Online Appendix to "The Origins of Firm Heterogeneity: A Production Network Approach"

Andrew B. Bernard* Emmanuel Dhyne† Glenn Magerman‡
Kalina Manova§ Andreas Moxnes¶

May 3, 2022

A Data Sources

A.1 Data sources

The empirical analysis draws on four micro-level datasets at the National Bank of Belgium (NBB), covering the period 2002-2014: (i) the NBB B2B Transactions Dataset, (ii) annual accounts from the Central Balance Sheet Office supplemented by (iii) VAT declarations, and (iv) the Crossroads Bank. Firms are identified by a unique enterprise number, which is common across all databases and allows for unambiguous merging.

Firm-to-firm relationships The confidential NBB B2B Transactions Dataset contains the value of yearly sales among all VAT-liable Belgian enterprises for the years 2002 to 2014, and is based on the VAT listings collected by the tax authorities. All firms that deliver goods or services as an economic activity, on a regular and independent basis, are VAT liable. These enterprises have to charge an ad valorem tax on their sales and can recover VAT paid on their purchases. This includes foreign companies with a branch in Belgium and firms whose securities are officially listed in Belgium. Enterprises that only perform financial

*Tuck School of Business at Dartmouth, CEP, CEPR & NBER; andrew.b.bernard@tuck.dartmouth.edu
†National Bank of Belgium & Université de Mons; Emmanuel.Dhyne@nbb.be
‡ECARES - Université libre de Bruxelles, CEPR & National Bank of Belgium; glenn.magerman@ulb.be
§University College London, CEPR & CEP; k.manova@ucl.ac.uk
¶University of Oslo & CEPR; andreamo@econ.uio.no

transactions, medical or socio-cultural activities such as education are exempt. Firms with sales less than ï¿œ15,000 can choose to be exempt from the VAT liabilities. The standard VAT rate in Belgium is 21%, but for some goods a reduced rate of 12% or 6% applies.[1]

At the end of every calendar year, all VAT-liable enterprises have to file a complete listing of their Belgian VAT-liable customers over that year. An observation in this dataset refers to the value of sales in euro by enterprise i to enterprise j within Belgium, excluding the VAT due on these sales. The reported value is the sum of invoices from i to j in a given calendar year. Whenever this aggregated value is ï¿œ250 or greater, the relationship has to be reported. Together with the VAT declarations, these listings are the basis for the VAT amount due per tax period. Fines for late or incomplete reporting ensure a very high quality of the data. Note that each relationship is directed, as the observation from i to j is different from an observation from j to i. The dataset thus covers both the extensive and the intensive margins of the Belgian production network. A detailed description of the collection and cleaning of this dataset is given in Dhyne et al. (2015). We refer to that paper for further details.

Firm characteristics We extract information on enterprises' annual accounts from the Central Balance Sheet Office at the NBB for the years 2002 to 2014. Enterprises above a certain size threshold have to file annual accounts at the end of their fiscal year. We retain information on the enterprise identifier (VAT ID), turnover (total sales in euro, code 70 in the annual accounts), input purchases (total material and services inputs in euro and net changes in input stocks, codes 60+61), labor cost (total cost of wages, social securities and pensions in euro, code 62), and employment (average number of full-time equivalent (FTE) employees, code 9087). We annualize all flow variables from fiscal years to calendar years by pro-rating the variables on a monthly basis.[2]

Enterprises below a size threshold can report abbreviated annual accounts. These firms still report labor cost and employment, but are not required to report turnover or input purchases. For these small enterprises, we supplement information on turnover and input expenditures from their VAT declarations. All VAT-liable enterprises have to file periodic VAT declarations with the tax administration. The VAT declaration contains the total sales value (including domestic sales and exports), the VAT amount charged on those sales (both to other enterprises and to final consumers), the total amount paid for inputs sourced (including domestic and imported inputs), and the VAT paid on those input purchases. This

[1] See ec.europa.eu/taxation_customs for a complete list of rates. These rates did not change over our sample period.

[2] In our data, 78% of firms have annual accounts that coincide with calendar years, while 98% of firms have fiscal years of 12 months.

declaration is due monthly or quarterly depending on firm size, and it is the basis for the VAT due to the tax authorities every period. We aggregate the VAT declarations to the annual frequency to match the annual accounts.

We obtain information on the main economic activity of each enterprise at the NACE 4-digit level from the Crossroads Bank of Belgium for the years 2002 to 2014. We concord NACE codes over time to the NACE Rev. 2 version to deal with changes in the NACE classification over our panel from Rev. 1.1 to Rev. 2. Table 1 lists industry groups at the NACE 2-digit level. Finally, we also retrieve the postal code of the main establishment of the firm from the Crossroads Bank of Belgium. There are 589 municipalities in Belgium, which correspond to the Local Administrative Units level 2.

Since total sales and network sales come from different datasets, we enforce:

$$S_i = Snet_i + fd_i$$

where final demand $fd_i \geq 0$. For observations where $Snet_i$ from the NBB B2B transactions dataset is larger than S_i, we replace $S_i = \max(S_i, Snet_i)$. This amounts to 6% of observations in our estimation sample. The median value of $Snet_i$ that overshoots S_i is 17% of the value reported in the annual accounts for these observations, before we correct them.

