

A Balls-and-Bins Model of Trade*

Roc Armenter and Miklós Koren[†]

February 2010

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. Typically the number of observations—that is, total shipments—is low relative to the number of possible classifications—e.g., countries and product codes. 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 model cannot reproduce the small fraction of exporters among U.S. firms. We discuss the implications for identifying the relevant model of the extensive margin in trade.

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 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,

*For useful comments we thank two anonymous referees and the editor, as well as George Alessandria, Arnaud Costinot, Alan Deardorff, Jonathan Eaton, 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, Princeton, the SED, the NBER Summer Institute, Michigan, Stanford, MIT, and the University of Zurich. We also thank Jennifer Peck for excellent research assistance. Much of this research was carried out while Koren was visiting the International Economics Section of Princeton University, and he gratefully acknowledges their hospitality. The views expressed here do not necessarily reflect the views of the Federal Reserve Bank of Philadelphia or the Federal Reserve System.

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

and how many products they send to how many destinations. This, in turn, has led 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 different years in several countries.

We argue that the sparse nature of trade data is crucial to understand these stylized facts. Data are sparse if the number of observations—that is, total shipments—is too low relative to the number of possible classifications—country and product code pairs. Sparse data have some distinctive features owing to the low number of observations: most categories have very few or no observations, and the distribution of the number of observations per category is unimodal at a low count.

Trade data are collected through customs forms, one for each export shipment. There were about 22 million export shipments originating in the U.S. in 2005. This may seem a number safe from small-sample problems. However, there are 229 countries and 8,867 product codes with active trade, so a shipment can have more than 2 million possible classifications. Most of the traded categories had only 1 shipment during the year, a clear sign that the data are sparse. There are too few shipments, partly because some products are indivisible, and partly because of the constraints of transportation technology.²

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

¹The 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.

²The typical shipment is rather small; the median shipment value is \$12,800. Bulky and valuable products are mostly shipped by themselves, hence shipment value is fully determined by the value of the product itself. Smaller and less valuable products are grouped together in batches. The value of these batches show some (quantitatively small) variation depending on the transportation technology used. Section 2 provides more details.

for the extensive and intensive margin, respectively. We can derive analytically the relevant moments. We show how to compute the prevalence of zeros and how it varies with the number of balls and 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. To this end 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).³ For the dimension of choice (product codes or destination countries) we construct the bin size distribution using category totals. 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, and exports to some countries more than others (gravity). However, 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. 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. In other words, given what we observe for aggregate trade flows (by country, by product), a large number of zeros is to be expected.

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 left tail 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.⁴ 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.

We must emphasize that in a dense dataset—i.e., with many observations relative to the number of categories—the balls-and-bins model would be unable to match any stylized fact on the extensive margin. Indeed all bins will be non-empty and the predictions for the extensive margin will be trivial.

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 if a fact cannot falsify the balls-and-bins model, it will fail to identify the right model of the extensive margin. For example, as long as a model correctly predicts the gravity

³Unfortunately 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.

⁴The 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 of total exporters	42%	43%
Share of total exports	0.4%	0.3%
Single-destination exporters		
Fraction of total exporters	64%	44%
Share of total exports	3.3%	0.3%
Single-destination, single-product exporters		
Fraction of total exporters	40%	43%
Share of total exports	0.2%	0.3%
Exporters in U.S. manufacturing		
Fraction of 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.

equation, it will also match the pattern of zeros across countries once the sparsity of the data is accounted for. Similarly any model that reproduces the distribution of export sales will be able to match the facts concerning multi-product and multi-destination exporters. The qualification is important: the balls-and-bins model embeds the economic determinants of the data used in the calibration. Other economic forces need not have played any role in shaping the outcomes.

Importantly, we 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. For a model to match the data it will be necessary to postulate a systematic relationship between firms, products, and markets. Hence it is in the split between exporters and non-exporters that we will be able to identify the relevant trade model of the extensive margin.

We view the balls-and-bins model as a useful statistical tool that can quantitatively discern the interesting facts in sparse datasets. 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.

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 discusses what we can and cannot learn about trade theories from sparse data. Finally, Section 9 concludes. The Appendix provides extensions to the main model, and describes in detail the datasets used in the cited papers.

2 Trade data are sparse

This Section explains what we mean by “sparse” data. We briefly lay out the typical features of a sparse dataset and argue that these are prevalent in trade data. We then briefly explore the causes of this sparsity.

In a categorical dataset, each observation is classified into one of several disjoint categories. We call a dataset sparse if the number of observations is too low relative to the number of categories. Statistical inference on the relative frequencies of categories requires a much larger sample than in non-categorical datasets. Moreover, the sample size needed grows very fast with the number of categories K . For example, the number of observations must be of order $O(K \log(K))$ for maximum likelihood estimates to exist.⁵

The distinctive feature of a sparse dataset is that most categories have very few or no observations. Moreover, the distribution of the number of observations per category is unimodal at a low count (zero to four observations). For example, categories with a single observation may be the most common, followed by categories with none or two observations, then categories with three observations, and so on.

Categories with no observations are typical of sparse data; yet they are not unique to them. For example, if the data generating process is censored at some minimum scale, we will observe plenty of empty categories even if there are many observations. However, the resulting data will have no or few categories with a low, but positive, number of observations; generally, the distribution of shipments per category will be bi-modal (at zero and at some large positive number above the minimum scale).

Our first observation is that disaggregated trade data is actually categorical. The Census Bureau collects trade data through customs forms, one for each export shipment.⁶ For each shipment, the export declaration records the destination country, the product being shipped, and the selling firm (among other variables), which are categorical variables. The units of observations are the shipments themselves, so the number of observations, at its most disaggregated level, equals the number of shipments.

According to the “U.S. Exports of Merchandise” published by the Census Bureau, there were 21.6 million export shipments in 2005. Whether this number of observations yields sparse or dense data depends also on the number of categories.⁷ 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. That makes about 2 million potential product–country categories, or about one for each 11 shipments. If we further break exports shipments by firm the number of categories exceeds the number of observations by orders of magnitude. The sparsity problem remains severe even if we accumulate several years of data.

