Online Appendix to "A Balls-and-Bins Model of Trade"

Roc Armenter and Miklós Koren*

October 2013

For online publication

A Data reference

Description of U.S. export data

Export data in the U.S. are based on Shipper's Export Declaration (SED) forms filed by exporters with the Customs and Border Protection and the Census Bureau. Filing a separate SED is mandatory for each shipment valued over $2,500. A *shipment* is defined as "all merchandise sent from one USPPI [firm] to one foreign consignee, to a single foreign country of ultimate destination, on a single carrier, on the same day."[1]

Each shipment is assigned a unique product code out of 8,988 potential "Schedule B" codes (of which 8,880 had positive exports in 2005). The Schedule B classification is based on the Harmonized System; the first six digits are HS codes. The remaining 4 digits are specific to U.S. exports. For convenience, we refer to these product codes in the paper as 10-digit HS codes.

We drop all 15 product codes in Chapter 98 (Special Classification Provisions). These categories are for products that are not identified by kind, either because of their low value, or some other reason.

There are 231 potential destination countries. Some of these entities are not countries but territories within countries (for example, Greenland has its own country code). We drop the country code 8220 (Unidentified Countries) and 8500 (International Organizations).

The Census Bureau publishes product–country aggregates based on this shipment-level dataset in "U.S. Exports of Merchandise." For each statistic, it also reports the number of SEDs (hence the number of shipments) that statistic is based on.

We calculate the average shipment size for a product–country pair as the total value of exports divided by the total number of shipments in 2005. For each product, we then take the median shipment size across destination countries.

Armenter: Federal Reserve Bank of Philadelphia. E-mail: roc.armenter@phil.frb.org. *Koren*: Central European University, MTA KRTK and CEPR. E-mail: korenm@ceu.hu

[1] "Correct Way to Complete the Shipper's Export Declaration," February 14, 2001 version.

Baldwin and Harrigan (2011)

Baldwin and Harrigan (2011) use data on U.S. imports and exports with all trading partners in 2005 in their analysis. This data comes from the U.S. Census, which reports value, quantity, and shipping mode for imports and exports and shipping costs and tariff charges for imports by trading partner and 10-digit HS commodity code. The Census does not report import trade values less than $250 for imports and $2,500 for exports, so small trade values are treated as zeroes. For imports, their dataset contains 228 trading partners (countries for which at least one good had a nonzero import value) for goods in 16,843 different 10-digit HS categories. For exports, there are 230 trading partners for goods in 8,880 different 10-digit HS categories (see Table 2).

Baldwin and Harrigan also use data on trading partner distance from the United States from Jon Haveman's website:

`http://www.macalester.edu/research/economics/PAGE/HAVEMAN /Trade.Resources/Data/Gravity/dist.txt`.

Macro variables (GDP, GDP per worker) are from the Penn World Tables.

Helpman, Melitz, and Rubinstein (2008)

Helpman, Melitz and Rubinstein (2008) use annual trade data on bilateral trade flows for 158 countries (see Table A1 for a list) from Feenstra's "World Trade Flows, 1970-1992" and "World Trade Flows, 1980-1997".

They also use data on population and GDP per capita from the Penn World Tables and the World Bank's World Development Indicators. They use data from the CIA World Factbook on whether a country is landlocked or an island, along with each country's latitude, longitude, legal origin, colonial origin, GATT/WTO membership status, primary language and religion.

Data from Rose (2000) and Glick and Rose (2002) is used to identify whether a country pair belonged to a currency union or the same FTA, and data from Rose (2004) to identify whether a country is a member of the GATT/WTO.

The variable capturing regulation costs of firm entry is derived from data reported in Djankov et al. (2002).

Bernard, Jensen, and Schott (2007)

Bernard, Jensen, and Schott (2007) use a dataset that links individual trade transactions to information on the U.S.-based firms involved in the transactions. Data on trade transactions for exports in 1993 and 2000 is collected by the U.S. Census Bureau, and includes information on export value, quantity, destination, date of transaction, port, and mode of transport at the 10-digit HS code level. Shipments data are collected for all export shipments above $2,500. Transaction-level data on imports are collected by U.S. Customs and Border Protection for all import shipments above $2,000. Detailed firm data comes from the Longitudinal Business Database of the Census Bureau. This dataset includes employment and survival information

for all U.S. establishments, though the linked dataset does not include establishments in industries outside the scope of the Economic Census.

Hummels and Klenow (2005)

Hummels and Klenow (2005) use data from the United Nations Conference on Trade and Analysis (UNCTAD) Trade Analysis and Information System (TRAINS) CD-ROM for 1995. This dataset consists of bilateral import data for 5,017 goods, 76 importing countries and all 227 exporting countries. Goods are classified by 6-digit HS code. They also use matching employment and GDP data for a subset of 126 exporters and 59 importers from Alan Heston et al. (2002). More detailed U.S. trade data comes from the "U.S. Imports of Merchandise" CD-ROM for 1995 from the U.S. Bureau of the Census. This dataset reports value, quantity, freight paid, and duties paid for 13,386 10-digit commodity classifications and 222 countries of origin, 124 of which have matching data on employment and GDP.

Bernard and Jensen (1999)

This paper uses firm-level data from the Longitudinal Research Database of the Bureau of the Census from 1984-1992. Their dataset includes all plants that appear in the Census of Manufactures for 1987 and 1992. For comparisons which involve more than one year, the set of firms is further restricted to those which also appear in the the Annual Survey of Manufactures for the inter-census years. The result is an unbalanced panel of between 50,000 and 60,000 plants for each year.

