A Balls-and-Bins Model of Trade*

Roc Armenter and Miklós Koren†

June 22, 2008

Abstract

A number of stylized facts have been documented about the extensive margin of trade—which firms export, and how many products they send to how many destinations. We argue that the sparse nature of trade data is crucial to understanding these stylized facts. Trade data are collected through customs forms, one for each export shipment, specifying the country of destination and the product code. Typically the number of observations—that is, total shipments—is low relative to the number of possible classifications—e.g., countries and product codes. Given the sparse data, we note that some of the reported facts would be expected to arise even if exports shipments were randomly allocated across classifications. These facts are thus not informative of the underlying economic decisions. We propose a statistical model to account for the sparsity of trade data. We formalize the assignment of shipments to categories as balls falling into bins. The balls-and-bins model quantitatively reproduces the prevalence of zero product-level trade flows across export destinations. The model also accounts for firm-level facts: as in the data, most firms export a single product to a single country but these firms represent a tiny fraction of total exports. In contrast, the balls-and-bins cannot reproduce the small fraction of exporters among U.S. firms, and overpredicts their size premium relative to non-exporters. We argue that the balls-and-bins model is a useful statistical tool to discern the interesting facts in disaggregated trade data from patterns arising mechanically through chance.

1 Introduction

International trade has long been concerned with aggregate patterns—what and how much countries trade—and their welfare implications. Recently, finely disaggregated trade data

*For useful comments we thank Arnaud Costinot, Jonathan Eaton, James Harrigan, Tom Holmes, László Mátyás, Marc Melitz, Virgiliu Midrigan, Esteban Rossi-Hansberg, Peter Schott, Adam Szeidl, Ayşegül Şahin, and seminar participants at the Federal Reserve Bank of New York, the Institute for Advanced Studies in Vienna, Central European University, UC San Diego, and Princeton University. We also thank Jennifer Peck for excellent research assistance. The views expressed here do not necessarily reflect the views of the Federal Reserve Bank of New York or the Federal Reserve System.

†*Armenter*: Federal Reserve Bank of New York. E-mail: roc.armenter@ny.frb.org. *Koren*: Princeton University. E-mail: mkoren@princeton.edu.

have become available and have had an enormous impact on the field. This has spurred a fast-growing research that documents the extensive margin in trade—which firms export, and how many products they send to how many destinations. This, in turn, has lead to new theories in international trade.

A number of stylized facts have been documented about the extensive margin of trade: (1) Most product-level trade flows across countries are zero; (2) the incidence of non-zero trade flows follows a gravity equation; (3) only a small fraction of firms export; (4) exporters are larger than non-exporters; (5) most firms export a single product to a single country; (6) most exports are done by multi-product, multi-destination exporters.[1] These facts have proven to be very robust across datasets from various years in various countries.

We argue that the sparse nature of trade data is crucial to understanding these stylized facts. Trade data are collected through customs forms, one for each export shipment, specifying the country of destination and the product code. Typically the number of observations—that is, total shipments—is low relative to the number of possible classifications—country and product code pairs. For example, there were about 24 million shipments originating in the U.S. in 2000. However, there are 229 countries and 8,867 product codes with active trade, so a shipment can have more than 2 million possible classifications. We should then not be surprised to observe empty categories, or to learn that the U.S. does not export all products to all countries.

Given the sparsity of the data, how do we interpret a missing trade flow? Take the example of "vessels for passenger and freight transport." Switzerland did not import a vessel from the United States in 2005. Being a landlocked country, it probably never will. At the same time, 130 of the 188 coastal countries did not import a vessel either: they have a positive demand for American vessels yet do not buy one every year.

In this paper we propose a statistical model to account for the sparsity of trade data. We formalize the assignment of shipments to categories as balls falling into bins. Each shipment constitutes a discrete unit (the ball), which, in turn, is allocated into mutually exclusive categories (the bins). This structure is inherent to disaggregate trade data: we observe a given number of shipments and each of them is classified into a unique category. Because we want an atheoretical account of the sparsity of the data, the model assigns balls to bins at random. That is, a ball falling in a particular bin is an independent and identically distributed random event whose probability distribution is determined solely by the distribution of bin sizes.

In spite of its simplicity, the balls-and-bins model has a rich set of predictions. After a number of balls, some bins may end up empty and some will not. Among the latter some will contain a large number of balls, some few. These are taken to be the model's predictions for the extensive and intensive margin, respectively. We can derive analytically the relevant moments. Given a number of balls and a bin size distribution, we show how to compute the

[1] The following is a necessarily incomplete list of references. Helpman, Melitz and Rubinstein (2007) and Baldwin and Harrigan (2007) for facts 1 and 2; Haveman and Hummels (2004) and Hummels and Klenow (2001, 2005) for fact 1; Bernard and Jensen (1999) and Bernard, Eaton, Jensen and Kortum (2003) for facts 3 and 4; Bernard, Jensen and Schott (2007) for facts 3 to 6; Bernard, Jensen, Redding and Schott (2007) for facts 2 to 6; and Eaton, Kortum and Kramarz (2004, 2007) for facts 5 and 6. See the main text and the Appendix for further discussion.

prevalence of zeros and how it varies with the number of balls and the skewness of the bin-size distribution. These are indeed all the model's systematic relationships between export flows and the extensive margin: the *assignment* of balls to bins is random.

We are interested, though, in a quantitative evaluation. For this we set the number of balls equal to the number of observed shipments in the trade flow of interest (for example, total trade between two countries or total exports by a given firm).[2] For the dimension of choice (product codes or destination countries) we construct the bin size distribution using aggregate flows. For example, there are 8,867 bins for the 10-digit Harmonized System product codes, with each bin size set to the corresponding share in total U.S. exports. The calibration accounts for the fact that the U.S. exports some products more than others, but it assumes no systematic differences across destination countries in the composition of exports.

The results are striking: the balls-and-bins model *quantitatively* reproduces many of the stylized facts on the extensive margin in trade. Table 1 summarizes our findings. For twelve statistics we report the data and the corresponding prediction by the model—the details on both are in the main text. Zero product-level trade flows are as prevalent in the model as in the data; indeed the pattern of zeros across export destinations is also the same. Despite the number of export shipments (24 million) exceeding the potential product–country pairs (about 2 million), the model makes clear that zeros are to be expected: in a random assignment, the first ball to fall in a non-empty bin comes very early—what is known as the birthday paradox—and thus empty bins are prevalent. Moreover, there is a large variation in the size of the trade flows and categories. Trade with most of the 229 countries is very small and most of the 8,867 traded HS codes are tiny. It is exactly for these country-product pairs that the trade flows are missing in the data. They go missing in the model as well: few balls and tiny bins make for many empty bins.

The model also accounts for firm-level facts: as in the data, most firms export a single product to a single country but these firms represent a very small fraction of total exports. The skewness in the distribution of exports across firms is essential to understand the success of the balls-and-bins model. Most exporters are tiny and are hence assigned only one ball in the model. They are thus predicted to be single-product, single-country exporters.[3] This finding suggests that once we account for the skewness of export sales, the incidence and relative size of single- and multi-product exporters follow.

What do we learn when the balls-and-bins model matches a particular fact? Surely we are not suggesting that firms actually ship their goods at random! Our view, instead, is that we cannot conclude *anything*: if a fact cannot falsify the balls-and-bins model, it will also fail to identify the relevant economic theory and thus should not be the basis to favor any model. Indeed, as long as a model correctly predicts aggregate flows, it will be able to match the stylized facts by introducing a small amount of discreteness and sufficient idiosyncratic variation to reproduce the sparse nature of the data.

[2]Unfortunately we do not have access to shipment data at the firm level. In this case we approximate the number of shipments by dividing the firm-level trade flows into balls of \$36,000 — the value of the average export transaction in the U.S. in 2000.

[3]The average exports of the bottom three quarters of all exporters are just \$75,000. By contrast, the top one quarter of exporters export \$20 million on average.

Description	Data	Balls-and-bins
HS10-level product×country U.S. export flows		
Share of zeros	82%	72%
OLS coefficient of nonzero flow on GDP	0.08	0.10
Firm×country U.S. export flows		
Share of zeros	98%	96%
Gravity equation for firms, GDP OLS coefficient	0.71	0.56
Single-product exporters		
Fraction over total exporters	42%	43%
Share of total exports	0.4%	0.3%
Single-destination exporters		
Fraction over total exporters	64%	44%
Share of total exports	3.3%	0.3%
Single-destination, single-product exporters		
Fraction over total exporters	40%	43%
Share of total exports	0.2%	0.3%
Exporters in U.S. manufacturing		
Fraction over total firms	18%	74%
Size-premium of exporters	4.4	34

Table 1: Summary of Findings
Details on sources, data and model are in the main text and in the Appendix.

We can also learn from the balls-and-bins model when it misses a data pattern. For example, we attempt to predict the share of exporters among manufacturing firms. In the balls-and-bins model 74 percent of firms will export — in contrast with 18 percent in the data. Surprisingly, the model also overpredicts the export size premium. This suggests that the split between exporters and non-exporters goes well beyond the difference in size.

We view the balls-and-bins model as a useful statistical tool that can quantitatively discern the interesting facts from the patterns arising mechanically through chance. It can be applied to any categorical dataset, such as the division of total exports by products, firms, or destination countries. These datasets contain a lot of information: it is crucial that we focus on the facts that will help us differentiate among competing trade theories as well as inform the development of new ones. We should emphasize that we believe there will be no shortage of interesting facts in the data.

