A Balls-and-Bins Model of Trade*

Roc Armenter and Miklós Koren†

March 2012

*For useful comments we thank the Editor, the referees 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, the University of Zurich, CEPR ERWIT, and the LSE. 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

A Balls-and-Bins Model of Trade

March 2012

International trade has long been concerned with aggregate patterns—what and how much countries trade with each other. The recent availability of finely disaggregated trade data 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. A number of stylized facts have emerged regarding the incidence and pattern of zero trade flows at the product and the firm level, the size and frequency of exporters and, among the latter, the size and frequency of multi-product and multi-destination exporters.[1]

We argue that several of these facts fail to identify the relevant theory of the extensive margin from a surprisingly large class of successful trade models. For example, trade models capable of reproducing the gravity equation and the sectoral composition of trade—even those without an explicit treatment of the extensive margin—can replicate the prevalence and pattern of zero trade flows as well. Similarly, provided the model matches the observed skewness in export sales, it will predict that most exports are done by relatively few, but large, multi-product, multi-destination exporters. The reason lies in the nature of trade data.

Our first observation is that disaggregate trade data are categorical by construction. We observe a finite number of shipments which constitute the basic units of observation.[2] Each shipment is then assigned a unique category in each of several classification systems, e.g., the shipment belongs to one of many product categories, can go to one of many destination countries, and may be sold by one of many firms in the economy.

Trade datasets are typically sparse, that is, the number of observations—total shipments in a given year—is too low relative to the number of possible classifications—country, firm and product codes—to ignore the categorical structure of the data. 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 country-product

[1] The following is a necessarily incomplete list of references: Helpman, Melitz and Rubinstein (2007), Baldwin and Harrigan (2007), Haveman and Hummels (2004) and Hummels and Klenow (2001, 2005) on zero trade flows; Bernard and Jensen (1999), Bernard, Eaton, Jensen and Kortum (2003), Bernard, Jensen and Schott (2007), Bernard, Jensen, Redding and Schott (2007) and Eaton, Kortum and Kramarz (2004, 2007) for firm-level facts. See the main text and the web appendix for further discussion.

[2] Trade data are collected through customs forms, one for each export shipment. See Appendix for details.

classifications. More than 40 percent of the traded country-product pairs had only 1 or 2 shipments during the year, a clear sign that the data are sparse.[3]

We propose a parsimonious statistical benchmark for working with categorical datasets. Our contribution is twofold. First, we show that several facts on the extensive margin are to be expected in a sparse dataset once some key facts on total flows are accounted for, such as the gravity equation across countries or the heterogeneity in size across sectors or firms. These statistics are then not sufficiently informative about the extensive margin to distinguish among structural models that successfully match total trade flows. Second, our framework can also guide the identification of the right model of the extensive margin, by either selecting more informative data moments or establishing a benchmark for a quantitative evaluation of structural models.

We formalize the assignment of observations to categories as balls falling into bins. Each observation constitutes a discrete unit (the ball), which, in turn, is allocated into mutually exclusive categories (the bins). This structure is inherent to categorical data, and thus to disaggregate trade datasets: we observe a given number of shipments and each of them is classified into a unique category. Because we want a parsimonious benchmark, 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.

The number of balls and the distribution of bin sizes are treated as parameters in the model. These can be calibrated directly from the data on total trade flows (i.e., across countries, products or firms), or from the corresponding predictions of a structural model.[4] As a parsimonious choice, we calibrate the bin sizes assuming no systematic relationship across classifications (e.g., across countries and products).

In spite of its simplicity, the balls-and-bins model makes a rich set of predictions. After a number of balls, some bins will end up empty and some will not. Among the latter some will contain a large number of balls, some few. These are taken to be the model's predictions for the extensive and intensive margin, respectively. We characterize the prevalence of zeros and how it varies with the number of balls and the bin-size distribution. These are indeed

[3]Statistical inference on categorical datasets requires a much larger sample size. Indeed, 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. See Section 9.8 of Agresti (2002) for a summary discussion of statistical inference in sparse categorical data.

[4]Trade models have been very successful at reproducing aggregate trade flows. Thus, as they say, our analysis stands on the shoulder of giants.

all the model's systematic relationships between export flows and the extensive margin: the *assignment* of balls to bins is random.