Bernard, Eaton, Jensen and Kortum (2003)

Bernard, Eaton, Jensen and Kortum (2003) use data from the 1992 U.S. Census of Manufactures in the Longitudinal Research Database of the Bureau of the Census. This dataset covers over 200,000 plants, and records the value of their shipments, production and non-production employment, salaries and wages, value-added, capital stock, ownership structure, and value of exports.

Bernard, Jensen, Redding and Schott (2007)

Bernard, Jensen, Redding and Schott (2007) use transaction-level U.S. data from the 2002 U.S. Census of Manufactures. This paper also looks at more detailed data from the Linked-Longitudinal Firm Trade Transaction Database, which is based on data collected by the U.S. Census Bureau and the U.S. Customs Bureau. The dataset reports the product classification, value and quantity shipped, data of shipment, trading partner, mode of transport, and participating U.S. firm for all U.S. trade transactions between 1992 and 2000.

Eaton, Kortum, and Kramarz (2004)

Eaton, Kortum, and Kramarz (2004) use French firm-level data on type and destination of exported goods from 1986. This dataset is constructed by merging customs data with tax-administration data sets from Bénéfices Réel Normal (BRN)-Système Unifié de Statistiques d' Entreprises (SUSE) data sources, and contains information on over 200 export destinations and 16 SIC industries.

Eaton, Kortum, and Kramarz (2011)

Eaton, Kortum, and Kramarz (2011) use sales data of over 200,000 French manufacturing firms to 113 markets in 1986. As in Eaton, Kortum, and Kramarz (2004), this dataset is constructed by merging customs data with tax-administration data sets from Bénéfices Réel Normal (BRN)-Système Unifié de Statistiques d' Entreprises (SUSE) data sources.

B Robustness analysis

We have explored a host of alternative calibrations. We detail here a selected few that shed light on the key determinants of our results. In the first set of calibrations we vary the total number of shipments (and thus observations) by assuming a counterfactual average shipment size. This illustrates how the model predictions depend on the sparsity of the data. Second we document the role of the skewness on trade flows, focusing on the calibration for firm-level facts.

B.1 Total number of shipments

We redo here our results for different numbers of shipments. We specify a "ball size" (in dollars) and convert trade flows into a discrete number of shipments by dividing the trade flow by the assumed "ball-size." We report results for ball sizes equal to \$2,500, \$18,000, \$36,000, \$100,000, and \$500,000. The smallest ball size is the lowest observed value of export transactions given the reporting rules of the Census Bureau, while \$36,000 is the actual average shipment size, and thus the value that ensures that the total number of shipments in the exercise is the same as in the data. Table 1 reports the implied number of shipments, with the corresponding number in the data highlighted in boldface.

Ball size	Number of shipments (10^6)
\$2,500	311.1
\$18,000	43.2
\$36,000	**21.6**
\$100,000	7.7
\$500,000	1.5

Table 1: Ball-size calibrations and number of shipments (in millions)

First we experiment with ball sizes equal and larger than the average size of an export. From an economic point of view, it may well be the case that the relevant decisions involve multiple transactions simultaneously and a calibration with a larger ball size would be appropriate. Table 2 shows our quantitative results for ball sizes between \$36,000 and \$500,000. We also included the corresponding data value for each of the stylized facts.

		Ball size		
Moment	Data	\$36k	\$100k	\$500k
HS10-level product×country U.S. export flows				
Share of zeros	82%	72%	80%	90%
OLS coefficient of nonzero flow on GDP	0.08	0.10	0.09	0.06
Firm×country U.S. export flows				
Share of zeros	98%	96%	98%	99%
Gravity for firms, GDP OLS coefficient	0.71	0.56	0.61	0.68
Single-product exporters				
Fraction over total exporters	42%	43%	57%	76%
Share of total exports	0.4%	0.3%	1.1%	7.4%
Single-destination exporters				
Fraction over total exporters	64%	44%	58%	77%
Share of total exports	3.3%	0.3%	1.1%	7.5%
Single-destination, single-product exporters				
Fraction over total exporters	40%	43%	57%	76%
Share of total exports	0.2%	0.3%	1.1%	7.4%
Exporters in U.S. manufacturing				
Fraction over total firms	18%	74%	61%	41%
Size premium of exporters	4.4	34	25	16

Table 2: Ball-size calibrations: \$36,000; \$100,000; and \$500,000

The changes in the magnitudes are intuitive. First, as we calibrate to a larger ball size, there are fewer balls overall and the incidence of empty product bins increases. This applies equally for zeros in trade or the fraction of single-product, single-destination exporters. Single-product and single-country exporters also increase their export share. With fewer shipments overall, most firms will end up with just one ball and would necessarily be single-product, single-country exporters. A larger ball-size calibration also reduces the fraction of exporting firms, closer to the one we see in the data. This is because if firms are taken to have fewer balls, it is less likely that any one of them comes from exports. However, even the \$500,000 ball-size calibration would predict significantly more exporters (41%) than in the data (18%). This suggests that economies of scale in deciding whether or not to export are rather strong.

Incidentally the column under the actual average shipment size, \$36,000, provides an additional check as it shuts down all the systematic variation in shipment size in trade flows. The resulting predictions from the model are virtually undistinguishable from those using the number of shipments per trade flow.

Table 3 shows our quantitative results for smaller ball sizes, between $36,000 and $2,500. These calibrations illustrate neatly the slow rate of convergence to a dense data set: the number of shipments under the smaller ball-size calibrations is orders of magnitude larger than the documented evidence yet sparsity still gives rise to zeros. The last column describes the limit as ball size shrinks to zero and the data is perfectly dense.

