A Balls-and-Bins Model of Trade[*]

Roc Armenter and Miklós Koren[†]

January 21, 2008

Abstract

A number of stylized facts have been documented about the extensive margin of trade—whether to export or not, and, if so, how many products to how many destinations. We note that some of the reported facts would be expected to arise if exports shipments were randomly allocated across categories (e.g., product codes, destination countries). They are, thus, not informative of the underlying economic decisions. We formalize the random assignment of shipments to categories as balls falling into bins, reproducing the structure inherent to disaggregate trade data. The balls-and-bins model quantitatively reproduces the prevalence of zero product-level trade flows across export destinations. The model also accounts for firm-level facts: as in the data, most firms export a single product to a single country but these firms represent a tiny fraction of total exports. In contrast, the balls-and-bins cannot match the small fraction of exporters among U.S. firms and overpredicts their size premium relative to non-exporters. We argue that the balls-and-bins model is a useful statistical tool to discern the interesting facts in disaggregated trade data from patterns arising mechanically through chance.

1 Introduction

International trade has long been concerned with aggregate patterns—what and how much countries trade—and their welfare implications. Finely disaggregated trade data have recently become available and have had an enormous impact on the field. It has spurred a fast-growing research that documents the extensive margin in trade—which firms export, and how many products they send to how many destinations. This, in turn, has lead to new theories.

A number of stylized facts have been uncovered about the extensive margin of trade. The following facts have proven to be very robust. (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

[*] Preliminary. Please do not circulate. For useful comments we thank Arnaud Costinot, James Harrigan, Tom Holmes, Virgiliu Midrigan, Esteban Rossi-Hansberg, Peter Schott, Adam Szeidl, Ayşegül Şahin, and seminar participants at the Federal Reserve Bank of New York and the Institute for Advanced Studies in Vienna. The views expressed here do not necessarily reflect the views of the Federal Reserve Bank of New York or the Federal Reserve System.

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

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]

We note that some of these data patterns would be expected to arise if export shipments were randomly allocated across categories. Consider for example the incidence of multi-product exporters. If each shipment is randomly assigned one product classification, then the more shipments an exporter has the more likely they are classified in multiple categories. Hence large exporters will tend to be multi-product.

We formalize the random assignment of shipments to categories as balls falling into bins. Individual shipments represent a discrete unit (the ball), which, in turn, is randomly allocated into mutually exclusive categories (the bins). This structure is inherent to disaggregate trade data: we observe a given number of shipments; each of them is classified into a unique category. In our model, a ball falling in a particular bin is an independent and identically-distributed random event whose probability distribution is determined solely by the relative size of the bins.

What do we learn when the balls-and-bins model matches a particular fact? Surely we are not suggesting that firms actually ship their goods at random! Our view, instead, is that we cannot conclude *anything*: if a fact cannot falsify the balls-and-bins model, it will also fail to identify the relevant economic theory and thus should not be the basis to favor any model (structural or else). Any theory will be able to account for such a fact once the model is properly augmented with the idiosyncratic heterogeneity and indivisibility inherent in the data.

The balls-and-bins model is thus a useful statistical tool that can discern the interesting facts from the patterns arising mechanically through chance. It can be applied to any categorical dataset, such as the division of total exports by products, firms, or destination countries. These datasets contain a lot of information: it is crucial that we focus on the facts that will help us differentiate among competing trade theories as well as inform the development of new ones.

In spite of its simplicity, the balls-and-bins model has a rich set of predictions. After a number of balls, some bins may end up empty and some will not. Among the latter some will contain a large number of balls, some few. These are taken to be the model's predictions for the extensive and intensive margin, respectively. Given a number of balls and a bin size distribution, we can analytically derive the prevalence of zeros and the fraction of balls sitting in one-ball bins. We can also show how zeros vary with the number of balls and the effect of an asymmetric bin size distribution. These are indeed all the model's systematic relationships between export flows and the extensive margin: the *assignment* of balls to bins is random.

We are interested, though, in a quantitative evaluation. For this we map the balls-and-bins model into the patterns of interest as follows. First we divide an observed trade flow (that is, total trade between two countries, or total exports of a firm) into balls of $36,000 — the value of the average export transaction in the U.S. in 2000 . For example, total exports between the U.S. and Argentina were $3.8 billion, and thus there are 105,000 balls. For the dimension of choice (product codes or destination countries) we construct a bin size distribution using flows. Keeping up with the example, we can construct about 9,000 bins for the 10-digit Harmonized System product codes, each bin of

[1] The following is a necessarily incomplete list of references. Helpman, Melitz and Rubinstein (2007) and Baldwin and Harrigan (2007) for facts 1 and 2; 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.

the size of the corresponding share in total U.S. exports. We then use the balls-and-bins model to predict the expected number of zero product-level trade flows between U.S. and Argentina.

