The Gravity Equation in International Trade: An Explanation

Thomas Chaney

NBER WORKING PAPER SERIES

THE GRAVITY EQUATION IN INTERNATIONAL TRADE: AN EXPLANATION

Thomas Chaney

Working Paper 19285
http://www.nber.org/papers/w19285

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 2013

I want to thank Fernando Alvarez, Michal Fabinger, Xavier Gabaix, Sam Kortum, Bob Lucas, Jim Tybout, Jon Vogel and seminar participants in Berkeley, Bilkent, Bocconi, Boston University, Chicago, Erasmus, Hitotsubashi, LBS, Louvain-CORE, LSE, the NY Fed, Oxford, Princeton, Rochester, Sciences Po, Toulouse, UBC Vancouver, Yale and Zurich for helpful discussions, and NSF grant SES-1061622 for financial support. I am indebted to Jong Hyun Chung, Stefano Mosso and Adriaan Ten Kate for their research assistance. During the last year, I have received compensation for teaching activities from the Toulouse School of Economics, as well a research grant from the National Science Foundation (SES-1061622), in excess of $10,000. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

The Gravity Equation in International Trade: An Explanation
Thomas Chaney
NBER Working Paper No. 19285
August 2013
JEL No. F1

ABSTRACT

The gravity equation in international trade is one of the most robust empirical finding in economics: bilateral trade between two countries is proportional to size, measured by GDP, and inversely proportional to the geographic distance between them. While the role of size is well understood, the role of distance remains a mystery. I propose the first explanation for the gravity equation in international trade, based on the emergence of a stable network of input-output linkages between firms. Over time, a firm acquires more suppliers and customers, which tend to be further away. I show that if, as observed empirically, (i) the distribution of firm sizes is well approximated by Zipf's law and (ii) larger firms export over longer distances on average, then aggregate trade is inversely proportional to distance. Data on firm level, sectoral, and aggregate trade support further predictions of the model.

Thomas Chaney
Department of Economics
University of Chicago
1126 East 59th Street
Chicago, IL 60637
and NBER
thomas.chaney@gmail.com

Introduction

Fifty years ago, Jan Tinbergen (1962) used an analogy with Newton's universal law of gravitation to describe the patterns of bilateral aggregate trade flows between two countries A and B as "proportional to the gross national[1] products of those countries and inversely proportional to the distance between them,"

$$T_{A,B} \propto \frac{(GDP_A)^{\alpha} (GDP_B)^{\beta}}{(Dist_{AB})^{\zeta}}$$

with $\alpha, \beta, \zeta \approx 1$. The so called "gravity equation" in international trade has proven surprisingly stable over time and across different samples of countries and methodologies. It stands among the most stable and robust empirical regularities in economics.

While the role of economic size ($\alpha, \beta \approx 1$) is well understood in a variety of theoretical settings, to this day no explanation for the role of distance ($\zeta \approx 1$) has been found. This paper offers such an explanation for the first time.

The empirical evidence for the gravity equation in international trade is strong. Both the role of distance and economic size are remarkably stable over time, across different countries, and using various econometric methods. Disdier and Head (2008) use a meta-analysis of 1,467 estimates of the distance coefficient ζ in gravity type equations in 103 papers. There is some amount of dispersion in the estimated distance coefficient, with a weighted mean effect of 1.07 (the unweighted mean is 0.9), and 90% of the estimates lying between 0.28 and 1.55. Despite this dispersion, the distance coefficient ζ has been remarkably stable, hovering around 1 over more than a century of data. If anything, Disdier and Head (2008) find a slight increase in the distance coefficient since 1950. The size coefficients α and β are also stable and close to 1. Anderson and van Wincoop (2003) show how to estimate gravity equations in a manner that is consistent with a simple Armington model, and how to deal especially with differences in country sizes.[2] Silva Santos and Tenreyro (2006), Helpman, Melitz and Rubinstein (2008) and Eaton, Kortum and Sotelo (2011) show how to accommodate zeros in the matrix of bilateral trade flows to estimate gravity equations.

[1] Since then, the empirical trade literature has typically used *GDP* as a measure of size rather than *GNP*. As the results are similar with both measures, I will use *GDP* throughout this paper.

[2] McCallum (1995) measures a very large negative effect of the US-Canada border. Anderson and van Wincoop (2003) show that the large difference in the size of the US and Canada explains this seemingly implausible border effect.

Existing theoretical models can easily explain the role of economic size in shaping trade flows, but none explains the role of distance. Krugman's (1980) seminal contribution was motivated in part by the empirical regularity of the gravity equation. His model explains how in the aggregate, trade flows are proportional to country size, and adversely affected by trade barriers. To the extent that distance proxies for trade barriers, his model can also explain why distance has a negative impact on trade flows in general, but it has nothing else to say about the precise role of distance. Several others have shown that the same type of predictions as Krugman can be derived in various other settings. Anderson (1979) derives a similar gravity equation under the Armington assumption that goods are differentiated by country of origin. Eaton and Kortum (2002) derive a similar gravity equation in a modern version of trade driven by Ricardian comparative advantages. Chaney (2008) extends the Melitz (2003) model to derive a similar gravity equation in a model with heterogeneous firms. Arkolakis, Costinot and Rodriguez-Clare (2012) show that the same gravity equation can be derived in many settings with or without heterogeneous firms.