We first show that, once we match the total trade flows, several patterns arise naturally in sparse data. We set the number of balls equal to the number of trade shipments observed in 2005 U.S. export flows, about 22 million. 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 results are striking: the balls-and-bins model *quantitatively* reproduces many of the 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; the pattern of zeros across export destinations is also the same. Indeed, we replicate facts regarding zeros as long as trade flows across countries follow a gravity specification, and the trade shares across HS codes are skewed. Trade with most of the countries is then very small and most of the 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. Most trade models in use—including those without any explicit extensive margin—can replicate gravity and accommodate the heterogeneity across products. Once we account for the categorical nature of the data, any of these models will replicate the prevalence and pattern of zeros across export destinations.

Balls and bins also matches several of the firm-level facts: in the model, 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. It is the left tail of the distribution of exports across firms which proves key to reproducing these firm-level facts. 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. Several models in the literature are able to reproduce the skewness in sales; the incidence and relative size of single- and multi-product exporters follows for all of them.

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* fact on the extensive margin. Indeed, all bins would be non-empty and the predictions for the extensive margin would be trivial.

The balls-and-bins model is also useful to spot statistics that *are* informative about models of the extensive margin. For example, we attempt to predict the share of exporters

Description	Data	Balls-and-bins
HS10-level product×country U.S. export flows		
Share of zeros	82%	72%
OLS coefficient of non-zero 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 web appendix.

among manufacturing firms. According to the balls-and-bins model, 74 percent of firms should export — in contrast with 18 percent in the data. Similarly, the balls-and-bins model underpredicts the fraction of single-destination exporters. Thus the split between exporters and non-exporters is an useful statistic to discern among structural models of the extensive margin, as the *systematic* relationship between firms and foreign market access is not driven by sparsity.

In Section 5, we illustrate why identifying theories is difficult when data are sparse. We introduce a simple extension of the model where some of the bins are closed, and will be empty with probability one—loosely capturing the extensive-margin implications of a wide array of models. These "fundamental zeros" are in addition to any potential "sample zeroes" predicted by balls and bins. To uncover the best model of the extensive margin we need to infer the pattern of fundamental zeroes from the observed, total number of zeroes. We show that whenever the balls-and-bins model matches the data, the total number of zeros hardly varies across many possible sets of fundamental zeros. The reason is that whenever the data are sparse we are trading fundamental zeros with sample zeros essentially one-to-one. In contrast, if the data are not sparse, then the number of total zeros increases rapidly with fundamental zeros; we would be closing bins with a high chance of being non-empty. Thus sparsity is the common cause of the non-identification problem and the success of balls and bins.

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 raises 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.

1 Balls and bins

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

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

The state of the system is given by the full distribution of balls across bins, $\{x_1, x_2, \cdots, x_K\}$. Clearly, this distribution is a random variable. Since we are primarily interested in the "extensive margin," that is, the split between empty and non-empty bins, we define d_i to be an indicator variable that takes the value of 1 if bin i is non-empty, $x_i > 0$, and 0 otherwise. The "intensive margin" will be given by the number of balls per non-empty bin.

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

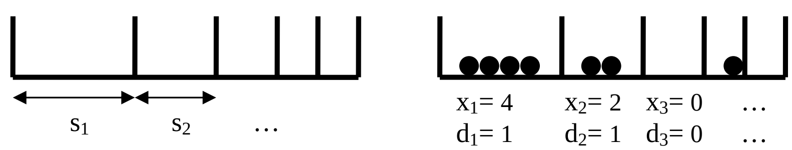

Fig. 1A **Fig. 1B**

Figure 1: Balls and bins

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

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

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

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

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

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

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

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

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

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

This is our key statistic out of the balls-and-bins model. We will use it to derive many of the 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. Perhaps more subtly, the relationship is not linear. 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.[5] Hence, the relationship flattens out and the number of filled bins increases slowly. The remaining balls can only add to the "intensive margin." Note that the model is very stark in its limiting predictions 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.

[5]The first ball falling into a non-empty bin 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.

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

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

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

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

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 converges to zero.

1.1 Multiple classification systems

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 many 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, respectively.[6] There is, conceptually, nothing different from the previous case: we can always re-arrange the classification system into a row of bins of length TK, so that s_{ij} denotes the likelihood of the ball falling into category i along the first dimension and category j along the second. Using (1) we can derive the expected number of zeros,

$$E(k|n) = \sum_{j=1}^{T} \sum_{i=1}^{K} [1 - (1 - s_{ij})^n], \tag{4}$$

and similarly for the remaining predictions.

