared ORBIS to data from Dun and Bradstreet (D&B). D&B contains over 100 million establishments, for which it lists locations and parent companies. We matched the treatment firms in ORBIS to D&B firms using fuzzy string matching and then compared the number of establishments each data set lists in any country in the world. The graph below (Figure A.3) shows the results of this comparison.

[7] Note that to create this figure, we matched not just on patent ID but also on firm name. As a consequence, changes in company names or mergers and acquisitions would lead to failed matches. This explains the discrepancy with the 61 percent match rate mentioned in the main text, which only merges on location and patent ID.

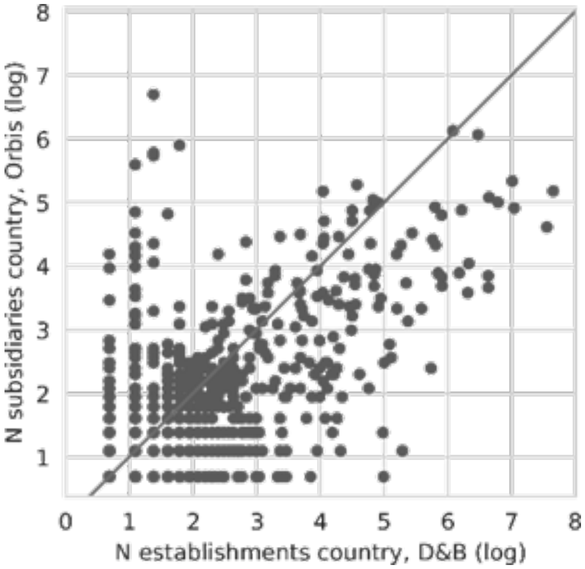

Figure A.3. Comparison of coverage ORBIS/D&B.

Although D&B has far from complete coverage, it lists about twice as many establishments as ORBIS (21.6 per country-firm cell against 10.3 in ORBIS) for our treatment firms. This means that at least 50 percent of the establishments (and therefore activities in foreign locations) are not covered by ORBIS.

Taken together, this suggest that a 61 percent match rate of treatments to ORBIS is very high. For these 61 percent, we were able to find concrete evidence of foreign establishments in the treatment locations. The fact that for 39 percent of treatments, we cannot identify a concrete investment in the region is likely due to the documented attrition rates, lack of coverage, and risk that patents have been sold on to new owners. Overall, our findings suggest that the bulk of treatments in the main text involve FDI and the opening of establishments abroad.

Appendix B: Technology Leaders by Technology Class

Tables B.1 to B.6 show the ten most innovative firms in the period 1975–1985 in terms of patenting output by technological category. Technology categories follow the NBER classification developed by Hall, Jaffe, and Trajtenberg(2001). The tables show how many patents a firm produced in the technology category over the 1975–1985 period, both in levels and as a share of the global total, as well as how many foreign investments the firm made in the period 1985–2002. A foreign investment is detected when a firm files a patent with inventors who reside abroad. From 1985 to 2007, there are 1,385 foreign investments in Chemicals, 1,768 in Computers & Communications, 1,136 in Drugs & Medical, 1,851 in Electrical & Electronics, 1,377 in Mechanical, and 1,544 in Other technologies (mainly consisting of low-tech classes such as Apparel & Textile, Heating Technology, Furniture & House Fixtures, and Agriculture & Food).

Table B.1

Chemical

Rank	Company	Patent Count	Share of Total	Foreign Investments	Share of Total
1	Bayer Aktiengesellschaft	3,496	2.32%	20	1.44%
2	Ciba-Geigy Corporation	2,672	1.78%	13	0.94%
3	E. I. Du Pont de Nemours and Company	2,558	1.70%	13	0.94%
4	The Dow Chemical Company	2,521	1.68%	12	0.87%
5	General Electric Company	2,256	1.50%	24	1.73%
6	BASF Aktiengesellschaft	2,137	1.42%	17	1.23%
7	Hoechst Aktiengesellschaft	2,135	1.42%	5	0.36%
8	Phillips Petroleum Company	2,089	1.39%	2	0.14%
9	Exxon Research & Engineering Co.	2,000	1.33%	3	0.22%
10	Mobil Oil Corporation	1,796	1.19%	1	0.07%

Table B.2

Computers and Communications

Rank	Company	Patent Count	Share of Total	Foreign Investments	Share of Total
1	International Business Machines	2,292	4.30%	137	7.75%
2	Canon Kabushiki Kaisha	1,539	2.89%	7	0.40%
3	Hitachi, Ltd.	1,390	2.61%	13	0.74%
4	U.S. Philips Corporation	1,223	2.29%	40	2.26%
5	Siemens Aktiengesellschaft	1,163	2.18%	26	1.47%
6	Bell Telephone Laboratories, Incorporated	1,119	2.10%	0	0.00%
7	RCA Corporation	975	1.83%	0	0.00%
8	Motorola, Inc.	913	1.71%	27	1.53%
9	Texas Instruments Incorporated	751	1.41%	36	2.04%
10	Sony Corporation	694	1.30%	0	0.00%