		Ball size			
Moment	Data	$36k	$18k	$2,500	none
HS10-level product×country U.S. export flows					
Share of zeros	82%	72%	66%	45%	0
OLS coefficient of nonzero flow on GDP	0.08	0.10	0.10	0.09	0
Firm×country U.S. export flows					
Share of zeros	98%	96%	94%	86%	0
Gravity for firms, GDP OLS coefficient	0.71	0.56	0.53	0.42	0
Single-product exporters					
Fraction over total exporters	42%	43%	35%	15%	0
Share of total exports	0.4%	0.3%	0.1%	0.0%	0
Single-destination exporters					
Fraction over total exporters	64%	44%	35%	15%	0
Share of total exports	3.3%	0.3%	0.1%	0.0%	0
Single-destination, single-product exporters					
Fraction over total exporters	40%	43%	35%	14%	0
Share of total exports	0.2%	0.3%	0.1%	0.0%	0
Exporters in U.S. manufacturing					
Fraction over total firms	18%	74%	81%	95%	100%
Size premium of exporters	4.4	34	67	337	n.a.

Table 3: Ball-size calibrations: $36,000; $18,000; and $2,500

As expected, smaller ball sizes imply fewer empty bins, both for products and for firms. Note, however, that even with a $2,500 ball size the majority of product and firm bins remain empty. This means that even a very small degree of indivisibility leads to a large number of empty bins. (In the limit, of course, there will be no empty bins.) Smaller balls also imply less action on the "extensive margin." Because most bins are filled, it is unlikely for new balls to fall in empty bins – hence the coefficient of country size on number of product or firm bins is smaller.

B.2 Skewness in export sales

The skewness in trade flows and categories plays an important role in our results. The gravity equation naturally generates skewness across destinations as some of them are large or small, close or far. Similarly heterogeneity in products (turnips or airplanes) generates a large variation across product categories.

We focus our robustness analysis on the skewness present in the distribution of export sales across exporters. The bottom line is that the balls-and-bins model matches the firm-level export patterns whenever the calibration accounts for the left-tail in the export distribution—in short, the small exporters.

In the first set we retain the use of the lognormal distribution and vary the parameter σ. By adjusting σ downwards we reduce the skewness of the distribution. The location parameter μ remains constant so we match the median exporter sales. Doing so preserves the left tail properties of the distribution; the skewness is reduced by thinning the right tail of the distribution. We maintain the calibration of the bin size distribution as detailed in the paper.

Table 4 reports the results. The first row is the calibration used in the paper. We focus on the fraction of exporters that are single-product and single-country exporters. These are 42 % and 60 % in the data. We explore parameters down to $\sigma = 2.3$, which is well below any estimate for the distribution of *domestic* sales.

		Fraction of Exporters	
σ	μ	Single-product	Single-country
3	11	43.3 %	44.1 %
2.7	11	42.6 %	43.5 %
2.5	11	42.0 %	43.0 %
2.3	11	41.3 %	42.4 %

Table 4: Lognormal distribution: Alternative calibrations.

As Table 4 makes clear, the model predictions are very robust. The predicted fractions decrease, but only very slowly. The fraction of single-product, single-country exporters (not reported) remains very close to 40 % for all calibrations.

The reason why the model is so robust is that the right tail of the distribution is irrelevant. Virtually all firms selling more than $100,000 are predicted to be multi-country, multi-product exporters. It does not really matter how exporters above this threshold are distributed.

It is instead the small exporters—the left tail—that drives our results. We can make this point explicitly by choosing a calibration with few small exporters. We reduce σ to 2.7 and *increase* the location parameter to 12.5 so we cut by half the number of exporters below $100,000. In this case, the fraction of single-product exporters falls to 31.7 %, and the fraction of single-country exporters falls to 32.6 %. The predicted fractions now fall sharply if we keep lowering σ and increasing μ, collapsing the distribution from both sides. With $\sigma = 2.3$ and $\mu = 13.5$ the fraction of single-product exporters is just 16 %.

We also experimented with alternative distributions as the Pareto distribution and the Yule-Simon distribution (drawing the number of shipments directly). The results reinforced the conclusion that matching the left-tail is a necessary and sufficient condition for the balls-and-bins model to match the facts.

C Aggregation and size premiums

C.1 Aggregate statistics

There is a total of T trade flows (countries, firms) in the dataset, each indexed by t and comprised of n_t shipments. The distribution of shipments across trade flows, $n_1, n_2, ..., n_T$, is taken as given. We find it useful to describe the distribution of shipments across trade flows as a probability distribution over $\mathbb{N}$, denoted π_n.[2] Each shipment can be classified into one of K categories.

The expected number of non-empty bins across all trade flows is given by

$$E(k|n_1, n_2, ..., n_T) = \sum_{n=1}^{N} \pi_n \sum_{i=1}^{K} [1 - (1 - s_i)^n] = \sum_{i=1}^{K} \sum_{n=1}^{N} \pi_n [1 - (1 - s_i)^n]. \tag{1}$$

Let $G(z)$ denote the *probability generating function* (PGF) corresponding to the distribution $\{\pi_n\}$:

$$G(z) = \sum_{n=1}^{N} \pi_n z^n.$$

Then the number of non-empty bins can be written as

$$E(k|n_1, n_2, ..., n_T) = \sum_{i=1}^{K} [1 - G(1 - s_i)].$$

Since $G(z)$ is strictly convex, uneven bin-size distributions will have a smaller expected number of non-empty bins.

What about the proportion of single-bin trade flows? For each trade flow of size n, the probability is $\sum_{i=1}^{K} s_i^n$. The conditional probability is then

$$\Pr(k = 1|n_1, n_2, ..., n_T) = \sum_{n=1}^{N} \pi_n \sum_{i=1}^{K} s_i^n = \sum_{i=1}^{K} \sum_{n=1}^{N} \pi_n s_i^n.$$

We can also express it in terms of the PGF as

$$\Pr(k = 1|n_1, n_2, ..., n_T) = \sum_{i=1}^{K} G(s_i).$$

It then becomes clear that the convexity of $G(z)$ also preserves the properties of each flow with respect to the fraction of single bins. In particular, we can now assert that more even bin-size distributions induce a lower fraction of single-bin flows.