B Additional results: Stylized facts

B.1 Distributions of firm size and number of connections

We provide additional moments on the distribution of firm sizes and the number of customers and suppliers in Table 2. Again, all variables are in logs, and demeaned by their NACE 4-digit industries.

We also report information on the raw (non-demeaned) distributions of these variables, confirming that massive heterogeneity in the production network is not driven by particular sectors, but instead is natural across all types of economic activity. Table 3 reports the distribution of firm size in the full sample, as well as separately for six aggregated industries for the year 2014: primary and extraction, manufacturing, utilities, construction, market services and non-market services. Tables 4 and 5 report summary statistics for the number of firm customers and the number of firm suppliers, respectively. While there is natural variation in average values across different types of economic activity, sizable dispersion is prevalent in all industries.

Table 1: NACE classification of industry groups (NACE Rev 2)

NACE Section	NACE Division	Description	Industry
A	NACE 01-03	Agriculture, forestry and fishing	Primary and Extraction
B	NACE 05-09	Mining and quarrying	Primary and Extraction
C	NACE 10-33	Manufacturing	Manufacturing
D	NACE 35	Electricity, gas, steam and air conditioning supply	Utilities
E	NACE 36-39	Water supply; sewage, waste management and remediation activities	Utilities
F	NACE 41-43	Construction	Construction
G	NACE 45-47	Wholesale and retail trade; repair of motor vehicles and motorcycles	Market Services
H	NACE 49-53	Transportation and storage	Market Services
I	NACE 55-56	Accommodation and food service activities	Market Services
J	NACE 58-63	Information and communication	Market Services
K	NACE 64-66	Financial and insurance activities	Market Services
L	NACE 68	Real estate activities	Market Services
M	NACE 69-75	Professional, scientific and technical activities	Market Services
N	NACE 77-82	Administrative and support service activities	Market Services
O	NACE 84	Public administration and defence; compulsory social security	Non-Market Services
P	NACE 85	Education	Non-Market Services
Q	NACE 86-88	Human health and social work activities	Non-Market Services
R	NACE 90-93	Arts, entertainment and recreation	Non-Market Services
S	NACE 94-96	Other service activities	Non-Market Services
T	NACE 97-98	Activities of households as employers; undifferentiated goods- and services-producing activities of households for own use	–
U	NACE 99	Activities of extraterritorial organizations and bodies	–

	$\ln S_i$	$\ln n_i^c$	$\ln n_i^s$
variance	1.73	1.87	0.60
skewness	0.57	0.13	-0.02
kurtosis	4.54	3.45	4.17
p1	-2.88	-3.18	-2.04
p5	-1.93	-2.22	-1.26
p10	-1.50	-1.71	-0.93
p25	-0.84	-0.89	-0.46
p50	-0.11	-0.01	-0.00
p75	0.72	0.87	0.46
p90	1.66	1.72	0.93
p95	2.34	2.25	1.26
p99	3.82	3.41	1.97

Table 2: Moments of distributions.
Notes: all variables are demeaned by their NACE 4-digit industries by regressing log variables on sector fixed effects and retaining the residuals.

Table 3: Total sales.

				Percentiles						
Industry	N	Mean	SD	p10	p25	p50	p75	p90	p95	p99
Primary and extraction (NACE 01-09)	2,818	12.9	451	.2	.4	.8	2.0	5.1	10.0	53.6
Manufacturing (NACE 10-33)	16,827	15.0	259	.3	.5	1.2	4.0	14.6	36.1	218.0
Utilities (NACE 35-39)	840	36.1	444	.4	.8	1.9	6.8	24.8	62.4	492.6
Construction (NACE 41-43)	19,008	2.4	14	.2	.3	.6	1.5	3.7	7.1	27.6
Market services (NACE 45-82)	53,532	6.3	86	.2	.4	1.0	2.5	7.4	15.6	73.8
Non-market services (NACE 84-96)	1,122	2.0	8	.2	.3	.5	1.1	2.8	6.8	32.0
All	94,147	7.5	155	.2	.4	.9	2.4	7.4	16.2	88.8

Note: values are reported in millions of euro.

Table 4: Number of customers.

				Percentiles						
Industry	N	Mean	SD	p10	p25	p50	p75	p90	p95	p99
Primary and extraction (NACE 01-09)	2,818	74	187	3	7	23	79	187	276	655
Manufacturing (NACE 10-33)	16,827	130	513	5	13	39	119	280	464	1,192
Utilities (NACE 35-39)	840	363	3,067	5	13	51	178	481	895	3,202
Construction (NACE 41-43)	19,008	50	246	4	7	16	37	94	174	533
Market services (NACE 45-82)	53,532	147	926	3	7	29	112	290	504	1,661
Non-market services (NACE 84-96)	1,122	66	433	2	4	10	40	105	187	591
All	94,147	123	797	3	8	26	92	245	429	1,299

Table 5: Number of suppliers.