Next we verify the symptoms of sparsity in U.S. export data. As it is well known, there are many potential trade flows (country-product pairs) for which no shipment was observed. But it is the categories with shipments that provide the stronger evidence. Table 2 reports the first few elements of the distribution of shipments per country–product category with observed shipments. More than one quarter of the categories had only one shipment in 2005.

⁵See Section 9.8 of Agresti (2002) for a summary discussion of statistical inference in sparse categorical data.

⁶See the Data Appendix for more detailed description of the Census export data.

⁷That the Census collects all export shipments in any given year is irrelevant: the data remains a sample of the underlying process. Clearly five minutes, or a week, would be deemed insufficient. Whether a quarter, a year, or five years are appropriate sampling periods depends on the resulting number of observations.

The second most frequent shipment number is two. Sixty percent of categories had five or fewer shipments in a year.

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

2.1 The causes of sparsity

We have seen that disaggregated trade data shows signs of sparsity, suggesting the number of observations is too low. Why are there no more shipments in a year’s worth of trade data? To understand this, we study the value of individual shipments and explore how it varies with a number of observable covariates.⁸ We are interested in what determines whether goods are shipped in small packages worth, say, \$1,000, or in bulky containers worth, say, \$500,000. If the typical shipment is large, there will be few observations, and the resulting dataset will be sparse.

There is substantial variation in the values of shipments. The typical shipment is rather small; the median shipment size is \$12,800. The vast majority, 94%, of products have a shipment size below \$50,000. 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. The smallest shipment is given by the reporting threshold of \$2,500.

We find two major determinants of shipment value: the physical characteristics of the product, and the transportation technology used. In fact, there is a clear distinction between products whose shipment value is given by the product’s characteristics and those for whose the transportation mode is important. Bulky and valuable products are mostly shipped by themselves, hence shipment value is fully determined by the value of the product itself. Smaller and less valuable products are grouped together in batches. The value of these batches show some (quantitatively small) variation depending on the transportation technology used.

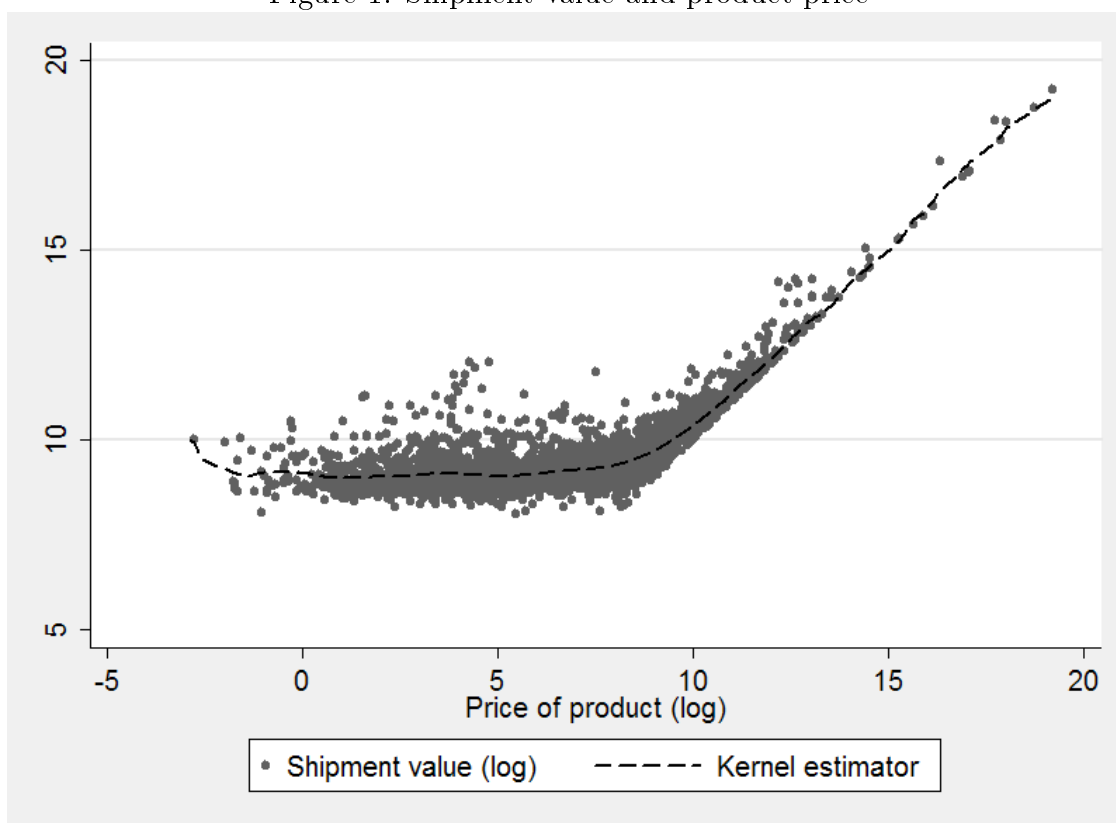
Much of the variation in shipment size is explained by product characteristics. A regressions of log shipment value on product fixed effects yields an R^2 of 0.37. Not surprisingly, some products are bulky by their very nature. The biggest shipments include aircraft (\$42 million), spacecraft (\$5 million), tanker ships (\$15 million) and floating drilling platforms (\$5

⁸We use the monthly “U.S. Exports of Merchandise” DVDs published by the Census Bureau in its original detail: broken down by districts of origin, month of shipment and mode of transport. See Appendix for details.

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.

When we look at countable products (the ones that report “numbers,” “pieces,” “pairs,” or “dozens” as the units of quantity), we can see how the price of the product is related to its typical shipment size.⁹ Figure 1 plots the median shipment size of each indivisible product (dollar per shipment, log scale) against the median price of the product (dollar per unit, log scale). The 45-degree line corresponds to the case when the product is shipped by itself: the typical value of the shipment is equal to the typical price tag of the product. The figure also reports a nonparametric curve estimating how shipment size depends on product price.