Finally we can also calculate the fraction of *balls* that have fallen into a single bin. This corresponds to, for example, the fraction of *sales* attributed to single-product firms.

$$\sum_{n=1}^{N} \pi_n n \sum_{i=1}^{K} s_i^n = \sum_{i=1}^{K} \sum_{n=1}^{N} \pi_n n s_i^n.$$

[2] To be precise, we assume that the support is bounded by some finite N.

With the use of the PGF notation,

$$\sum_{n=1}^{N} \pi_n n s_i^n = G'(s_i)s_i.$$

And we can easily have the average size of trade flows that all fall in bin i is

$$\frac{\sum_{n=1}^{N} \pi_n n s_i^n}{\sum_{n=1}^{N} \pi_n s_i^n} = \frac{G'(s_i)s_i}{G(s_i)}.$$

It is important to note that, unless the number of trade flows is infinite, the actual fractions will be a random variable. Since all distributions are known it is actually possible to derive the actual distribution for each moment. It is, however, often unpractical to do so and one can use Monte Carlo methods to derive the distribution as needed.

C.2 Exporter's size premium

Let π_n be the unconditional size distribution of firms. The firm-size distribution conditional on not exporting is

$$\Pr(n|\text{no export}) = \frac{\Pr(\text{no export}|n)\pi_n}{\Pr(\text{no export})}.$$

The average sales (number of balls) of non-exporters is

$$E(n|\text{no export}) = \sum_{n=1}^{\infty} \frac{\pi_n n(1-s)^n}{\Pr(\text{no export})}.$$

The average sales for the population of firms is

$$E(n) = \sum_{n=1}^{\infty} \pi_n n.$$

We can express the expected sales of non-exporters in terms of the probability generation function $G(z)$ of the firm size distribution.

$$E(sales|\text{no export}) = \frac{(1-s)G'(1-s)}{G(1-s)},$$

the elasticity of G evaluated at $1-s$. Note that G is differentiable. The unconditional mean is given by the same formula but evaluated at $z = 1$:

$$E(sales) = \frac{1G'(1)}{G(1)}.$$

A sufficient condition for non-exporters being smaller than the average if the elasticity of G is increasing in z.

To see how the skewness in the firm size distribution leads to a large exporter premia, we parameterize the distribution as a *zeta distribution*. This is the discrete analogue to Pareto distribution, and its probability mass function is

$$\pi_n = \frac{n^{-\alpha}}{\zeta(\alpha)}.$$

Here α is the tail exponent, and is estimated to be about 2.06 by Axtell (2001). The probability generating function of the zeta distribution is

$$G(z) = \frac{\mathrm{Li}_\alpha(z)}{\zeta(\alpha)},$$

where Li_α is the (non-analytic) polylogarithm function. By properties of polylogarithm, the elasticity of $G(z)$ is given by

$$\frac{zG'(z)}{G(z)} = \frac{\mathrm{Li}_{\alpha-1}(z)}{\mathrm{Li}_\alpha(z)}.$$

With $\alpha = 2.06$, this implies that exporters are about 18 times as big as non-exporters. If we lower α closer to 2, we are putting more mass of the distribution on its upper tail. For $\alpha = 2.02$, exporters are 27 times as big as non-exporters.

D Nesting balls-and-bins in a structural model

In this Appendix we show how to nest the balls-and-bins framework in a structural model. The model of choice is a slightly modified version of Helpman, Melitz, and Rubinstein (2008), extended to encompass product heterogeneity as well as comparative-advantage forces. The model also features firm-level fixed costs of exporting to each country. We briefly derive the key equations concerning firm-level and aggregate (country-product) trade flows.

The first step is to generate finite-sample data predictions from the model. The model's market shares inform the likelihood that a shipment/observation belongs to a particular category. Following the same steps as in the main text, the probability distribution over categories can be used to compute the moments of interest, like the expected number of empty country-product trade flows or the share of exporters among all firms, given a number of observations. In the model presented here, some categories may have zero probability due to the fixed costs and a bounded productivity distribution: we call these fundamental zeros.

We also show that the structural model nests nicely our baseline calibration for the balls-and-bins framework, where we assumed no systematic relationship between countries, products, and firms. In short, our baseline calibration corresponds to parametrizations such that the model's trade volumes are multiplicatively separable in country, product, and/or firm fixed-effects. It is typically possible, and quite intuitive, to calibrate a trade model this way. For example, for country-product trade flows to be multiplicatively separable in the model presented here, relative wages across sectors must be equal across countries—this amounts to shutting down traditional comparative advantage forces—and the productivity distribution must be unbounded—isolating aggregate trade flows from firm-level fixed costs.

In addition we also ask whether the fixed costs and the upper bound on the productivity distribution—the key parameters behind the extensive margin at the firm and aggregate level, respectively—are identified in a sparse dataset. We show the fixed costs are readily pinned down by matching the share of exporters observed in the data—a fact that the balls-and-bins model missed. In contrast, the mechanism leading to empty country-product trade flows is poorly identified in a sparse dataset—even if we seek identification within the strict confines of a single structural model.

D.1 A simple trade model with economies of scale

We present here a modified version of Helpman, Melitz, and Rubinstein (2008), who in turn took the model's key features from Melitz (2003) and Chaney (2008). The main modification is to introduce product heterogeneity so the model has predictions for the full product-country set of trade flows. Since we check the model exclusively against U.S. export data, we abridge the model description along several dimensions and focus on the equations concerning the foreign demand for U.S. goods. The reader can refer to Helpman, Melitz, and Rubinstein (2008) for a complete description of the model.

