B Data Appendix

In this Appendix, I provide more details on the data sources, refinements of the data, and some details on the estimation techniques associated with the empirical work in Chapters 5 and 8. The entire dataset and Stata program codes are available at http://scholar.harvard.edu/antras/books. I often refer to this url as the 'book's website'.

B.1 Raw U.S. Import and Export Data

The basis for the empirical work conducted in this book is the *U.S. Related-Party Trade* database collected by the U.S. Bureau of Customs and Border Protection and managed by the U.S. Census Bureau. This dataset can be downloaded from the following U.S. census website: http://sasweb.ssd.census.gov/relatedparty/. The data are available at different levels of industry aggregation, but the most disaggregated level available online is the six-digit North American Industry Classification System (NAICS). At the time of writing this Appendix, the data are available for the period 2002-2012. In the empirical work I instead used data for the period 2000-2011. Data for 2000 and 2001 are available as part of the entire dataset on the book's website.

Throughout the data construction, non-manufacturing sectors were dropped. In addition, industries that are not reported in the Related-Party Trade data were also dropped. In the U.S. Related-Party Trade data, five manufacturing industries are reported at the 5-digit NAICS: 31131X (Alumina and Aluminum Production and Processing), 31181X (Bread and Bakery Product Manufacturing), 31511X (Hosiery and Sock Mills), 33631X (Motor Vehicle Gasoline Engine and Engine Parts Manufacturing), and 33641X (Aerospace Product and Parts Manufacturing). This last industry 33641X is somewhat different from the rest. First, it is the only five-digit industry with zero import flows. Furthermore, it is the only one of these industries for which six-digit industries with the same initial five-digits (in particular, 336411, 336412, 336413, 336414, 336415, and 336419) are reported in the dataset. For these reasons, this five-digit sector 33641X was dropped from the dataset. All other NAICS-level industry variables described below were also constructed for the four surviving synthetic five-digit industries.

The U.S. Related-Party Trade database reports imports and exports associated with each of 233 foreign countries. The data for South Sudan are, however, only available for 2011, so this country was dropped from the sample. The raw data were rectangularized by treating any missing values as zero import or export flows. Overall, we work with data for 390 industries, 232 countries, and 12 years, for a total of 1,085,760 observations.

The data define related-party import transactions involving parties "with various types of relationships including any person directly or indirectly, owning, controlling or holding power to vote, 6 percent of the outstanding voting stock or shares of any organization." On the other hand, a related-party export transaction is one "between a U.S. exporter and a foreign consignee, where either party owns, directly or indirectly, 10 percent or more of the other party." Although these ownership requirements are very low, I argued in Chapter 1 that BEA data suggest that intrafirm trade is generally associated with one of the entities having a controlling stake in the other entity. The dataset also contains data on non-related imports and exports, which involve parties that "have no affiliation with each other or who do not meet the relevant equity requirements" for related-party trade. Although in principle an indicator of whether or not a transaction involves related parties is required for *all* import or export transactions recorded by the U.S. Bureau of Customs and Border Protection, in practice that information is missing in some cases. The dataset labels those volumes of trade as "not reported."

Table B.1 presents descriptive statistics for the key variables in the U.S. Related-Party Trade database. A few features of the table are worth highlighting. First, related-party imports account for 51.6% of overall U.S. manufacturing imports over 2000-11, and for 31.8% of overall U.S. manufacturing exports. Second, the average (unweighted) related-party import and export shares are, however, much lower (23.8% for imports and 10.4% for exports). Third, the large number of zeros in the data implies that the number of observations with a well-defined intrafirm import share is less than one third the total number of observations (i.e., country and six-digit NAICS combinations in the data). Conversely, on the export side, more than fifty percent of the observations feature positive exports and well-defined intrafirm export shares. Fourth, non-reported import transactions account for a negligible 0.04% of U.S. imports (that percentage goes up to 3.71% for exports, but I do not employ that export share in the empirical analyses in the book).