Figure 1: Shipment value and product price



The clear pattern emerging from Figure 1 is that products that are bulky and or highly valuable (worth more than about \$12,000) are generally shipped individually. The right-hand side of the plot lines up very much with the 45-degree line.¹⁰ Among these bulky items, the value of the shipment is fully determined by the value of the product itself.

There is a substantial variation of shipment value even within narrowly defined product categories. These categories mask a tremendous amount of heterogeneity in the products

⁹Such 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.

¹⁰When we sort by weight instead of prices, we get very similar results.

included. Take, for example, “self-propelled combined harvesters,” a 10-digit product category. As bulky and expensive machines (the median price tag is \$155,000), harvesters are typically shipped one by one. Yet, there is a substantial variation in the unit value of these individual shipments, because the price tags of the machines themselves vary greatly. The interquartile range of unit values is \$45,000 to \$320,000 for the previous example.

It is important to note that the typical “bulky” product is much smaller than a tanker ship, a drilling platform or a combined harvester. The \$12,000 threshold that is apparent from Figure 1 is surpassed by practically all cars, most industrial equipment, and numerous other products, accounting for a total of 45 percent of exports of countable products.

What happens to products worth less than \$12,000? Products that are worth less than \$12,000 are shipped in batches of around \$12,000, probably because it is not profitable to send low-value shipments. For these products, on the left-hand side of Figure 1, we see little variation in shipment value. That variation can be explained by the physics and economics of transportation. For example, men’s shirts (an indivisible, but low-value product) are typically shipped in containers, and it is not worth to ship a container that is half empty. In fact, the same product is shipped in bigger batches if shipped by sea than if shipped by air.

We explored how shipment size depends on the mode of transportation, and some basic characteristics of the destination country (its size, level of development and distance to the U.S.). Shipments by air are 36 percent smaller than shipments by ground (the reference category). In contrast, shipments by sea are 36 percent larger. Given the physical capacities of vessels, trucks, and airplanes, these differences are not surprising. Shipment sizes are not significantly correlated with GDP per capita of the destination country, nor its distance to the U.S. There is a weak correlation with country size, but is quantitatively very small.^{11,12}

Our overall assessment is that, while logistics play a role in determining the size of shipments, the variation generated by differences in transportation and storage are small and uncorrelated with typical economic variables of interest. That is, for most questions in international trade, the specifics of the transportation technology can be safely ignored.¹³ More importantly for our purposes, because the economic environment matters so little for shipment size, we are confident that trade datasets from other countries or other years will suffer from the same sparsity problem.

¹¹As an illustration, we consider shipments to Germany and Belgium. Conditional on the type of product and the mode of transport, Germany receives only 4 percent bigger shipments. For the median product, which has a shipment value of \$12,800, the difference is only \$500.

¹²The results are very similar if we measure the size of shipments by their weight rather than their value. The only difference is that countries adjacent to the U.S. (Canada and Mexico) receive lower-weight shipments.

¹³Logistics, though, can play a role at the aggregate level. Alessandria, Kaboski, and Midrigan (forthcoming) show how inventory building affects the response of trade to large devaluations. Logistics also play a crucial role for a number of other questions, such as timeliness (e.g., Evans and Harrigan, 2005 and Djankov, Freund, and Pham, 2006) and market power of intermediaries (Hummels, Lugovskyy and Skiba, 2009).

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 $i \in \{1, 2, \dots, K\}$. 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, $\{x_1, x_2, \dots, x_K\}$. 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 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 2 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 2: Balls and bins

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

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

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 $n \rightarrow \infty$, 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 $k \in \{1, 2, \dots, K\}$. Since the number of non-empty bins is a sum of random variables, we easily obtain that

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

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*

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

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 3 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.¹⁴ Eventually, all

¹⁴The 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.

bins get filled, and the relationship flattens out. The remaining balls can only add to the “intensive margin.” More formally, as $n \rightarrow \infty$, the number of non-empty bins converges to K .

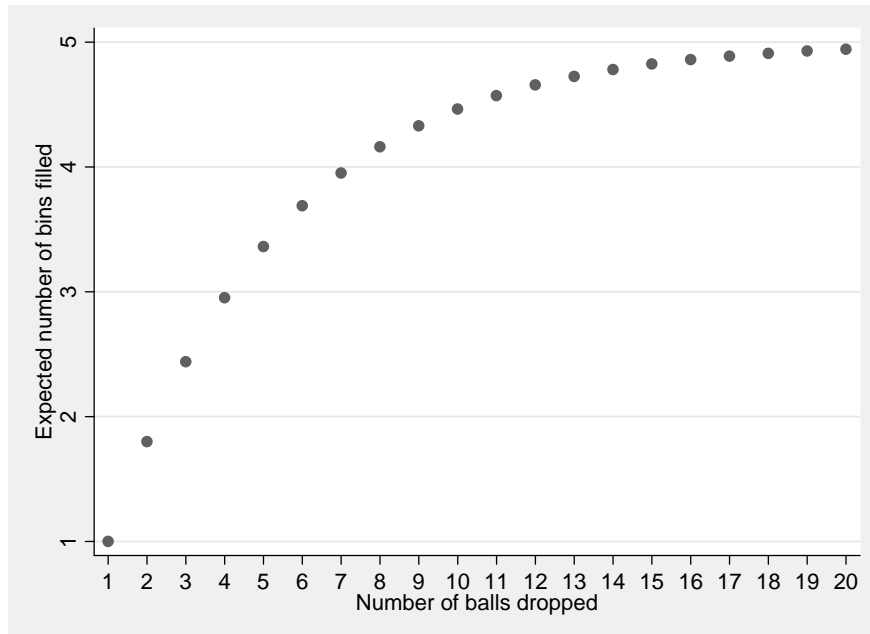


Figure 3: The extensive margin

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 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.¹⁵ Visually, one can think of throwing balls over a T by K grid of bins as in Figure 4. Each classification system comes with its size distribution, v_1, v_2, \dots, v_T and s_1, s_2, \dots, s_K , which in Figure 4 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.