None of these models however can explain the precise role played by distance. The fact that the distance elasticity of trade has remained stable around -1 over such a long time and over such diverse countries is almost a direct rejection of these models. In all of these models, granted that trade costs increase with distance in a log-linear way, the distance elasticity of trade is the product of some deep parameters of the model[3] with the distance elasticity of trade barriers. To explain why the distance coefficient is close to -1, those models need some mysterious alignment of those deep parameters. Even if that magical alignment were to happen in a particular year, for a particular sector and a particular country, it is hard to understand how it could survive beyond that point for more than a century. The technology of transportation, the political impediments to trade, the nature of the goods traded, as well as the relative importance of the countries trading these goods all have undergone some tremendous changes over the course of the last century. In other words, all the deep parameters identified by the various existing trade theories have been evolving over time, while the empirical distance coefficient in the gravity equation has remained essentially constant.

This paper offers the first explanation that is immune to this critique. I explain not only the role of economic size, which is straightforward, but also the role of distance. This explanation is based on the emergence of a stable network of importers and exporters. I assume that there are

[3]The demand elasticity in the Krugman and Armington models, the dispersion of productivities across firms in the Eaton and Kortum model, and a combination of both in the Melitz-Chaney model.

two ways for firms to circumvent the barriers associated with international trade. The first one is to pay a direct cost for creating a foreign contact. This cost is in essence similar to the trade cost assumed in all existing trade models. The second one is to "talk" with one's existing contacts, and learn about the contacts of one's contacts. This second channel requires direct interaction. While advances in the technology for transportation or communication will surely affect the first type of cost, and may even affect the frequency of the second type of interaction, it does not change the need for direct interaction. In my model, the geographic distribution of any one firm's exports does depend on how distance affects the direct cost of creating contacts. But in the aggregate, the details of this distance function vanish, and the gravity equation emerges. This is the main contribution of this paper: even if technological, political or economic changes affect the particular shape of firm level exports, in the aggregate, the gravity equation remains essentially unaffected.

I derive those results in a stylized model of a dynamic network of input-output linkages between firms. This model is most closely related to Oberfield (2013). In Oberfield, firms decide which supplier they want to purchase inputs from, but they only use inputs from a single supplier; moreover, firms are exogenously assigned to a set of potential suppliers. In my model on the other hand, firms decide endogenously how many suppliers they want to buy inputs from; moreover, firms in my model use inputs from various suppliers at the same time. The key difference between both models is that while Oberfield considers explicitly production chains made up of heterogeneous firms, I only characterize a much simpler case where all firms along a production chain are symmetric. I allow firm heterogeneity between production chains, but not within. I also introduce geographic space explicitly into my model, and analyze the patterns of trade, which are absent from Oberfield. In that sense, both models are complementary.

Specifically, I offer three main theoretical contributions.

The first contribution is to build a tractable model of vertical production chains. Firms combine capital and labor with intermediate inputs supplied by upstream firms. I show how firms optimally decide to enter a market, and subsequently, how they optimally invest into acquiring new upstream suppliers. I characterize a balanced growth path in this model where growth is driven both by the entry of new firms, and by the growth of existing firms. I derive explicit expressions for aggregate production, productivity and welfare, as well as the size distribution of firm sizes and of shipment sizes.

The second contribution is to characterize analytically how the geographic distributions of a

firm's suppliers and customers evolve over time. Over time, a firm meets the suppliers (customers) of its existing suppliers (customers). The locations of a firm's initial suppliers are drawn from some exogenous probability distribution. The location of a firm's new suppliers depends in a complex fashion on where the firm's existing suppliers are located, but also on where the suppliers of its suppliers are located... etc. Within a cohort of firms, all those geographic distributions evolve jointly as a complex dynamic system. To characterize this dynamic system, I adopt a probabilistic approach. I solve not for the geographic location of an individual firm's suppliers and customers, which is governed by the luck of the draw, but rather the probability distribution of those locations within a large cohort. Using Fourier analysis, I solve for this dynamic system. Over time, as a firm acquires more upstream suppliers and downstream customers, those suppliers and customers tend to be located further and further away. So the imports and exports of larger firms are shipped over longer distances on average.

The third contribution is to solve for the patterns of aggregate trade flows between countries in this complex dynamic system. From the previous characterization of firm level shipments, I know that as a firm's network of suppliers and customers grows, its efficiency increases, it gets larger, it trades more, and towards more distant trading partners. I show that this dynamic system reaches a stable steady state for any initial condition that satisfies some weak conditions. Because larger firms trade over longer distances, ultimately, whether a country exports little or a lot towards distant places depends on whether there are relatively many or few large firms. In particular, I show explicitly that if the stationary distribution of firm sizes is close to Zipf's law, then aggregate flows are inversely proportional to geographic distance. This result holds asymptotically for any initial condition that govern where small firms export. The model also predicts that systematic deviations away from Zipf's law should be associated with deviations away from the -1 elasticity of aggregate trade with respect to distance.

While the primary contribution of this model is to shed light on the structure of aggregate trade and in particular the role of distance, it also offers tools to address various questions in an economy characterized by complex vertical production chains. For instance, I use this model to quantify the aggregate efficiency loss from trade disruptions, both upon impact and over time.

The remainder of the paper is organized as follows. In section 1, I present a stylized dynamic model of input-ouput linkages between firms. This model offers useful tools for analyzing a dynamic network of firm linkages. This is however not the main contribution of the paper, and the hasty

reader may skip this section so as to focus her attention on the next section. Section 2 contains the main contribution of the paper, namely the characterization of the patterns of trade at the firm level and in the aggregate. I derive general conditions under which the gravity equation holds, i.e., aggregate trade is proportional to country size and inversely proportional to geographic distance. In section 3, I analyze the static and dynamic costs of trade disruptions in this model. In section 4, I bring some of the main theoretical predictions of the model to the data. I relegate to Appendix A all mathematical proofs, and to Appendix B the description of the data.