Industry	N	Mean	SD	Percentiles p10	p25	p50	p75	p90	p95	p99
Primary and extraction (NACE 01-09)	2,818	75	83	22	34	54	86	138	197	411
Manufacturing (NACE 10-33)	16,827	110	154	25	42	70	126	225	313	670
Utilities (NACE 35-39)	840	152	280	25	45	80	153	286	456	1,508
Construction (NACE 41-43)	19,008	76	97	23	34	52	85	140	204	466
Market services (NACE 45-82)	53,532	69	87	17	30	50	80	132	184	363
Non-market services (NACE 84-96)	1,122	63	71	19	29	47	71	116	156	335
All	94,147	79	108	20	32	53	88	152	218	464

B.2 Market share and number of customers: seller-buyer sector-pair demeaning

A potential concern in interpreting the stylized facts in Section 2.2, in particular the relationship between average market shares and number of customers, is that firms with many customers might be selling different types of products compared to firms with few customers, even within the same 4-digit industry of the seller. In that case, average market shares might depend on seller-buyer pair characteristics, rather than characteristics of the seller only. For example, there may be both niche and broad-market sellers within the same industry, where the broad seller has more customers than the niche seller. Presumably, buyers of niche and broad-market products to some extent also belong to different industries: e.g. broad-market coffee roasters may sell to grocery stores, while niche coffee roasters may sell to coffee shops.

To take this into account, we reproduce the correlation between customer market shares and the number of customers in Fact 2, using fixed effects for the seller-buyer industry pair. By also demeaning by the customer sector, we account for heterogeneity in input requirements across sectors *within* the sector of the supplier. Specifically, for every seller i, we calculate the average market share and the number of customers in each 4-digit buyer industry k, $\bar{\delta}_{ik}$ and n_{ik}, respectively. The average market share is calculated as in the main text: the weighted geometric mean of m_{ij}/M_j^{net} across customers j belonging to buyer industry k. We then regress $\bar{\delta}_{ik}$ on n_{ik} (in logs), including seller-buyer industry pair fixed effects. The OLS slope coefficient remains negative and significant at -.03. While part of the variation is indeed absorbed by sector-pair-specific components, this shows that, even when looking at sales within a very detailed cell of the input-output table, e.g. coffee roasters selling to coffee shops, sellers with many customers obtain systematically lower market shares of their buyers' input purchases.

Dep. var: $\ln S_i^{net}$	(1) Baseline	(2) No fringe buyers (firm-level)	(2) No fringe buyers (global)
$\ln n_i^c$	.77***	.81***	.91***
	(.00)	(.01)	(.00)
Seller industry fixed effects	Yes	Yes	Yes
N	94,147	80,224	80,156

Table 6: Fringe buyers, 2014.

Note: Regressions comparing the elasticity of sales with respect to the number of customers across (1) baseline, (2) after dropping fringe buyers (bottom quartile) within each firm, and (3) after dropping fringe buyers (bottom quartile) across all firms. Significance: * $<$ 5%, ** $<$ 1%, *** $<$0.1%.

B.3 Fringe buyers

Another potential concern is that sellers with many connections have relatively more fringe buyers than firms with fewer connections, i.e. less important customers in terms of bilateral sales m_{ij}. This pattern would imply that firms with more customers have lower average sales per customer and lower market shares in these customers, but higher sales to each customer at each customer rank, rejecting the model in the main text.

We investigate this possibility by looking at the relationship between firm sales and the number of customers after dropping potential fringe buyers. We identify and drop fringe buyers in two ways. First, we drop transactions m_{ij} that are below the first quartile of a firm's transactions m_{ij} (i.e., the quartile is firm-specific).[3] Alternatively, we drop transactions that are below the global 1st quartile across all m_{ij} in the sample. In each case we recalculate the adjusted number of customers n_i^c and network sales S_i^{net}. Table 6 repeats the baseline analysis in this modified sample. The first column reproduces the baseline results for the regression of $\ln S_i^{net}$ on $\ln n_i^c$, while columns 2 and 3 show the results after removing fringe buyers. The slope coefficient increases slightly, but the main message of Fact 2 remains, namely that average sales (and market shares) are smaller for firms with many customers.

[3] This means that the analysis is restricted to firms with 5 or more customers.

C Theory Appendix

C.1 Variance decomposition

C.1.1 Identities

We decompose bilateral sales m_{ij} into i, j and ij components:

$$m_{ij} \overset{def}{=} G\psi_i\theta_j\omega_{ij} \tag{1}$$

where G is a mean across all ij, and without loss of generality, ψ_i, θ_j and ω_{ij} are defined relative to that mean. Summing over all customers j of seller i, $j \in \mathcal{C}_i$, on both sides:

$$\begin{aligned} S_i^{net} \overset{def}{=} \sum_{j\in\mathcal{C}_i} m_{ij} &= \sum_{j\in\mathcal{C}_i} G\psi_i\theta_j\omega_{ij} \\ &= G\psi_i \sum_{j\in\mathcal{C}_i} \theta_j\omega_{ij} \end{aligned} \tag{2}$$

where S_i^{net} is the total network sales of i to all its customers j. Total sales of i can be expressed as an identity:

$$\begin{aligned} S_i &= S_i^{net}\frac{S_i}{S_i^{net}} \\ \Rightarrow \ln S_i &= \ln S_i^{net} + \ln \beta_i \end{aligned}$$