In order to keep the benchmark as parsimonious as possible, we assume each ball is randomly and *independently* allocated across classification systems. There is thus no systematic relationship across categories. For example, destination countries would buy the same basket of products in exactly the same proportions; or all exporters are equally likely to sell a given product to a given market. The independence across classifications suits our pursuit of a parsimonious benchmark, but the model remains tractable if we want to introduce a richer specification.[7]

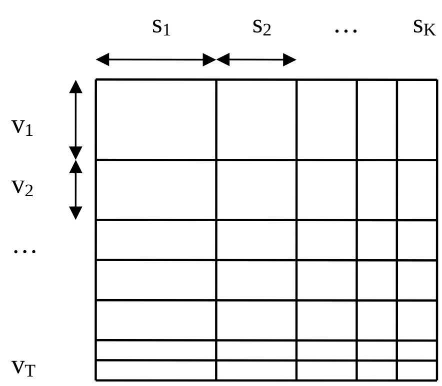

Figure 2: Balls and bins: T by K case

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

[7]Trade models typically make predictions about trade flows, X_{ij}, as a function of the model's parameters and observable variables, like distance, GDP, factor abundance, etc. To map the model to categorical data, we can equate the likelihood of one shipment to belong to category i, j to be $s_{ij} = \frac{X_{ij}}{\bar{X}}$, where $\bar{X}$ is the U.S. total exports as predicted by the model. In Section 5 we provide an example.

Visually, one can think of throwing balls over a T by K grid of bins as in Figure 2. Each classification system comes with its size distribution, $v_1, v_2, ..., v_T$ and $s_1, s_2, ..., s_K$, which in Figure 2 pins down the size of rows and columns, respectively. The probability of a given ball falling in the bin (i, j) is then $v_i s_j$.

An additional advantage of approaching the dataset as a whole is that allows working with *conditional* moments, for example, the number of empty product bins for a given country. For each realization of ball throws there will be a number of balls in each row and in each column, denoted $n_1, n_2, ..., n_T$ and $m_1, m_2, ..., m_K$, respectively. (Note that n_i or m_j may be zero.) We can then ask the distribution of balls across columns $1, 2, ..., K$ within a given row with n_j balls.

More interestingly, we can compute the statistics of interest given a distribution of balls $n_1, n_2, ..., n_T$ across rows. This will allow us, for example, to derive how the fraction of zero product-level bilateral flows varies across U.S. export destinations using the actual aggregate export flows. Since the classification in each system is independent, 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):

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

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

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

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

Similarly, we can compute the probability of a single non-empty bin for each row using (3). Then we can derive the proportion of rows which are expected to contain a single non-empty bin. This will allow us, for example, to derive how the fraction of single-product exporters 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.

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

In practice, we will sometimes approximate the distribution of balls across rows with some parametric distribution. The web appendix 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.

2 Zeros in trade flows

2.1 Product-level zeros

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 opportunity 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.

2.2 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.[8] 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.[9]

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

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

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

Other levels of aggregation lead to a similar incidence of zeros. Table 2 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 2: The incidence of zeros under different classifications

Baldwin and Harrigan (2007) then report how the incidence of zeros relate to the size of the importer and its distance to the U.S. Larger countries that are closer buy a larger variety of products. 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.*

2.3 From the data to the model

In order to map the balls-and-bins model to the data, we proceed as follows. The trade flow of interest is the total U.S. exports to a given country, that is, we will have as many trade flows as destination countries (229). We measure the number of shipments going to a country to calibrate the number of balls. For example, Canada (the biggest importer) received 7.4 million shipments in 2005. Equatorial Guinea, the median buyer of U.S. exports, had 2,641 shipments. A natural question is why there are no more trade shipments in a year. In the web appendix, we document the determinants of the number and size of shipments in more detail. In summary, goods are bulky—shipped rarely and by themselves—for about half of the product categories. The other half has many more shipments, but are still constrained by the physics of shipping: it is not worth sending a container half empty.

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

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

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

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

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

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

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

Note that we are computing the expectation conditional on the number of export shipments from the U.S. to each country. To retrieve the incidence of zeros we only need to 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 229×8,867 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.