The results are striking: the balls-and-bins model *quantitatively* reproduces many of the data patterns listed above. Let us first return to the previous example: the balls-and-bins not only accurately predicts how common are zeros in the U.S. product level bilateral trade flows, but it also reproduces the pattern of zeros across destination countries. To understand how such a simple random model can replicate the data we first note that the actual number of export shipments (24 million) is low relative to the number of potential product-country pairs (about 2 million). Second, there is a very large variation in the size of the trade flows and categories. Trade with most of the 200 countries is very small and most of the 9,000 traded HS codes are tiny. It is exactly for these that the trade flows are missing in the data. They go missing in the model as well: few balls and tiny bins make for many empty bins.

The success of the balls-and-bins model extends to firm-level facts. We find that single-product and single-destination exporters are as numerous in the balls-and-bins model as they are in the data. Exporters that sell one product to one country account for 40 percent of total exporters in the data — the corresponding number is 43 percent in the balls-and-bins model. These firms, however, account for a minuscule 0.2 percent of total exports in the data — and balls-and-bins predicts their export share to be 0.3 percent.

Once again the large dispersion in flows and categories 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.[2] Because balls are indivisible, these tiny exporters are predicted to be single-product, single-country exporters. This finding suggests that it is important to understand the sources of skewness in the distribution of exports across firms. Once that skewness is accounted for, the incidence and relative size of single- *vs* multi-product exporters follow.

The balls-and-bins model, though, also tells us a lot when it misses key data patterns. 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. Hence exporters are fewer than we would expect. Surprisingly, the model also overpredicts the export size premium: in the data exporters are much smaller than we would expect just from randomness. The model's miss indicates that there is a fundamental difference between small and large firms beyond their different scale.

We hope the successes and failures of the balls-and-bins model leads us to a reappraisal of the stylized facts in the extensive margin in trade. We should emphasize that we do not imply that there are no interesting facts in the data. The balls-and-bins model is a tool to recognize the key deviations from randomness in the data, and these are the facts we believe one should focus on when building models.

There are, of course, many possible random models. And with enough ad-hoc meddling one would be able to come up with a random model that fits a particular set of moments. We have thus to argue for our choice of the balls-and-bins model. There are two key distinct elements in our

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

model: first, an indivisibility at the transaction level; second, independence across balls, i.e., where a ball falls is independent of the distribution of other balls.

The indivisibility follows from the discreteness of the underlying trade data. The 2000 Linked-Longitudinal Firm Trade Transaction Database of the U.S., for example, is built from 24 million individual export transactions.[3] The average exporting firm has only about 140 export transactions in the data. Given the enormous skewness in firm sizes, the median firm may have much fewer transactions. We can no longer ignore the discrete nature of the data. Alternatively, the indivisibility can also be interpreted as a constraint of the environment. Goods must be traded in boxes of $36,000, either because they are physically indivisible or because it is not economically profitable to divide them. We believe that the $36,000 number represents a small degree of indivisibility that can easily be justified either way.

The independence across balls is a natural assumption. We could, for example, have assumed that balls have a higher chance to fall into bins that are already full — and thus impose a force for specialization across firms. But this is exactly what we would like to avoid in order to provide a neutral null hypothesis and let the data speak otherwise.

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 mechanic degree of concentration arising from randomness. Such an index is more difficult for trade facts, which do not focus on a particular dimension.

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 next section describes the setup of the balls-and-bins model and characterizes some of its properties. Section 3 presents the empirical facts on missing product-level trade flows and discusses how the the balls-and-bins model matches these facts. Section 4 conducts the same exercise for firm-level trade flows. Section 5 looks at whether the balls-and-bins model can predict the number and size of exporters. Section 6 discusses the extensive margin of products and destination countries at the firm level. Section 7 offers some extensions. Finally, Section 8 concludes.

[3] Bernard, Jensen and Schott (2007), Table 20.

2 A model of balls and bins

We characterize a trade flow (such as total exports from the U.S. to Argentina, or total exports of a given firm) with a number of indivisible units, or "balls," denoted by n. The trade flow is then partitioned into K disjoint categories (such as the 15,000 10-digit Harmonized System product classifications). We call these categories "bins" and index them by subscript $i \in \{1, 2, \cdots, K\}$.

We then formalize the random assignment of export shipments to categories as balls falling into bins. The probability that a given ball lands in bin i is given by the bin size s_i, such that $0 < s_i \leq 1$ and $\sum_{i=1}^{K} s_i = 1$. Thus where a ball lands is an independent and identically-distributed random variable.

We are primarily interested in the "extensive margin," that is, how many of the bins remain empty after throwing the n balls. The "intensive margin" will be given by the number of balls per non-empty bin. The model has a known probability distribution for both margins. The number of balls in each bin follows a multinomial distribution with parameters n, $s_1, s_2, ..., s_K$. The joint probability distribution of a ball distribution $\{n_1, n_2, ..., n_K\}$ with $\sum_{i=1}^{K} n_i = n$ is

$$\Pr(n_1, n_2, ..., n_K) = \frac{n!}{n_1! \cdots n_K!} s_1^{n_1} \cdots s_K^{n_K}.$$

Obviously, n_i and n_j are not independent given a total number of balls n, as a ball falling in bin i reduces the expected number of balls in bin j.