Given our results, it is natural to ask why trade data are sparse. A look at average shipment sizes across products suggests that indivisibilities are important. The largest shipments observed include aircraft, spacecraft, and tanker ships. Some goods are divisible but storage and transportation limitations make it unpractical to do so. For example, the median shipment size of enriched uranium is $13 million.[4]

[4]As Hummels, Lugovskyy and Skiba (2008) document, minimum scale requirements are also paramount in maritime trade.

We must emphasize, though, that we treat shipments as discrete because they are the finest unit of observation possible. While trade theories predict a stable system of flows, datasets consist of a finite number of transactions for a given interval of time. The Census dataset for a given year should thus be treated as a sample of the underlying system, even if it constitutes the universe of shipments in that particular year.[5]

A paper close to us in spirit is Ellison and Glaeser (1997). They ask whether the observed levels of geographic concentration of industries are greater than would be expected to arise randomly. To this end they introduce a "dartboard" model of firm location. In contrast with our results, the "dartboard" model reaffirms the previous results on geographic concentration. Ellison and Glaeser (1997) are also able to provide a new index for geographic concentration which takes a value of zero under the dartboard model and thus controls for the mechanical degree of concentration arising from randomness. Such an index is more difficult for trade facts, which do not focus on a particular dimension.

The questions sparsity brings are similar to the debate about the theoretical content of the gravity equation for bilateral trade flows. The gravity equation is hugely successful in predicting trade flows, yet it may be of limited use in distinguishing trade theories. Deardorff (1998) argues that "just about any plausible model of trade would yield something very like the gravity equation," hence the gravity equation should not be the basis for favoring one theory over another. Evenett and Keller (2002) and Haveman and Hummels (2004) also show that the gravity equation is consistent with both complete and incomplete specialization models.

Our paper is also related to a large literature that tests the robustness of empirical findings through Monte Carlo techniques or sensitivity analysis. To our knowledge these tests have not been commonplace in international trade. An early exception is the analysis on trade-related international R&D spillovers in Keller (1998). There has also been some work on the robustness of gravity equation models. Ghosh and Yamarik (2004) use Leamer extreme bounds analysis to construct a rigorous test of specification uncertainty and find that the trade creation effect associated with regional trading arrangements is fragile. Anderson, Ferrantino, and Schaefer (2004) use Monte Carlo experiments to explore alternative specifications of the gravity model and find coefficient bias to be pervasive.

The paper is organized as follows. The next section presents some new evidence that illustrates that trade datasets are sparse. Section 3 describes the setup of the balls-and-bins model and characterizes some of its properties. Section 4 presents the empirical facts on missing product-level trade flows and discusses how the balls-and-bins model matches these facts. Section 5 conducts the same exercise for firm-level trade flows. Section 6 discusses the extensive margin of products and destination countries at the firm level. Section 7 looks at whether the balls-and-bins model can predict the number and size of exporters. Section 8 offers some extensions. Finally, Section 9 concludes. The Appendix provides extensions to the main model, and describes in detail the datasets used in the cited papers.

[5]See Appendix C on how to map trade models into discrete datasets.

2 Trade data are sparse

According to the "U.S. Exports of Merchandise" published by the Census Bureau, there were 21.6 million export shipments in 2005.[6] Each shipment is assigned a unique product code out of 8,988 potential codes (of which 8,867 had positive exports in 2005) and one out of 229 destination countries.[7] That makes about 2 million potential product–country categories, or about one for each 11 shipments.

The average number of shipments per category masks enormous skewness in the number of shipments across categories. Looking at destination countries first, Canada had the most shipments, 7.35 million. Equatorial Guinea, the median buyer of U.S. exports, had only 2,641 shipments. There is a similar skewness across product categories. The product category with the most shipments is "parts and accessories, for motor vehicles of headings 8701 to 8705, n.e.s.o.i.," with 386,619 shipments. The median product category had only 480 shipments.

Shipments are even sparser once we divide them up by *both* destination countries and products. Of all the positive product–country export flows, the median only consists of 4 shipments, and around 29% of product–country categories have only 1 shipment. Table 2 summarizes the distribution of the number of shipments across product–country categories.

Number of shipments	Frequency
1	28.7%
2	12.8%
3	7.8%
4	5.4%
5	4.1%
6-9	9.9%
10 and above	31.4%

Table 2: Number of shipments across product–country categories

That even positive flows have so few observations clearly indicates that the dataset is too sparse to make strong inference from zeros. In other words, the large fraction of ones should make us doubt that the zeros are observationally different.

Let us take the opportunity to explore further why the trade data are sparse. A look at shipment sizes suggests that they are the result of the indivisibility of the product. The typical shipment is rather small; the median shipment size is $12,800. As Table 3 shows, 94% of products have a shipment size below $50,000. There is, however, substantial variation in shipment sizes. The biggest shipment is a single shipment of "cargo aircraft of an unladen weight exceeding 15,000 kg" to Singapore, in the amount of $245 million.

Some products are bulky by their very nature. The biggest shipments include aircraft ($42 million), spacecraft ($5 million), tanker ships ($15 million) and floating drilling platforms ($5

[6] We focus on U.S. data which are widely used in the above-mentioned empirical studies. As we argue below, we expect such sparsity to be a prevalent feature of all transaction-level trade data. See the Data Appendix for more detailed description of the Census export data.

[7] Some of these entities are not really countries but are small territories. Results do not change substantially if one restricts the analysis to the 191 actual countries.

Shipments size	Frequency
less than $5,000	2.3%
$5–10,000	34.3%
$10–20,000	40.4%
$20–50,000	16.9%
$50,000–1 million	5.7%
above $1 million	0.6%

Table 3: Shipment sizes across product categories

million). Some products are inherently divisible but storage and transportation limitations make it unpractical to do so. For example, the median shipment size of enriched uranium is $13 million.

More formally, product dummies explain 40% of the variance of log shipment sizes. By contrast, destination-country dummies only explain 4% of the variation. Distance to the destination country is not significantly correlated with shipment sizes.

We can further explore the effect of physical indivisibility on shipment sizes by looking at the 2,374 products that report "numbers" as the units of quantity. These are classified to be indivisible by the Census Bureau.[8] Overall, the weight of the product explains most of the variation in shipment size; the rank correlation between the two is 0.72. Products with the 100 biggest shipment size have a median weight of 6.5 metric tons. By contrast, the ones with the 100 smallest shipment size have a median weight of 1.2 kilograms.

To summarize, there is suggestive evidence that the indivisibility present in this dataset is mostly related to product characteristics rather than the economic environment. We thus expect sparsity to arise in other trade datasets as well.

3 A model of balls and bins

We model the assignment of export shipments to categories as balls falling into bins. The balls-and-bins model reproduces the structure inherent in disaggregate trade data. A trade flow (such as total exports from the U.S. to Argentina, or total exports of a given firm) is composed of a finite number of shipments, each of them a discrete unit of observation (the balls). Every shipment has been classified into mutually exclusive categories, for example, into one of the 10-digit Harmonized System product classifications (the bins).

Formally, let $n \in \mathbb{N}$ be the number of balls (observations). Let $K \in \mathbb{N}$ be the number of bins (categories), each of them indexed by subscript $i \in \{1, 2, \cdots, K\}$. The probability that any given ball lands in bin i is given by the bin size s_i, with $0 < s_i \leq 1$ and $\sum_{i=1}^{K} s_i = 1$. Thus where a ball lands is independent of the number and location of the other balls.

The state of the system is given by the full distribution of balls across bins, $\{x_1, x_2, \cdots, x_K\}$. Clearly, this distribution is a random variable. Since we are primarily interested in the "extensive margin," that is, the split between empty and non-empty bins, we define d_i to be an

[8]Such products include, among others, bulky machinery and transportation equipment, but also smaller items such as valves, integrated circuits and other parts; books, apparel, and live animals.

indicator variable that takes the value of 1 if bin i is non-empty, $x_i > 0$, and 0 otherwise. The "intensive margin" will be given by the number of balls per non-empty bin.

Figure 1 shows that the balls-and-bins model looks as simple as it sounds. Figure 1A depicts five bins, ordered by size. Figure 1B shows a particular realization after throwing seven balls. Bins 3 and 5 are empty and thus we have $d_3 = d_5 = 0$.

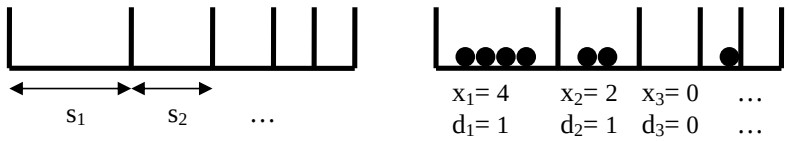

Fig. 1A **Fig. 1B**

Figure 1: Balls and bins

We can derive the key moments of the model analytically. For given bin sizes $\{s_1, s_2, ..., s_K\}$, the joint probability of a number of balls $\{x_1, x_2, ..., x_K\}$, is given by the multinomial distribution,

$$\Pr(x_1, x_2, ..., x_K) = \frac{n!}{x_1! x_2! \cdots x_K!} s_1^{x_1} s_2^{x_2} \cdots s_K^{x_K},$$

where $n = \sum_{i=1}^{K} x_i$. Note that, given a total number of balls n, the particular number of balls in two given bins, x_i and x_j, are not independent random variables. A ball falling in bin i is a ball less falling elsewhere, so it reduces the expected number of balls in bin j.