Table B.1. Some Descriptive Statistics from the NAICS Related-Party Trade Database

Variable (in $ except shares)	Mean	Std. Dev	Min	Max	N
Total Imports	14,712,628	282,669,054	0	44,917,394,621	1,085,760
a) Related-Party	7,587,177	220,242,612	0	44,134,184,241	1,085,760
b) Non-related-Party	7,119,158	112,745,454	0	20,981,735,046	1,085,760
c) Not Reported	6,294	567,136	0	269,396,613	1,085,760
Related-Party Share a/(a+b)	0.2380	0.3265	0	1	312,884
Total Exports	8,594,996	113,419,707	0	19,996,871,796	1,085,760
Related-Party	2,736,289	61,075,666	0	13,174,432,899	1,085,760
Non-related-Party	5,539,536	67,292,534	0	14,757,989,972	1,085,760
Not Reported	319,172	7,736,500	0	1,574,623,834	1,085,760
Related-Party Share a/(a+b)	0.1039	0.2150	0	1	565,145

The U.S. Related-Party Trade database is also available at the finer six-digit Harmonized System (HS) industrial disaggregation. This dataset is not publicly available but can be purchased from the U.S. Census. Although I have not used it in the tests performed in this book, I have made it available for download at http://scholar.harvard.edu/antras/books. Over the period 2000-11, this more detailed dataset contains information on U.S. imports (related, non-related and non-reported) for 5,705 products and for 238 countries and territories. Despite the fact that this finer disaggregation generates more than 16 million potential observations on U.S. imports, only for about 10 percent of these cases (1,572,949 to be precise) are U.S. imports positive. This in turn leads to an overall number of 1,568,711 intrafirm trade shares in the data, exceeding by a factor of five the 312,884 available shares when using the publicly available NAICS dataset.[27] Still, the mean and variance of intrafirm trade shares are very similar to those reported in Table B.1 for the six-digit NAICS data.

B.2 U.S. Import and Export Data at IO2002 Level

In many of the empirical tests presented in this book, offshoring is correlated with variables that, because of their nature and characteristics, can only be computed with Input-Output data (more on this below). For this reason, the natural industry classification to work with in those cases is the I-O commodity code classification. More specifically, I use the 2002 Input-Output industrial classification, or IO2002 for short.

[27] For 4,238 observations, the data indicate positive imports that are entirely recorded as "Non-Reported". Because I define the share of intrafirm trade as related-party imports divided by the sum of related-party imports and non-related party imports, I cannot compute a well-defined intrafirm trade share in those cases.

Following Antràs and Chor (2013), the raw data in NAICS industry codes were mapped to six-digit IO2002 industries using a correspondence provided by the Bureau of Economic Analysis (BEA) as a supplement to the 2002 U.S. Input-Output Tables.[28] This concordance is a straightforward many-to-one mapping for the manufacturing industries (NAICS first digit = 3). Two industries required a separate treatment, as the NAICS data were at a coarser level of aggregation than could be mapped into six-digit IO2002 codes. A synthetic code 31131X was created to merge IO 311313 (Beet sugar manufacturing) and 31131A (Sugar cane mills and refining), while a separate code 33641X merged IO 336411, 336412, 336413, 336414, 33641A (all related to the manufacture of aircraft and related components). This approach is somewhat distinct from the one I used to handle these aircraft subsectors in the NAICS dataset, where I simply dropped 33641X rather than merging it into a single category with all other five digit sectors starting with 33641. Because the NAICS sector 33641X features no U.S. imports, this small divergence should have little impact on the results. All other industry variables described below were also constructed for these two synthetic IO2002 codes 31131X and 33641X.

Overall, the IO2002 identifies U.S. imports and exports (related, non-related, and non-reported) for 253 sectors, 232 countries and 12 years of data, for a total of 704,352 observations, of which again many are zeroes. Table B.2 provides some basic descriptive statistics from this related-party IO2002 trade data.

Table B.2. Some Descriptive Statistics from the IO2002 Related-Party Trade Database

Variable (in $ except shares)	Mean	Std. Dev	Min	Max	N
Total Imports	22,679,545	376,502,066	0	44,917,395,456	704,352
a) Related-Party	11,695,648	281,267,307	0	44,134,184,241	704,352
b) Non-related-Party	10,974,196	166,941,030	0	20,981,735,046	704,352
c) Not Reported	9,702	749,284	0	269,396,613	704,352
Related-Party Share a/(a+b)	0.2438	0.3265	0	1	227,829
Total Exports	13,549,842	175,761,983	0	27,862,790,144	704,352
Related-Party	4,259,789	86,715,148	0	13,174,432,899	704,352
Non-related-party	8,797,599	103,436,155	0	15,297,237,562	704,352
Not Reported	492,455	12,228,037	0	2,585,156,012	704,352
Related-Party Share a/(a+b)	0.1068	0.2121	0	1	416,933

[28] See, for instance, http://www.bea.gov/industry/xls/2002DetailedItemOutput.xls.