In the remainder of the section we derive analytically some of the key properties of the model.

2.1 The extensive margin

Let d_i be an indicator variable that takes the value of 1 if bin i is empty, and 0 otherwise. After dropping n balls the expected value of d_i is the probability that bin i receives none of those:

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

Each ball has a $(1 - s_i)$ probability of landing elsewhere. Since where a ball lands is an independent event, the probability that none of n balls fall in a given bin i is $(1 - s_i)^n$. We denote the total number of empty bins (or zeros) by k,

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

Clearly, for $n \geq 1$ $k \in \{K - n, K - n + 1, ..., K - 1\}$, as at least one bin has to be non-empty but no more than n can be filled.

We thus obtain the expected number of empty bins

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

As $(1 - s_i) \leq 1$ for all i and $(1 - s_i) < 1$ for at least one i, we clearly have that the expected number of empty bins decreases in n. Quite trivially it also increases with K.

The expected number of 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 will be needed to hit the small bin.

Formally, the expected number of 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 empty bins decreases.

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

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

for $\alpha \in [0,1]$. Then the expected number of empty bins under $\{\tilde{s}_i\}$ is at most as large as than under $\{s_i\}$ for all $n \geq 2$.

The symmetric distribution $\{1/K\}$ is more even than any asymmetric distribution. At the other extreme, if $s_1 = 1 - \varepsilon/K$ and $s_i = \varepsilon/K$, we have

$$E(k|n) = (\varepsilon/K)^n + (K-1)(1-\varepsilon/K)^n,$$

which tends to $(K-1)$, the maximum number of empty bins, as ε tends to zero.

The extensive margin is just the number of non-empty bins,

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

This is clearly increasing in the number of balls, n. The following figure plots the expected number of non-empty bins against the number of balls for 5 symmetric bins.

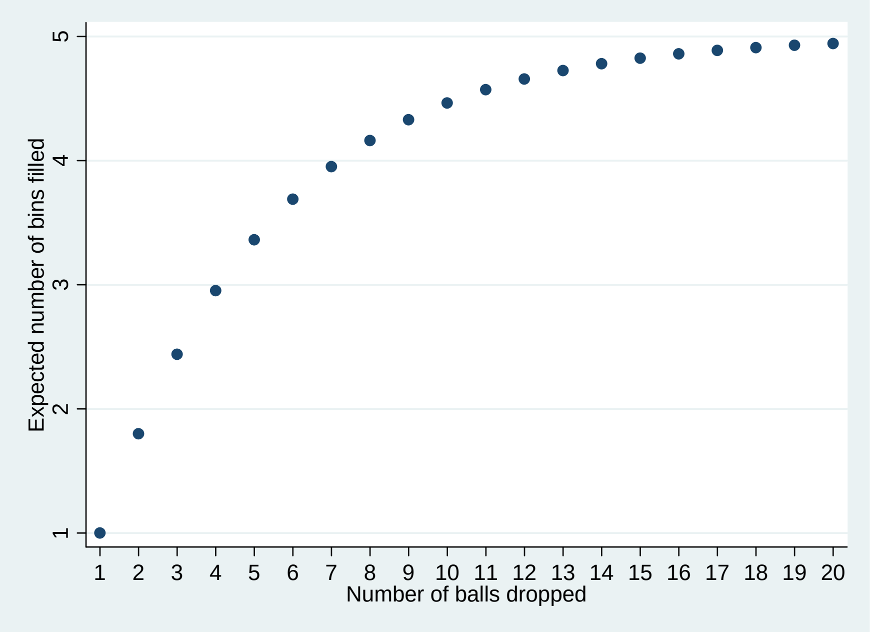

Figure 1: Balls and bins

The first few balls fall into separate bins almost surely. Because of that, as long as balls are few, the number of filled bins is close to the number of balls. The relationship is close to linear. Most adjustment is on the "extensive margin." Then balls are getting more and more likely to fall in non-empty bins, and the number of filled bins will fall behind the number of balls.[4] Eventually, all bins get filled, and the relationship flattens out. The remainder of balls can only increase the "intensive margin."

2.2 Ones

Next we look at the probability that a single bin contains all the balls. This is the analogue of a firm that sells only one product or to only one country.

What is the probability that a particular bin i contains all the n balls? Each ball had s_i probability of falling into bin i, so this probability is s_i^n. For any particular bin, the probability that it is the single non-empty bin is falling in n. What is the probability that exactly one bin is non-empty? This can be any of the K bins, giving a probability of

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

Obviously the probability of a single non-empty bin decreases with the number of balls, n. More interestingly, the probability of a single non-empty bin increases with the dispersion of bin sizes.