1 A model of input-output linkages between plants

In this section, I develop a simple model of the formation of a stable network of input-output linkages between firms. The model is an extension of the Krugman (1980) model of international trade in differentiated goods subject to matching frictions similar to the Chaney (2013) model of trade networks.

This model is purposefully simple, and is meant to illustrate how the proposed dynamic model of firm trade can be derived in a classical trade setting. I also use this model to shed new light on the aggregate welfare gains from trade. The hasty reader may skip this section so as to focus her attention on the formation of a stable network of exporters in the following section 2.

1.1 Set-up

There are two types of goods: final goods and intermediate inputs. Final goods are produced by combining differentiated intermediate inputs available locally. Intermediate inputs are themselves produced by combining differentiated inputs and labor, so that the economy features roundabout production. Final goods are sold locally to consumers on a perfectly competitive market. Intermediate inputs are produced and distributed worldwide by monopolistically competitive firms. Since only intermediate inputs are traded, I will focus most of my attention on the production and trade of these intermediate goods. Due to matching frictions, intermediate input firms source their inputs from, and sell their output to a subset of producers only.

Final goods.— Final goods are sold on a competitive market by atomistic firms that share the same constant returns production function,

$$Q_{\text{final}} = \left(\int_{i \in \mathcal{K}} q_{\text{final}}(i)^{\frac{\sigma-1}{\sigma}} \, di \right)^{\frac{\sigma}{\sigma-1}} \tag{1}$$

where $\mathcal{K}$ is the set of intermediate goods available locally for final assembly, $q_{\text{final}}(i)$ is the quantity of inputs sold by firm i for final assembly, and $\sigma > 1$ is the elasticity of substitution between inputs.

Intermediate inputs.— Firm i buys intermediate inputs from a continuum of suppliers $k \in \mathcal{K}_i$ and sells its output to a continuum of customers $j \in \mathcal{J}_i$ as well as to the final good producers on a monopolistically competitive market. Both sets $\mathcal{K}_i$ and $\mathcal{J}_i$ will be endogenously determined dynamically below. Firm i combines L_i units of labor with $q_i(k)$ units of differentiated intermediate inputs from each supplier k to produce Q_i units of output,

$$Q_i = \left(\int_{k \in \mathcal{K}_i} q_i(k)^{\frac{\sigma-1}{\sigma}} dk \right)^{\alpha \frac{\sigma}{\sigma-1}} L_i^{1-\alpha} \tag{2}$$

with $0 < \alpha < 1$ the share of intermediate inputs in production and $\sigma > 1$ the elasticity of substitution between any two intermediate inputs.

Free entry.— There is a perfectly elastic fringe of potential entrepreneurs ready to start an intermediate input producing firms. Starting up a business entails a constant fixed entry cost of f_E units of labor. This free entry of new firms ensures that in equilibrium, all profits are dissipated.

Workers and consumers.— In every location, there is a competitive labor market with an equilibrium wage w, and a measure L_t of consumers/workers at time t. Population grows at a constant rate γ.

1.2 The static problem of the firm

Consider what happens within period t. For the moment, I drop the time t index.

Firm demand.— Each period, firm i faces the same iso-elastic demand from any customer, whether they are another intermediate good producer of a final good producer, $j \in \{\mathcal{J}_i \cup \text{final}\}$,

$$p_j(i)\, q_j(i) = \frac{p_j(i)^{1-\sigma}}{\int_{k \in \mathcal{K}_j} p_j(k)^{1-\sigma} dk} X_j \tag{3}$$

with $p_j(i)$ the price charged by i to customer j, and X_j the total spending on intermediate inputs by j.

Firm pricing.— Given these iso-elastic demands, firm i charges all its customers the same constant mark-up, $\frac{\sigma}{\sigma-1}$, over its marginal cost,

$$p_j(i) = p_i = \frac{\sigma}{\sigma-1} w^{1-\alpha} \left(\int_{k \in \mathcal{K}_i} p_i(k)^{1-\sigma} dk \right)^{\frac{\alpha}{1-\sigma}} \tag{4}$$

with w the competitive wage rate.

Symmetry assumption.— For simplicity, I will consider a symmetric equilibrium where all firms within a cohort have the same number of suppliers and customers[4] and therefore charge the same price: denoting $K = \|\mathcal{K}\|$ the measure of set $\mathcal{K}$, a symmetric equilibrium will be such that $K_j = K$ and $p_j(k) = p$ for any $j, k \neq i$.

This symmetry assumption is strong. It simplifies greatly the analysis of the optimal strategies of firms, of the equilibrium, and of the patterns of firm level and aggregate trade flows. Relaxing this assumption forces me to keep track of the joint distribution of suppliers and customers, in every location, and renders the analysis complex. I show in the Appendix how to relax this assumption. The most important propositions (2 and 3) of this paper are preserved, but at a great cost in terms of analytical complexity. While the simplicity of the analysis it affords is the primary justification for this simplifying assumption, it is also a somewhat appropriate approximation of reality: Atalay, Hortacsu, Roberts and Syverson (2011) show for the US a tendency towards positive assortative matching, whereby firms with many suppliers tend to be connected to suppliers with many suppliers themselves; Acemoglu, Carvalho, Ozdaglar and Tahbaz-Salehi (2012) show a similar property at the level of industrial sectors.

Firm sales.— Given this symmetry assumption, the demand equation (3) and the pricing equation (4), the total sales of firm i only depend on its number of suppliers and of customers,

$$p_i Q_i = \int_{j \in \mathcal{J}i} p_i q_j(i)\, dj + p_i q_{\text{final}}(i) = K_i^{\alpha} D_i \tag{5}$$

where D_i is a demand shifter that does not depend on i's price.