B.3 Isolating the Intermediate Input Component of U.S. Imports and Exports

In this section, I provide more details on the Wright (2014) methodology for isolating the intermediate input component of trade flows. The key input for this data correction is a list of End-Use industrial categories available from the U.S. Bureau of Economic Analysis. The BEA uses these end-use codes to allocate goods to their final use, within the National Income and Product Accounts. Importantly, U.S. imports and exports at the ten-digit Harmonized System level are similarly allocated to end-use codes. Foreign Trade Statistics distinguish six one-digit end-use categories: (0) Foods, feeds, and beverages; (1) Industrial supplies and materials; (2) Capital goods, except automobiles; (3) Automotive vehicles, parts and engines; (4) Consumer goods (nonfood), except auto; and (5) Other merchandise. Apart from these six principal end-use categories, the classification is further subdivided into about 140 broad commodity groupings. Wright (2014) advocates dropping all products with an end-use code equal to 0, 4 or 5, as well as a subset of the commodity groupings in the other three end-codes. The full list of dropped BEA commodity groupings can be found in Table 7 of Wright (2014).

The practical implementation of this correction consists of four steps:

Step 1. We begin by mapping each ten-digit HS product to a BEA end-use code. In order to maximize such a mapping, we put together multiple years of concordance tables published by the Census Bureau. The Census website provides tables for recent years from 2008 to 2013. We also downloaded older tables for years from 1993 to 1997 from Jon Haveman's trade data website (http://goo.gl/5pyijB). Technically, import and export HS codes are administered by different federal agencies (Export codes, known as Schedule B, is administered by the Census, and import codes, known as Harmonized Tariff System (HTS), is administered by the U.S. International Trade Commissions (USITC)). As such, a complete mapping requires putting together concordance tables for both import codes and export codes. We first map each HS product to end-use code, using the most recent concordance table published in 2013, and move to the previous year's table, if the mapping is still incomplete.

Step 2. We then use detailed ten-digit HS U.S. import and export data by foreign country for 2000-11 available from Peter Schott's website at Yale University and drop all ten-digit flows consisting of finished goods, as dictated by the concordance constructed in step 1. See Schott (2008) for more details on the ten-digit U.S. import and export data.

Step 3. Next, we aggregate the ten-digit HS U.S. import and export data back to the IO2002 level using a concordance between ten-digit HS codes and

IO2002 codes available from the U.S. Bureau of Economic Analysis website at http://www.bea.gov/industry/xls/HSConcord.xls. We do so both for total flows as well as for only the intermediate input component of these flows (after dropping final goods). Comparing these two flows we obtain an IO2002-country-year specific 'discount factor' by which overall imports and exports need to be multiplied to obtain intermediate input imports and exports.

Step 4. We then apply these discount factors to the Related-Party IO2002 trade data constructed based on the NAICS Related-Party trade data. Note that the applied discount factor varies across IO2002 codes, countries and years, and is also distinct for imports and exports. When the discount factor is 0, this implies that particular IO2002-country-year observation does not contain any intermediate input flows. Whenever an IO2002 sector features zero aggregate imports of intermediate inputs in each year we treat that sector as a final-good sector and drop it from the sample in all Wright-adjusted regressions. We also drop IO2002 sectors 325411 ('Medicinal and botanical manufacturing') and 33299B ('Arms, ordinance, and accessories') because they feature extremely low input import flows, and only for 2000 and 2001, and their recorded offshoring shares for those years are negative. Overall we entirely drop 39 industries, which are listed in Table B.3.

Apart from these dropped sectors, many other industries feature small aggregate volumes of intermediate input flows, and many zero flows for imports from particular countries. Even though these observations are not dropped, their associated discount factor will be tiny or equal to zero (and in the latter case they will be dropped from log-linear specifications). Table B.3 presents basic descriptive statics comparing total and intermediate input U.S. imports and exports. The table indicates that, overall, intermediate input flows account for 53.1% of total U.S. imports and 67.8% of total U.S. exports. Although not reported in the table, one can also compute the share of intermediate inputs in related-party imports and exports. These turn out to be only marginally higher than for overall trade, equalling on aggregate for 54.4% of intrafirm imports and 68.4% of intrafirm exports.