Proposition 2. *Let $\{s_i\}$ be a bin size distribution and let*

$$\{\tilde{s}_i\} = \alpha\{s_i\} + (1 - \alpha)1/K \tag{4}$$

for $\alpha \in [0, 1]$. Then the probability of a single non-empty bin under $\{\tilde{s}_i\}$ is at most as large as than under $\{s_i\}$ for all $n \geq 2$.

This follows from the convexity of (3) in s_i. Again, at the extreme, if $s_1 = 1 - \varepsilon/K$ and $s_i = \varepsilon/K$, we have

$$\Pr(k = 1|n) = (1 - \varepsilon/K)^n + (K - 1)(\varepsilon/K)^n,$$

which tends to 1 as ε tends to zero.

2.3 Aggregation

So far we have only looked at a single trade flow. Often, however, we are interested in some aggregate statistic, such as the total number of empty product categories across all countries, or the fraction of exporters among all firms.

Suppose there is a total of M trade flows (countries, firms) in aggregate data, each indexed by m. Trade flow m is comprised of n_m balls, and we can characterize that trade flow conditional on n_m using the tools above.

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

Let π_n denote the fraction of flows with exactly n balls. This defines a probability distribution over $\{1, 2, ..., N\}$, where N is the size of the largest flow.

The average number of empty bins across all trade flows is given by

$$E(k) = \sum_{n=1}^{N} \pi_n \sum_{i=1}^{K} (1 - s_i)^n = \sum_{i=1}^{K} \sum_{n=1}^{N} \pi_n (1 - s_i)^n. \tag{5}$$

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 empty bins can be written as

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

Since $G(z)$ is strictly convex, uneven bin-size distributions will have a larger expected number of empty bins. That is, aggregation preserves the aforementioned properties.

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 unconditional probability is

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

We can also express it in terms of the PGF as

$$\Pr(k = 1) = \sum_{i=1}^{K} G(s_i).$$

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

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

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

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, extremely unpractical to do so and we will instead use Monte Carlo methods to derive the distribution as needed.

3 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.

3.1 The facts

Baldwin and Harrigan (2007) recently reported that most potential destination country product combinations are missing in 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.[5] In 2005, the U.S. exported 8,877 different 10-digit Harmonized System categories to 230 different countries.[6] Of these 2,041,710 potential trade flows, 1,677,213 (or 82%) were missing. In other words, the average country only bought 18% of the 8,877 products the U.S. exports.

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 similar patterns, the incidence of zeros only decreases significantly 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 1: 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 more 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.

[5] Hummels and Klenow (2005) also look at the product-margin of aggregate exports. They have a different measure of the extensive margin, something we plan to analyze later.

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

	Non-zero trade flow
Real GDP	0.073*** (0.007)
Real GDP per capita	0.022** (0.009)
Distance = 0	0.300*** (0.055)
0 < distance < 4000km	0.236*** (0.025)
4000 < distance < 7800	omitted
7800 < distance < 14000	0.006 (0.030)
Distance > 14000	0.049 (0.033)
Observations	965,151
Clusters	99
R^2	0.34

Table 2: Non-zero flows and gravity – *The data (Baldwin and Harrigan, 2007)*

Table 2 reports the results.[7] Larger countries are more likely to import any given product. The same is true for richer countries. The incidence of non-zero 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 nonzero exports decreases with destination-country size and increases with distance.*

3.2 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 (230). We assign each ball a constant dollar value of $36,000, which is the average size of export shipments in 2000.[8] We then convert the total value of the trade flow into the number of balls by dividing by $36,000 and rounding up to the next integer. For example, exports to Canada (the biggest importer) were $168 billion (in 2000 dollars), which corresponds to 4.7 million balls. Exports to Argentina (a median importer) were $3.8 billion, corresponding to 105,000 balls. To keep the empirical applications comparable, we report all values in 2000 dollars.

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

[8] We take this number from Bernard, Jensen and Schott (2007). Table 20 reports the total number of export shipments (above $2,500) as 23.9 million. The total value of these shipments was $855 billion. The average shipment is hence $36,000 in 2000 dollars.

The bins correspond to the 8,877 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 total exports of a given HS code with aggregate merchandise exports.

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

$$k_c = \sum_{i=1}^{8877} (1 - s_i)^{n_c},$$

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

$$k = \sum_{c=1}^{230} k_c.$$

3.3 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 73%, surprisingly close to the data (82%). Using HS6 codes results on 69% of the potential product-level bilateral flows being zero for 79% in the data.

Moreover the model matches quantitatively the pattern of zeros across flows in the data. To show this, we plot the fraction of product-level zeros for each country against total U.S. exports to that country in Figure 2. The dots represent the actual fraction of zeros in the data, the line is the predicted number of empty bins for each country. We already know that the balls-and-bins model somewhat underpredicts zeros, but the shape of the relationship to total exports is strikingly similar.[9]

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 can 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.[10] 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 3 reports the results. For convenience, the first column repeats the regression on non-zero flows in the data. Bigger and closer countries are more likely to have a non-zero flow under the balls-and-bins model, just as in the data. Moreover, the magnitudes of the coefficients are surprisingly similar. The only exception are the two countries bordering the U.S. ("distance= 0"), Canada and Mexico. These seem to import more HS codes in the data than under the balls-and-bins model.