where $\beta_i \overset{def}{=} \frac{S_i}{S_i^{net}} \geq 1$. Next, we can further decompose the j-specific term in eq.(2) into components that are observable in the data:

$$\sum_{j\in\mathcal{C}_i} \theta_j\omega_{ij} = n_i^c\bar{\theta}_i\bar{\omega}_i\frac{1}{n_i^c}\sum_j \frac{\theta_j}{\bar{\theta}_i}\frac{\omega_{ij}}{\bar{\omega}_i}$$

where n_i^c is the number of business customers j of firm i, and we choose $\bar{\theta}_i \overset{def}{=} \left(\prod_j \theta_j\right)^{\frac{1}{n_i^c}}$ and $\bar{\omega}_i \overset{def}{=} \left(\prod_j \omega_{ij}\right)^{\frac{1}{n_i^c}}$ to be geometric means of ω_{ij} and θ_j respectively for each i. Taking logs on both sides:

$$\ln\left(\sum_{j\in\mathcal{C}_i}\theta_j\omega_{ij}\right) = \ln n_i^c + \ln\bar{\theta}_i + \ln\bar{\omega}_i + \ln\left(\frac{1}{n_i^c}\sum_{j\in\mathcal{C}_i}\frac{\theta_j}{\bar{\theta}_i}\frac{\omega_{ij}}{\bar{\omega}_i}\right) \tag{3}$$

$$= \ln n_i^c + \frac{1}{n_i^c}\sum_{j\in\mathcal{C}_i}\ln\theta_j + \frac{1}{n_i^c}\sum_{j\in\mathcal{C}_i}\ln\omega_{ij} + \underbrace{\ln\left(\frac{1}{n_i^c}\sum_{j\in\mathcal{C}_i}\frac{\theta_j}{\bar{\theta}_i}\frac{\omega_{ij}}{\bar{\omega}_i}\right)}_{\stackrel{def}{=}\ln\Omega_i^c} \tag{4}$$

Combining all the above leads to the decomposition of sales for every i:

$$\begin{aligned}\ln S_i &= \ln S_i^{net} + \ln\beta_i \\ &= \ln G + \ln\psi_i + \ln n_i^c + \ln\bar{\theta}_i + \ln\bar{\omega}_i + \ln\Omega_i^c + \ln\beta_i \end{aligned} \tag{5}$$

The residual term $\ln\bar{\omega}_i$ is obtained from the 2-way FE OLS regression, and by the properties of OLS, $\ln\bar{\omega}_i = 0$.

C.1.2 Properties of the Decomposition

Consider first a two component case: $\ln S^{net} = \ln\psi + \ln\xi$. Then $Var(\ln S^{net}) = Var(\ln\psi) + Var(\ln\xi) + 2Cov(\ln\psi, \ln\xi)$. When we run an "inverse" regression of components on sales ($\ln\psi = \alpha + \beta_1\ln S^{net} + \varepsilon$), OLS estimation implies $\beta_1 = \frac{Cov(\ln S^{net},\ln\psi)}{Var(\ln S^{net})}$.

Note that

$$\begin{aligned} Cov(\ln S^{net}, \ln\psi) &= Cov(\ln\psi + \ln\xi, \ln\psi) \\ &= Cov(\ln\psi, \ln\psi) + Cov(\ln\xi, \ln\psi) \\ &= Var(\ln\psi) + Cov(\ln\xi, \ln\psi) \end{aligned} \tag{6}$$

where the penultimate equation follows from applying a linear combination of two random variables. Hence $\beta_1 = \frac{Cov(\ln S^{net},\ln\psi)}{Var(\ln S^{net})} = \frac{Var(\ln\psi)}{Var(\ln S^{net})} + \frac{Cov(\ln\xi,\ln\psi)}{Var(\ln S^{net})}$. Similarly for $\beta_2 = \frac{Cov(\ln S^{net},\ln\xi)}{Var(\ln S^{net})} = \frac{Var(\ln\xi)}{Var(\ln S^{net})} + \frac{Cov(\ln\xi,\ln\psi)}{Var(\ln S^{net})}$. Combining both:

$$\beta_1 + \beta_2 = \frac{1}{Var(\ln S^{net})}\left(Var(\ln\psi) + Var(\ln\xi) + 2Cov(\ln\xi, \ln\psi)\right) = 1 \tag{7}$$

From eq.(6), we see that these covariances are attributed equally to each component. I.e. the OLS coefficient thus captures the share of total variance in $\ln S^{net}$ explained by the variance in $\ln\psi$ plus any covariances across components.

The general case for the sum of $K \geq 2$ general random variables $X = \sum_{k=1}^{K} X_k$ is a straightforward extension, so that

$$\sum_{k=1}^{K} \beta_k = \frac{1}{Var(X)} \left(\sum_k Var(X_k) + \sum_{i \neq k} Cov(X_i, X_k) \right) = 1$$

Since we perform an exact decomposition (from the identities above), the sum of the variance shares sum to one, explaining all of the variance in X. Each OLS coefficient for X_k is equivalent to the share of the variance explained in X.

C.2 Seller Fixed Effect And Average Market Share

This section shows the relationship between seller fixed effects and average market shares.