2.4 The model's predictions

We find that indeed most of the 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 3 reports the predicted fraction of zeros for other levels of sectoral aggregation. The model's predictions track the observed incidence of zeros rather 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 3. 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

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 3: The incidence of zeros under different classifications

zeros, hence overpredicts the number of exported products, but the shape of the relationship to total exports is strikingly similar.

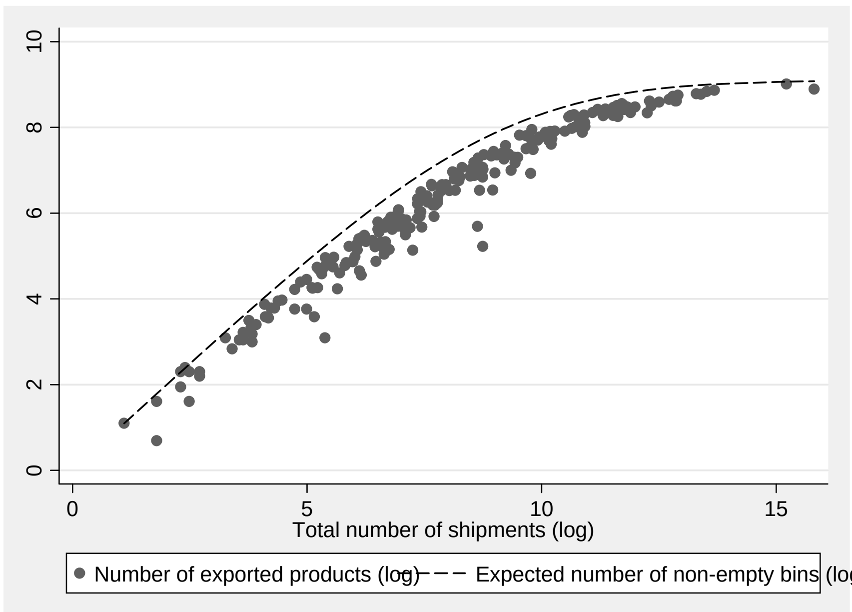

Figure 3: The number of shipments and the number of products

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. 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 4 reports the results for the balls-and-bins model. For convenience, we replicate the regression in Baldwin and Harrigan (2007) and report the resulting coefficients in the first column.[11] 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.

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

Table 4: 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) that 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

[11] 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. using a linear probability model, so coefficients can be understood as marginal effects. Distance is divided in the same categories as in Baldwin and Harrigan (2007). 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.

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 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 is most important for the calibration.

On a more positive note, we show how 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.

2.5 Firm-level zeros

Zeros in firm-level trade flows display a remarkably similar pattern in the data. The average exporting firm in 2000 shipped goods to only 3.5 countries from a total of 229.[12] In other words, 98 percent of potential firm–country trade flows are zero. Again, the zero trade flows follow a well-defined spatial pattern, with zeros being more frequent for small, distant countries.

We can calibrate the balls-and-bins model to study zeros in firm-level trade. The number of balls per destination country are again taken by counting the shipments going to that country. The key difference is that now we need to create bins for *firms* as opposed to product categories. The total number of bins equals the number of exporting firms, 167,217.[13] The size distribution of firm bins is calibrated as follows. We approximate the distribution of exports with a lognormal distribution with mean $\mu = 11$ and standard deviation $\sigma = 3$. This specification matches the mean exports of $5.11 million and has a median exports of $59,300, and does a good job in matching the Lorenz curve reported in Bernard, Jensen and Schott (2007). As it is well known, there is a striking skewness in the distribution of exports across firms.[14]

The balls-and-bins model predicts that 96 percent of the potential firm times 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 5 presents the results. For convenience, we reproduce the regression estimate by Bernard, Jensen, Redding and Schott (2007) in the first column. 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 fewer bins will be left empty. 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.[16]

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

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

[14] Note that this is conditional on having positive exports. 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 4.

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

[16] We calibrated firm bins to the distribution of overall sales in manufacturing, 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. We also explored alternative distribution, like a Pareto, with similar results.

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

Table 5: Exporting firms and gravity – *Balls and bins*

3 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.[17] This way we can use the balls-and-bins model to understand these moments and abstract, for now, from the split between exporters and non-exporters.

The key facts about the extensive margin at the firm level are that while most firms export a single product to a single country, the bulk of exports is done by multi-product, multi-destination exporters.[18]

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.*

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

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

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 minuscule 0.2 % of total exports.