The model has a known probability distribution for the extensive margin. After dropping n balls the expected value of d_i is the probability that bin i receives at least one of those:

$$E(d_i|n) = 1 - \Pr(x_i = 0|n) = 1 - (1 - s_i)^n.$$

Each ball has a $(1-s_i)$ probability of landing elsewhere. Where a ball lands is an independent event, therefore the probability that none of n balls fall in a given bin i is $(1-s_i)^n$. Obviously, as the number of balls increases, it is less and less likely that any given bin remains empty. In the limit, as $n \to \infty$, the probability $\Pr(x_i = 0|n)$ is zero for all bins $i \in K$.

We denote the total number of non-empty bins by k,

$$k = \sum_{i=1}^{K} d_i.$$

Clearly, k is a random variable itself with $k \in \{1, 2, ..., K\}$. Since the number of non-empty bins is a sum of random variables, we easily obtain that

$$E(k|n) = \sum_{i=1}^{K} \left[1 - (1 - s_i)^n\right]. \tag{1}$$

This is our key statistic out of the balls-and-bins model. We will use it to derive many of the stylized facts on the extensive margin, both at the country and at the firm level. The comparative statics with respect to the number of balls are as one would expect: more shipments increase the expected number of non-empty bins. Note that the model is very stark in its prediction as the number of shipments grows large: the number of empty bins converges almost surely to zero.

The expected number of non-empty bins also depends on the distribution of bin sizes. Two bins of equal size fill up very fast: toss a coin ten times and with almost absolute certainty the coin will have turned heads some times and tails some others. But if a bin is, say, 10 times the size of the other, then a lot of balls may be needed to hit the small bin. This property of the model will play an important role later, as in many of the quantitative exercises the distribution of bin sizes is particularly skewed.

Formally, the expected number of non-empty bins (1) is convex in s_i for all $n \geq 2$. This implies that as we even out a bin-size distribution the expected number of non-empty bins increases.

Proposition 1. *Let $\{s_i\}$ be a bin size distribution and let*

$$\{\tilde{s}_i\} = \alpha\{s_i\} + (1-\alpha)1/K \tag{2}$$

for $\alpha \in [0,1]$. Then for all $n \geq 2$ the expected number of non-empty bins under $\{\tilde{s}_i\}$ is not less than under $\{s_i\}$.

Figure 2 plots the expected number of non-empty bins against the number of balls for 5 symmetric bins. The first few balls fall into distinct bins almost surely. Because of that, as long as balls are few, the number of filled bins is close to the number of balls and the relationship is essentially linear. In other words, most adjustment is on the "extensive margin." As the number of balls increases, it is more and more likely that balls fall in non-empty bins, and the number of filled bins trails the number of balls.[9] Eventually, all bins get filled, and the relationship flattens out. The remaining balls can only add to the "intensive margin." More formally, as $n \to \infty$, the number of non-empty bins converges to K.

In some occasions we will focus not on the extensive margin but on zeros, that is, the number of empty bins. It is, of course, trivial to derive the corresponding statistic:

$$K - E(k|n) = \sum_{i=1}^{K} (1-s_i)^n.$$

This is clearly decreasing in the number of balls, n.

We are also interested in the proportion of firms that sell only one product or serve only one country. To this end we derive the probability that a single bin contains all the balls or, equivalently, that exactly one bin is non-empty. Each ball had s_i probability of falling into

[9] The first ball falling to a non-empty bins comes very early, roughly in proportion to the square root of the number of bins, $\sqrt{K}$. This is sometimes known as the "birthday paradox:" it takes only 23 balls before any one of 365 equal-sized bins will contain two or more balls with probability 1/2.

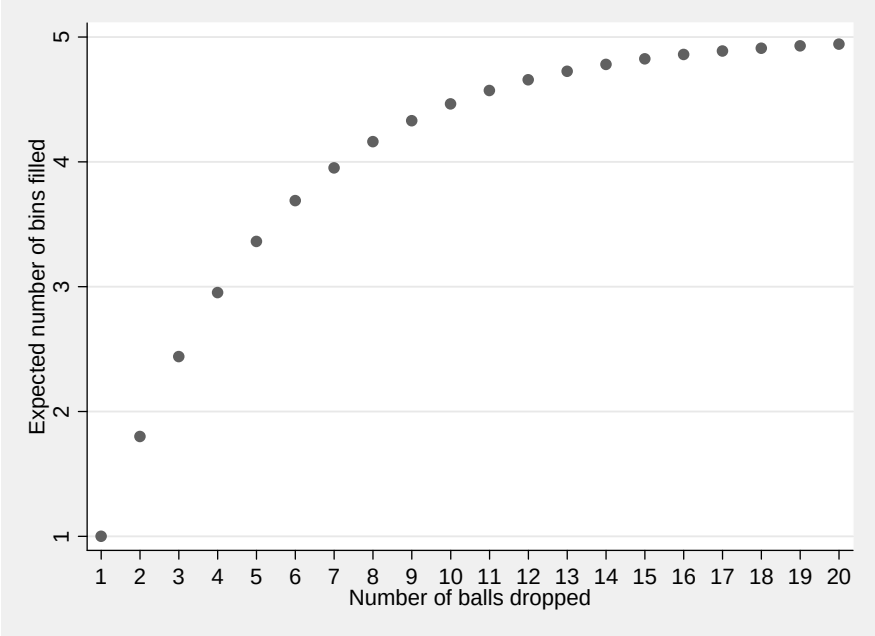

Figure 2: The extensive margin

bin i, so with probability s_i^n all balls fell in bin i. Of course this could happen to any of the K bins, but they are mutually exclusive events. Hence,

$$\Pr(k = 1|n) = \sum_{i=1}^{K} s_i^n. \tag{3}$$

The probability of a single non-empty bin decreases with the number of balls, n, and increases with the dispersion of bin sizes. Again, the model becomes degenerate as the number of balls grows: the probability of a single non-empty bin rapidly converges to zero.

3.1 Aggregate Statistics

So far we have derived the relevant moments for a single trade flow. Often, however, we will be interested in aggregate statistics that involve many trade flows. For example, we will look at the fraction of empty product categories for total U.S. exports as well as how this fraction varies across destinations.

In order to derive aggregate statistics we need to work with the dataset as a whole. The key difference is that each shipment is now classified along many dimensions. For example, in a dataset containing all U.S. export each shipment is given one HS code as well as one export destination out of 229 different countries.

We introduce a two-dimensional version of the balls-and-bins model, where each shipment is randomly assigned a classification in two systems, with T and K categories.[10] Visually,

[10] It is also easy to extend the model to higher-dimensional classification systems.

one can think of throwing balls over a T by K grid of bins as in Figure 3. Each classification system comes with its size distribution, $v_1, v_2, ..., v_T$ and $s_1, s_2, ..., s_K$, which in Figure 3 pin down the size of rows and columns, respectively. The probability of a given ball falling in the bin (i, j) is $v_i s_j$ so the ball is randomly and *independently* allocated across classification systems.

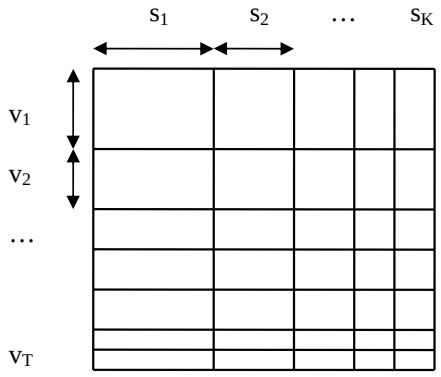

Figure 3: Balls and bins: T by K case

There is, conceptually, nothing different from the previous case: we can always re-arrange the grid into a row of bins of length TK. We can thus use the formulas derived before. For example, if we are interested in the expected total number of non-empty bins after throwing n balls, we have that

$$E(k|n) = \sum_{j=1}^{T} \sum_{i=1}^{K} [1 - (1 - v_i s_j)^n]. \tag{4}$$

The advantage of the two-dimensional version is that it allows us to easily work with *conditional* moments, for example, the number of empty product bins for a given country. For each realization of ball throws there will be a number of balls in each row and in each column, denoted $n_1, n_2, ..., n_T$ and $m_1, m_2, ..., m_K$, respectively. (Note that n_i or m_j may be zero.) Figure 4 illustrates. We can then ask the distribution of balls across columns $1, 2, ..., K$ within a given row with n_j balls. Since the classification in each system is independent, this is equivalent to the exercise we started the section with. Highlighted in Figure 4 is row $j = 4$. It is the same as in Figure 1, we only need to substitute n by n_4.