There are $j = 0, 1, \ldots, J$ countries with the U.S. indexed by $j = 0$. There is a continuum of firms, each indexed by $\omega \in \Omega$, adding up to a mass N_j in country j and producing each a differentiated commodity with non-mobile labor. We assume each firm's differentiated good belongs to one out of G different product categories or "sectors," with the distribution of firms across sectors given by $\{\mu^g\}$, with $\sum_G \mu^g = 1$. Firms are heterogeneous in their labor productivity, φ_i, distributed according to a truncated Pareto distribution,

$$\Psi(\varphi) = \frac{1 - \varphi_l^k \varphi^{-k}}{1 - \varphi_l^k \varphi_h^{-k}}$$

on the support $[\varphi_l, \varphi_h]$, allowing for the possibility that the support is unbounded above, $\varphi_h = \infty$. The wage rate in country j for sector g is denoted w_j^g and is taken as given. We allow the wage distribution across sectors to vary across countries, introducing comparative advantage as a possible source of trade.

The demand for firm ω, belonging to sector g, located in country i and selling in country j is given by

$$y_{ij}^g(\omega) = \frac{\alpha^g Y_j}{P_j^g} \left(\frac{p_{ij}^g(\omega)}{P_j^g} \right)^{-\sigma} \tag{2}$$

where Y_j is the country j income, α^g is the share of expenditures in good g, $\sigma > 1$ is the elasticity of substitution across goods, and the price index P_j^g is given by

$$P_j^g = \left(\sum_{i \in J} \int_{\omega \in \Omega_{ij}^g} \left(p_{ij}^g(\omega) \right)^{1-\sigma} d\omega \right)^{1/(1-\sigma)}, \tag{3}$$

where Ω_{ij}^g is the set of firms from country i, in product classification g, that sell in country j. This demand system is standard in trade and can be derived from well-known preferences.

To ship a good to country i from country j a firm must incur on "iceberg" trade costs $\tau_{ij} \geq 1$, with $\tau_{ii} = 1$. Firms operate under monopolistic competition, resulting in the familiar markup-over-marginal-cost pricing,

$$p_{ij}^{g}(\omega) = \frac{\sigma}{\sigma - 1} \frac{\tau_{ij} w_i^g}{\varphi(\omega)} \tag{4}$$

if firm ω sells in country j. If it does, its revenues are

$$r_{ij}^{g}(\omega) = p_{ij}^{g}(\omega) y_{ij}^{g}(\omega) = \alpha^g Y_j \left(\frac{p_{ij}^{g}(\omega)}{P_j^g} \right)^{1-\sigma}. \tag{5}$$

Subtracting the variable costs, the firm would make net revenues equal to $r_{ij}^{g}(\omega)/\sigma$.

But not all firms export to all destinations.[3] There is a fixed cost F for exporting to each destination. The decision rules for export participation are quite simple. Firm ω will export to country j if and only if

$$r_{ij}^{g}(\omega) - \sigma F \geq 0.$$

The export-participation decisions can be solved for in terms of productivity thresholds, which in turn makes it quite simple to solve for the set of firms exporting a product type to a destination. Since the only idiosyncratic attributes of a firm in country i are its productivity level and the product category it belongs to, we can rewrite the (potential) revenues in terms of a simple function $r_{ij}^{g}(\varphi)$, where we substitute the optimal price (4) in the revenue equation (5). This can be used to characterize the vector of thresholds φ_{ij}^{g} for each good g as

$$r_{ij}^{g}\left(\varphi_{ij}^{g}\right) = \sigma F. \tag{6}$$

If $\varphi_{ij}^{g} < \varphi_l$, then all firms in product g in country i export to country j. Conversely, if $\varphi_{ij}^{g} > \varphi_h$ then no firm in product g in country i export to country j. For threshold values in the interior of the support, then only a fraction of the firm export, namely those with productivity above the threshold, $\varphi \geq \varphi_{ij}^{g}$.

To characterize bilateral trade volumes, let

$$V_{ij}^{g} = \int_{\varphi_{ij}^{g}}^{\varphi_h} \varphi^{\sigma-1} d\Psi(\varphi)$$

with the understanding that $V_{ij}^{g} = 0$ if $\varphi_{ij}^{g} > \varphi_h$. For the predicted export flows in good g from country i to country j we just aggregate the sales revenues across the firms active in that market,

$$X_{ij}^{g} = \int_{\varphi \geq \varphi_{ij}^{g}} r_{ij}^{g}(\varphi)\, d\Psi(\varphi). \tag{7}$$

With some simple algebra we obtain that

$$X_{ij}^{g} = \left(\frac{\sigma \tau_{ij} w_i^g}{(\sigma - 1) P_j^g} \right)^{1-\sigma} \alpha^g Y_j \mu^g N_i V_{ij}^{g} \tag{8}$$

[3]We assume there are no fixed costs associated with selling domestically, so all firms are active in their own country.

where the price index for good g in country j is

$$P_j^g = \left(\sum_{i \in J} \left(\frac{\sigma \tau_{ij} w_i^g}{(\sigma - 1)} \right)^{1-\sigma} \mu^g N_i V_{ij}^g \right)^{\frac{1}{1-\sigma}}. \tag{9}$$

Note that the aggregate bilateral trade flow between country i to j in the product classification g will be zero if no firm exports to that market, that is, $V_{ij}^g = 0$ or $\varphi_{ij}^g > \varphi_h$.

To complete the model we just need income-expenditure equality conditions in each country. However, we will not solve the general equilibrium and instead focus on the model's implications for U.S. firms and exports. Empirically, several equilibrium variables will simply be captured by country or product fixed-effects, as in Helpman, Melitz, and Rubinstein (2008) and other empirical applications.