Sales from i to j is $m_{ij} = e^G \psi_i \theta_j \omega_{ij}$. Therefore network purchases can be written as $M_j^{net} = e^G \theta_j \sum_{i \in \mathcal{S}_j} \psi_i \omega_{ij}$. Rearranging, we get

$$\theta_j = \frac{M_j^{net}}{e^G \sum_{i \in \mathcal{S}_j} \psi_i \omega_{ij}}. \tag{8}$$

Seller i's market share in j's network purchases is

$$\frac{m_{ij}}{M_j^{net}} = \frac{e^G \psi_i \theta_j \omega_{ij}}{M_j^{net}}.$$

The average of log market shares (i.e., the log of the geometric mean) across i's customers is

$$\begin{aligned} \ln MktSh_i &= \frac{1}{n_i^c} \sum_{j \in \mathcal{C}_i} \ln \frac{m_{ij}}{M_j^{net}} \\ &= \frac{1}{n_i^c} \sum_{j \in \mathcal{C}_i} \ln \frac{e^G \psi_i \theta_j \omega_{ij}}{M_j^{net}} \\ &= G + \ln \psi_i + \frac{1}{n_i^c} \left(\sum_{j \in \mathcal{C}_i} \ln \frac{\theta_j \omega_{ij}}{M_j^{net}} \right). \end{aligned}$$

Substituting θ_j with equation (8) above, we obtain

$$\begin{aligned}\ln MktSh_i &= \ln \psi_i + \frac{1}{n_i^c}\left(\sum_{j\in\mathcal{C}_i} \ln \frac{\omega_{ij}}{\sum_{i'\in\mathcal{S}_j}\psi_{i'}\omega_{i'j}}\right)\\ &= \ln \psi_i - \frac{1}{n_i^c}\sum_{j\in\mathcal{C}_i}\ln \tau_j\end{aligned}$$

where $\tau_j \equiv \sum_{i'\in\mathcal{S}_j}\psi_{i'}\omega_{i'j}$ and we used the fact that $\sum_{j\in\mathcal{C}_i}\ln\omega_{ij} = 0$ by the properties of OLS.

C.3 Existence and Uniqueness

We prove existence and uniqueness by showing that the fixed network equilibrium belongs to the class of models analyzed by Allen et al. (2016). They consider the following system of equations:

$$\prod_{h=1}^{K}\left(x_i^h\right)^{\gamma_{kh}} = c_i^k + \sum_{j=1}^{N} K_{ij}^k \prod_{h=1}^{K}\left(x_j^h\right)^{\beta_{kh}},$$

where $i, j \in \{1, .., N\}$ are firms/sectors, x_i^h is the type h equilibrium variable, c_i^k is a constant and K_{ij}^k are exogenous linkages between i and j. With $K = 1$ this reduces to

$$x_i^\gamma = c_i + \sum_{j=1}^{N} K_{ij} x_j^\beta. \tag{9}$$

The backward fixed point in equation (6) can be written in the form of equation (9) with $\gamma = 1$, $c_i = 0$, $\beta = 1-\alpha$ and $K_{ij} = z\left(\lambda'\right)^{\sigma-1}$. Using their notation, $\mathbf{A}$ is simply $1-\alpha$ and therefore the maximum eigenvalue of $\mathbf{A}^p$ is also $1-\alpha < 1$. According to their Theorem 2(i), there exists a unique and strictly positive solution to the backward fixed point.

The forward fixed point in equation (7) can be written in the form of equation (9) with $\gamma = 1$, $c_i = \mu^{1-\sigma} z\left(\lambda\right)^{\sigma-1} P\left(\lambda\right)^{(1-\sigma)(1-\alpha)} X/\mathcal{P}^{1-\sigma}$, $\beta = 1$ and $K_{ij} = \left(1-\alpha\right)/\left(\mu P\left(\lambda'\right)^{1-\sigma}\right)$. Using their notation, $\mathbf{A}$ is 1 and therefore the maximum eigenvalue of $\mathbf{A}^p$ is also 1. According to their Theorem 2(ii.a) there exists at most one strictly positive solution to the forward fixed point.

D Additional results: Decomposition

D.1 Results by sector

The main results pool the within-sector results across all sectors. To evaluate heterogeneity across sectors, we also perform the firm size decomposition separately by NACE 2-digit sector. To deal with possible incidental parameters, we drop sectors with fewer than 5 observations. Results are in Table 7. Table 8 shows the average value of the components across sectors, the standard deviation, and the coefficient of variation. All coefficients of variation are smaller than one, indicating relatively small differences in the firm size decomposition across sectors.

D.2 Results by year

Table 9 performs the firm size decomposition separately for each year in the sample. The results are virtually identical across all years.