More interestingly, we can compute the statistics of interest given a distribution of balls $n_1, n_2, ..., n_T$ across rows. This will allow us, for example, to derive how the fraction of zero product-level bilateral flows varies across U.S. export destinations using the actual aggregate export flows. As discussed above, the conditional statistics for any given row are as in the first version of the model. Let k_t denote the number of non-empty bins in row t. We can thus easily construct the distribution of the expected number of non-empty bins per category

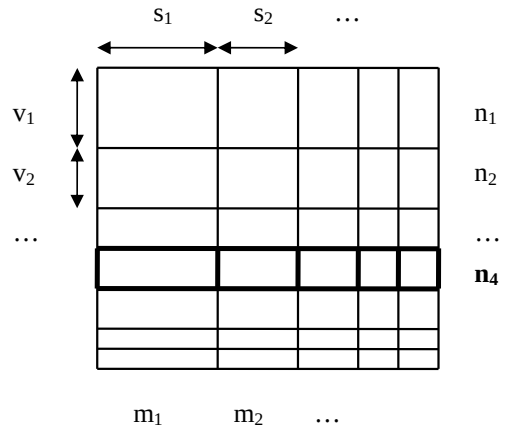

Figure 4: Balls and bins: T by K case

$t \in T$ using (1):

$$E(k_t|n_t) = \sum_{i=1}^{K}[1-(1-s_i)^{n_t}], \tag{5}$$

for $n_t \in \{n_1, n_2, ..., n_T\}$. The expected total number of non-empty bins given $\{n_1, n_2, ..., n_T\}$ is thus

$$E(k|n_1, n_2, ..., n_T) = \sum_{j=1}^{T}\sum_{i=1}^{K}[1-(1-s_i)^{n_j}]. \tag{6}$$

It is important to note that, since $\{n_1, n_2, ..., n_T\}$ is a random variable, conditional aggregate statistics will not coincide with the corresponding unconditional expectation $E(k|n)$ with $n = \sum_{j=1}^{T} n_j$.

Similarly, we can compute the probability of a single non-empty bin for each row using (3). Then we can derive the proportion of rows which are expected to contain a single non-empty bin. Since the number of empty bins is independent across rows,

$$\Pr(k_t = 1|n_1, n_2, ..., n_T) = \sum_{j=1}^{T}\sum_{i=1}^{K} s_i^{n_j}.$$

In practice we will sometimes approximate the distribution of balls across rows $\{n_1, n_2, ..., n_T\}$ with some parametric distribution. Appendix A shows how to compute aggregate statistics in this case. The Appendix also describes how to compute the fraction of balls that are expected to fall into single non-empty bin rows: this is useful when we want to derive the fraction of exports originated in single-product or single-destination exporters.

4 Zeros in product-level trade flows

The first data pattern we explore is the prevalence of product-level zeros (i.e., missing trade flows) in country-level exports. In other words, we look at the extensive margin of products when the units of observation are countries. We later discuss firm-level evidence.

We also take the chance to carefully describe how we map the data to the balls-and-bins model and back. The methodology is essentially the same for every exercise in the paper.

4.1 The facts

Baldwin and Harrigan (2007) recently reported that most potential destination-country product combinations are missing in U.S. exports. In 2005, the U.S. exported 8,867 different 10-digit Harmonized System categories to 229 different countries. Of these 2,030,543 potential trade flows, 1,666,046 (or 82%) were missing.[11] In other words, the average country only bought 18% of the 8,877 products the U.S. exports. Helpman, Melitz and Rubinstein (2007) look at the country-level zeros in the gravity equation. Of all potential country pairs, only about 50% have positive trade in either direction.[12]

Empirical regularity 1. *Most of the potential product-country export flows are zero — 82% of them in the U.S.*

Other levels of aggregation lead to a similar incidence of zeros. Table 4 reports the incidence of zeros for four classification levels. Zeros only stop being prevalent at the very broad, 2-digit level.

Classification	Number of bins	Incidence of zeros
10-digit	8,877	82%
6-digit	5,182	79%
4-digit	1,244	66%
2-digit	97	36%

Table 4: The incidence of zeros under different classifications

Baldwin and Harrigan (2007) then report how the incidence of zeros relate to the size of the importer and its distance to the U.S. Larger countries that are closer buy a larger variety of products. Here we replicate a regression close to their specification. For the top 99 trading partners of the U.S., we regress the incidence of a positive export flow on real GDP of the importer, real GDP per capita, and the distance of the importer from the U.S. Distance is divided in the same categories as in Baldwin and Harrigan (2007). We use a linear probability model, so coefficients can be understood as marginal effects.

[11] Haveman and Hummels (2004) report a similar incidence of zeros for imports.

[12] Hummels and Klenow (2005) also look at the product-margin of aggregate exports. They have a different measure of the extensive margin.

	Non-zero trade flow
Real GDP	0.081***
	(0.007)
Real GDP per capita	0.025**
	(0.009)
Distance = 0	0.330***
	(0.060)
0 < distance < 4000km	0.259***
	(0.027)
4000 < distance < 7800	omitted
7800 < distance < 14000	0.006
	(0.033)
Distance > 14000	0.054
	(0.037)
Observations	877,833
Clusters	99
R^2	0.24

Table 5: Non-zero flows and gravity – *The data (Baldwin and Harrigan, 2007)*

Table 5 reports the results.[13] Larger countries are more likely to import any given product. The same is true for richer countries. The incidence of non-zero flows decreases with distance: closer countries have more non-zero flows than farther countries (the omitted category is the intermediate distance).

Empirical regularity 2. *The incidence of non-zero product exports increases with destination-country size and decreases with distance.*

4.2 From the data to the model

In order to map the balls-and-bins model to the data, we proceed as follows. The trade flow of interest is the total U.S. exports to a given country, that is, we will have as many trade flows as destination countries (229). We measure the number of shipments going to a country to calibrate the number of balls. For example, Canada (the biggest importer) received 7.4 million shipments in 2005. Equatorial Guinea, the median buyer of U.S. exports, had 2,641 shipments.

The bins correspond to the 8,867 10-digit HS categories in which the U.S. exports at all. The size of each bin (s_i) is the share of each HS code in *total* U.S. exports in 2005. That

[13] Standard errors are clustered at the country level. These results are comparable to Table 4 of Baldwin and Harrigan (2007). The coefficients are similar, but not identical, potentially due to somewhat different real GDP measures.

is, we divide the number of export shipments in a given HS code with the total number of shipments (21.6 million).[14]

We then calculate the expected number of non-empty bins for each country using the previous formula (1),

$$E(k_c|n_c) = \sum_{i=1}^{8867}[1-(1-s_i)^{n_c}],$$

where n_c is the number of balls for country c and k_c is the number of non-empty HS categories in exports to country c. The expected number of non-empty bins overall is then

$$E(k|n_1, n_2, ..., n_{229}) = \sum_{c=1}^{229} k_c.$$

Note that we are computing the expectation conditional on the number of export shipments from the U.S. to each country. To retrieve the incidence of zeros we only need to substract from and divide by the appropiate number of categories; 8,867 if we are looking at the zeros for a particular trade flow, or 229×8,867 for overall U.S. exports.

The assumption underlying this calibration is that each destination country would buy the same basket of American products in exactly the same proportions. The only difference across countries is that smaller countries (such as Equatorial Guinea) have a smaller sample of shipments—drawn from the same distribution—than larger ones (such as Canada). Most trade theories are concerned with the differences in the structure of trade across countries: our calibration provides a neutral, atheoretical benchmark.

4.3 The model's predictions

We find that indeed most of potential product-level bilateral flows are zero in the model. The expected share of zeros is 72%, surprisingly close to the data (82%). That is, seven out of every eight zeros are to be expected given the sparsity of the data. Table 6 reports the predicted fraction of zeros for other levels of sectoral aggregation. The model's predictions track the observed incidence of zeros pretty well at all levels.

Classification	Number of bins	Data	Balls and bins
10-digit	8,867	82%	72%
6-digit	5,182	79%	68%
4-digit	1,244	66%	52%
2-digit	97	36%	23%
Section	21	16%	10%

Table 6: The incidence of zeros under different classifications

[14] We ignore the 121 HS codes for which we did not observe any shipment in 2005. It is possible to account for the missing bins with a simple specification: if anything, ignoring the missing bins reduces the expected fraction of zeros in the model.

Moreover the model matches quantitatively the pattern of zeros across flows in the data. To show this, we plot the number of exported products for each destination country against the total number of export shipments to that country in Figure 5. The dots represent the actual number of products in the data, the line is the predicted number of non-empty bins for each country. We already know that the balls-and-bins model somewhat underpredicts zeros, hence overpredicts the number of exported products, but the shape of the relationship to total exports is strikingly similar.[15]

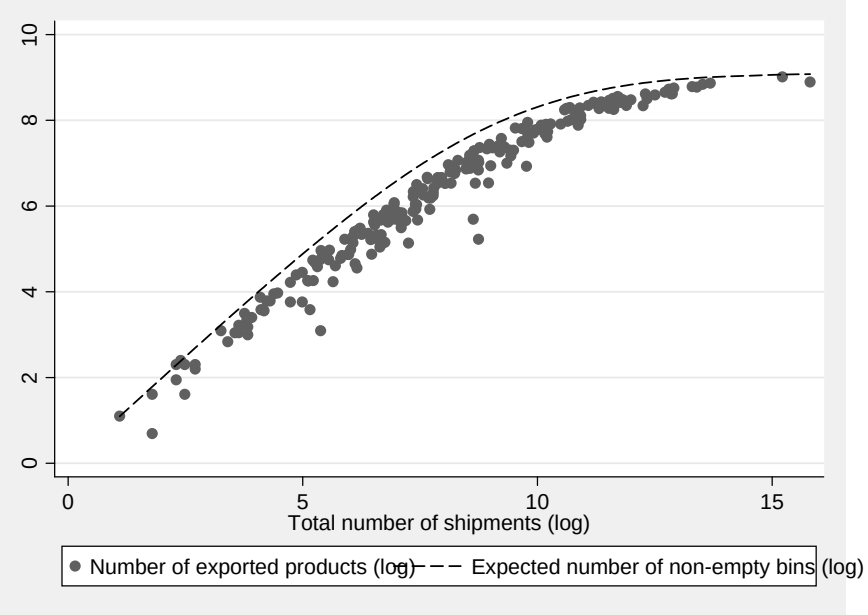

Figure 5: The number of shipments and the number of products

Zeros are more likely to occur in small export flows (those with few balls). This already suggests that non-zero flows may follow a gravity equation, as total export flows are well known to adhere to gravity. We then try to replicate the gravity specification in Baldwin and Harrigan (2007). We take the predicted probability of a non-zero flow $(1-(1-s_i)^{n_c})$ and regress it on the gravity variables such as country size and distance.[16] We emphasize that the balls-and-bins model has nothing to say about gravity, but given that the total number of balls (n_c) is highly correlated with the gravity variables, we may find some significant correlations.