Apart from this Wright (2014) correction, in some of the empirical tests in the book I follow Nunn and Trefler (2013*b*) in restricting the sample to the set of countries for which it is more plausible that U.S. intermediate input purchases are associated with U.S. headquarters purchasing inputs from abroad (rather than foreign headquarters exporting inputs to U.S. suppliers). The details of this correction appear in the main text of Chapter 5, so I will not repeat them here. Quantitatively, this correction removes five countries from the sample accounting for a mere 3.18% of U.S. imports. As mentioned in the main text, I have also experimented with a more extensive correction based on Nunn and Trefler (2013*b*) that drops 18 countries accounting for 32.52% of U.S. imports. These corrected flows are available from the files in the book's website.

Table B.3. IO2002 Sectors Excluded from the Sample by the Wright (2014) Correction

311111	Dog and cat food manufacturing	314110	Carpet and rug mills
311119	Other animal food manufacturing	315100	Apparel knitting mills
311210	Flour milling and malt manuf.	315230	Women's and girls' cut & sew apparel
311230	Breakfast cereal manufacturing	321991	Manufactured mobile home manuf.
31131X	Sugar Manufacturing	322291	Sanitary paper product manufacturing
311320	Chocolate & confectionery manuf	325411	Medicinal and botanical manufacturing
311340	Nonchocolate confectionery ma.	325620	Toilet preparation manufacturing
311410	Frozen food manufacturing	331314	Secondary smelting/alloying of alum.
311513	Cheese manufacturing	33299B	Arms, ordinance, and accessories
31151A	Fluid milk and butter manuf.	33461A	Software, audio, & video media reprod.
311520	Ice cream and frozen dessert ma.	335221	Household cooking appliances
311615	Poultry processing	335222	Household refrigerator & freezers
311820	Cookie, cracker, and pasta ma.	336213	Motor home manufacturing
311910	Snack food manufacturing	336991	Motorcycle, bicycle, & parts
311920	Coffee and tea manufacturing	336992	Military armored vehicle & tanks
311930	Flavoring syrup & concentrate ma.	337110	Wood kitchen cabinet & countertops
312110	Soft drink and ice manufacturing	337121	Upholstered household furniture
312120	Breweries	33721A	Office furniture manufacturing
312130	Wineries	337910	Mattress manufacturing
312140	Distilleries		

Table B.4. Descriptive Statistics Illustrating the Effects of the Wright (2014) Correction

Variable (in $)	Mean	Std. Dev	Min	Max	N
Total Imports	22,679,545	376,502,066	0	44,917,395,456	704,352
Wright-Adjusted Input Imports	12,042,991	216,171,921	0	44,206,174,208	704,352
Total Exports	13,549,842	175,761,983	0	27,862,790,144	704,352
Wright-Adjusted Input Exports	9,180,273	140,980,583	0	27,860,740,096	704,352

B.4 Computing Offshoring Shares

The construction of offshoring shares requires data not only on U.S. imports and exports, but also on U.S. domestic shipments. Ignoring the Wright adjustment for intermediate inputs, industry-level offshoring shares are simply computed as the ratio of U.S. imports to the sum of U.S. shipments plus U.S. imports minus U.S. exports in that sector. The analogous offshoring share for a given foreign country is computed as U.S. imports from that particular country divided by the same industry-level denominator, so the sum of country-industry-level offshoring shares corresponds to the aggregate industry-level offshoring share.

Data on U.S. shipments were obtained from the NBER-CES Manufacturing database for the period 2000-09 and from the Annual Survey of Manufacturing (ASM) for 2010 and 2011. Both of these data sources are available at the six-digit NAICS level, which facilitates their merging with the six-digit NAICS U.S. import and export data. The mapping between the trade data and the NBER-CES database is quite clean, but merging the ASM data required a series of small adjustments to deal with the fact that 88 industries are reported at the more aggregated five-digit NAICS level in the ASM dataset. In order to minimize the loss of industry categories associated with adding those two last years of data, we imputed shipment values for all six-digit sectors available in the trade and NBER data for which only the more aggregated five-digit industry was available in the ASM data. This imputation was based on breaking up the 2010 and 2011 ASM total values of shipments at the five-digit level into six-digit values based on the relative weights of the different six-digit segments in the NBER-CES data over the period 2005-09.