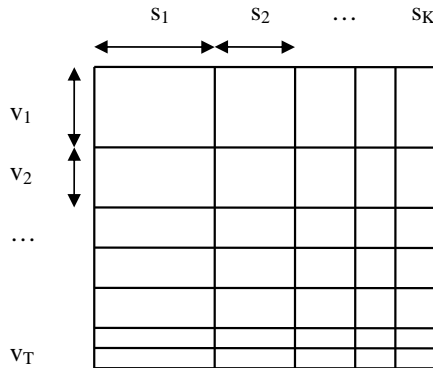


Figure 4: Balls and bins: T by K case

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

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

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

¹⁵It is also easy to extend the model to higher-dimensional classification systems.

column, denoted n_1, n_2, \dots, n_T and m_1, m_2, \dots, m_K , respectively. (Note that n_i or m_j may be zero.) Figure 5 illustrates. We can then ask the distribution of balls across columns 1, 2, \dots , 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 5 is row $j = 4$. It is the same as in Figure 2, we only need to substitute n by n_4 .

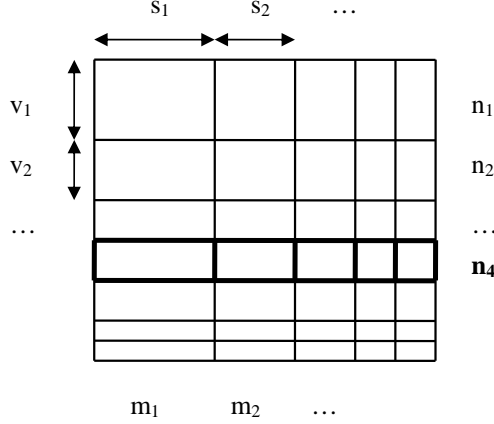


Figure 5: Balls and bins: T by K case

More interestingly, we can compute the statistics of interest given a distribution of balls n_1, n_2, \dots, 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 $t \in T$ using (1):

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

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

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

It is important to note that, since $\{n_1, n_2, \dots, n_T\}$ 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,

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

In practice we will sometimes approximate the distribution of balls across rows $\{n_1, n_2, \dots, n_T\}$ 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.¹⁶ 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.¹⁷

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 3 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 3: 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

¹⁶Haveman and Hummels (2004) report a similar incidence of zeros for imports.

¹⁷Hummels and Klenow (2005) also look at the product-margin of aggregate exports. They have a different measure of the extensive margin.

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.

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

Table 4 reports the results.¹⁸ 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

¹⁸Standard 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.

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 is, we divide the number of export shipments in a given HS code with the total number of shipments (21.6 million).¹⁹

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

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

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 subtract from and divide by the appropriate number of categories; 8,867 if we are looking at the zeros for a particular trade flow, or $229 \times 8,867$ 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 5 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.

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 6. 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.

¹⁹We 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.

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 5: The incidence of zeros under different classifications

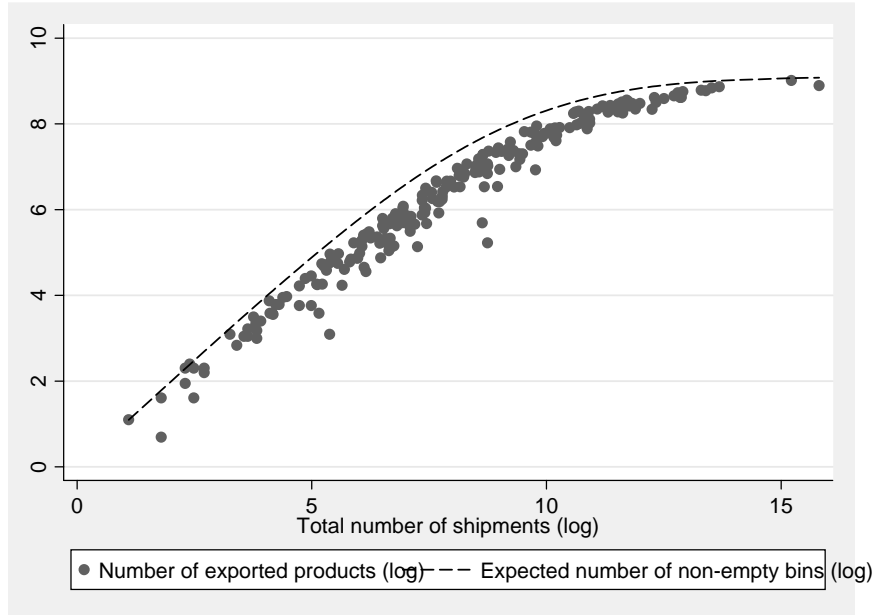


Figure 6: The number of shipments and the number of products

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.²⁰ 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 6 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

²⁰We take the distance categories from Table 3 of Baldwin and Harrigan (2007). Real GDP is taken from the World Development Indicators.

more HS codes in the data than under the balls-and-bins model.

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

Table 6: 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.

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. For comparison, we find 53% zeros if we assume that all 8,867 HS codes have the same size.

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.

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.

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. In order to shut down any shipment size variation across destinations, we computed the fraction of zeros by dividing export flows (in dollars) by the average shipment value, \$36,000. The fraction and pattern of zeros are virtually unchanged.

We also replace the actual trade flows with the trade flows *predicted* by the gravity equation in Table 4. We find 66% zeros, the number being slightly lower than the baseline result mainly due to the reduced country sample. This exercise also allows us to pin the key determinant of the skewness in trade flows. Assuming distance has no effect on trade flows reduces the number of zeros only slightly to 64%. In contrast, assuming all countries have identical size brings the fraction of zeros down to 30%. Thus it is the skewness in country size, through its impact on export flows, that it is most important for the calibration.