The second column of Table 7 reports the results. For convenience, the first column repeats the regression on non-zero flows in the data. Bigger and closer countries are more likely to have a non-zero flow under the balls-and-bins model, just as in the data. Moreover, the magnitudes of the coefficients are surprisingly similar. The only exception are the two countries bordering the U.S. ("distance= 0"), Canada and Mexico. These seem to import more HS codes in the data than under the balls-and-bins model.

[15] In fact, in section 8, we show that a small change in the size of the ball achieves a perfect fit.

[16] We take the distance categories from Table 3 of Baldwin and Harrigan (2007). Real GDP is taken from the World Development Indicators.

	Non-zero trade flow	B+B model
Real GDP	0.081***	0.100***
	(0.007)	(0.008)
Real GDP per capita	0.025**	0.036***
	(0.009)	(0.010)
Distance = 0	0.330***	0.210***
	(0.060)	(0.032)
0 < distance < 4000km	0.259***	0.275***
	(0.027)	(0.032)
4000 < distance < 7800	omitted	omitted
7800 < distance < 14000	0.006	-0.014
	(0.033)	(0.035)
Distance > 14000	0.054	0.045
	(0.037)	(0.048)
Observations	877,833	877,833
Clusters	99	99
R^2	0.24	0.46

Table 7: Non-zero flows and gravity – *Balls and bins*

Quantitatively, the dispersion in flow and bin sizes plays an important role. In both cases the distribution is skewed, that is, some product categories and U.S. trade partners are very large, but the vast majority of product categories and trade partners are very small. It is precisely for the combination of latter (small country export for a small product category) than we have the missing trade flows in the data. And it is precisely for smaller bins and fewer balls that the model predicts the most zeros. For comparison, we find 53% zeros if we assume that all 8,867 HS codes have the same size.

Let us start with the distribution of bin sizes. The size of the average bin is $1/8867 = 1.13 \times 10^{-4}$. However, the size distribution across bins is rather skewed. The size of the median bin is 2.2×10^{-5}, about five times smaller than the average.

What is the source of this skewness across product categories? Category sizes may partly reflect the export specialization of the U.S., as higher exports of a product make that product category bigger. However, they are also affected by the nature of the classification system. As an illustration, we flag all product categories that contain either of the words "parts," "other," and "n.e.s.o.i." (for "not elsewhere specified or included") as *catch-all* categories. These are probably heterogeneous aggregates of various products. Of the 100 biggest categories, 69 are such catch-all. In contrast, only 8 of the 100 smallest categories are catch-all.

The skewness of trade flows is also important. Canada alone accounts for more than one fifth of total U.S. exports; the top five U.S. trade partners account for more than a half of the total.

It is important to emphasize that it is the dispersion in bin sizes, and not some particular bins being large and other small, that leads the balls-and-bins to predict so many zeros. To

check for this we re-run the model with the bin-size distribution calibrated to the HS shares of U.S. exports to Canada and Mexico only. These two trade flows contain very few zeros and so the size distribution of bins would not be affected by the large incidence of zeros in the data. The predicted fraction of zeros under these bin sizes is 76%. We find similar predictions if we use the shares of other countries or some exogenous bin-size distribution with skewness. In Appendix C, we show how to use theories to pin down bin sizes. We also show that zeros will be prevalent irrespective of the particular model used to calibrate bin sizes.

5 Zeros in firm-level trade flows

We can also ask about zeros in firm-level trade flows: we find a remarkably similar pattern. Bernard, Jensen and Schott (2007) report that the average exporting firm in 2000 shipped goods to only 3.5 countries from a total of 229.[17] In other words, 98 percent of potential firm–country trade flows are zero.

Again, the zero trade flows follow a well-defined spatial pattern. Firm-level export zeros are more frequent for small, distant countries. In other words, the number of firms exporting to a particular destination increases with country size and decreases with distance.

Table 8 reproduces column 2 of Table 6 from Bernard, Jensen, Redding and Schott (2007). The log number of exporting firms is regressed on log GDP of the destination country and its log distance from the U.S.

	Log number of exporting firms
Log GDP	0.71^{***}
	(0.04)
Log distance	-1.14^{***}
	(0.16)
Observations	175
R^2	0.74

Table 8: Exporting firms and gravity – *The data (Bernard, Jensen, Redding and Schott, 2007)*

We can calibrate the balls-and-bins model similarly to the previous exercise. The key difference is that now we need to create bins for *firms* as opposed to product categories. We take the number and sizes of exporting firms as given. In other words, we only try to explain the *allocation* of exporting firms across destination markets, we do not analyze the question of which firms export. That is done in Section 7.

The number of balls per destination country are again taken by counting the shipments going to that country. The total number of bins equals the number of exporting firms,

[17] Bernard, Jensen and Schott (2007), page 11.

167,217.[18] Because there are many more firm bins than we had product bins, we already expect that many more bins remain empty.

The size distribution of firm bins is calibrated as follows. We take the size distribution of firm-level export flows from Bernard, Jensen and Schott (2007). Their Table 3 contains a Lorenz curve of exports: What fraction of exports is accounted for by the top 1, 5, 10, 25, and 50% of exporters? The following table reports the fraction of firms and the average exports in each of these percentile bins.

Export percentile	Fraction of firms	Average exports
$99-100$	0.01	\$413 million
$95-99$	0.04	\$15.5 million
$90-95$	0.05	\$3.37 million
$75-90$	0.15	\$886,000
$50-75$	0.25	\$184,000
$0-50$	0.50	\$20,500
Total	1.00	\$5.11 million

Table 9: The distribution of firm-level exports – *Bernard, Jensen and Schott (2007)*

There is a striking skewness in the distribution of exports across firms. While the average firm exports \$5.11 million, the bottom half of *exporters* export only \$20,500.[19] The top 1% of exporters account for 80.9% of total exports.

We approximate the distribution of exports with a lognormal distribution with mean $\mu = 11$ and standard deviation $\sigma = 3$. This matches the mean exports of \$5.11 million and has a median exports of \$59,300. The lognormal distribution does a good job in matching the Lorenz curve reported in Bernard, Jensen and Schott (2007).[20] The size distribution of bins will then inherit this lognormal distribution with the additional normalization that the bin sizes add up to one.

The underlying assumption here is that all countries could be served by all the exporting firms, only that small countries draw a smaller sample of shipments and may end up with fewer firms. We assume no systematic sorting of firms into destination markets, hence this exercise provides a natural benchmark.

The balls-and-bins model predicts that 96 percent of the potential firm×country trade flows is going to be zero. This is very close to the 98 percent we see in the data. What about the distribution of firm zeros across destinations? For each country, we can calculate the expected number of non-empty firm bins. We can then regress (the log of) this number on GDP and distance.[21]

[18] Bernard, Jensen and Schott (2007), Table 2.

[19] Note that this is conditional on having positive exports. A large fraction of firms have zero exports and are omitted from this analysis.

[20] A Pareto distribution does similarly well and leads to similar results.

[21] We take GDP (in current-price USD) from the World Development Indicators. We take distance from the bilateral distance dataset of CEPII.

Table 10 presents the results. For convenience, we reproduced the regression estimate by Bernard, Jensen, Redding and Schott (2007) in the first column.[22] The coefficient estimates in the simulated regression are similar to the ones in the actual data. Just as in the data, bigger, closer countries are served by more exporters: the more balls are thrown, the less bins will be left empty.

	Log number of exporting firms	Log number of non-empty bins
Log GDP	0.71***	0.56***
	(0.04)	(0.03)
Log distance	−1.14***	−0.95***
	(0.16)	(0.13)
Observations	175	181
R^2	0.74	0.75

Table 10: Exporting firms and gravity – *Balls and bins*

Interestingly, given that there are so many firm bins, the skewness in firm exports does not play as big a role as it did for product bins; most firm bins are going to remain empty anyway. We also calibrated firm bins to the distribution of overall sales in manufacturing (Table 12), which resulted in 93% of firm–country bins remaining empty and a 0.60 elasticity of the number of firms exporting to a country with respect to country size. When using 167,217 symmetric firm bins, we got 82% empty bins and an elasticity of 0.72. The results seem to be driven by the fact that the number of exporting firms is far larger than the number of shipments for a typical country. (Recall that the median country received only 2,641 shipments.)

Again, this does not imply that the assignment of firms to destination markets is indeed random. The only conclusion we can draw is that the variation in market size is so huge given the sparsity of the data that any model that accounts for both can match the gravity equation of firms — *even* if the assignment of firms is random.