As mentioned in Chapter 5, for a small percentage of industries and years the recorded value of total shipments falls short of the value of U.S. exports. This is true for years prior to 2009 so it is not explained by the adjustments described in the previous paragraph. These observations are typically dropped in the empirical tests in this book.

So far, I have described the construction of offshoring shares for the NAICS case and without any adjustment for intermediate input trade. In order to compute offshoring shares at the IO2002 level, we simply repeated the steps above but using as the basis U.S. imports and exports at the IO2002 level, as well as U.S. shipments at the same industry classification. The latter series was obtained by filtering the constructed NAICS shipment series described above through the same BEA concordance table used to transition from NAICS to IO2002 trade flows. Wright-adjusted offshoring shares were computed in an analogous manner based on Wright-adjusted U.S. imports, exports and shipments. To isolate the intermediate input component of U.S. shipments, I applied a discount factor to overall shipments equal to the average of the import and export 'Wright' discount factors applied to trade flows in that industry over the period 2000-11. Again, in some cases, the resulting value of input shipments fell short of the value of input exports, and these observations are dropped from most regressions.

This concludes our discussion of the construction of the main dependent variables in the empirical tests in the book. I now turn to describing the explanatory variables used in those tests.

B.5 Industry-Country-Level Covariates

Freight Costs. Sectoral and exporter-specific measures of freight costs associated with U.S. imports were downloaded from Peter Schott's website (see Schott, 2010,

for further documentation). More specifically, freight costs are computed as the ratio of CIF imports to FOB imports for a given product and origin country for the period 2000-05. Although this variable varies year by year, to salvage observations for 2006 through 2011, we construct a time-invariant measure of freight costs equal to the average of the ratio of Cost Insurance and Freight (CIF) import volumes to Free On Board (FOB) import values for a given country and product over the period 2000-05. We then assign this average measure to all 12 observations associated with a given exporting country and sector. The data are originally available at the six-digit NAICS level but we also constructed them at the IO2002 level using the same BEA concordance table employed throughout the book.

U.S. Tariffs. U.S. tariffs corresponds to U.S. applied tariffs from the World Integrated Trade Solution (WITS) database maintained by the World Bank. We again construct a measure at the exporter-sector level based on the average of this variable over multiple years, and in this case we do so for 2000-10 given data availability. The data are originally available at the six-digit HS level. We used the concordance in Pierce and Schott (2009) to transition from six-digit HS codes to six-digit NAICS codes. To construct IO2002 tariff levels, we used the same BEA concordance from HS6 to IO2002 employed when applying the Wright intermediate input correction, as described in step 3 of section B.3 above.

B.6 Industry-Level Covariates

Trade Costs. Industry-level freight costs and tariffs were computed based on the industry-country series we have just described in section B.5 but averaging them over all exporting countries. In both the NAICS and IO2002 cases, data on freight costs and tariffs were missing for some industry codes. For those sectors, we imputed a value equal to the weighted freight costs and tariffs of the sectors with which the industry shared the same first four digits, or (if the value was still missing) the same first three digits, using industry shipment values as weights.

R&D Intensity. We build on Nunn and Trefler (2013*b*), who calculated R&D expenditures to total sales on an annual basis for the period 1998-2006 using the U.S. firms in the Bureau van Djik's Orbis dataset. Their original data are reported for IO1997 industries. To obtain IO2002 values, we follow Antràs and Chor (2013) and construct a crosswalk from IO1997 to IO2002 through the NAICS industry codes. More specifically, the R&D intensity for each IO2002 industry was calculated as the weighted average value of $\log(0.001 + R\&D/Sales)$ over that of its constituent IO1997 industries over the years 2000-2005, using the industry output values in the 1997 U.S. I-O Tables as weights. There remained 13 IO2002 industries without R&D intensity values after the above procedure. A similar procedure to that described above for trade costs was used to obtain the R&D intensity for the remaining 13 IO2002 codes (based on the R&D intensity of the IO2002 codes with which the industry shared the same first four or three digits).