D.3 Long differences

We also perform the decomposition in changes, or long differences from 2002 to 2014. First, we estimate equation (1) on two cross-sections, the baseline year 2014 ($t = 1$) and year 2002 ($t = 0$). We then calculate the change in every demeaned variable in the decomposition:

$$\Delta \ln S_i = \Delta \ln G + \Delta \ln \psi_i + \Delta \ln n_i^c + \Delta \ln \bar{\theta}_i + \Delta \ln \Omega_i^c + \Delta \ln \beta_i$$

where $\Delta \ln S_i$ denotes the log difference in S_i from $t = 0$ to $t = 1$ etc. We then demean all variables at the NACE 4-digit level. Finally, we regress each log-differenced component on $\Delta \ln S_i$. Since this is a within-firm analysis, we can only estimate the variance components in changes on firms that are active in both 2002 and 2014. Results are in Table 10. The main finding, that the number of customers dominates the decomposition, remains.

D.4 Controlling for seller location

A potential concern is that firms sort into different locations based on size and productivity: Large firms might be located in large cities, where there are more potential customers but also more suppliers, simultaneously creating competition for market shares in each customer. Large firms might then have both more customers and lower market shares in each customer for reasonsunrelated to our baseline model.

To assess this possibility, we perform the firm size decomposition controlling for the location of the seller.[4] In particular, we demean all variables by both the 4-digit NACE

[4] We thank an anonymous referee for this suggestion.

NACE Sector	N	$\ln\psi_i$	$\ln n_i^c$	$\ln\bar{\theta}_i$	$\ln\Omega_i^c$	$\ln\beta_i$	NACE Sector	N	$\ln\psi_i$	$\ln n_i^c$	$\ln\bar{\theta}_i$	$\ln\Omega_i^c$	$\ln\beta_i$
1	2,292	0.26	0.49	0.02	0.29	-0.05	49	2,939	0.07	0.53	0.02	0.35	0.02
2	298	0.16	0.51	0.04	0.31	-0.02	50	101	0.09	0.28	0.07	0.23	0.32
3	48	0.71	-0.04	0.05	0.36	-0.08	51	32	0.29	0.25	0.04	0.30	0.11
7	5	-0.11	0.47	-0.14	0.51	0.27	52	884	0.07	0.47	0.04	0.30	0.12
8	174	0.16	0.42	0.06	0.32	0.04	53	102	-0.12	0.88	0.02	0.13	0.09
9	20	0.27	0.22	0.03	0.28	0.20	55	836	0.11	0.75	0.07	0.21	-0.15
10	2,720	0.41	0.49	0.07	0.39	-0.36	56	4,304	0.11	0.57	0.05	0.19	0.08
11	150	0.36	0.30	0.04	0.36	-0.05	58	430	0.19	0.48	0.04	0.27	0.01
12	18	0.07	0.20	-0.06	0.24	0.55	59	356	0.22	0.35	0.05	0.34	0.04
13	661	0.22	0.17	0.05	0.26	0.30	60	21	0.30	0.27	0.05	0.37	0.01
14	267	0.26	0.41	0.03	0.20	0.09	61	149	0.11	0.37	0.07	0.24	0.21
15	69	0.20	0.50	-0.02	0.21	0.11	62	1,962	0.24	0.41	0.10	0.18	0.08
16	1,081	0.12	0.73	0.04	0.23	-0.12	63	182	0.25	0.39	0.10	0.22	0.04
17	252	0.34	0.17	0.06	0.30	0.13	64	1,137	0.16	0.39	0.06	0.23	0.16
18	1,131	0.22	0.43	0.06	0.25	0.04	65	11	0.36	0.11	0.14	0.26	0.12
19	22	0.09	0.56	-0.03	0.33	0.05	66	336	0.19	0.37	0.11	0.22	0.12
20	590	0.29	0.17	0.10	0.29	0.15	68	1,305	0.20	0.43	0.09	0.19	0.10
21	156	0.27	0.37	0.05	0.33	-0.03	69	4,065	0.02	0.67	0.08	0.09	0.14
22	620	0.24	0.31	0.04	0.26	0.15	70	1,726	0.28	0.31	0.10	0.21	0.09
23	1,023	0.27	0.49	0.06	0.27	-0.08	71	1,382	0.22	0.39	0.12	0.25	0.03
24	359	0.25	0.17	0.06	0.29	0.22	72	58	0.13	0.01	0.10	0.24	0.52
25	3,294	0.17	0.44	0.06	0.26	0.07	73	650	0.29	0.31	0.09	0.31	0.01
26	462	0.17	0.30	0.07	0.25	0.21	74	177	0.11	0.43	0.08	0.33	0.05
27	458	0.20	0.43	0.08	0.26	0.03	75	101	0.19	0.96	0.01	0.15	-0.32
28	940	0.15	0.38	0.07	0.22	0.18	77	302	0.23	0.53	0.05	0.20	-0.02
29	249	0.30	0.07	0.09	0.37	0.16	78	255	0.14	0.75	-0.00	0.14	-0.02
30	81	0.25	0.17	0.06	0.33	0.19	79	236	0.17	0.52	0.00	0.25	0.06
31	899	0.13	0.66	0.04	0.20	-0.02	80	98	0.02	0.64	0.04	0.32	-0.02
32	648	0.43	0.39	0.05	0.20	-0.07	81	839	0.11	0.64	0.09	0.25	-0.09
33	755	0.14	0.54	0.06	0.29	-0.03	82	288	0.21	0.38	0.12	0.26	0.03
35	123	0.11	0.38	0.05	0.43	0.03	84	16	0.29	0.36	0.11	0.23	0.00
36	53	-0.03	0.50	0.03	0.32	0.18	85	139	0.07	0.94	0.06	0.15	-0.22
37	119	0.16	0.44	0.04	0.50	-0.15	86	39	0.10	0.08	0.13	0.15	0.54
38	519	0.20	0.35	0.05	0.32	0.07	87	13	0.13	0.31	0.05	0.21	0.31
39	38	0.18	0.50	0.05	0.28	-0.02	90	84	0.07	0.55	0.05	0.22	0.10
41	5,596	0.15	0.50	0.07	0.31	-0.04	91	5	-0.04	1.15	0.09	-0.07	-0.13
42	1,253	0.16	0.51	0.10	0.34	-0.11	92	25	-0.35	0.22	0.07	0.16	0.90
43	12,159	0.08	0.76	0.06	0.25	-0.15	93	265	0.18	0.62	0.03	0.26	-0.08
45	5,582	0.18	0.56	0.02	0.29	-0.05	94	5	0.42	-0.51	0.26	0.19	0.65
46	12,724	0.29	0.52	0.03	0.23	-0.06	95	29	-0.20	0.76	0.03	0.57	-0.16
47	10,012	0.04	0.48	0.00	0.18	0.29	96	526	0.16	0.65	0.07	0.26	-0.13