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

Let us turn now to the calibration of our benchmark. 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.[19] 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 6: 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 **??**. As we did earlier, we approximate the distribution of exports with a lognormal distribution with $\mu = 11$ and $\sigma = 3$. 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.

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

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, 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 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 that 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 understanding 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.[20] 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 on 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.[21] 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 the firm-level facts discussed here.

[20] The web appendix contains several alternative calibrations of firm size skewness.

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

4 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 in number 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.[22] Other studies have likewise 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, 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.*

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). Using publicly available data from the Statistics of U.S. Businesses of the Census for year 2002, 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. 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.[23] 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.[24]

[22] See Table 2 in Bernard, Jensen, Redding and Schott (2007). The data is from the 2002 Economic Census.
[23] We also experimented with fitting a Pareto distribution with similar results.
[24] 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.

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.[25] 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

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

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 of ending 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.

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.[26]

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

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

[26]In the web appendix we formally derive the exporter's size premium and include a parametric example.

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.

5 The implications of balls and bins for trade theories

What can an empirical observation about the extensive margin tell us about trade theories? We have claimed that if a fact is matched by balls and bins, then there is a wide class of theoretical models which are also consistent with that fact. In this section we formalize this argument by delineating this class of models.

We present a simple extension of the balls and bins model, which combines *fundamental zeros* (bins closed by a theoretical trade model) with *sample zeros* (bins not receiving any balls by chance). Theories of the extensive margin typically talk about fundamental zeros: trade flows which economic forces prevent from ever taking place. For example, the Melitz model predicts that small firms will never export (making the export bin of a small firm a fundamental zero); while neoclassical models predict that goods in which the country does not have a comparative advantage is never exported (making these country-product bins fundamental zeros). We ask here whether it is possible to make inference about the set of fundamental zeros, and thus to test a theory, from the set of observed zeros, that is, the sum of fundamental and sample zeros.

As in Section 1, there are n shipments that will be classified into K mutually-exclusive categories or bins, which we index in increasing order by size. Let $\Theta \subset \{1, 2, ..., K\}$ denote a set of fundamental zeros that remain empty with probability one. Different models imply different sets of fundamental zeros, $\Theta \neq \Theta'$. If we set $\Theta = \emptyset$, we recover the baseline balls-and-bins model, with no fundamental zeros.

For a given Θ, the expected number of bins that will be empty is just the sum of fundamental and sample zeros,

$$\psi(\Theta) = \sum_{i \in \Theta} 1 + \sum_{i \notin \Theta} (1 - s_i)^n.$$

The second term is the expected number of sample zeros: among open bins there is still a chance that no ball falls in a particular bin. While closed bins count one-to-one towards total zeros, an open bin counts only on the probability it becomes a sample zero, $(1 - s_i)^n$. Note that the expected number of sample zeros is strictly decreasing with the set of fundamental zeros: if $\Theta \subset \Theta'$ then clearly less bins remain to be closed by chance.[27]

The total number of zeros will be informative about theories only to the extent that the expected number of total zeros ψ varies sufficiently as a function of the set of fundamental zeros Θ. The distance between the predictions of two distinct theories, $\psi(\Theta) - \psi(\Theta')$, is given by

$$\psi(\Theta) - \psi(\Theta') = \sum_{i \in \Theta \setminus \Theta'} [1 - (1 - s_i)^n] - \sum_{j \in \Theta' \setminus \Theta} [1 - (1 - s_j)^n] .$$

Intuitively, two theories can be only distinguished on the basis of bins that are closed under one theory but not the other. We can bound the distance between two given theories using the distance from the balls-and-bins model,

$$|\psi(\Theta) - \psi(\Theta')| \leq \psi(\Theta \cup \Theta') - \psi(\emptyset), \tag{7}$$

where $\Theta \cup \Theta'$ is the theory that combines both sets of fundamental zeros. Simply put, the difference in expected zeros cannot be more than the difference between combining all the bins closed in either model and having no bin closed.[28]

Under what conditions is the class of theories similar to balls and bins large? The difference in expected zeros between any theory and the balls-and-bins model is given by

$$\psi(\Theta) - \psi(\emptyset) = \sum_{i \in \Theta} [1 - (1 - s_i)^n].$$

Recall the balls-and-bins model can match the zeros in the data in so far there are many bins that are very likely to be empty, that is, $(1 - s_i)^n$ is close to 1 for many bins. This implies that fundamental and sample zeros are traded virtually one-to-one when shutting down most bins, with little change in expected zeros. Thus if the balls-and-bins prediction is close to the data, even theories that close a majority of bins will not have distinct predictions. By bound (7) we will not be able to properly identify any underlying pattern of fundamental zeros among them.