This yielded a complete series for R&D intensity of the 'selling' sector. In many specifications we instead use a measure of the R&D intensity of the 'average buyer' of an industrial good. This buyer version of the variable was computed as a weighted average of the R&D intensity of the industries that purchase the good in question (call it good v), with weights equal to these buying sectors' input purchase values of good v as reported in the 2002 U.S. I-O Tables. The construction of R&D intensity at the six-digit NAICS level was analogous, although in imputing values for missing observations we also made use of four-digit and three-digit measures of R&D intensity kindly provided by Heiwai Tang. Late in the production of this book, Davin Chor alerted me to the fact that, in the IO2002 dataset, there is one industry (IO334411, "Electron Tube Manufacturing"), with a huge ratio of R&D expenditures over sales. This ratio is equal to 660 and is a clear outlier relative to other sectors. It should thus be treated with caution. Fortunately, all the results presented in this book are virtually unaffected when excluding this industry from the analysis.

Capital and Skill Intensities. These were obtained from the NBER-CES Manufacturing Industry Database (Becker et al., 2009). Skill intensity is the log of the number of non-production workers divided by total employment. Physical capital intensity is the log of the real capital stock per worker. Equipment capital intensity and plant capital intensity are respectively the log of the equipment and plant capital stock per worker. The NBER-CES data are originally available at the six-digit NAICS level, so matching them to the Related-Party Trade dataset only required minor adjustments related to trade data for five manufacturing industries being reported at the 5-digit NAICS (as explained in B.1 above). These industries were in turn mapped to IO2002 codes using the same procedure described in section B.2 above for the related-party trade data. For each factor intensity variable, a simple average of the annual values from 2000-2005 was taken to obtain the seller industry measures. The factor intensities for the average buyer were then calculated using the same procedure as described for the average buyer R&D intensity.

Detailed Capital Equipment Intensities. In some specifications we report results that break capital equipment intensity into the separate effects of expenditures on (i) automobiles and trucks for highway use, (ii) computers and peripheral data processing equipment, and (iii) all other machinery and equipment computers. These were obtained from the Annual Survey of Manufactures (2002-2010) which reports the data at the six-digit NAICS level. As mentioned before when discussing data on shipments from 2010 and 2011 (which also originate in the ASM), for 88 industries capital expenditures are reported at the more aggregated five-digit NAICS level. In order to impute those values to the six-digit sectors within each of these 88 industries we followed the same approach as in the case for industry shipments, but using overall equipment expenditures from the NBER-CES dataset as weights. The final measures of auto, computer, and other equipment intensity were obtained by dividing these types of capital expenditures by the

wage bill and taking logarithms. Values for these variables at the IO2002 level were obtained by filtering those variables through the BEA concordance described above, while 'average buyer' versions of these variables were also computed using the same approach as with R&D intensity and capital and skill intensity.

Productivity Dispersion. As in Antràs and Chor (2013), we build on Nunn and Trefler (2008), who constructed dispersion for each HS6 code as the standard deviation of log exports for its HS10 sub-codes across U.S. port locations and destination countries in the year 2000, from U.S. Department of Commerce data. We associated the dispersion value of each HS6 code to each of its HS10 sub-codes. These were mapped into IO2002 industries using the IO-HS concordance, taking a trade-weighted average of the dispersion value over HS10 constituent codes; the weights used were the total value of U.S. imports for each HS10 code from 1989-2006, from Feenstra, Romalis and Schott (2002). A similar procedure to that described above for trade costs and R&D intensity was used to obtain the dispersion measure for the remaining 13 IO2002 codes. The construction of a productivity or size dispersion measure at the six-digit NAICS level was analogous to that of the R&D intensity and required using data from encompassing five- or four-digit sectors to impute values to industries that otherwise would have been left with missing values.

Demand Elasticities. U.S. demand elasticities at the IO2002 level were computed as in Antràs and Chor (2013). We begin with the U.S. import demand elasticities for HS10 products computed by Broda and Weinstein (2006). This was merged with a comprehensive list of HS10 codes from Pierce and Schott (2009). For each HS10 code missing an elasticity value, we assigned a value equal to the trade-weighted average elasticity of the available HS10 codes with which it shared the same first nine digits. This was done successively up to codes that shared the same first two digits, to fill in as many HS10 elasticities as possible. Using the IO-HS concordance provided by the BEA with the 2002 U.S. I-O Tables, we then took the trade-weighted average of the HS10 elasticities within each IO2002 category. At each stage, the weights used were the total value of U.S. imports by HS10 code from 1989-2006, calculated from Feenstra, Romalis and Schott (2002). U.S. demand elasticities at the six-digit NAICS level were computed in an analogous way based on the same HS10 elasticities but using the HS-NAICS concordance in Pierce and Schott (2009) to compute six-digit NAICS averages. In both cases, there remained industries without elasticity values after the above procedures. Values for these sectors were imputed following the same approach as for other variables above, using data from encompassing five- or four-digit sectors. Finally, in order to compute 'average buyer' demand elasticities, we took a weighted average of the elasticities of industries that purchase the input in question, with weights equal to these input purchase values as reported in the 2002 U.S. I-O Tables. This is the same approach used to construct buyer versions of R&D, capital and skill intensity.