4.4 What have we learnt?

What do we conclude for this exercise? First, the results suggest that the economic forces shaping the distribution of trade across products and destinations are *sufficient* to explain the prevalence and patterns of zero in U.S. exports. Thus these facts are not useful to identify theories specifically about the extensive margin. Section 8 discusses two models with radically different implications for the number of zeros under dense data, that is, a very large number of observations. Both models are capable to match the aggregate trade patterns, and are calibrated to do so. We show then how these two models have virtually identical predictions regarding zeros under sparse data. It is worth recalling that the balls-and-bins model predicts no zeros at all for dense data. Unfortunately the sparsity problem does not go away by combining some years of data or looking at six digits HS codes.

On a more positive note, we note that a quantitative evaluation of the models could elicit some identification. We underpredict the fraction of zeros as well as the impact of distance. Both effects are relatively small so we would need trade models capable of matching the data with precision. We also believe there are other “cuts” of the data that will be more robust to sparsity. Later we will have an example of this for firm-level data.

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 229.²¹ In other words, 98 percent of potential firm–country trade flows are zero.

²¹Bernard, Jensen and Schott (2007), page 11.

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 7 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 7: 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, 167,217.²² 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). We take a look at a Lorenz curve of exports: What fraction of exports is accounted for by the top 1, 5, 10, 25, and 50% of exporters? Table 8 reports the fraction of firms and the average exports in each of these percentile bins.

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 \$20,500.²³ 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).²⁴ The size distribution of

²²Bernard, Jensen and Schott (2007), Table 2.

²³Note that this is conditional on having positive exports. A large fraction of firms have zero exports and are omitted from this analysis.

²⁴A Pareto distribution does similarly well and leads to similar results.

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

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.²⁵

Table 9 presents the results. For convenience, we reproduced the regression estimate by Bernard, Jensen, Redding and Schott (2007) in the first column.²⁶ 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*** (0.04)	0.56*** (0.03)
Log distance	-1.14*** (0.16)	-0.95*** (0.13)
Observations	175	181
R^2	0.74	0.75

Table 9: Exporting firms and gravity – *Balls and bins*

²⁵We take GDP (in current-price USD) from the World Development Indicators. We take distance from the bilateral distance dataset of CEPII.

²⁶Because 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.

Interestingly, the skewness in firm exports does not play as big a role as it did for product bins: given that there are so many, most firm bins are going to remain empty anyway. We calibrated firm bins to the distribution of overall sales in manufacturing (Table 11), 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 frequency and pattern of zeros in firm-level trade flows.

6 Firm-level export patterns

We now 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.²⁷ 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).

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.²⁸

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.

²⁷Though export datasets can be merged with domestic data such as in Bernard, Jensen, and Schott (2007) and Eaton, Kortum and Kramarz (2004).

²⁸The following facts are for U.S. merchandise trade in 2002, reported in Bernard, Jensen, Redding and Schott (2007), Table 4.

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.²⁹ 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 10: 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 8. 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 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.

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 possibly 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, and the small fraction of exports that they account

²⁹The assumption here is that the structure of aggregate exports did not change too much between 2002 and 2005.

for. 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.

Our results suggest that the skewness of the exporter distribution is key to understand the split between single-destination, single-product firms and the rest. In particular, the left tail of the export distribution—the small exporters—is what enables the balls-and-bins model to match the data. This property of the distribution is not specific to exporters. For example, our results do not change when we calibrate the model to match the observed skewness in *domestic* sales for the U.S. In contrast, the balls-and-bins model underpredicts the data once we censor the left tail. Interestingly, the right tail properties of the distribution have little bearing in the results as virtually all firms selling more than \$100,000 are predicted to be multi-country, multi-product exporters. We thus conclude that trade models capable of matching the fraction of small exporters in the data will also be able to reproduce the firm-level export patterns discussed here.³⁰ As we shall see in the next Section, the split between exporters and non-exporters is not due to sparsity. There are thus strong economic forces, yet to be fully understood, shaping the distribution of exporters and thus the firm-level facts discussed here.

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 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.³¹ 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.

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

³⁰Most of the literature has not paid much attention to small exporters, with the exception of Arkolakis (2009).

³¹See Table 2 in Bernard, Jensen, Redding and Schott (2007). The data is from the 2002 Economic Census.

That exporters are few and they are larger than non-exporters have been confirmed in other datasets, in other settings, and with other measures of size.

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. 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 Statistics of U.S. Businesses of the Census for year 2002, though, reports the number and total sales of firms in each of eight size bins (see Table 11).

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 11: 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 the distribution of firm sales by a lognormal distribution with $\mu = 13.2$ and $\sigma = 2.66$. 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. In the 2002 Economic Census, there were 297,873 manufacturing firms. As before we obtain the number of balls n per firm by dividing its total sales by \$36,000 and rounding up.³²

To distinguish between exporters and non-exporters we only need two bins: one for domestic sales, the other for foreign sales. Total receipts amounted to \$3.94 trillion for manufacturing firms in the 2002 Economic Census. Exports of manufactured goods amounted to \$545 billion in 2002.³³ 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.

We find that exporters are much less common in the data than in the model: 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

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

³²In 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.

³³Bureau of the Census, FT-900, “International Trade in Goods and Services.” We converted all figures to 2000 dollars.

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. 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. Almost half of the firms with a paltry \$100,000 of total sales should export. A median firm has a 95% chance to export. It is clear that this is not the case in the data: exporting is a more unlikely event than the balls-and-bins model 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.

It is reassuring that the balls-and-bins misses: we firmly believe that the decision to export is the key extensive margin in the trade data, and reflect its very own economic determinants. We can assess the economic significance of the departure from the balls-and-bins predictions by introducing a minimum scale requirement for exporters.³⁴ More precisely, we proceed to shut down the foreign-sales bins for which the balls-and-bins model predicts fewer shipments than a threshold k^* . Note that where each shipment ends up is no longer independent of the distribution of existing shipments: the minimum scale implies that whenever we observe a shipment going abroad we expect this firm to have a “disproportionate” share of export shipments.