Perhaps a more concrete example is helpful. Consider a theory Θ such that bins below a certain size are closed. While admittedly stylized, the specification naturally captures the extensive-margin implications of a wide array of models with economies of scale. The theory

[27] Indeed, ψ is a modular function, that is, for any $\Theta \subset \Theta'$ and $x \notin \Theta'$, $\psi(\Theta \cup x) - \psi(\Theta) = \psi(\Theta' \cup x) - \psi(\Theta')$.
[28] Technically, the bound is an implication of the lattice structure of subsets and the modularity of ψ.

specifies a threshold $t \geq 0$ such that if a bin is smaller than t, that bin will remain empty.[29] Let θ denote the number of fundamental zeros; the number of bins that are smaller than t, $\theta = \max\{i : s_i \leq t\}$, Within this class of models the parameter θ fully describe the theory Θ.

We compare the model θ with the balls-and-bins baseline and obtain that the difference in total zeros is given by

$$\psi(\theta) - \psi(0) = \sum_{i=1}^{\theta}[1 - (1 - s_i)^n].$$

The formula has a simple interpretation: it is the number of bins among $\{1, 2, \ldots, \theta\}$ that are expected to be non-empty in *the balls-and-bins baseline.* That is, across a large range of thresholds that possibly close many bins there will be almost no change in total zeros as we are closing the smaller bins that were expected to be empty anyway. But this is, in a nutshell, the reason that the balls-and-bins model is capable of generating a large number of total zeros.

Note that if the data were dense, that is, number of shipments would be very large, total zeros would increase one-to-one with the number of fundamental zeros, as $\lim_{n\to\infty} \psi(\theta) = \theta$. In this case, total zeros would be informative about fundamental zeros, and there would be no problem identifying the relevant model of economies of scale.

We briefly illustrate the discussion here for product-level zeros and the prevalence of exporting firms, using the threshold specification for Θ. Let us start with product-level zeros. The balls-and-bins model predicted that 72.2 percent of product-country pairs would be empty, that is, $\psi(0) = 0.722$. If we set θ to close *half* of all bins, the prediction for total zeros barely budges to 72.5 percent.[30] That is, even though the model closes more than a million country-product pairs, the total number of zeros only changes by 0.3 percent. Simply put, the balls-and-bins model matches the data because the vast majority of bins should be expected to be empty: even the median bin has less than one chance in a hundred to be non-empty. Closing these bins has virtually no impact on the total number of product-level zeros, and thus the latter cannot be used to identify θ precisely, or, more broadly, an estimate of the fixed costs of exporting.

[29] The threshold is stated directly in terms of bin size, but we can always scale the threshold units to shipments by multiplying t by the number of shipments n, and then to dollars by multiplying it with the average shipment value.

[30] As in Section 2.1, we condition on the flow of shipments per country. The results are very similar if we do not condition or even after re-calibrating bin sizes to respect the aggregate distribution of sales over products.

The conclusion is drastically different in the exercise regarding exporting firms. As soon as we start closing exporting bins, the share of exporters drops very fast. Recall that under the benchmark balls and bins calibrations, 74 percent of firms were exporters, that is, only 26 percent were nonexporters. Only by shutting down one tenth of the exporting bins the share of exporters drop below 70 percent. From then on, the share of exporters drops virtually one-to-one with the share of fundamental zeros. For example, closing 20 percent more bins we find the fraction of exporters drops below half. In stark contrast with the previous exercise, the model predictions react sharply to threshold values, so we can take the fraction of exporters as an informative moment of the magnitude of the underlying economies of scale.

6 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 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.

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.[31] 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.

[31] 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 (2009) for Hungary; Eaton, Eslava, Kugler and Tybout (2007) for Colombia.

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

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

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

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

[17] 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.

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

[19] 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.

[20] Halpern, L., Koren, M. and Szeidl, A.: 2009, Imported Inputs and Productivity, Working Paper.

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

[22] 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.

[23] 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.

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

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

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

[27] Krugman, P.: 1980, Scale Economies, Product Differentiation, and the Pattern of Trade, *American Economic Review* **70**, 950-959.

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

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