We will also be interested in firm-level facts, like the fraction of exporters or their relative size. For example, the predicted revenues of a U.S. firm with productivity φ in sector g are

$$r^g(\varphi) = \sum_{j=0}^{J} 1\left[\varphi \geq \varphi_j^g\right] r_j^g(\varphi)$$

where $1\left[\varphi \geq \varphi_j^g\right]$ is the indicator function and whenever we omit the subscript for the country of origin the latter is understood to be the United States or $j = 0$. It is also straightforward to compute the total sales distribution. The fraction of exporters among U.S. firms in product sector g is simply $1 - \Psi\left(\min_{j \geq 1} \varphi_j^g\right)$. It is also quite trivial to characterize the distribution of domestic and foreign sales among exporters, and the share of multi-destination exporters.

We note that the predicted flow for good g to country j may be zero if and only if (i) the productivity distribution is bounded above, since then it is possible that the threshold φ_j^g is outside the support for the labor productivity, $\varphi_j^g > \varphi_h$, and (ii) there are fixed costs at the firm level, $F > 0$. This is actually the key mechanism in Helpman, Melitz, and Rubinstein (2008).

D.2 Finite-sample data predictions

The previous model, and indeed most trade models, takes the form of a set of continuous trade flows. That is, if we were to evaluate the models at different frequencies, the predicted export flows would just scale up or down proportionately with the frequency. The set of traded good-country destinations, for example, will be invariant. In this sense trade in the models is similar to oil flowing through a pipeline at a constant rate.

The data, however, consists of a finite number of observations, corresponding to the transactions in a given time period, usually a year. We bridge the gap between the theory and the data by sampling the model, mapping the model's market shares into the likelihood that a given transaction belongs to a particular category. In doing so, we effectively nest the balls-and-bins framework with the underlying structural model.

Let $n \in \mathbb{N}$ be the number of observations. This can be set to reproduce the number of observations in the data but we can also explore the model implications in a dense dataset

by letting n tend to infinity. It is also possible to "discretize" the revenues flows in the model (or the data) assuming a shipment size, say ξ. This is particularly useful when the number of shipments is unknown or the researcher does not want to use the observed shipment distribution.

The key step is to derive the likelihood a shipment belongs to a given category from the model-implied market shares. The probability that a shipment from the firm has country j as a destination is given by

$$s_j^g(\varphi) = \frac{1\left[\varphi \geq \varphi_j^g\right] r_j^g(\varphi)}{r^g(\varphi)}.$$

That is, the likelihood of a shipment sent to country j is given by the share of that destination within the firm's total sales. The probability then that *any* shipment from the firm reaches destination j, that is, the probability that the firm is exporting to country j, is simply

$$1 - \left(1 - s_j^g(\varphi)\right)^{n^g(\varphi)},$$

where $n^g(\varphi)$ is the number of shipments/observations assigned to the firm. The formulas from the main text carry on, making it possible to compute the expected number of destinations a firm will serve, expected relative size, and so on.

In order to obtain predictions for the aggregate country-product flows, we also need to characterize the probability that a shipment belongs to a given firm i. Again, this is simply the firm's share as predicted by the model,

$$s^g(\varphi) = \frac{r^g(\varphi)}{\bar{R}}$$

where $\bar{R}$ is the U.S. total sales as predicted by the model. Now adding up all firms, we would obtain that the probability that a single shipment belongs to a given country-product pair is

$$s_j^g = \frac{X_j^g}{\bar{R}}.$$

The above probability distribution is defined for all $j = 0, 1, 2, ..., J$ countries, including the home country. By contrast, most of our facts concern only export flows. This is not a problem, as shipments are assumed to be i.i.d., then the distribution conditional on the shipment being shipped abroad is simply

$$s_j^g = \frac{X_j^g}{\bar{X}}. \tag{10}$$

where $\bar{X}$ is total U.S. exports as predicted by the model.

Summarizing, we let the structural model determine the bin-size distribution and replicate the number of shipments or observations in the data. The market shares in the underlying model become the likelihood that a single shipment belongs to the category of interest. A market share can be zero in the model: a fundamental zero as the particular category will not be observed to receive shipments. These fundamental zeros, though, will coexist with sample zeros if the data are sparse. If instead the data are dense, the realized frequency of

shipments across categories will converge almost surely to the probability distribution, and thus we recover exactly the predicted market shares in the structural model. In particular, only fundamental zeros will remain.

We should emphasize that it is possible to create a finite-sample data set prediction for *any* structural model simply following the steps presented here.

D.3 Nesting the baseline balls-and-bins calibration

In the main text, we constructed our baseline calibration for the balls-and-bins by matching the distribution across product classifications and countries—the marginal distributions—and constructing the probability of a particular country-good pair as the product of probabilities, that is, abstracting from any systematic relationship between product and countries and, by extension, firms.

We nest this baseline assumption in the structural model presented above as a special case. From (10) it should be clear that we can write $s_j^g = s_g s_j$, where s_g and s_j are the probabilities of a shipment being of product g and being sent to destination j, respectively, if and only if we can express the trade flow in the underlying model as

$$X_j^g = d_g d_j$$

for all g and j. That is, the trade flow is multiplicatively separable in two "fixed effects," one for the product and one for the country of destination.[4] The fixed effects clearly allow the model to capture perfectly the marginal distributions of trade across products and countries.