Table 7: Firm size decomposition by NACE 2-digit industry (2014)

NACE Sector	lnβ_i	lnψ_i	$\ln n_i^c$	$\ln \bar{\theta}_i$	$\ln \Omega_i^c$
mean	0.07	0.17	0.43	0.06	0.26
st. dev.	0.20	0.14	0.24	0.05	0.09
CV	2.70	0.81	0.55	0.82	0.34

Table 8: Variation in the firm size decomposition across NACE 2-digit industries (2014)

Year	N	lnβ_i	lnψ_i	$\ln n_i^c$	$\ln \bar{\theta}_i$	$\ln \Omega_i^c$
2002	81,254	0.05	0.17	0.49	0.04	0.25
2003	83,678	0.05	0.17	0.49	0.04	0.25
2004	85,030	0.04	0.18	0.49	0.04	0.25
2005	86,474	0.04	0.17	0.49	0.04	0.25
2006	88,581	0.04	0.17	0.50	0.04	0.25
2007	91,027	0.03	0.18	0.50	0.04	0.25
2008	92,280	0.03	0.18	0.50	0.04	0.25
2009	92,333	0.04	0.17	0.50	0.04	0.25
2010	92,713	0.03	0.17	0.50	0.04	0.25
2011	94,093	0.03	0.18	0.50	0.05	0.25
2012	95,375	0.03	0.18	0.50	0.05	0.25
2013	94,135	0.02	0.18	0.50	0.05	0.25
2014	94,147	0.01	0.18	0.51	0.05	0.25

Table 9: Firm szie decomposition by year (2002-2014)

N	lnβ_i	lnψ_i	$\ln n_i^c$	$\ln \bar{\theta}_i$	$\ln \Omega_i^c$
40,974	0.10***	0.09***	0.57***	0.01***	0.23***

Note: Significance: * $<$ 5%, ** $<$ 1%, *** $<$0.1%.

Table 10: Firm size decomposition, long differences(2002-2014)

	N	$\ln\beta_i$	$\ln\psi_i$	$\ln n_i^c$	$\ln\bar{\theta}_i$	$\ln\Omega_i^c$
$NACE + NUTS3$	94,147	0.02***	0.17***	0.51***	0.05***	0.25***
$NACE \times NUTS3$	80,328	0.01***	0.16***	0.54***	0.05***	0.25***

Note: Significance: * < 5%, ** < 1%, *** <0.1%.

Table 11: Firm size decomposition by NACE 4-digit and NUTS3 (2014)

sector and the NUTS3 location of the seller. There are 44 NUTS3 regions in Belgium called *arrondissements*. We present two versions of the decomposition in Table 11. The first version demeans all variables by seller industry (NACE 4-digit) and seller location (NUTS3) fixed effects. The second version demeans all variables by seller industry-location pair fixed effects. As before, to account for potential incidental parameters, we drop cells with fewer than 5 observations. The results are almost identical to the baseline. This strongly suggests that the contribution of the different variance components is not significantly driven by agglomeration effects.