The success of the balls-and-bins model may be perplexing to the reader. However it is just indicating that the heterogeneity underlying in the data is so large that, in the aggregate, it is as if every export shipment were randomly classified into one product category.

[9] In fact, in section 7, we show that a small change in the size of the ball achieves a perfect fit.

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

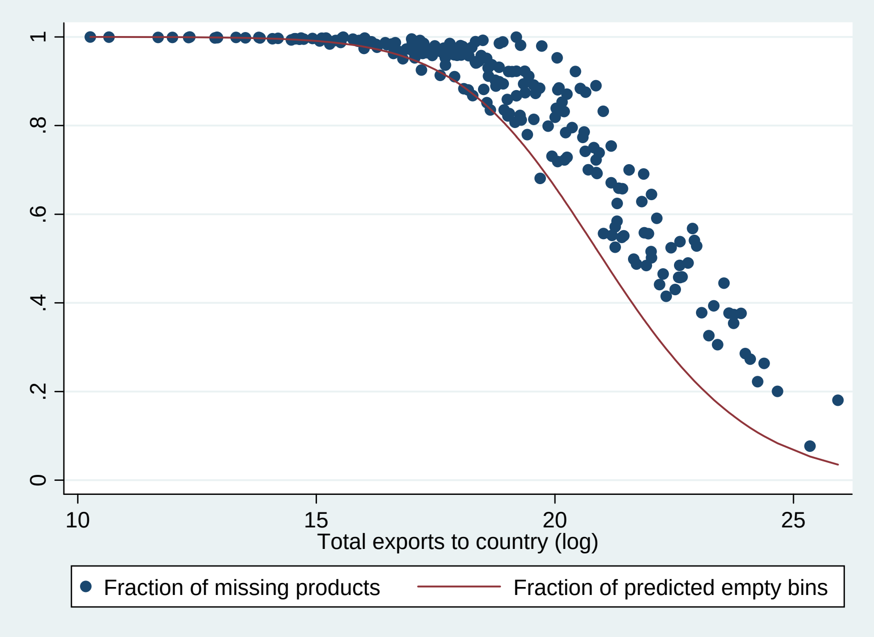

Figure 2: The incidence of zeros and total exports

	Non-zero trade flow	B+B model
Real GDP	0.073*** (0.007)	0.098*** (0.006)
Real GDP per capita	0.022** (0.009)	0.018** (0.009)
Distance = 0	0.300*** (0.055)	0.154*** (0.021)
0 < distance < 4000km	0.236*** (0.025)	0.205*** (0.027)
4000 < distance < 7800	omitted	omitted
7800 < distance < 14000	0.006 (0.030)	-0.024 (0.028)
Distance > 14000	0.049 (0.033)	0.012 (0.043)
Observations	965,151	965,151
Clusters	99	99
R^2	0.34	0.34

Table 3: Non-zero flows and gravity – *Balls and bins*

Quantitatively, the dispersion in flow and bin sizes plays a key 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 than the model predicts the most zeros.

Let us start with the distribution of bin sizes. The size of the average bin is $1/8877 = 1.13 \times 10^{-4}$. However, the size distribution across bins is rather skewed. The size of the median bin is 1.7×10^{-5}, about seven times smaller than the average. The following tables list the five biggest and the five smallest bins.

HS code	Description	Share
8802.40.00.40	Airplanes exceeding 15,000 kg for passenger transport	0.023
8542.21.80.05	Monolithic integrated circuits, for other then HDTV, silicon	0.021
8473.30.00.00	Parts and accessories of computers	0.015
8803.30.00.10	Other parts of airplanes or helicopters	0.013
8708.99.80.75	Parts and accessories of tractors	0.012

Table 4: The five largest HS codes

The biggest HS category is airplanes (a lumpy category indeed). This reflects the comparative advantage of the U.S. to produce complex machinery such as airplanes. Other large categories include "catch-all" categories of parts and accessories. Their being large probably only reflects that these categories are broad aggregates (even at the 10-digit level).

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.

HS code	Description	Share
0706.10.40.00	Turnips	3.11×10^{-9}
5208.21.20.90	Woven fabric of cotton...	3.26×10^{-9}
5210.51.60.20	Woven fabric of cotton, mixed with fibers	3.32×10^{-9}
0910.20.00.00	Saffron	3.68×10^{-9}
3825.41.00.00	Waste organic solvents (halogenated)	3.86×10^{-9}

Table 5: The five smallest HS codes

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 relative shares of HS codes calibrated in different ways. First, we take 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%. Second, we look at an even distribution of all the 8,877 HS codes. This would lead to 53% zeros.