When is it possible to express trade flows as $X_j^g = d_g d_j$ in the underlying model? By simple inspection of the demand function (8) it is clear that most of the terms depend on the product or destination, but not both. The exceptions are the price level of the good g in country j, P_j^g, and the composition term, V_j^g.[5] Inspecting the price level, the composition term reappears, now for every country pair, V_{ij}^g, and we now have the vector of wages for good g across origin countries, w_i^g. Note that the former term is closely related to the economies of scale in exporting, while the latter is tied to comparative advantage forces. In order to obtain the desired separability in trade flows $X_j^g = d_g d_j$, we require a condition on each:

1. The relative wages across sectors are identical in every country, i.e.,

$$w_i^g = \omega_g \omega_i.$$

2. The productivity distribution is unbounded, $\varphi_h = \infty$.

The first condition is not surprising: if some countries can produce certain products (relatively) cheaper than other countries, that is, have a comparative advantage on these goods, there will be a systematic relationship between destinations and products. Mathematically,

[4]The decomposition does not need to be unique, as some terms may not be indexed to either product or country. The choice of decomposition does not matter as the resulting distribution of probabilities s_j^g will be invariant. A similar exercise can be done for firm-level flows.

[5]Recall we have fixed the U.S. as the country of origin and omitted the corresponding subscript.

it is clear that the unit cost must be itself multiplicatively separable if we want the predicted flows to satisfy $X_j^g = d_g d_j$. With $w_i^g = \omega_g \omega_i$, the price index can be re-written as

$$P_j^g = \omega_g \left(\mu^g\right)^{\frac{1}{1-\sigma}} \left(\sum_{i \in J} \left(\frac{\sigma \tau_{ij} \omega_i}{(\sigma - 1)} \right)^{1-\sigma} N_i V_{ij}^g \right)^{\frac{1}{1-\sigma}},$$

and we are left only with V_{ij}^g as the only joint country-product term.

The second condition is closely related to the underlying model's ability to generate zero trade flows *in the aggregate*. It does not impose any condition on the fixed cost parameter, F, so it is possible to have firms that do not export yet the condition be satisfied. Formally, the steps are as follows. A key property of the Pareto distribution when $\varphi_h = \infty$ is that

$$V_{ij}^g = \varphi_{ij}^g \frac{k}{1 + k - \sigma}.$$

From the entry condition $r_{ij}^g \left(\varphi_{ij}^g\right) = \sigma F$ we can solve for the threshold using the revenue function (5). The latter is multiplicatively separable in origin, destination, and product but for the price index itself, which allows us to express the threshold as such as well, $\varphi_{ij}^g = \rho_g \rho_i \rho_j \left(P_j^g\right)^{-1}$. Substitute back in (9) and there are no joint product-country terms left.

We should note that the second condition is sufficient, but not necessary. For example, it is possible to have $\varphi_h < \infty$ and $X_j^g = d_g d_j$ if there are no fixed costs, $F = 0$. This alternative, though, is not really satisfactory since it completely shuts down the extensive margin in the model.

The structural model allows many extensions and yet our baseline balls-and-bins calibration remains nested as a special case. For example, fixed costs can be specified to depend on the country of origin, country of destination, and product. Once again, imposing that the heterogeneity is multiplicatively separable, $F_{ij}^g = \phi_g \phi_i \phi_j$, would recover the property behind our baseline calibration. Similarly it would be possible to encompass differences in the productivity distribution across sectors or countries, additional factors of production, or sector-specific trade costs.

D.4 Identification in a sparse dataset

In the main text we showed the difficulties at distinguishing different model classes in a sparse data set. Unfortunately, there are also difficulties within a structural model when it comes to identify its parameters. As an example, we document the model's prediction, evaluated for a sparse data set, for the share of exporters and positive country-product trade flows across a series of calibrations. We start from the baseline balls-and-bins case, i.e., a calibration satisfying conditions 1 and 2 above. We then explore how the model's predictions vary with the fixed costs, F. Finally we work with a truncated Pareto distribution and explore different values for the upper bound parameter, φ_h.

D.4.1 Set parameters

We refer to the literature to pin down the elasticity of substitution σ to 6. The lower bound for the support φ_l can be normalized to 1 and the slope k is set to match the estimates for the tail distribution of firm size, $k\left(\sigma-1\right)^{-1}=1.06$, as it is common in the literature.

We use product and country fixed-effects to capture the heterogeneity across products and countries in the model, respectively. Note that, it is in principle possible, and perhaps even preferable, to complete the calibration by using GDP or similar for each country's income Y_j, distance and other proxies for the matrix of trade costs, τ_{ij}, PPP exchange rates for relative wages, as well as proxies for product market shares. However, for the purposes here, the fixed-effect approach is better suited since it ensures the distribution of trade across products and countries—the marginal distributions—exactly matches the data. It has the additional advantage that the fixed-effects can be readily computed for most cases by simply equating them to the respective product and country market shares in U.S. exports. The only difficulty is in the case of a truncated Pareto distribution, which requires to solve a large non-linear system to obtain the fixed-effects. We do find, however, that for the range of parameters explored below the fixed-effects vary little and actually simply track the observed country and product shares in total U.S. exports very closely. Finally, evaluating the model under a sparse data requires deciding on the the number of shipments observed in the data, about 22 million.

We are thus left with the two parameters of interest: the fixed cost of exporting to one destination, F, as well as the upper bound of the support, φ_h. These govern fundamental zeroes at the firm and country-product level, respectively. Thus the natural empirical targets the fraction of traded country-good pairs and the fraction of exporters among all firms.

D.4.2 Identifying fixed costs of exporting at the firm level

We start from the balls-and-bins framework, with $F=0$ and an unbounded productivity support, as well as no comparative advantage trade due to differences in relative wages across countries. As discussed above, the structural model does not introduce any systematic relationship between countries and products under these conditions.

The results are, not surprisingly, as in the main text. The model generates 72 percent empty trade flows in the country-product matrix and greatly overstates the share of exporters, close to 75 percent in the model versus 18 percent in the data. Exporters in the model, despite being more frequent than in the data, are also larger than their data counterparts.