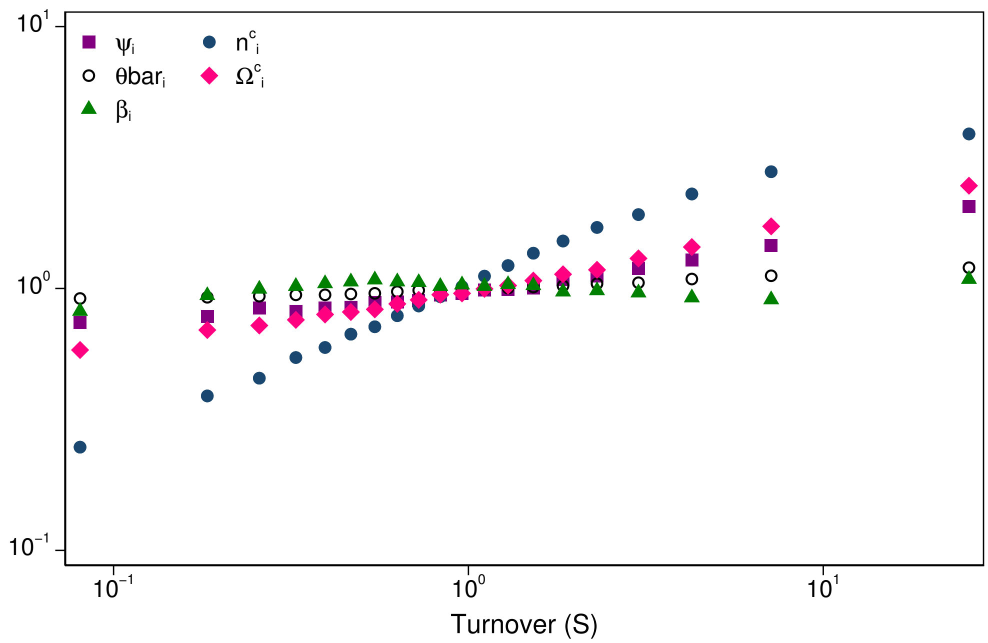
D.5 Non-parametric results

We test whether the decomposition results are stable across the firm size distribution using a simple non-parametric approach. Figure 1 groups log sales into 20 equal-sized bins, computes the mean of log sales and the components $\ln\psi_i$, $\ln n_i^c$, $\ln\bar{\theta}_i$, $\ln\Omega_i^c$ and $\ln\beta_i$ within each bin, and then creates a binned scatterplot of these data points. The result is a non-parametric visualization of the conditional expectation function, where the sum of the components on the vertical axis equals log sales on the horizontal axis. Overall, the slopes of each margin are close to linear, suggesting that the OLS decomposition results in the main text are applicable for both small and large firms.

D.6 Business Groups

The VAT identity number is the unique firm identifier we use across all datasets in this paper. This ID refers to the legal entity of the firm, and is the standard interpretation of a company in micro-level datasets. It is also the level at which a company reports annual accounts, firm-to-firm network data, and other typical firm characteristics. Still, there might be concerns that the legal entity does not constitute the economic entity of the company. In particular, multiple VAT identities may be owned by the same company. Intra-firm trade between these commonly owned entities might then exhibit non-market behavior in terms of the existence of particular buyer-supplier relationships and intra-firm pricing, affecting not only the value of sales relationships m_{ij} and the number of customers n_i^c, but potentially all

Figure 1: Firm size decomposition: Binned scatterplot.



Note: This binned scatterplot groups firms into 20 equal-sized bins by log sales, computes the mean of log sales and the components $\ln \psi_i$, $\ln n_i^c$, $\ln \bar{\theta}_i$, $\ln \Omega_i^c$ and $\ln \beta_i$ within each bin, and graphs these data points. The result is a non-parametric visualization of the conditional expectation function.

Table 12: Business Group Size Decomposition ($\ln S_i$).

Relative final demand	$\ln\beta_i$	.01	(.00)
Upstream	$\ln\psi_i$	.16	(.00)
# Customers	$\ln n_i^c$	.54	(.00)
Avg Customer Capability	$\ln\bar{\theta}_i$	.04	(.00)
Customer Interaction	$\ln\Omega_i^c$	.24	(.00)

Note: The table reports coefficient estimates from OLS regressions of a firm size margin (as indicated in the row heading) on total firm sales. All variables are first demeaned by their 4-digit NACE industry average. Standard errors in parentheses.

components in the firm size decomposition.

We therefore perform the firm size decomposition on business groups instead of VAT identities, as in Tintelnot et al. (2021). In particular, VAT identities are grouped into a single firm if the same parent company owns at least 50% of their shares. Turnover, inputs, employment and labor costs are summed across subsidiaries to the group level, after subtracting within-group transactions from turnover and inputs to avoid double counting. The NACE code of the firm with the largest turnover is assigned to the group. We report results for the decomposition for 2014 in Table 12. The number of observations is now 86,245. The results are almost identical to the baseline.

E Bootstrapped Standard Errors

The standard errors from the structural estimation in Section 5 are bootstrapped. The bootstrapping procedure is performed as follows. First, for each bootstrap sample, we randomly draw individual observations with replacement until we obtain the same sample size as the main dataset. Then, for each bootstrap sample, we create the empirical moments used in the simulated method of moments estimation. We create 200 bootstrap moments in total. Finally, we estimate the parameters of the model 200 times using the 200 different bootstrap moments. The standard deviations of the estimated parameters are the standard errors reported in Table 5.

References

Allen, Treb, Costas Arkolakis, and Xiangliang Li, "On the Existence and Uniqueness of Trade Equilibria," Technical Report, Working paper 2016.

Dhyne, E., G. Magerman, and S. Rubinova, "The Belgian Production Network 2002-2012," *National Bank of Belgium Working Paper*, 2015, *288.*

Tintelnot, Felix, Ken Kikkawa, Magne Mogstad, and Emmanuel Dhyne, "Trade and Domestic Production Networks," *Review of Economic Studies*, 2021, *88* (2), 643–668.