4 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 exported to only 3.5 countries from a total of about 230.[11] In other words, 98 percent of potential firm-country pair trade flows are zero.

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

Table 6 reproduces column 2 of Table 6 from Bernard, Jensen, Redding and Schott (2007). The log number of exporting firms are 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 6: 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 the next section.

The number of balls per destination country are again taken by dividing the total exports to that country by $36,000. The total number of bins equals the number of exporting firms, 167,217.[12] Because there are many more firm bins (167,217) than we had product bins (8,877), we already expect that many more bins remain empty.

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

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.[13] The top 1% of exporters account for 80.9% of total exports.

We approximate the distribution of exports with a lognormal distribution with $\mu = 10.99$ and $\sigma = 2.99$. 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

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

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

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

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

and Schott (2007).[14] The size distribution of bins will then inherit this lognormal distribution with the additional normalization the bin sizes add up to one.

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.[15]

Table 8 presents the results. For convenience, we reproduced the regression estimate by Bernard, Jensen, Redding and Schott (2007) in the first column.[16] The coefficient estimates in the simulated regression are strikingly 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^{***}	0.67^{***}
	(0.04)	(0.04)
Log distance	-1.14^{***}	-0.89^{***}
	(0.16)	(0.17)
Observations	175	181
R^2	0.74	0.68

Table 8: Exporting firms and gravity – *Balls and bins*

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 that any model that accounts for that can match the gravity equation of firms - *even* if the assignment of firms is random.

A direct consequence of both the product and the firm counts following so strong a gravity equation is that the "intensive margin," that is, the average amount exported *per product per firm*

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

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

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

will follow an inverse gravity equation, as reported in Table 6 of Bernard, Jensen, Redding and Schott (2007). Larger, closer countries will buy *less* per product per firm. This can be easily understood within the balls-and-bins model. A country that only has a single export transaction necessarily buys only one product from one firm. The intensive margin is 1 ball per product per firm. If a country has two transactions, those two transactions are very likely to come from two distinct firms and correspond to two distinct product categories (it may be a computer from Dell and a case of wine from Kendall-Jackson). The intensive margin is then 2 balls per 2 firms per 2 products, 1/2, less than for the single-ball country. Larger countries buy less per product per firm. In fact, the gravity equation for our simulated intensive margin is very close to the one reported by Bernard, Jensen, Redding and Schott (2007): $-0.20 \times \text{GDP} + 0.37 \times \text{distance}$.

5 Exporting firms

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

According to the survey by Bernard, Jensen, Redding and Schott (2007), only 18% of manufacturing firms export at all. The fraction drops to about 3% when all firms outside manufacturing are included.[17] 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 3. *Exporters are few — only 18% of manufacturing firms export in the U.S.*

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

Empirical regularity 4. *Exporters are large — among U.S. manufacturing firms, exporters sell 4.4 times more than non-exporters.*

We follow essentially the same steps as before to map the model to the data. The key difference is that now the output flow will be originated by firms, not countries, and will include total sales, not only exports. As before we obtain the number of balls n per firm by dividing its total sales by \$36,000 and rounding up.[18]

We thus need data on total sales per firm in order to construct the distribution of balls (π_n). Unfortunately we do not have direct access to this data for the U.S. The 2002 Statistics of U.S. Businesses of the Census, though, reports the number and total sales of firms in each of eight size bins (see Table 9).

As 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

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

[18] In the previous section we used evidence on the average shipment value to pin down the "ball size." We have no direct equivalent for total sales. In Section 7 we document the results for different balls sizes.

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 9: The distribution of firm sales in manufacturing – *Census*

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

To distinguish between exporters and non-exporters we only need two bins: one for domestic sales, the other for foreign sales. In the 2002 Economic Census, there were 297,873 manufacturing firms. Their total receipts amounted to $3.94 trillion. Exports of manufactured goods amounted to $545 billion in 2002.[19] That is, 13.9% of manufacturing receipts come from exports. This pins down the size of the domestic bin at 0.861 and the size of the export bin at 0.139.

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

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

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

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

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

The model's prediction for the exporter's size premium is also off. Surprisingly, though, the model overpredicts the size of exporters. That is, despite exporters being four fifths of total firms in the model for one fifth in the data, the model predicts that exporters are 34 times larger than

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

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.

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

To derive the size-exporting relationship formally, let π_n be the unconditional size distribution of firms. The size distribution conditional on no exports 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 then

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

We recover the notation for the probability generation function $G(z) = \sum_{n=1}^{\infty} \pi_n z^n$ of the firm size distribution. We can then express the expected sales of non-exporters as

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

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 $z = 1$:

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

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 large exporter premia, we parametrize the distribution as the *zeta distribution*. This is the discrete analogue to Pareto distribution, and its probability mass function is

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

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

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

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

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

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

Summarizing, what do we learn from the balls-and-bins miss? First, the split between exporters and non-exporters is not just a matter of chance: there is some economic force that makes the two types of firms quite different. Second, the data has a weak sorting of exporters by size: exporters are smaller, not larger, than expected.[20]

6 Firm-level export patterns

We then turn to evidence on the extensive margin at the level of individual exporting firms. In this section we ask how many products firms export and how many destinations they serve. Note that the universe of interest now is the set of *exporting firms*, because the empirical facts are usually reported only for firms that have some exports.[21] This way we can use the balls-and-bins model to understand these moments without having to factor that the split between exporters and non-exporters is very different from random.