In the main text we claimed that the share of exporters is indeed a useful stylized fact to identify the underlying theory of the extensive margin of interest, that is, export participation at the firm level. In the simple model presented here, the extensive margin is clearly tied to the fixed costs of exporting, so we explore whether parameter values $F>0$ help the model match the share of exporters.

Indeed they do. As we increase F, we reduce the share of exporters, as expected. The relationship is sharp, and we easily obtain a point estimate such that the model predicts about

18 percent of the firms exporting, as in the data.[6] Increasing or decreasing the fixed cost by ten percent drives the share of exporters in the model to 16 and 20 percent, respectively. The exporter size premium, though, remains substantially above the data, about 16 versus $4-5$, respectively.

D.4.3 The extensive margin at the country-product level

In contrast, the fraction of traded country-product pairs does not bulge at all as F is increased. As argued earlier, economies of scale at the firm level do not introduce a systematic relationship between countries and products. Thus the baseline calibration for the balls-and-bins model remains in place and as long as the fixed-effects are set to match the aggregate distribution of trade across products and countries, we will obtain the exact same predictions for any parameter value $F > 0$. Because of this, we can target the two stylized facts separately, as we did in the main text.

What are the effects of the upper bound in the productivity distribution, $\varphi_h < \infty$? The structural model only predicts zero-probability trade flows when there is an upper bound. The model with unbounded support underpredicts the share of zeros in country-product flows, so there is room for improvement. However, we find that imposing an upper bound, even a tight one, has a muted impact on the model's predictions. As we reduce φ_h we start "closing" trade flows that were very small to start with, and thus predicted to be zero with very high probability anyway in the sparse dataset.

Of course, the support for the productivity distribution can be set small enough that eventually the predicted fraction of empty trade flows increases. Unfortunately, it is impossible to do so without shutting down trade all together in many product categories as well as to several destinations. Decreasing the upper bound φ_h also decreases the share of exporters, in a rather mechanical way: as the support is reduced, there is a larger mass of firms close to the lower bound—which are less likely to export. If we decrease F then to keep the share of exporters constant at 18 percent, the model actually cannot increase the fraction of empty trade flows beyond about 75 percent.

References

[1] Agresti, Alan: 2002, *Categorical Data Analysis*, Second Edition, John Wiley and Sons. Hoboken, NJ.

[2] Alessandria, G., Kaboski, J., and Midrigan, V.: 2003, "Inventories, Lumpy Trade, and Large Devaluations," *American Economic Review,* forthcoming.

[3] Axtell, R. L.: 2001, "Zipf Distribution of U.S. Firm Sizes," *Science* **293**(5536), 1818–1820.

[6] The fixed cost parameters comes close to $1 million, substantially above estimates elsewhere using a richer model of export participation.

[4] Baldwin, R. and Harrigan, J.: 2011, "Zeros, Quality, and Space: Trade Theory and Trade Evidence", *American Economic Journal: Microeconomics*, 3(2):60-88, May.

[5] Bernard, A. B., Eaton, J., Jensen, J. B. and Kortum, S.: 2003, "Plants and Productivity in International Trade", *American Economic Review* 93(4):1268–1290, September.

[6] Bernard, A. B. and Jensen, J. B: 1999, "Exceptional Exporter Performance: Cause, Effect, or Both?", *Journal of International Economics* 47(1):1–25, February.

[7] Bernard, A. B., Jensen, J. B., Redding, S. J. and Schott, P. K.: 2007, "Firms in International Trade", *Journal of Economic Perspectives* 21(3):105–130, Summer.

[8] Bernard, A. B., Jensen, J. B. and Schott, P. K.: 2007, "Importers, Exporters and Multinationals: A Portrait of Firms in the U.S. that Trade Goods", *in* Dunne, J.B. Jensen and M.J. Roberts (eds.), Producer Dynamics: New Evidence from Micro Data.

[9] Djankov, Simeon Freund, Caroline Pham, Cong S., 2006. "Trading on time," Policy Research Working Paper Series 3909, The World Bank.

[10] Eaton, J., Eslava, M., Kugler, M. and Tybout, J.: 2007, "Export Dynamics in Colombia: Firm-Level Evidence", NBER Working Paper No. 13531.

[11] Eaton, J., Kortum, S. and Kramarz, F.: 2004, "Dissecting Trade: Firms, Industries, and Export Destinations", *American Economic Review* 94(2):150–154, May.

[12] Eaton, J., Kortum, S. and Kramarz, F.: 2011, "An Anatomy of International Trade: Evidence from French Firms", *Econometrica*, 79(5):1453–1498, September.

[13] Evans, Carolyn L. and James Harrigan, 2005. "Distance, Time, and Specialization: Lean Retailing in General Equilibrium," *American Economic Review*, American Economic Association, vol. 95(1), pages 292-313, March.

[14] Johson, N. L., Kepm, A. W., and Kotz, S.: 2005, *Univariate Discrete Distributions*, John Wiley & Sons.

[15] Helpman, E., Melitz, M. and Rubinstein, Y.: 2008, "Estimating Trade Flows: Trading Partners and Trading Volumes", *Quarterly Journal of Economics* 123(2):441-487, May.

[16] Haveman, J. and Hummels, D.: 2004, "Alternative hypotheses and the volume of trade: the gravity equation and the extent of specialization", *Canadian Journal of Economics* 37(1):199–218, April.

[17] Hummels, D., Klenow, P. J.: 2005, "The Variety and Quality of a Nation's Exports", *American Economic Review* 95(3):704–723, June.

[18] Hummels, David and Lugovskyy, Volodymyr and Skiba, Alexandre, 2009. "The trade reducing effects of market power in international shipping," *Journal of Development Economics*, vol. 89(1), pages 84-97, May.