6 Firm-level export patterns

We then turn to evidence on the extensive margin at the level of individual exporting firms. In this section we ask how many products firms export and how many destinations they serve. Note that the universe of interest is the set of *exporting firms*, because the empirical facts are usually reported only for firms that have some exports.[23] This way we can use the balls-and-bins model to understand these moments despite the split between exporters and non-exporters being very different from random (as we will see in the next section).

[22]Because we may have used somewhat different data sources, especially for distance, we have 181 destination countries in contrast to the 175 countries of Bernard, Jensen, Redding and Schott (2007). The differences in coverage, however, are likely very small.

[23]Though export datasets can be merged with domestic data such as in Bernard, Jensen, and Schott (2007) and Eaton, Kortum and Kramarz (2004).

The key stylized facts about the extensive margin at the firm level are that while most firms exports a single product to a single country, the bulk of exports is done by multi-product, multi-destination exporters.[24]

To start with, 42% of the firms export only a single product, defined by the 10-digit HS code. While being a little less than half of the total firms, they account for a tiny fraction of total exports, 0.4%.

Empirical regularity 3. *42% of firms export a single product (defined as a 10-digit HS code). These firms account for only 0.4% of exports.*

A similar pattern exists for firms that export to a single country. These firms account for a little less than two thirds of the total, but still amount to a small fraction of total exports.

Empirical regularity 4. *64% of firms export to a single country. These firms account for only 3.3% of exports.*

But perhaps the most striking fact corresponds to the fraction of firms that export a single product to a single country. These firms represent 40% of the total exporters yet account only for a miniscule 0.2 % of total exports.

Empirical regularity 5. *40% of firms export a single product to a single country. These firms account for only 0.2% of total exports.*

We use the same bin sizes as for the aggregate flows to calibrate the bins. The 10-digit HS codes are calibrated to the aggregate export share of each HS code in total U.S. exports in 2005. The size of each country bin is calibrated to the share of that country in total U.S. export flows.[25] The following table lists the five biggest country bins.

Country	Share
Canada	0.341
Mexico	0.189
Japan	0.041
United Kingdom	0.035
Germany	0.030

Table 11: The five biggest country bins

We assume each firm has a different number of export balls. Because we do not have data on the number of shipments at the firm level, we calibrate the number of balls to the distribution of exports across firms, reported in Table 9. We approximate the distribution of exports with a lognormal distribution with $\mu = 11$ and $\sigma = 3$. This matches the mean exports of \$5.11 million and has a median exports of \$59,300. Corresponding to the average

[24]The following facts are for U.S. merchandise trade in 2002, reported in Bernard, Jensen, Redding and Schott (2007), Table 4.

[25]The assumption here is that the structure of aggregate exports did not change too much between 2002 and 2005.

size of export shipments in 2000, we take each $36,000 of export sales to represent one ball, rounding up. Because of the extreme skewness in the distribution of exports by firm, many firms will end up with just one export ball.

The predicted fraction of single-product exporters is 43%. This is very close to the actual fraction in the data (42%). The predicted fraction of exports coming from single-product producers is 0.3%, close to the actual 0.4%.

Let us see how the balls-and-bins model manages to reproduce the fraction of single-product exporters with such precision. In the model practically all single-product exporters have only one ball. This is because with 8,867 HS codes, the second ball is very likely to fall into an HS category different from the first one. Only 0.3% of two-ball exporters are single-product exporters. The key to understanding the incidence of single-product exporters is that there are plenty of very small exporters, who export $36,000 or less.

The model underpredicts the data with respect to the fraction of single-country exporters, 44% in the model for 64% in the data. The reason is that the fraction of single-country exporters falls sharply with firms with the second and third balls. For example, the model predicts that only 11% of firms with two shipments export both of them to Canada (and less than 4% to Mexico). We conjecture that the fraction of relatively-large exporters that export only to Canada (and possibly Mexico) is significantly higher in the data than in the model, indicating possible large market or proximity effects.

Last but not least, the balls-and-bins is right on the spot with respect to the fraction of single-product, single-country exporters.

Note that a fraction of 40% of single-product, single-country exporters implies that most single-product exporters are also single-country exporters, and vice versa. Is this surprising? The balls-and-bins model makes it clear the fact follows from the presence of many small exporters. Almost all single-product exporters have only one ball, and these are all going to be single-country exporters. And this exactly what we see in the data. The conditional probability of single-country exporters among single-product exporters is 99.9% in the model, close to the 96% in the data.

We conclude that the split between single-destination, single-product firms and the rest is very much in line with what we would expect given the skewness of the exporter distribution.

Of course, this does not mean there are no interesting facts in the data! First, without all the reported facts we would have not been able to establish the importance of the skewness of the export distribution. Second, there are interesting deviations from randomness. We have already pointed to the fact that exporters to NAFTA countries exhibit some differences: they are more likely to export multiple products and are larger than expected.

7 Exporting firms

We now move on to the differences between exporting and non-exporting firms. It is a well-established fact that exporters are few and they are significantly larger than non-exporting firms.

According to the survey by Bernard, Jensen, Redding and Schott (2007), only 18% of manufacturing firms export at all. The fraction drops to about 3% when all firms outside manufacturing are included.[26] Other studies have confirmed the scarcity of exporters. Plant-level statistics also fall in the same pattern. For the quantitative exercise, we stay with the fraction of exporters among U.S. manufacturing firms.

Empirical regularity 6. *Exporters are few — only 18% of manufacturing firms export in the U.S.*

The second fact is that exporters sell significantly more than non-exporters — about 4.4 times more than non-exporters according to Bernard, Jensen, Redding and Schott (2007). Again, firms outside manufacturing and plant-level evidence reveal similar patterns. That exporters are few and they are larger than non-exporters have been confirmed in other datasets, in other settings, with other measures of size.

Empirical regularity 7. *Exporters are large — among U.S. manufacturing firms, exporters sell 4.4 times more than non-exporters.*

We follow essentially the same steps as before to map the model to the data. The key difference is that now the output flow will include total sales, not only exports. As before we obtain the number of balls n per firm by dividing its total sales by \$36,000 and rounding up.[27]

We thus need data on total sales per firm in order to construct the distribution of balls (π_n). Unfortunately we do not have direct access to this data for the U.S. The 2002 Statistics of U.S. Businesses of the Census, though, reports the number and total sales of firms in each of eight size bins (see Table 12).

Size bin	Fraction of firms	Average sales
0–\$100,000	0.145	\$55,600
\$100,000–\$500,000	0.305	\$257,000
\$500,000–\$1 million	0.144	\$718,000
\$1–5 million	0.257	\$2.26 million
\$5–10 million	0.060	\$6.84 million
\$10–50 million	0.063	\$19.3 million
\$50–100 million	0.010	\$56.4 million
over \$100 million	0.015	\$670 million
Total	1.000	\$13.2 million

Table 12: The distribution of firm sales in manufacturing – *Census*

As it is well known, there is enormous skewness in the size distribution of firms. Whereas 59% of firms sell less than \$1 million, the average firm sells \$13.2 million. We approximate

[26] See Table 2 in Bernard, Jensen, Redding and Schott (2007). The data is from the 2002 Economic Census.

[27] In the previous section we used evidence on the average shipment value to pin down the "ball size." We have no direct equivalent for total sales. In Section 8 we document the results for different balls sizes.

the distribution of firm sales by a lognormal distribution with $\mu = 13.4$ and $\sigma = 2.44$. This corresponds to median sales of \$680,000 and average sales of \$13.2 million. We also experimented with fitting a Pareto distribution with similar results.

To distinguish between exporters and non-exporters we only need two bins: one for domestic sales, the other for foreign sales. In the 2002 Economic Census, there were 297,873 manufacturing firms. Their total receipts amounted to \$3.94 trillion. Exports of manufactured goods amounted to \$545 billion in 2002.[28] That is, 13.9% of manufacturing receipts come from exports. This pins down the size of the domestic bin at 0.861 and the size of the export bin at 0.139.

Our finding here is that exporters are much less common in the data than they would be if sales were randomly allocated between the domestic and abroad market: 74% of the manufacturing firms should be exporting according to the balls-and-bins model, compared to 18% in the data.

It is easy to see why the model overpredicts the fraction of exporters. The probability that a firm with n balls of total sales does not export is

$$(1 - s)^n = 0.86^n.$$

Because where each ball ends up is independent of the distribution of existing balls, each \$36,000 has quite a high chance to end up going to a foreign market. Among the smallest firms, that is, with one ball, 14% of them export. This is already a very high number given that only 18% of total manufacturing firms export. It obviously gets worse. Almost half of the firms with a paltry \$100,000 of total sales should export. It is clear that this is not the case in the data: exporting is a more unlikely event than the random assignment of sales across markets would indicate.

The unconditional probability of exporting is convex in the fraction of exports, s, so if there is heterogeneity across industries, the aggregate economy will contain fewer exporters than predicted by the average s. However, at the 3-digit level, this heterogeneity is rather small, and does not change the exporting probability substantially.

The model's prediction for the exporter's size premium is also off. Surprisingly, though, the model *overpredicts* the size of exporters. That is, despite exporters being four fifths of total firms in the model for one fifth in the data, the model predicts that exporters are 34 times larger than non-exporters on average, while in the data they are "only" 4.4 times larger. In terms of the exporter size premium, in log sales, the difference in the model is 3.53, for 1.48 in the data.[29]

To understand why exporters are larger under balls-and-bins than in the data, note that balls-and-bins implies that the largest firms export with a probability close to one. Even the median firm that has \$660,000 dollars in sales, corresponding to 18 balls, exports with probability 0.93. The skewness of the firm sales distribution then implies that the average firm in the top half of the distribution is much larger than any of the non-exporters, who mainly come from the bottom half. The fact that the size premium is smaller in the data

[28] Bureau of the Census, FT-900, "International Trade in Goods and Services." We converted all figures to 2000 dollars.