We find that in order to have a share of 18% exporters we need to set $k^* = 24$, that is, firms predicted to have \$850,000 or less in foreign sales must be taken to have *zero* export sales. This is a very large number that leaves no doubt that there are significant barriers to export participation.³⁵

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.³⁶

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 suggests the data has a weak sorting of exporters by size: exporters are smaller, not larger, than expected. In other words, there have to be a substantial fraction of very large firms that do not export – in contrast with the model.

³⁴Clearly we have economies of scale and Melitz (2003) in mind.

³⁵For comparison purposes, we repeat the exercise for country-product trade flows. In order to increase the share of zeros from 72%—the balls-and-bins prediction—to 82%—the data—we only need to shut down trade flows below \$40,000.

³⁶In Appendix A we formally derive the exporter’s size premium and include a parametric example.

8 Weak identification

Throughout the paper we have claimed that a stylized fact will fail to identify the relevant economic theory if it cannot falsify the balls-and-bins model. In order to back this claim up, we discuss two very simple models that differ radically in their theoretical implications for the extensive margin. We then derive the models' fraction of unobserved (product-country) trade flows under sparse data. In doing so, the mechanics of the balls-and-bins model will become apparent. It is then no surprise that we find that both models have nearly identical predictions.

8.1 A simple Krugman model

We start with a simple model with differentiated country varieties, loosely following Krugman (1980) and Helpman and Krugman (1985). There are J distinct products, which we will identify with HS-10 product codes in the data. Consumers in each country value consumption bundles according to a Cobb-Douglas utility function

$$U = \sum_{j=1}^J (C^j)^{\alpha_j}$$

where $\alpha_j > 0$ for all j and C^j is a composite of all countries' varieties of product j , given by a constant elasticity of substitution (CES) subutility function,

$$C^j = \left(\sum_{l=1}^L (C_l^j)^\rho \right)^{1/\rho}.$$

The number of countries is L , and $\rho = 1 - \frac{1}{\sigma}$ is the CES parameter governed by the elasticity of substitution, σ . We assume the elasticity of substitution is constant across products and countries.

Let p_l^j be the price of the U.S. variety of good j that country l faces. The familiar CES demand formulation for U.S. exports of product j by country l is

$$\log X_l^j = \log \alpha_j + \log Y_l - (\sigma - 1) (\log p_l^j - \log P_l),$$

where Y_l and P_l are the country l income and ideal price index respectively. We assume a very simple specification for trade costs, $\tau_l = \tau_0 d_l^\gamma$ (where d_l denotes the distance to country l), and predict trade flows with

$$\log X_l^j = \kappa_l + \log \alpha_j + \log Y_l - (\sigma - 1)\gamma \log d_l \tag{7}$$

where κ_l collects the constant terms across products and import destinations.

The specification (7) embodies the gravity equation, so we know that the model can match the pattern of U.S. trade across export destinations. We set the coefficient on distance, $(\sigma - 1)\gamma$, to 0.8 as reported by Anderson and van Wincoop (2003). We use real GDP for

country income—see Section 4.3 for details with the data. For ease of exposition we exclude multilateral resistance terms.³⁷ The overall demand for product j is shifted by $\log \alpha_j$, so the model can replicate the skewness in trade flows across product categories. We treat α_j as a fixed effect and let the sector composition of U.S. exports pin down the parameter values.

In the model all possible product-country trade flows will be positive: the assumption of differentiated products imply there are no perfect substitutes, and the CES demand is positive for all finite prices.

8.2 A simple model with fixed costs

We now introduce a fixed cost of exporting, so the model can generate zeros. The economies of scale imply that the U.S. will export to a given destination only if the demand for the particular product is large enough. We start by characterizing the *potential* U.S. exports in product j to country l , \tilde{X}_l^j . The demand structure is the same as in the previous model and potential exports are then given by (7),

$$\log \tilde{X}_l^j = \kappa + \log \alpha_j + \log Y_l - (\sigma - 1)\gamma \log d_l.$$

There is a fixed cost associated with each product-country trade flow, ϕ . As a result a U.S. firm will find it profitable to export product j to country l only if $\tilde{X}_l^j \geq \phi/(1 - \rho)$. If this is the case, then actual exports are equal to the potential flow, $X_l^j = \tilde{X}_l^j$. Otherwise, there is no trade in that particular category, $X_l^j = 0$. A richer model, as Melitz (2003), would allow to calibrate the parameter ϕ to some dollar values. For illustrative purposes, we just set ϕ such that half of the possible U.S. export flows are zero. We maintain the previous values for all remaining parameters, ensuring that if we set $\phi = 0$, the fixed cost model boils down exactly to the Krugman model.

8.3 Model predictions with sparse data

The previous models, and indeed most trade models, take the form of a set of continuous trade flows. That is, if we were to evaluate the models at different frequencies, the predicted export flows would just scale up or down proportionately with the frequency. In this sense trade in the models is similar to oil flowing through a pipeline at a constant rate.

The data, however, consists of a finite number of observations, corresponding to the transactions in a given time period, usually a year. We bridge the gap between the theory and the data by re-interpreting the model's predictions as the likelihood that a given transaction is included in a finite sample. According to the model, the probability that a sample of one shipment is a product j going to country l is

$$\pi_l^j = \frac{X_l^j}{\tilde{X}}.$$

³⁷Including multilateral resistance terms, as well as a dummy for FTA countries (Canada, Mexico), did not bring any substantial change in the results.

where \bar{X} is the U.S. total exports as predicted by the model. This probability is a function of the model’s parameters and observable variables, like distance and GDP. In the Krugman model, all probabilities are positive since $X_l^j > 0$. Under the fixed cost model, though, there is zero probability to observe a transaction in certain categories: we call a trade flow for which $\pi_l^j = 0$ a *fundamental zero*. However, not all unobserved flows will be fundamental zeros. For a sample of n shipments, the probability of not observing a shipment in a particular category is given by

$$\Pr(\hat{X}_l^j = 0|n) = (1 - \pi_l^j)^n \tag{8}$$

where \hat{X}_l^j is the realized trade flow. If the model predicts a fundamental zero for exports in product j to destination l , then clearly $\Pr(\hat{X}_l^j = 0|n) = 1$ for all n . But we would need an infinite amount of observations to be completely sure that an unobserved flow is a fundamental zero, that is,

$$\lim_{n \rightarrow \infty} \Pr(\hat{X}_l^j = 0|n) > 0 \text{ iff } \pi_l^j = 0.$$

In a sparse data set, though, this asymptotic result is of little use.