It is clear from the previous Equation (5) that the number of suppliers and customers increases output, sales, and ultimately profits which are a constant fraction $1/\sigma$ of total sales. The mass of suppliers, or "contacts," K_i, plays a role equivalent to capital, or to a productivity shifter. I will use the "capital" analogy, and denote by I_i the "investment" in acquiring "information" about new contacts. The notion that the diversity of intermediate inputs plays a role similar to capital has been explored since at least Romer (1990).[5] The mass of customers, J_i, plays a role equivalent to a proportional demand shifter. A firm is willing to pay for information about new upstream and downstream contacts, as well as sell the information it has about its existing contacts. Before analyzing this dynamic decision of the firm, I first describe the general equilibrium of this economy

[4] In such a symmetric equilibrium, a lot the complexity of the input-output structure of the economy is assumed away. See Carvalho, Acemoglu, Ozdaglar and Tahbaz-Saleh (2012), Atalay, Hortacsu, Roberts and Syverson (2011) or Oberfield (2013) for models with a more complex structure.

[5] See among many examples the theoretical model of di Giovanni and Levchenko (2010) or the empirical evidence of Halpern, Koren and Szeidl (2011) for two recent applications of this notion in trade.

in each period.

1.3 The static general equilibrium

Having characterized the behavior of each firm individually, I can now solve for the general equilibrium of the model at each period.

Equilibrium firm price.— Each firm charges a price to any of its consumer that is constant mark-up $\frac{\sigma}{\sigma-1}$ over its marginal cost of production. The firm's marginal cost itself depends on the prices charged by its suppliers, which in turn depend on their marginal cost... etc. The distribution of prices is therefore the solution to a potentially complex fixed point problem. Under the symmetry assumption above, this fixed point problem is substantially simplified. A firm with K suppliers only buys intermediate inputs from other firms with K suppliers. In this symmetric equilibrium, that firm charges a price to its consumers that is the same as the price it pays to its suppliers. Using the pricing expression in Equation (4), I can solve for the unit price p_K of a firm with K suppliers,

$$p_K = \frac{\sigma}{\sigma-1} w^{1-\alpha} \left(\int_{\mathcal{K}} p_K^{1-\sigma} dk \right)^{\frac{\alpha}{1-\sigma}} \Rightarrow p_K^{1-\sigma} = K^{\frac{\alpha}{1-\alpha}} \tag{6}$$

where I have normalized the competitive wage w so that $\frac{\sigma}{\sigma-1} w^{1-\alpha} = 1$. A firm with K suppliers sells its output to other intermediate input producers (each with K suppliers themselves), as well as to the final good producer. When selling to an intermediate good producer, that firm competes against K other intermediate good producers (each charging the same price for their own inputs). When selling to the final good producer on the other hand, that firm competes against all local intermediate input producers, some with more, some with fewer suppliers. The total sales of firm i with K suppliers is then,

$$p_i Q_i = p_i^{1-\sigma} \int_{\mathcal{J}_i} \frac{X_j}{\int_{\mathcal{K}_j} p_k^{1-\sigma} dk} dj + p_i^{1-\sigma} \frac{y}{P^{1-\sigma}}$$

where X_j is the total spending on intermediate inputs by firm j, P is the ideal price index for all intermediate goods purchased by the local final good producer, and y is the total spending on intermediate inputs by this final good producer.[6]

Equilibrium aggregate prices.— Having solved for the price of any firm in Equation (6), if there

[6] As the final goods market is competitive, the total spending on intermediate inputs by the final goods producer, y, is equal to the income of the local consumers. Because of free entry, all profits are dissipated, and aggregate income equals wL.

is a measure M of firms in each location, then the ideal price index is given by,

$$P = \left(\int_{K_0}^{\infty} p_K^{1-\sigma} M dF(K) \right)^{\frac{1}{1-\sigma}} = \left(M \mathbb{E}_F \left[K^{\frac{\alpha}{1-\alpha}} \right] \right)^{\frac{1}{1-\sigma}} \tag{7}$$

where M is the measure of firms and F is the cumulative distribution of firms with K suppliers in the population. I derive an explicit solution for the distribution F and prove that it is time invariant in Section 2. I solve endogenously for the measure of firms M in Section 1.5 where I characterize the optimal entry decision of firms.

Several observations are in order. First, the aggregate price index decreases with the measure of firms, as in any model with love for variety. Second, more efficient firms (lower price firms) contribute more to lowering the aggregate price index, as in any model with heterogeneous firms. The novel aspect of this model is that those firms with lower prices are typically firms with many suppliers, which are part of more complex production chains. More complex and longer production chains contribute to more efficient production, and a lower aggregate price index. Those more complex production chains are more fragile in the sense that they involve a larger number of firms; on the other hand, to the extent that intermediates are substitutable ($\sigma > 1$), firms in those complex production chains also have more options to substitute away from a failing supplier.