Table B.3

Drugs and Medical

Ran k	Company	Patent Count	Share of Total	Foreign Investments	Share of Total
1	Merck & Co., Inc.	878	2.38%	4	0.35%
2	Bayer Aktiengesellschaft	838	2.27%	5	0.44%
3	Ciba-Geigy Corporation	617	1.67%	3	0.26%
4	Eli Lilly and Company	497	1.35%	1	0.09%
5	American Cyanamid Company	429	1.16%	3	0.26%
6	Pfizer Inc.	423	1.14%	0	0.00%
7	E. R. Squibb & Sons, Inc.	402	1.09%	2	0.18%
8	Beecham Group Limited	388	1.05%	0	0.00%
9	The Upjohn Company	343	0.93%	0	0.00%
10	Hoffmann-La Roche Inc.	332	0.90%	1	0.09%

Table B.4

Electrical and Electronic

Rank	Company	Patent Count	Share of Total	Foreign Investments	Share of Total
1	General Electric Company	3,575	3.52%	46	2.49%
2	RCA Corporation	2,864	2.82%	0	0.00%
3	Westinghouse Electric Corp.	2,571	2.53%	5	0.27%
4	Hitachi, Ltd.	2,366	2.33%	11	0.59%
5	U.S. Philips Corporation	2,342	2.30%	40	2.16%
6	Siemens Aktiengesellschaft	2,335	2.30%	37	2.00%
7	International Business Machines	1,575	1.55%	38	2.05%
8	Tokyo Shibaura Denki Kabushiki Kaisha	1,173	1.15%	0	0.00%
9	Xerox Corporation	1,160	1.14%	8	0.43%
10	Motorola, Inc.	1,093	1.08%	38	2.05%

Table B.5

Mechanical

Rank	Company	Patent Count	Share of Total	Foreign Investments	Share of Total
1	General Motors Corporation	1,699	1.31%	3	0.22%
2	Nissan Motor Co., Ltd.	1,696	1.31%	0	0.00%
3	Robert Bosch GmbH	1,461	1.13%	26	1.89%
4	Caterpillar Tractor Co.	1,237	0.96%	0	0.00%
5	General Electric Company	1,187	0.92%	15	1.09%
6	Toyota Jidosha Kogyo Kabushiki Kaisha	1,173	0.91%	0	0.00%
7	Honda Giken Kogyo Kabushiki Kaisha	1,153	0.89%	0	0.00%
8	Hitachi, Ltd.	900	0.70%	4	0.29%
9	Westinghouse Electric Corp.	771	0.60%	0	0.00%
10	Canon Kabushiki Kaisha	755	0.58%	3	0.22%

Table B.6

Other Technologies

Rank	Company	Patent Count	Share of Total	Foreign Investments	Share of Total
1	General Electric Company	969	0.88%	13	0.84%
2	The Singer Company	578	0.53%	4	0.26%
3	Minnesota Mining and Manufacturing	524	0.48%	7	0.45%
4	Mobil Oil Corporation	496	0.45%	3	0.19%
5	Phillips Petroleum Company	478	0.44%	1	0.06%
6	E. I. Du Pont de Nemours and Company	443	0.40%	5	0.32%
7	Caterpillar Tractor Co.	435	0.40%	0	0.00%
8	General Motors Corporation	431	0.39%	3	0.19%
9	Nippon Gakki Seizo Kabushiki Kaisha	418	0.38%	0	0.00%
10	Halliburton Company	406	0.37%	3	0.19%

Appendix C: Citation Analysis

In this appendix, we provide details about the matching algorithm used in the patent citation analysis. We ask if patents in a treated region-technology cell cite the treatment patent more often than similar patents elsewhere do.[8] If knowledge spillovers from treatment patents to other firms in the region exist, we would expect that patents in treated cells cite these treatment patents more often than otherwise similar patents, but outside treated cells do. Sincecitations added by patent examiners do not reflect technological spillovers, we ignore all such citations in this analysis.

First, we collect all patents in treated region-technology cells filed after the treatment patent. Next, we match these patents to similar patents outside the treated cell, using a mix of propensity score and exact matching. First, we match exactly on patent class (using the 1,112 USPC main classes, assigned at the time that the patent was granted) and macroregion. Next, we refine these matches using propensity score matching on citation counts, number of inventors, and year of application of the patent. To ensure that matches are sufficiently similar, we impose a 0.0002 caliper. Finally, we prohibit certain matches. First, matched patents may not come from the same company, inventor, or country as the treatment patent(s). This prevents that citation patterns reflect national, cultural, or linguistic preferences, or firm- or inventor-specific knowledge. Second, we exclude matched patents filed over twenty years after the treatment patent(s). Figure C.1 summarizes the process.