We now proceed to compute the previous models’ predictions for zeros with a sample of $n = 22 \times 10^6$ observations, roughly the number of shipments in a year worth of U.S. exports. For each model we derive the likelihood π_l^j for all trade flows and then compute the expected fraction of unobserved trade flows. Table 12 collects the results of both models, both for dense data ($n = \infty$) and sparse data ($n = 22 \times 10^6$). For reference we also include the predicted fraction of zeros for the balls-and-bins model.³⁸

	Dense data	Sparse data
Krugman model	0 %	66 %
Fixed costs	50 %	67 %
<i>Balls and bins</i>	0 %	63 %

Table 12: Predicted fraction of unobserved product-level trade flows

As Table 12 makes clear, the abundance of zero trade flows in a sparse data set does not provide a basis to favor the fixed cost model: the Krugman model, with no fundamental zeros, has nearly identical predictions regarding the fraction of unobserved trade flows. The fraction of zeros in the data is around 66%. Simply put, for the vast majority of unobserved trade flows, we cannot tell whether it is a fundamental zero or just a trade flow that we should not expect to observe in a given year. In the Krugman model, the gravity equation predicts that most trade flows are tiny as the importing countries are small and distant. Hence these trade flows have a tiny, albeit positive, probability of being observed. Not surprisingly then, most go unobserved in any given year. Why does the fixed cost model not predict many

³⁸The fraction of zeros is smaller than reported in Section 4 mainly due to the reduced country sample.

more zeros? The reason is that the model’s fundamental zeros are exactly the trade flows that we should not expect to observe in the Krugman model either.

Changing the share of fundamental zeros in the fixed cost model does very little to the predicted share of unobserved flows. If we bring the fixed cost parameter down to zero, the predicted fraction of zeros barely decreases one percentage point as the fixed cost model boils down to the Krugman model. Increasing the share of fundamental zeros to 60% gets the fixed cost model to predict 69% unobserved flows.³⁹

It is important to emphasize the role played by the gravity equation and the skewness in product shares. Table 13 documents three alternative calibrations for both models. In the first exercise we set $\gamma = 0$, so distance has no bearing on trade volumes. We then go one step further, and completely shut down the gravity equation: all import destinations are assumed to be of the same size and to be at the same distance from the U.S. Finally, in the third calibration we impose symmetric product shares across U.S. exports.

Distance does not have a big impact in either model: the dispersion in country trade flows is reduced somewhat and the predicted fraction of unobserved flows falls by 3 percentage points in both models. Table 13 indicates that completely shutting down gravity does change the results. In particular, the fraction of unobserved flows in the Krugman model falls to 30%, as there are fewer trade flows predicted to be tiny. In the fixed cost model only fundamental zeros go unobserved. Hence in this case there is enough separation between the two models so we could use the data to favor one or the other. Of course, this is not of much use, because in this case neither model would match a key aggregate fact about trade – the gravity equation. We could have easily rejected both models simply by looking at the pattern of trade across destinations. Assuming symmetric product shares has a similar effect, although in this case the differences in the predicted fraction of zeros are still small.

	Dense data	Sparse data
<i>without distance $\gamma = 0$</i>		
Krugman model	0 %	63 %
Fixed costs	50 %	64 %
<i>symmetric countries</i>		
Krugman model	0 %	30 %
Fixed costs	50 %	50 %
<i>symmetric products</i>		
Krugman model	0 %	46 %
Fixed costs	50 %	50 %

Table 13: Predicted fraction of zeros for product-level trade flows for alternative calibrations

³⁹Both models have slightly higher fraction of zeros than the balls-and-bins model. The main reason is that the gravity equation predicts country trade flows that are somewhat more skewed than in the data.

More observations help to improve the identification across models. Unfortunately, they do so at a painfully slow rate. With twenty years worth of U.S. export shipments, we would still expect about one third of trade flows to go unobserved in the Krugman model, while the fixed cost model would predict 50% of zeros. Such long data spans may not be available to researchers. In addition, it is then necessary to consider dynamics explicitly in the models.

We have discussed just two basic models, but it is straightforward to generalize the procedure to any model and for different moments in the data. When can we generalize the results as well? We return to the balls-and-bins model to answer this question.

The reader would have recognized that the mechanics of the sampling process (8) is the same as those of the balls-and-bins model. We can indeed view the balls-and-bins results as the sampling of a quite special trade model with two defining properties:

1. It matches total category flows, e.g., total exports per country, per product, per firm...
2. It assumes no systematic relationship between firms, products or countries.

For example, for product-country zeros, trade flow probabilities (bin sizes) were calibrated to match total trade flows to each of the destination countries, as well as the product division of exports. However, there was no relationship between products and countries – every country had the same chance of buying any of the products. Similarly, for the single-product exporter application, we took as given the total trade flows by firm. However, we did not assume any specific matching between firms and products.

If the balls-and-bins model fits a particular fact on the extensive margin, we can be confident that any model that satisfies property 1) will match the fact as well. Any model that satisfies the gravity equation will be consistent with the pattern of product-country zeros. Any model that generates the observed skewed firm size distribution will be consistent with the fraction and size of single-product exporters.

Most trade theories will not share property 2) with the balls-and-bins model.⁴⁰ In fact, most trade theories are concerned with systematic *differences* in the pattern of trade across countries, across firms etc. If the balls-and-bins model fits a particular fact, this implies that we can learn little about these systematic differences based on this particular fact.

Whenever the balls-and-bins model does *not* match a particular fact, however, models will be able to do so by introducing some systematic relationship between firms, products, or countries. For example, the balls-and-bins model is unable to explain the small number of exporters given the firm size distribution. The model in Melitz (2003) is able to do so by postulating a particular relationship between firm size and access to foreign markets.