Equilibrium sales and shipments.— Any firm j charges a constant mark-up $\frac{\sigma}{s-1}$ over marginal cost, it spends a constant fraction α on intermediates inputs, so its total spending on inputs is a constant fraction of its total sales, $X_j = \alpha \frac{\sigma-1}{\sigma} p_j Q_j$. Furthermore, given the symmetry of the equilibrium, $p_i Q_i = p_j Q_j$, $J_i = K_j = K$ and $p_i = p_k = p_K$ for any firm i, j or k with Ksuppliers. I can then solve for the total sales of a firm with K suppliers,

$$\text{Sales}(K) = K^{\frac{\alpha}{1-\alpha}} \left(\frac{wL/P^{1-\sigma}}{1 - \alpha \frac{\sigma-1}{\sigma}} \right) \tag{8}$$

and for the value of the shipment of intermediate inputs from a firm with K suppliers to any one of its customer,

$$\text{Shipment}(K) = K^{\frac{1-2\alpha}{1-\alpha}} \left(\frac{wL/P^{1-\sigma}}{\frac{\sigma}{\alpha(\sigma-1)} - 1} \right) \tag{9}$$

Firms with more suppliers are more efficient at producing, they sell to more other intermediate input producers, they charge a lower price and are able to capture a larger market share of the final demand. The higher is the share of intermediate inputs α, the stronger is the multiplier effect of having more suppliers on a firm's productivity, and the larger the sales of firms with many

suppliers compared to firms with few. Whether a firm's shipment of intermediate inputs increases or decreases with its efficiency depends in a subtle way on the share of intermediate inputs in production. A firm with K suppliers competes against other firms with K suppliers when selling to any of its customers. Whether the efficiency of the one firm or that of its competitors dominates in determining the market share of each depends on whether the share of intermediate inputs if smaller of larger than that of labor.

In the special case where $\alpha = \frac{1}{2}$, the total sales of a firm with K suppliers (and K customers) is simply proportional to K, and all shipments have the exact same value. This special case is of course knife-edge, but it will prove a useful benchmark for the trade model I develop below. Under this assumption, the total volume of trade between two countries, whether for one or for many firms, is simply proportional to the number of shipments between the two countries. The analysis below considers this special case where all trade volumes are only driven by the extensive margin of shipments. Relaxing the simplifying assumption $\alpha = \frac{1}{2}$ would introduce an intensive margin of shipments.

1.4 The dynamic problem of the firm

Having characterized the static optimization of the firm, as well as solved for the static general equilibrium the firm faces each period, I can now turn to the dynamic problem of the firm.

The market for information.— A firm is born with a mass $K_0 = J_0$ of suppliers and customers. Once born, a firm expands its set of suppliers and customers. While a priori, firm i may buy and sell information about both suppliers and customers, I assume that i actively buys information about new suppliers from its existing suppliers, actively sells information about suppliers to its existing customers, but passively waits to be contacted by downstream firms. In a symmetric equilibrium, this simplification is innocuous since firms have as many suppliers as customers on average: if i is a supplier to j then j is a customer of i. The sequence of contact formation is depicted in Figure 1.

I assume that a firm who wants to buy information about a supplier always has the option of directly searching for suppliers on its own. This outside option technology offers new names at a given constant marginal cost of p_I units of labor. The seller of information makes a take-it-or-leave-it offer to any potential buyer of information. Facing the threat of this outside option, firm i sets a constant price $p_I w$ to reveal the name of one of its suppliers to its existing customers. The price $p_I w$ is set just low enough to prevent firm $j \in \mathcal{J}_i$ to look for contact directly instead. Just as

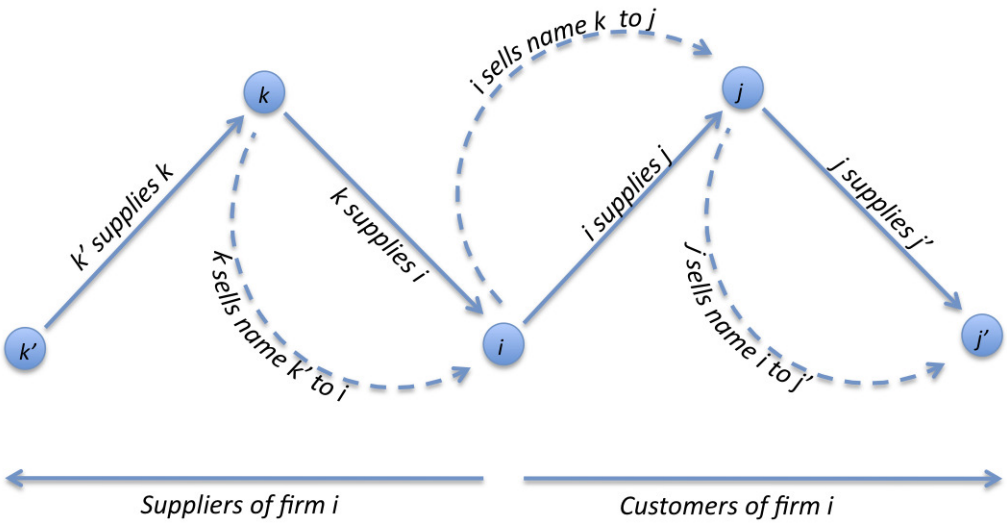

Figure 1: Firms buy and sell information about suppliers and customers.

Notes: The straight solid arrows represent input-output linkages: e.g. firm k supplies intermediate inputs to firm i. The curvy dotted lines represent information linkages: e.g. firm k sells to firm i information about a new supplier k'. After these information exchanges take place, firm i has a new supplier, k', and a new customer, j'.

i sells information about its suppliers, it also buys at the same price $p_I w$ information about new suppliers from its existing suppliers.

In addition to the direct cost $p_I w$ of buying information, firm i faces a convex adjustment cost $wC(I_i, K_i)$ denominated in units of labor. The adjustment cost function is assumed increasing and convex in I, $C_I, C_{II} > 0$, decreasing in K, $C_K < 0$, and homogenous of degree one in I and K. This convex adjustment cost function captures the idea that bringing a new supplier into a firm's production process entails some cost: those inputs have to be customized to fit into the production process, and the production process itself has to be adapted to be combined with this new source of inputs. It is analogous to the adjustment cost assumed in the classical theory of investment, as in Lucas (1967) or Hayashi (1982). As in the investment literature, I assume that the more suppliers a firm already has (K_i), the more efficient it is at acquiring new suppliers (I_i), in such a way that the adjustment cost C is proportional to K_i for a given investment share I_i/K_i. As in Lucas (1967), this homogeneity assumption will warrantee that Gibrat's law holds, in the sense that the growth rate of a firm is independent of its size.