We repeat this procedure, once for patents assigned to domestic firms and once for patents assigned to foreign firms in treated regions. This yields two matched samples. Next, we

[8] Whenever there are multiple patents associated with a treatment—which happens in about a third of our treatments—we consider all citations to any of these treatment patents.

split these samples into two parts: the first contains patents in cells treated by technology leaders (T5 firms) and their statistical twins. The second contains patents and matched counterparts in cells treated by lower-ranking firms, which we define as firms in the bottom ninety-five percentiles of the cumulative patenting distribution between 1975 and 1985 (B95 firms). Our quantity of interest is the following citation ratio (referred to as T/C ratio in Table 6):

$$r_{is} = \frac{c_{r\theta s}(\,F^{i}_{r\theta} = 1)}{c_{r'\theta' s}(\,F^{i}_{r'\theta'} = 0)}$$

where i is the rank of the treatment firm(s) (either T5 or B80) and s the status of the citing patents (domestic, foreign or all patents). $c_{r\theta s}(\,F^{i}_{r\theta} = 1)$ represents the number of citations $c_{r\theta s}$ from patents in treated cells ($F^{i}_{r\theta} = 1$) to the treatment patent(s). The denominator counts citations from matched patents. The larger this ratio, the greater the local spillovers are compared to the baseline scenario captured by the control patents. To determine which foreign firms generate the largest knowledge spillovers in treated regions, we compare the citation ratios in T5-treated (r_{T5s}) to those in B95-treated cells (r_{B95s}).

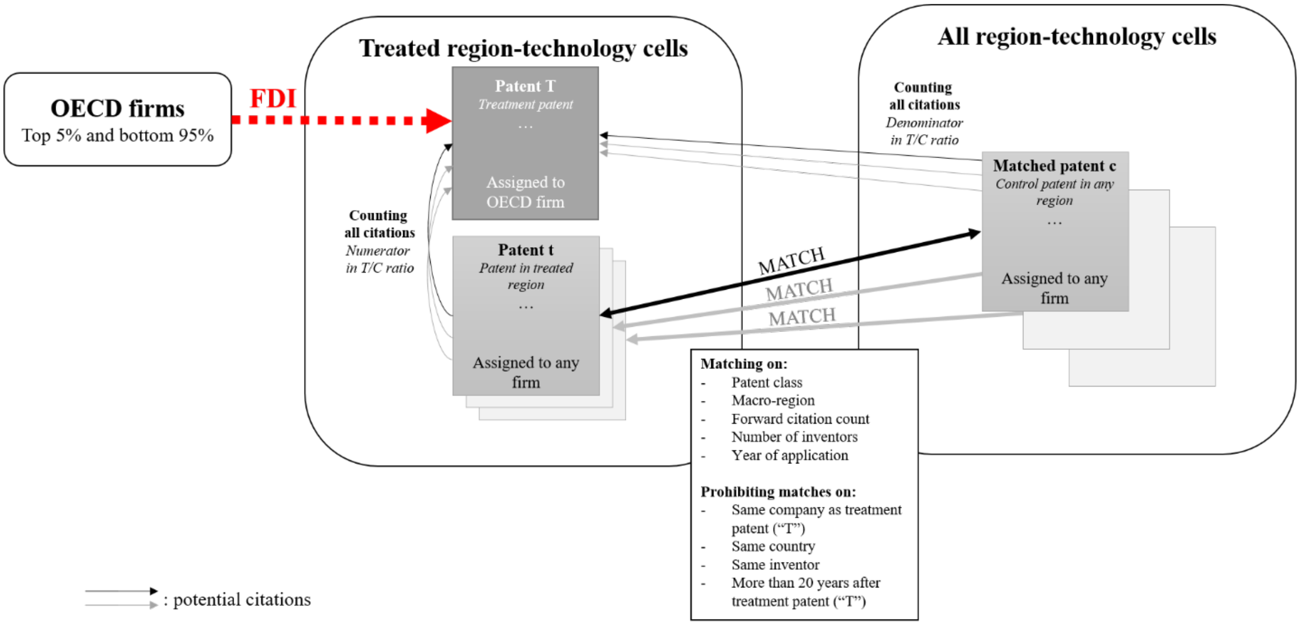

Figure C.1. Citation analysis.

References

Gennaioli, N., La Porta, R., De Silanes, F. L., and Shleifer, A. 2014. Growth in regions. *Journal of Economic Growth* 19 (3): 259–309.

Hall, B. H., Jaffe, A. B., and Trajtenberg, M. 2001. The NBER patent citation data file: Lessons, insights and methodological tools (No. w8498). Cambridge, MA: National Bureau of Economic Research.

Picard, R. 2015. GEOINPOLY: Stata module to match geographic locations to shapefile polygons, Statistical Software Components S458016, Boston College Department of Economics, revised 16 Aug 2015.