Input Substitutability. We begin with the import demand elasticities estimated by Broda and Weinstein (2006), but this time we use their estimates at the SITC Revision 3 three-digit level (rather than ten-digit HS level). As documented in their paper (see in particular their footnote 22), these elasticities were estimated in part off the substitution seen across HS10 product codes that fall under each SITC three-digit heading. These estimates would contain information on the degree of substitution across inputs under the assumption that the constituent HS10 products in each SITC three-digit category are typically used together as inputs in production. The three-digit SITC elasticities were mapped into IO2002 codes by first assigning them to HS codes using the concordance in Feenstra, Romalis and Schott (2002) and then using the HS to IO concordance provided by the BEA.

Specificity. This measure is borrowed from Antràs and Chor (2013), who in turn build on Rauch (1999) and Nunn (2007). For each IO2002 industry, it is calculated as the fraction of HS10 constituent codes classified by Rauch (1999) as neither reference-priced nor traded on an organized exchange, under Rauch's "liberal" classification. The original Rauch classification was for SITC Rev. 2 products; these were associated with HS10 codes using a mapping derived from U.S. imports in Feenstra, Romalis and Schott (2002). A higher value of this share is interpreted as the industry producing more differentiated goods, which in an input setting we associate with specificity.

Contractibility. We experiment with four measures of contractibility at the IO2002 industrial level. 'Nunn contractibility' was computed by Antràs and Chor (2013) from the 2002 U.S. I-O Tables following the methodology of Nunn (2007). We begin with the Rauch-Nunn sectoral measure of specificity above. For each IO2002 industry, we then calculate a weighted average specificity of the inputs used by that industry, where the weights correspond to each input's share in the overall input purchases of the industry in question. We took one minus this value as a measure of the 'Nunn contractibility' of each IO2002 industry. Levchenko and Costinot contractibility were obtained from Chor (2010), who in turn built them following the methodology of Levchenko (2007) and Costinot (2009), respectively. These two measures were normalized so that higher levels imply higher contractibility or lower dependence on formal contract enforcement. In particular, Levchenko contractibility is computed as the Herfindahl index of intermediate input use – rather than minus the Herfindahl as in Levchenko (2007) –, while Costinot contractibility is equal to the negative of the measure of complexity in Costinot (2009). Chor (2010) computed the Levchenko and Nunn measures at the 1987 Standard Industry Classification (SIC) level. We used a concordance from the U.S. census to map them into NAICS codes, and then the NAICS-IO2002 BEA concordance used for several measures above to obtain these variables at the IO2002 level.[29] Finally, BJRS contractibility corresponds to the measure of

[29] See http://www.census.gov/eos/www/naics/concordances/concordances.html.

intermediation in Bernard, Jensen, Redding and Schott (2010), who calculated this from U.S. establishment-level data as the weighted average of the wholesale employment share of firms in 1997, using the import share of each firm as weights. We use the IO2002 version of this variable constructed by Antràs and Chor (2013) (see their Data Appendix for more details).

We also compute 'average buyer' contractibility for each of these four contractibility measures using the same procedure described for computing average buyer R&D, capital and skill intensities.

Financial and Labor Contractibility. The Rajan and Zingales (1998) external dependence measure, the Braun (2002) asset tangibility measure, and the Cuñat and Melitz (2012) sales volatility measure are all borrowed from Chor (2010), who computed them at the 1987 SIC level. As with the Levchenko and Costinot contractibility variables discussed above, we used a concordance from the U.S. census to map them into NAICS codes, and then the NAICS-IO2002 BEA concordance used for several measures above to obtain these variables at the IO2002 level.