As in classical investment models, firm i has two reasons for accumulating more suppliers, i.e., "investing" in K_i: first, a higher K_i increases its productivity and profits; but it also lowers the future cost of "investment" in acquiring new suppliers, I_i. However, while firm i sells information

about its suppliers to its customers, having more suppliers does not change firm i's future prospect for selling more information: the price firm i sets for selling information about suppliers, $p_I w$, is set by an arbitrage condition, and the number of requests for names firm i receives depends on the decisions of its customers, $j \in \mathcal{J}_i$. At each point in time t, firm i receives $I(t)$ requests for names, where $I(t)$ depends on the "investment" decision of downstream firms, which is beyond firm i's control.

Finally, I assume that firm i's existing contacts are lost at some exogenous rate δ.

The dynamic problem of the firm.— Firm i solves the following dynamic optimization problem,

$$\max_{I_i(t)} V(0) = \int_0^{+\infty} e^{-rt} \left(K_i(t)^{\alpha} \frac{D_i(t)}{\sigma} - p_I w(t) I_i(t) + p_I w(t) I(t) - wC(I_i(t), K_i(t)) \right) dt \tag{10}$$

$$\text{s.t. } \dot{K}_i = I_i - \delta K_i$$

Firm i maximizes a discounted stream of profits, with a constant discount rate r. The first term represents per period profits, net of spending on intermediate inputs and labor, but before spending and receipts on information acquisition. It is a fraction $1/\sigma$ of the aggregate sales derived in Equation (5). In addition, firm i purchases information about I_i new suppliers at a price $p_I w$ each, and it sells information about I suppliers at a price $p_I w$ each. Finally, firm i pays the convex adjustment cost wC to integrate new suppliers into its production process.

The solution to this classical problem is such that the "investment" rate is independent of the stock of "capital" (Gibrat's law). In other words, firm i increases its number of suppliers K_i at a rate that is independent of K_i,

$$I_i(t) = \beta(p_I, r, \delta; t) \times K_i(t)$$

where the function β summarizes the contributions of the production function and the adjustment cost function that are relevant for the optimal investment decision. I derive formally in Appendix A the optimal investment policy function that solves (10), as well as explicit conditions under which the investment rate is constant over time and across all types of firms.

Because all firms are charging the same price $p_I w$ per contact information, firm i has no reason to direct its search for new suppliers to any particular $k \in \mathcal{K}_i$. To break this indeterminacy, I assume that the I_i new names are randomly drawn (uniformly) from the set of all existing suppliers $\mathcal{K}_i$. This means that any one of the existing suppliers $k \in \mathcal{K}_i$ reveals one of the names

of its suppliers, $k' \in \mathcal{K}_k$, with a probability βdt over a small time interval dt. To break the indeterminacy of which name $k' \in \mathcal{K}_k$ gets revealed by firm k, I simply assume that k draws k' at random among all its existing $\mathcal{K}_k$ contacts.

1.5 Entry decision and balanced growth path

Finally, I solve for the entry choice of new firms. After entry, monopolistic firms are able to extract a surplus from their consumers. This stream of profits attracts new entrants as long as its discounted value exceeds the fixed entry cost.

In this model, as in any model where productivity growth is driven by the accumulation of one factor of production (here the diversity of suppliers, K) combined with labor under constant returns to scale, growth is ultimately driven by population growth, as in Solow (1956). I will now characterize the entry decision of firms along a balanced growth path. Along this balanced growth path, population growth induces a continuous entry of new firms.

Free entry condition.— Each period, the symmetric equilibrium described in Section 1.3 holds, and all firms "invest" in acquiring new suppliers at the constant rate β as described in Section 1.4.

Consider a potential entrant that contemplates entering at time t. Its per period profits would be a fraction $1/\sigma$ of its sales, given in Equation (8). In a symmetric equilibrium, the firm would spend each period exactly as much in acquiring information about new suppliers as it would earn selling information about its consumers. The value of entry at time t, $V(t)$, is the discounted sum of those per period profits. Firms will enter in period t as long as it exceeds the cost of entry wf_E.

The perfectly elastic supply of potential entrants implies that the value of entry must exactly equal the cost of entry in every period,

$$V(t) \equiv \int_t^{+\infty} e^{-r(s-t)} \left(\frac{K(s)^{\frac{\alpha}{1-\alpha}} wL(s) / P(s)^{1-\sigma}}{\sigma - \alpha(\sigma - 1)} - wC(I(s), K(s)) \right) ds = wf_E, \forall t \qquad (11)$$

I will now use this series of free entry conditions to solve for the entry of firms along a balanced growth path.

Balanced growth path equilibrium.— After a firm enters, its measure of suppliers grows at a rate $\beta - \delta$, so that its measure of suppliers at age $s \geq 0$ is $K(s) = K_0 e^{(\beta-\delta)s}$. The "investment" rate is constant, so that $I(s) = \beta K(s)$. The homogeneity of the adjustment cost function C implies that $C(I(s), K(s)) = K(s) C(I(s)/K(s), 1) \equiv K(s) c(\beta)$.

The population grows at the exogenous rate γ so that $L(s) = L_0 e^{\gamma s}$. Along a balanced growth

path, the measure of firms grows at some constant endogenous rate γ_M. The ideal price index given in Equation (7) is therefore such that $P(s)^{1-\sigma} = M_0 \mathbb{E}_F \left[K^{\frac{\alpha}{1-\alpha}} \right] e^{\gamma_M s}$.