[29] In Appendix A we formally derive the exporter's size premium and include a parametric example.

suggests that the sorting of exporters and non-exporters by size is not as strong as predicted by the model. In other words, there have to be a substantial fraction of very large firms that do not export – in contrast with the model.

Summarizing, what do we learn from the balls-and-bins miss? First, the split between exporters and non-exporters is not just a matter of chance: there is some economic force that makes the two types of firms quite different. Second, the data has a weak sorting of exporters by size: exporters are smaller, not larger, than expected.[30]

8 Robustness

In our analysis of zeros in product-level trade we had access to the observed number of shipments for each flow of interest. Unfortunately shipment data is not always available at the desired level. For example in Section 6 we had to approximate the number of shipments by dividing the firm-level trade flow into balls of $36,000 (the value of the average export shipment in the U.S. in 2000). In this case we want to know how sensitive are results with respect to the ball-size calibration as well as how to account for ball-size heterogeneity.

In this Section we discuss our results for different ball-size calibrations as well as a useful specification for ball-size heterogeneity. We focus on product-level trade flows since in this case we can compare the results under the calibration with the results with the actual number of shipments.

8.1 Different ball-size calibration

If shipment data are not available, we can specify a ball-size and convert a trade flow into a discrete number of balls. Using the actual average size of export shipments ensures the total number of shipments in the exercise is the same as in the data.

We redo here our results for different ball-size calibrations. First we experiment with ball sizes equal and larger than the average size of an export: $36,000, $100,000, and $500,000. 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. Second, we report results for smaller ball sizes $18,000 and $2,500 (the lowest observed value of export transactions given the reporting rules of the Census Bureau). These calibrations illustrate neatly the slow rate of convergence to the continuous model: 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.

Table 13 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.

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. The

[30]Note that a fixed cost model, with a simple cut-off rule, has a very strong sorting of exporters by size. Indeed, were it to match a 18% exporter fraction, exporters would be orders of magnitude larger than non-exporters.

Moment	Data	Ball size $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 13: Ball-size calibrations: $36,000; $100,000; and $500,000

fraction of single-product and single-country exporters increases both in number and in 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.

For the prevalence of zeros in product-level trade flows we note that a slightly higher ball-size calibration would lead to an almost perfect fit of the data. Figure 6 replicates Figure 5 but instead of using the actual number of shipments we divide flows into balls according to ball-size calibrations $36,000, $100,000, and $500,000. A small increase in the ball size not only increases the overall incidence of product-level zeros to match the one in the data, but also achieves a perfect fit in terms of the relationship of zeros and total export.

Table 14 shows our quantitative results for smaller ball sizes, between $36,000 and $2,500. The last column describes the limit as ball size shrinks to zero and the indivisibility becomes negligible. This is to illustrate how the model would work without indivisibilities.

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

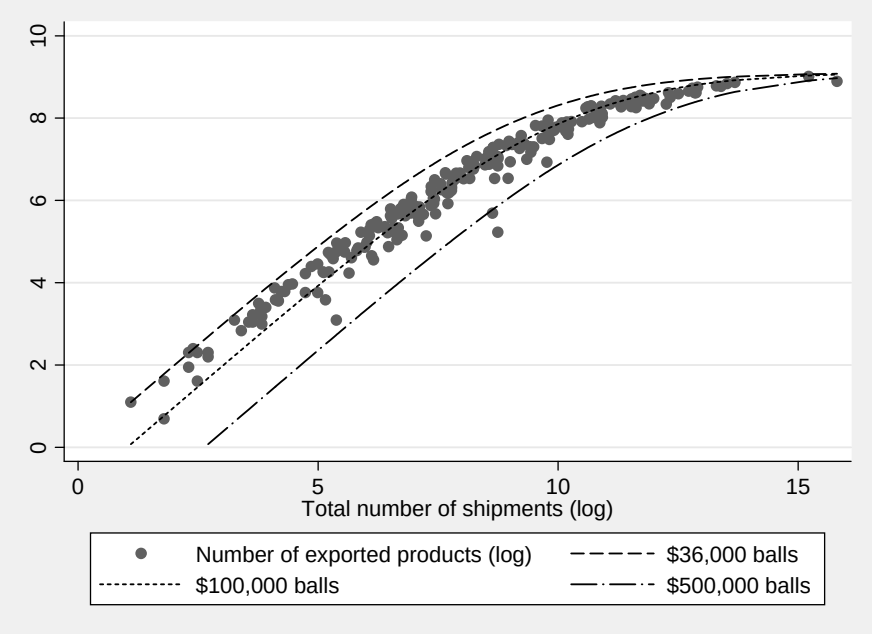

Figure 6: The incidence of zeros with different ball sizes

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

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.

With smaller balls, single-product and single-country exporters become much less prominent. This is because the more balls make it less likely that a firm will only have balls in one product bin or one country bin.

Finally, the fraction of firms that export increases as the ball size decreases. In the limit, without indivisibilities, all firms are expected to export in the balls-and-bins model.

8.2 Heterogeneity in ball sizes

Another concern with the lack of shipment data is that we may be missing variation in shipment size across destinations or products. However, if the shipment size does not vary systematically with the category of interest, then the analysis is very robust. Suppose the allocation of balls to bins is the same as described in Section 3, but now each ball has a random size, z, drawn from a common distribution $F(z)$ independently for each ball. We assume the distribution has support $(0, Z]$ (with Z being arbitrarily large), mean μ_z and a finite variance σ_z^2. How do we convert dollars into balls and vice versa?

Given the number of balls in a particular bin, x_i, the dollar value of the trade flow in that bin is

$$Y_i = \sum_{n=1}^{x_i} z_n,$$

which is the sum of independent and identically distributed random variables, where the limit of the summation is also random. This is a stopped-sum distribution that has support $[0, nZ]$, where n is the maximum number of balls.[31] Below we derive some key moments of this random variable.

Firstly, and perhaps most obviously, the probability of the flow being zero equals the probability of the bin containing no balls,

$$\Pr(Y_i = 0) = \Pr(X_i = 0) = (1 - s_i)^n.$$

This because all balls have a positive size.

The mean trade flow is just the mean number of balls times the mean ball size,

$$E(Y_i) = E\left[E(Y_i|x_i)\right] = E\left[x_i E(z_n)\right] = n s_i \mu_z.$$

This is again independent of σ_z^2, the heterogeneity in ball sizes.

The variance of the trade flow can be similarly derived as

$$\mathrm{Var}(Y_i) = \mathrm{Var}(x_i)\mu_z^2 + E(x_i)\sigma_z^2 = n s_i (1 - s_i)\mu_z^2 + n s_i \sigma_z^2.$$

[31] See Chapter 9 of Johnson, Kemp and Kotz (2005) on stopped-sum distributions.

The variance increases in the variance of the number of balls, $ns_i(1-s_i)$, but also in the variance of ball size, σ_z^2. Intuitively, the heterogeneity in ball sizes is another source of uncertainty about the total size of a trade flow.

An important consequence of the heterogeneity in ball sizes is that, with positive probability, some of the non-zero trade flows will be smaller than the average ball size. These correspond to bins with one or a few balls that are smaller than the average. This suggests that the lower tail of trade flows is not necessarily informative about the average ball size.

9 Conclusion

Sparse datasets *do* contain a lot of information. Ignoring the sparsity, however, can easily lead to mistake patterns arising mechanically for systematic stylized facts. Nowhere this problem is more acute than in the analysis of the extensive margin. A zero in the data is no more than an absence of observations for a particular category. A zero in the data must thus be evaluated against the overall number of observations and the relevance of the category of interest. While we would easily find many fundamental reasons why the U.S. did not export enriched uranium to North Korea in 2005, there were no observed U.S. shipments of enriched uranium to U.K. either—which is not really surprising given that there were only 59 shipments overall.

The balls-and-bins model provides a parsimonious and, more importantly, atheoretical account of the sparsity in the data. The structure of the model parallels that of the data: there is a given number of observed shipments, and each of them will be classified into a unique category; some trading partners are larger than others, and some products are traded more often than others. This is indeed all the structure in the model. From there the assignment of a shipments to a category is an independent and identically-distributed random event. Independence also governs the construction of the bin sizes. For example, the probability of a given country–product pair is just the product of the respective shares in aggregate trade.

Thus whenever the balls-and-bins model matches a particular fact we will fail to identify the relevant economic decisions driving trade. In other words, if the zeros in the data are explained by the sparsity of the dataset, then they should not be the basis to favor any fundamental reason for the lack of shipments.

Importantly, the balls-and-bins model also works in the opposite direction: whenever the model fails to reproduce a fact we know that strong economic forces are at play. In the paper we have discussed the incidence of exporters among domestic firms in some detail. Moreover, the balls-and-bins model provides a *quantitative* benchmark so we can better evaluate models against the data. For example, the data shows excess zeros in U.S. exports, so there is a demand for models that put some structure on the top of sparsity.