9 Conclusion

Categorical datasets *do* contain a lot of information, even if they are sparse. Ignoring the sparsity, however, can lead one to mistake sampling zeros for structural zeros. Nowhere is

⁴⁰The simple Krugman model does, but even the very simple fixed cost model does not.

this problem more acute than in the analysis of the extensive margin in trade. We argued that trade data are sparse, and we should expect many sampling zeros.

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 shipment 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 theory of the extensive margin in trade among those models that match the data used in the calibration (e.g., gravity or export sales). 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. We expect that a richer structural model that accounts for sparsity could do an even better job at matching the empirical patterns.

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.⁴¹ 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. Indeed the sparsity problem is so severe that it would not go away even if it becomes possible to combine several years of data. Instead we advocate to account for the sparsity and then focus on deviations—like the split between exporters and non-exporters. 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] Agresti, Alan: 2002, *Categorical Data Analysis*, Second Edition, John Wiley and Sons. Hoboken, NJ.
- [2] Anderson, M. A., Ferrantino, M. J. and Schaefer, K. C.: 2004, Monte Carlo Appraisals of Gravity Model Specifications, Working Paper.

⁴¹Bernard, 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 (2009) for Hungary; Eaton, Eslava, Kugler and Tybout (2007) for Colombia.

- [3] Anderson, J.E. and van Wincoop, E.: 2003, Gravity with Gravitas: A Solution to the Border Puzzle, *American Economic Review* **93**(1), 170-192.
- [4] Alessandria, G., Kaboski, J., and Midrigan, V.: 2003, Inventories, Lumpy Trade, and Large Devaluations, *American Economic Review*, forthcoming.
- [5] Arkolakis, C.: 2009, Market Penetration Costs and the New Consumers Margin in International Trade, *NBER Working Paper*, 14214.
- [6] Axtell, R. L.: 2001, Zipf Distribution of U.S. Firm Sizes, *Science* **293**(5536), 1818–1820.
- [7] Baldwin, R. and Harrigan, J.: 2007, Zeros, Quality and Space: Trade Theory and Trade Evidence, NBER Working Paper No. 13214.
- [8] 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.
- [9] Bernard, A. B. and Jensen, J. B.: 1999, Exceptional Exporter Performance: Cause, Effect, or Both?, *Journal of International Economics* **47**(1), 1–25.
- [10] 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.
- [11] 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*.
- [12] Damijan, J. P., Polanec, S. and Prasnikar, J.: 2007, Outward FDI and Productivity: Micro-evidence from Slovenia, *World Economy* **30**(1), 135–155.
- [13] 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.
- [14] Djankov, Simeon, Freund, Caroline and Pham, Cong S., 2006. "Trading on time," Policy Research Working Paper Series 3909, The World Bank.
- [15] Eaton, J., Eslava, M., Kugler, M. and Tybout, J.: 2007, Export Dynamics in Colombia: Firm-Level Evidence, NBER Working Paper No. 13531.
- [16] Eaton, J., Kortum, S. and Kramarz, F.: 2004, Dissecting Trade: Firms, Industries, and Export Destinations, *American Economic Review* **94**(2), 150–154.
- [17] Eaton, J., Kortum, S. and Kramarz, F.: 2007, An Anatomy of International Trade: Evidence from French Firms, Working Paper.
- [18] 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.

- [19] Carolyn L. Evans and James Harrigan, 2005. "Distance, Time, and Specialization: Lean Retailing in General Equilibrium," *American Economic Review*, American Economic Association, vol. 95(1), pages 292-313, March.
- [20] Evenett, S. and Keller, W.: 2002, On Theories Explaining the Success of the Gravity Equation, *Journal of Political Economy* **110**(2), 281–316.
- [21] 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.
- [22] 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.
- [23] Ghosh, S. and Yamarik, S.: 2005, A Sensitivity Analysis of the Gravity Model, *International Trade Journal* **19**(1), 83–126.
- [24] Halpern, L., Koren, M. and Szeidl, A.: 2009, Imported Inputs and Productivity, Working Paper.
- [25] Helpman, E., Melitz, M. and Rubinstein, Y.: 2008, Estimating Trade Flows: Trading Partners and Trading Volumes, *Quarterly Journal of Economics* **123**, 441-487.
- [26] Hummels, David and Lugovskyy, Volodymyr and Skiba, Alexandre, 2009. "The trade reducing effects of market power in international shipping," *Journal of Development Economics*, vol. 89(1), pages 84-97, May.
- [27] 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.
- [28] Hummels, D., Klenow, P. J.: 2005, The Variety and Quality of a Nation's Exports, *American Economic Review* **95**(3), 704–723.
- [29] Johson, N. L., Kepm, A. W., and Kotz, S.: 2005, *Univariate Discrete Distributions*, John Wiley & Sons.
- [30] Keller, W.: 1998, Are International R&D Spillovers Trade-Related? Analyzing Spillovers among Randomly Matched Trade Partners, *European Economic Review* **42**(8), 1469–1481.
- [31] Krugman, P.: 1980, Scale Economies, Product Differentiation, and the Pattern of Trade, *American Economic Review* **70**, 950-959.
- [32] Mayer, T. and Ottaviano, G.: 2007, The Happy Few: The Internationalization of European Firms, Bruegel Blueprint Series. Volume III.
- [33] 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, \dots, 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 .⁴² 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

$$E(k|n_1, n_2, \dots, n_T) = \sum_{n=1}^N \pi_n \sum_{i=1}^K [1 - (1 - s_i)^n] = \sum_{i=1}^K \sum_{n=1}^N \pi_n [1 - (1 - s_i)^n]. \quad (9)$$

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

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

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

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

We can also express it in terms of the PGF as

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

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.

⁴²To 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.

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

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 $z = 1$:

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

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

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

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

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

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

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

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

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 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."⁴³

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.

⁴³"Correct Way to Complete the Shipper's Export Declaration," February 14, 2001 version.

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

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

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

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

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.