Combining those elements with the free entry condition (11), the following condition must hold each period,

$$\int_0^\infty e^{-rs} \left(\frac{K_0^{\frac{\alpha}{1-\alpha}} e^{\frac{\alpha}{1-\alpha}(\beta-\delta)s} w e^{\gamma(s+t)}}{(\sigma - \alpha(\sigma-1)) M_0 \mathbb{E}\left[K^{\frac{\alpha}{1-\alpha}} \right] e^{\gamma_M (s+t)}} - K_0 e^{(\beta-\delta)s} c(\beta) \right) ds = w f_E, \forall t \tag{12}$$

For this equation to hold for all t, it is apparent that $\gamma_M = \gamma$ is necessary.[7] In other words, the measure of firms in every location grows at the same constant rate γ as the population. Once this condition is imposed, Equation (12) pins down M_0 and the measure of firms each period. The existence and uniqueness of a balanced growth path equilibrium is warranted.

Here is a recap of the conclusions from the above model. Firms are continuously born at a rate γ. A firm is born with an initial mass K_0 of suppliers. Subsequently, contacts are randomly created at a rate β and lost at a rate δ, with each new contact coming from the suppliers of the firm's existing suppliers. In addition, and under the simplifying assumption that the share of intermediate inputs is a half, a firm's size is proportional to its number of suppliers, and shipment sizes are the same for all firms.

The next section characterizes explicitly the dynamic evolution of firm level and aggregate trade flows, i.e., trade between the suppliers and customers of this model.

2 Firm level and aggregate trade flows

This section contains the main contribution of this paper. It explains the stable role played by geographic distance in shaping aggregate trade flows. The central results are given in Propositions 2 and 3. I introduce space explicitly in the model of input-output linkages of the previous section. I first analyze the patterns of trade at the firm level in section 2.1, before characterizing aggregate trade flows in section 2.2. I present some robustness checks in section 2.3.

2.1 The geography of firm level trade

In this section, I spell out a dynamic model of firm level trade flows that incorporates the key results derived from the economic model in the previous section. All the parameters introduced in

[7] The above free entry condition is of the form, constant + $\frac{e^{\gamma t}}{e^{\gamma_M t}}$constant = constant. This can hold for any t only if $\gamma_M = \gamma$.

this section $(K_0, \beta, \delta, \gamma)$ are the same as the ones in the economic model above. I treat the arrival rate of new firms, γ, and of new contacts, β, as parameters, knowing from the model above that they are the solutions to the dynamic optimization problem of the firm and the entry decision of new firms. The contacts in this section are the customers of the above model. The key addition of this section is the introduction of an explicit notion of geographic space.

Heuristically, the model is as follows.

New firms are continuously born. When a firm is born, it randomly contacts a geographically biased mass of firms over the entire world. After this initial period, contacts are randomly lost and created. Old contacts are lost to i.i.d. shocks. New contacts are created in the following way: each period, with some probability, a firm receives names from the contact lists of its existing contacts. In other words, a firm gradually meets the contacts of its contacts, who themselves acquire contacts in a similar way, etc.

Formally, the model is as follows.

Space.— Firms are uniformly distributed over an infinite one-dimensional continuous space represented by $\mathbb{R}$. Each coordinate along that line can be thought of as representing a city, and countries can be thought of as an arbitrary partition of that space, where a country is then a collection of cities, or an interval of the real line.

Time.— Time is continuous. In every location, new firms are born continuously, with the population of firms in each location growing at a constant rate γ, where γ stands for "growth". At time t, there is the same density of firms $e^{\gamma t}$ in every location, where I normalize the population at $t = 0$ to 1. As the model is perfectly symmetric, I will focus my attention on a firm located at the origin.

Birth of a firm.— When a firm is born, it samples a mass K_0 of contacts, distributed geographically according to the p.d.f. $g_0(\cdot)$. So the mass of contacts it acquires in the interval $[a, b]$ is $K_0 \int_a^b g_0(x)\, dx$. I assume that g_0 is symmetric and has a finite variance, but can take any arbitrary shape otherwise. For simplicity, I assume that when a firm is born, it samples contacts only among other firms of the same age: firms within each cohort gradually get connected to each other. While this simplifying assumption is strong, it simplifies the analysis greatly. The main set of predictions from the model still hold under the much weaker assumption that when a firm is born, it samples its K_0 new contacts among all existing firms, in such a way that those firms are at a finite average (squared) distance between from their own contacts. While this relaxed

assumption is much weaker, it forces me to keep track of the entire system of connections between all firms simultaneously. This renders the model analytically complex.[8]

Death of a firm.— I assume that firms are infinitely lived. This assumption is innocuous, and all results would carry through if firms are hit by random Poisson death shocks. A positive death rate for firms would simply be added to the death rate of contacts below.

Birth of contacts.— New contacts are continuously created as follows. At any point in time, each existing contact may reveal one of its own contacts according to a Poisson process with arrival rate β, where β stands for "birth".

Death of contacts.— Existing contacts are continuously lost according to a Poisson process with arrival rate δ, where δ stands for "death".

I assume $\gamma > \beta - \delta > 0$. While the second assumption $\beta - \delta > 0$ is not required to derive my results, it would generate counter-factual predictions, such as an infinitely long tail of infinitesimally small firms and firm sizes shrinking on average. The first assumption, $\gamma > \beta - \delta$ is required for a stable equilibrium to emerge, as it prevents older firms from becoming "too" large compared to the rest of the economy.