To summarize, we hope that our approach can be used in future empirical work using massive micro-level trade datasets. Recent transaction-level datasets are very detailed, and trade flows are typically broken down by firms, 8 or 10-digit product codes, and destination

countries.[32] By their very nature, these datasets are *sparse* in the sense that the number of observations is low with respect to the number of categories of interest. The balls-and-bins model provides a natural benchmark for working with sparse datasets, and can be easily adapted to any empirical application.

References

[1] Anderson, M. A., Ferrantino, M. J. and Schaefer, K. C.: 2004, Monte Carlo Appraisals of Gravity Model Specifications, Working Paper.

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

[3] Baldwin, R. and Harrigan, J.: 2007, Zeros, Quality and Space: Trade Theory and Trade Evidence, NBER Working Paper No. 13214.

[4] 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.

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

[6] 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.

[7] 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.

[8] Damijan, J. P., Polanec, S. and Prasnikar, J.: 2007, Outward FDI and Productivity: Micro-evidence from Slovenia, *World Economy* **30**(1), 135–155.

[9] Deardorff, A. V.: 1998, "Determinants of Bilateral Trade: Does Gravity Work in a Neoclassical World?," *in* The Regionalization of the World Economy, by Jeffrey Frankel (ed). University of Chicago Press.

[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.

[12] Eaton, J., Kortum, S. and Kramarz, F.: 2007, An Anatomy of International Trade: Evidence from French Firms, Working Paper.

[32] Bernard, Jensen and Schott (2007) describe the customs dataset of the U.S.; Eaton, Kortum and Kramarz (2004) for France; Mayer and Ottaviano (2007) for Belgium; Damijan, Polanec and Prasnikar (2004) for Slovenia; Halpern, Koren and Szeidl (2007) for Hungary; Eaton, Eslava, Kugler and Tybout (2007) for Colombia.

[13] Ellison, G. and Glaeser, E. L.: 1997, Geographic Concentration in U.S. Manufacturing Industries: A Dartboard Approach, *Journal of Political Economy* **105**(5), 889–927.

[14] Evenett, S. and Keller, W.: 2002, On Theories Explaining the Success of the Gravity Equation, *Journal of Political Economy* **110**(2), 281–316.

[15] Ghosh, S. and Yamarik, S.: 2004, Are Regional Trading Arrangements Trade Creating? An Application of Extreme Bounds Analysis, *Journal of International Economics* **63**(2), 369–395.

[16] Ghosh, S. and Yamarik, S.: 2004, Does Trade Creation Measure Up? A Reexamination of the Effects of Regional Trading Arrangements, *Economics Letters* **82**(2), 213–219.

[17] Ghosh, S. and Yamarik, S.: 2005, A Sensitivity Analysis of the Gravity Model, *International Trade Journal* **19**(1), 83–126.

[18] Halpern, L., Koren, M. and Szeidl, A.: 2007, Imports and Productivity, Working Paper.

[19] Helpman, E., Melitz, M. and Rubinstein, Y.: 2007, Estimating Trade Flows: Trading Partners and Trading Volumes, *Quarterly Journal of Economics*, forthcoming.

[20] Hummels, D., Lugovskyy, V. and Skiba, A.: 2008, The Trade Reducing Effects of Market Power in International Shipping, *Journal of Development Economics* forthcoming.

[21] 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.

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

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

[24] Keller, W.: 1998, Are International R&D Spillovers Trade-Related? Analyzing Spillovers among Randomly Matched Trade Partners, *European Economic Review* **42**(8), 1469–1481.

[25] Mayer, T. and Ottaviano, G.: 2007, The Happy Few: The Internationalization of European Firms, Bruegel Blueprint Series. Volume III.

[26] Melitz, M. J.: 2003, The Impact of Trade on Intra-industry Reallocations and Aggregate Industry Productivity, *Econometrica* **71**(6), 1695–1725.

Appendix

A Aggregation

In this subsection we formally derive the aggregate statistics given a set of trade flows. To be precise, suppose 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.[33] As in Section 3, each shipment can be classified into one of K categories.

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

$$E(k|n_1, n_2, ..., n_T) = \sum_{n=1}^{N} \pi_n \sum_{i=1}^{K} [1 - (1 - s_i)^n] = \sum_{i=1}^{K} \sum_{n=1}^{N} \pi_n [1 - (1 - s_i)^n]. \tag{7}$$

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. That is, aggregation preserves the properties discussed in Section 3.

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.

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

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.

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

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.

B Deriving the exporter's size premium

We now derive the size-exporting relationship formally. 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 parametrize 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.

C Mapping models into sparse data

In the main text we claimed that a stylized fact fails to identify the relevant economic theory if it cannot falsify the balls-and-bins model. In this appendix we elaborate further on this claim and show how to use balls and bins to map trade models into sparse data.

A trade theory makes predictions about the trade flow between firm i and destination country c in product s during a time interval of length Δt:

$$Y_{ics,\Delta t} = g(\theta, \mathbf{X}_{ics}, \Delta t),$$

where θ is a vector of model parameters, $\mathbf{X}_{ics}$ is a vector of firm, country and product characteristics (such as GDP, capital abundance, trade costs, productivity, trade barriers etc).[34]

[34]Some of these characteristics may be unobservable, for example, productivity.

Most trade theories take the form of a set of continuous flows, that is, as we reduce the period length Δt, the trade flow scales down proportionally. In this sense, all trade is similar to oil flowing through a pipeline. If we measured oil imports for 1 minute instead of 2 minutes, we would see half as much oil flowing through.

In practice, however, we observe discrete shipments. That is, over a time period, trade data consist of a finite list of transactions. In order to use the data to evaluate the model, we need to map the continuous flows into observables.

Suppose we observed just one shipment. Under the model, the probability that it is a product s going from firm i to country c is

$$\Pr(i, c, s|\theta, \mathbf{X}) = \frac{Y_{ics}}{\sum_{i'} \sum_{c'} \sum_{s'} Y_{i'c's'}}.$$

This probability is a function of parameters and observables,[35]

$$\Pr(i, c, s|\theta, \mathbf{X}) = \pi_{ics}(\theta, \mathbf{X}).$$

It is possible that π_{ics} takes the value of zero. In that case there is zero probability that we ever observe a shipment in this category. We call a trade flow (i, c, s) for which $\pi_{ics} = 0$ a *fundamental zero*.

Of course, if the observed shipment is bananas, it does not mean that all trade is in bananas only: we cannot equate all unobserved flows with fundamental zeros. As we collect data on more shipments, we must evaluate the likelihood that a particular trade flow remains unobserved.

Suppose then that we collect data on n shipments. We can calculate the probability that none of the n observations contains trade flow (i, c, s) as

$$\Pr\left[\neg(i, c, s)|\theta, \mathbf{X}\right]^n = \left[1 - \pi_{ics}(\theta, \mathbf{X})\right]^n. \tag{8}$$

Here we take the n shipments to represent an iid random sample from the system described by the model $\{\pi_{ics}\}$.

What if we had an infinite amount of data, $n \to \infty$? Then the observed share of trade flow (i, c, s) would converge in probability to the "true" share, π_{ics}. Only fundamental zeros would remain unobserved:

$$\lim_{n\to\infty} \left[1 - \pi_{ics}(\theta, \mathbf{X})\right]^n > 0 \text{ iff } \pi_{ics}(\theta, \mathbf{X}) = 0.$$

In a sparse dataset, though, this asymptotic result is of little use. Instead, we should use the likelihood function (8) with n equal to the number of observations.

Notice the similarity to balls and bins. Say we equate the number of balls to the number of observations, and use the theory to construct the bin sizes, $\{\pi_{ics}\}$.[36] The resulting likelihood is exactly the one given by equation (8).

[35] Because the function g is homogeneous of degree one in Δt, the ratio does not depend on Δt.

[36] In the main text, our objective was to provide an atheoretical account of data sparsity. To this end, we assumed that there is no systematic association between firms, products and countries, so that $\pi_{ics}(\theta, \mathbf{X}) = \alpha_i \beta_c \gamma_s$. Under this independence hypothesis all firms export all products to all countries in the same proportion.

As long as we account for the sparsity, all models will share the same likelihood function (8). Different models can still have different predictions for missing trade flows, as bin sizes, $\{\pi_{ics}\}$, vary across models.

However, sparsity is the key to matching the zeros in the data. The precise bin size distribution does not substantially affect the prevalence of zeros. To see this, consider an alternative to the exercise in Section 4. For each destination country, we take the number of shipments observed and fill up as many product categories as possible. That is, each ball is forced to fall into an empty bin. This allocation clearly results in the lowest possible number of empty bins, irrespective of the bin sizes. We find that more than half of the bins remain empty. For example, the number of shipments to the median country is less than one third of the total number of products, so more than two thirds of the product bins remain empty.

Therefore zeros will be prevalent across all models that match aggregate trade flows across countries. For example, the Helpman–Krugman and the Eaton–Kortum models are both succesful in matching aggregate trade flows as they both predict the gravity equation. Hence both models are consistent with the large number of zeros we see in the data.

The prevalence of zeros hence fails to identify the true model. Having said that, we believe there is scope for quantitative analysis. For example, the balls-and-bins model underpredict zeros by 10 percentage points (72% vs 82%). It would be very interesting to know which structural models can outperform balls and bins.

D 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."[37]

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).

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

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.

Baldwin and Harrigan (2007)

Baldwin and Harrigan (2007) 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 (2007)

Helpman, Melitz and Rubinstein (2007) 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 (2007)

Eaton, Kortum, and Kramarz (2007) 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.