Downstreamness. This variable is calculated based on data from the 2002 U.S. I-O Tables, as described in Antràs and Chor (2013). It corresponds to a weighted index of the average position in the value chain at which an industry's output is used (i.e., as final consumption, as direct input to other industries, as direct input to industries serving as direct inputs to other industries, and so on), with the weights being given by the ratio of the use of that industry's output in that position relative to the total output of that industry. I next provide some more specific details on this measure for the interested reader. To build intuition, recall the basic input-output identity:

$$Y_i = F_i + Z_i,$$

where Y_i is total output in industry i, F_i is the output of i that goes toward final consumption and investment ("final use"), and Z_i is the use of i's output as inputs to other industries (or its "total use" as an input). In a world with N industries, this identity can be expanded as follows:

$$Y_i = F_i + \underbrace{\sum_{j=1}^{N} d_{ij}F_j}_{\text{direct use of } i \text{ as input}} + \underbrace{\sum_{j=1}^{N}\sum_{k=1}^{N} d_{ik}d_{kj}F_j + \sum_{j=1}^{N}\sum_{k=1}^{N}\sum_{l=1}^{N} d_{il}d_{lk}d_{kj}F_j + ...}_{\text{indirect use of } i \text{ as input}}, \tag{B.1}$$

where d_{ij} for a pair of industries (i, j), $1 \leq i, j \leq N$, is the amount of i used as an input in producing one dollar worth of industry j's output. Building on this identity, Antràs and Chor (2013) suggest computing the (weighted) average position of an industry's output in the value chain, by multiplying each of the

terms in (B.1) by their distance from final use plus one and dividing by Y_i:

$$\begin{aligned} U_i = 1 \cdot \frac{F_i}{Y_i} + 2 \cdot \frac{\sum_{j=1}^{N} d_{ij} F_j}{Y_i} \\ +3 \cdot \frac{\sum_{j=1}^{N} \sum_{k=1}^{N} d_{ik} d_{kj} F_j}{Y_i} \\ +4 \cdot \frac{\sum_{j=1}^{N} \sum_{k=1}^{N} \sum_{l=1}^{N} d_{il} d_{lk} d_{kj} F_j}{Y_i} + ... \end{aligned} \tag{B.2}$$

It is clear that $U_i \geq 1$ and that larger values are associated with relatively higher levels of upstreamness of industry i's use. Although computing (B.2) might appear to require computing an infinite power series, notice that provided that $d_{ij} < 1$ for all (i, j) (a natural assumption), the numerator of the above measure equals the i-th element of the $N \times 1$ matrix $[I - D]^{-2} F$, where D is an $N \times N$ matrix whose (i, j)-th element is d_{ij} and F is a column matrix with F_i in row i.[30] In order to obtain a measure of downstreamness (rather than upstreamness), Antràs and Chor (2013) simply take the reciprocal of U_i, which necessarily lies in the interval $[0, 1]$. Antràs, Chor, Fally and Hillberry (2012) discuss additional appealing features of this downstreamness measure.

B.7 Country-Level Covariates

Relative Factor Abundance. Physical capital abundance corresponds to the log of the physical capital per worker averaged over 2000-2005. Physical capital was constructed by Davin Chor based on investment data from the Penn World Tables (version 7.1) using the perpetual inventory method. Skill abundance is measured as the average years of schooling at all levels (primary, secondary, and tertiary) averaged over 2000 and 2005 based on the Barro and Lee (2013) dataset.

Rule of Law. Country rule of law is obtained from the Worldwide Governance Indicators (see also Kaufmann et al., 2010). The annual index ranges from -2.5 to 2.5 and it was averaged over the period 2000-05.

Financial Development. Computed as private credit provided by banking sector is measured as the percentage of GDP, averaged over 2000-05, based on data from the World Bank's World Development Indicators.

Labor Market Flexibility. This corresponds to the country labor market flexibility index for the year 2004 used by Cuñat and Melitz (2012). It was originally constructed by the World Bank building on the work of Botero, Djankov, Porta, Lopez-de Silanes and Shleifer (2004).

[30] Because $Y = [I - D]^{-1} F$, this numerator also equals the i-th element of the $N \times 1$ matrix $[I - D]^{-1} Y$, where Y is a column matrix with Y_i in row i.

GDP per capita. Computed as the log of Real GDP per capita in constant 2005 dollars from the Penn World Tables (version 7.1), averaged over the period 2000-05.