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.[22]

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 5. *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.

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

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

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

Empirical regularity 6. *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 0.2 % of total exports.

Empirical regularity 7. *40% of firms export a single product to a single country. These firms account for only 0.2% of total exports.*

To calibrate bins, we use the same bin sizes as for the aggregate flows. 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.[23] The following table lists the five biggest country bins.

Country	Share
Canada	0.228
Mexico	0.127
Japan	0.064
China	0.048
United Kingdom	0.042

Table 10: The five biggest country bins

We assume each firm has a different number of export balls. The number of balls can be calibrated to the distribution of exports across firms, reported in Table 7. We approximate the distribution of exports with a lognormal distribution with $\mu = 10.99$ and $\sigma = 2.99$. This matches the mean exports of $5.11 million and has a median exports of $59,300. 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 8,877 HS codes, the second ball is very likely to fall into an HS category different from the first one. Only 0.3% of two-ball exporters are single-product exporters. The key to understanding the incidence of single-product exporters is that there are plenty of very small exporters, who export $36,000 or less.

With respect to the fraction of single-country exporters the model underpredicts the data, 44% in the model for 64% in the data. The relationship between number of balls and number of bins is somewhat less mechanical for destination countries. There is a very large bin (Canada) so there is a fair chance that the second ball goes in there too. Overall, 8.4% of two-ball firms are single-country

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

exporters. For three balls, this fraction is down to 1.5%. In the data, though, we conjecture that there exists relatively large exporters that export only to Canada (and possibly Mexico).

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

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

We conclude that the split between single-destination, single-product firms and the rest is very much in line with what we would expect given the skewness of the exporter distribution. Once again the balls-and-bins points to the very special split between exporters and non-exporters as the key fact behind most of the patterns on the extensive margin of trade.

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

7 Extensions

We have calibrated the size of the ball to the average size of export shipments, $36,000. Given that lumpiness plays a big role in our analysis, we experiment with other ball sizes, as well. Because individual export transactions are the fundamental units of observations, we take the average size, $36,000, as a *lower bound* on the ball size. This can easily be explained by some small frictions and indivisibilities in transportation, such as container shipping, the administrative burden of customs clearance etc. However, it may well be the case that the relevant decisions concern multiple transactions at the same time, that is, the ball size is larger.

The following table shows our quantitative results for ball sizes between $36,000 and $1 million. The latter represents such a big indivisibility that around 60% of all manufacturing firms would only be given one ball. We only report it to illustrate how the balls-and-bins model works with so few balls – we think such an indivisibility is hard to defend with economics.

The changes in the magnitudes are intuitive. First, as balls get bigger, the incidence of empty product bins increases. Fewer balls make for more empty bins. Bigger balls also reduce the fraction of exporting firms, closer to the one we see in the data. This is because if firms are made of fewer balls, it is less likely that any one of them comes from exports. However, even the $1 million ball would predict significantly more exporters (33%) than in the data (18%). This suggests that economies of scale in deciding whether or not to export are rather strong. The fraction of single-product and single-country exporters increases both in number and it their export share. Again, with larger balls, most firms will end up with just one ball and would be called a single-product, single-country exporter.

Figure 3 replicates Figure 2 for different ball sizes, $36,000, $100,000, and $500,000. A small

		Ball size			
Moment	Data	$36k	$100k	$500k	$1m
Fraction of empty bins	0.821	0.728	0.809	0.904	0.932
Relation to total export	−0.063	−0.086	−0.070	−0.042	−0.031
Fraction of firms that export	0.180	0.737	0.610	0.406	0.331
Exporter premium in log sales	1.48	3.53	3.23	2.80	2.55
Fraction of 1-product, 1-country exporters	0.40	0.43	0.57	0.76	0.83
Exports by 1-product, 1-country exporters	0.002	0.003	0.01	0.07	0.16

Table 11: The stylized facts with different ball sizes

increase in the ball size not only increases the overall incidence of product-level zeros to match the one in the data, but also achieves a perfect fit in terms of the relationship of zeros and total export.

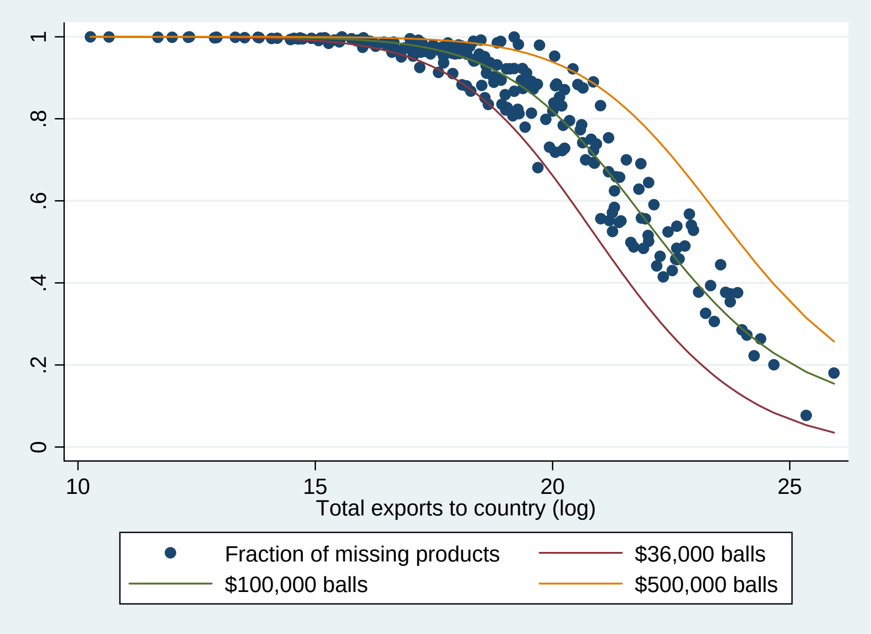

Figure 3: The incidence of zeros with different ball sizes

8 Conclusion

Our findings suggest directions for building new theories as well as amending existing ones. Because the balls-and-bins model features shipments as the units of observation, it does not allow for any economies of scale in export behavior. A large firm is modeled as a collection of many balls and is hence very similar to a collection of small firms. In contrast, if one introduces economies of scale in exporting (such as a fixed cost), large firms (who have paid the fixed cost) will be fundamentally different from a collection of small firms (who have not paid the fixed cost). The fact that the balls-and-bins model greatly misses the incidence of exporters suggests that economies of scale in exporting are indeed very strong.

The empirical patterns on other margins of export behavior (how many products and where to export) are well matched by the balls-and-bins model. This suggests that a model that has enough heterogeneity to match the enormous size dispersion of exporters will also match the facts about multi-product and multi-country exporters, even in the absence of (further) economies of scale. The challenge for building such a model is to link the dispersion of exports (which is much bigger than the dispersion of sales or employment) to fundamentals.

We want to emphasize that some of the heterogeneity in the data is necessarily omitted from any model (even if it allows for some heterogeneity in observables) and would thus be relegated to an error term. For example, the Melitz (2003) model introduces heterogeneity in firm productivity and derives that more productive firms sort into exporting. In the data however, the export decision is not fully determined by productivity alone. The decision to export is affected by myriad of other factors, such as ownership or management. It would be a mistake to reject the Melitz (2003) model just because two firms with the same productivity behave differently.[24] The relevant question instead is how much of the variation in export behavior is due to firm productivity and how much due to omitted heterogeneity.

An alternative way of accounting for all the heterogeneity in the data is to build a fully structural model that has enough heterogeneity in its parameters to be consistent with a set of moments. The benefit of doing so is that one can attach labels to the observed heterogeneity. For example, if some firms are bigger than others, it may be because those firms are more *productive*. Or if some firms export, while others of the same size do not, it may be because those firms have lower *fixed costs* of market access.[25] Then one can check whether the estimated parameters conform to our priors or estimates from other studies. Having a fully specified model is also a constraint, however. The model will only explain the set of moments it was designed to explain. Even the slightest change in the empirical question can cause problems. For example, it is difficult to compare product-level models estimated on datasets with different product classifications. In contrast, the balls-and-bins model can be easily adapted to any empirical application and is not sensitive to changes in statistical classifications, for example. We view the balls-and-bins model as more suitable for explorative data analysis.

We hence hope that our approach can be used in future empirical work using massive micro-level trade datasets. Recent transaction-level datasets are very detailed,[26] and trade flows are typically broken down by firms, 8 or 10-digit product codes, and destination countries. By their very nature, these datasets are *sparse* in the sense that most of the firm-product-country trade flows are missing. The balls-and-bins model provides a natural benchmark for working with such sparse datasets, and can be easily adapted to any empirical application.

[24] Similarly, it would be a mistake to reject a Ricardian model because countries import a particular product category from more than one country. If Belgium imports 99% of its red wine from France and 1% from Argentina, that is still fundamentally consistent with the predictions of the Ricardian model.

[25] Eaton, Kortum and Kramarz (2007), for example, build and calibrate a fully structural model of exporting firms that matches the key stylized facts in French firm-level export data.

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

References

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

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

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

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

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

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

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

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

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

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

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

[12] 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.

[13] 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.

[14] 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.

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

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

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

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

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

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

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