I will now define two concepts: the function f_t describes the geographic distribution of the contacts of a firm of age t, and the variable K_t describes the total mass of contacts of this firm,

$$f_t : \mathbb{R} \to \mathbb{R}^+ \text{ and } K_t \equiv \int_{\mathbb{R}} f_t(x)\, dx \tag{13}$$

$f_t(x)$ is the density of contacts a firm of age t has in location x. In other words, the mass of contacts a firm of age t has in the interval $[a, b]$ is $\int_a^b f_t(x)\, dx$. The total mass of contacts a firm of age t has worldwide is then K_t. Note that as f_t does not sum up to 1, it is not a probability density. The normalized f_t/K_t on the other hand is a well defined p.d.f.

The distribution of contacts evolves recursively according to the following Partial Differential Equation,

$$\frac{\partial f_t(x)}{\partial t} = \beta \int_{\mathbb{R}} \frac{f_t(x-y)}{K_t} f_t(y)\, dy - \delta f_t(x) \tag{14}$$

with the initial condition $f_0(x) = K_0 g_0(x)$.

I multiply both sides of the equation by dx for a rigorous interpretation. The left hand side of Equation (14) corresponds to the net creation of new contacts in a neighborhood dx of x. It

[8]The proof of the main proposition under the relaxed assumption is presented in the online appendix. See `https://sites.google.com/site/thomaschaney/Distance_Appendix.pdf`.

is equal on the right hand side to the gross creation of new contacts minus the destruction of old contacts. The gross creation of contacts can be decomposed into four components: β, $\frac{f_t(x-y)}{K_t}dx$, $f_t(y)\,dy$ and the integral sign $\int_{y\in\mathbb{R}}$. The first component, β, corresponds to the Poisson arrival of new information from a firm's contacts. With a probability βdt over a small time interval dt, any one of a firm's contact in location y will reveal the name of one of her own contacts. The second component, $\frac{f_t(x-y)}{K_t}dx$, corresponds to the probability that conditional on a contact in location y revealing the name of one of her contacts, this contact happens to be in a neighborhood dx of x.[9] Note here that I impose the simplifying assumption that a firm of age t only meets other firms in the same cohort, who themselves have the same distribution f_t. I show in the online appendix how this strong assumption can be substantially relaxed. The third component, $f_t(y)\,dy$, corresponds to the fact that a firm of age t has potentially several contacts in a neighborhood dy of y (exactly $f_t(y)\,dy$ of them), each of whom can potentially release the name of one of its contacts in x. The fourth component, $\int_{y\in\mathbb{R}}$, corresponds to the fact that the information about new contacts in x can potentially be intermediated via contacts in any location $y \in \mathbb{R}$. The second term with the minus sign on the right hand side of Equation (14) corresponds to the destruction of old contacts. Any one of the existing $f_t(x)\,dx$ contacts of a firm of age t in a neighborhood dx of x may be destroyed with the same probability δdt over a small time interval dt.

The Partial Differential Equation (14) admits an explicit analytical solution, which I relegate to Appendix A in the interest of conciseness. While the mathematically less inclined reader may skip the derivation of this solution, it contains a number of analytical tools that may prove useful in a variety of economic settings. The analytical solution to the geographic distribution of contacts f_t allows me to derive closed-form solutions for the number of contacts of an individual firm, its distribution within the population, and the geographic location of these contacts. Formal proofs of all results are provided in Appendix A.

First, the model predicts that as a firm ages, the number (mass) of its contacts increases,

$$K_t = K_0 e^{(\beta-\delta)t} \tag{15}$$

The total number of a firm's contacts grows at a constant rate equal to the net birth rate of contacts (birth rate β minus death date δ).

[9] Since the distribution f_t sums up to K_t, the normalized $\frac{f_t}{K_t}$ is a well defined p.d.f. that sums up to one. Moreover, the distribution of contacts for a firm located in y is the same as for a firm located in the origin ($y = 0$), where all coordinates are simply shifted by the constant $-y$: $f_{0,t}(x) = f_{y,t}(x-y)$.

Second, as both the number of a firm's contacts and the number of firms grow exponentially, the model predicts that the distribution of the number (mass) of contacts within the population is Pareto distributed. The fraction $F(K)$ of firms with K or fewer contacts is given by,

$$F(K) = 1 - \left(\frac{K}{K_0}\right)^{-\frac{\gamma}{\beta-\delta}} \text{ for } K \geq K_0 \tag{16}$$

From Equation (15), young firms have fewer contacts than old ones. The larger is the growth rate of the population as a whole, γ, the more young firms relative to old ones, the fewer firms with a large number of contacts, and the thinner the upper tail of the Pareto distribution of the number of contacts. From Equation (15) also, the higher is the growth rate of a firm's contacts, the larger the mass of contacts of old firms relative to young ones. The larger is the net birth rate of new contacts, $\beta - \delta$, the more firms with many contacts, and the fatter the upper tail of the Pareto distribution of the number of contacts.

If, as is approximately verified in the data, the cross-sectional distribution of the sizes of exporters is close to a Zipf's law, then we should expect the Pareto shape parameter to be close to 1, $\frac{\gamma}{\beta-\delta} \approx 1^+$.[10] While deviations from this stationary benchmark are to be expected in the data, these deviations ought not to be too large.

Third, the model predicts that as a firm ages, not only does it acquire more contacts, but those contacts become increasingly dispersed over space. Let me denote by f_K the geographic distribution of contacts of a firm with K contacts.[11] The average (squared) distance from the contact of a firm with K contacts, $\Delta(K)$, increases with its number of contacts,

$$\Delta(K) \equiv \int_{\mathbb{R}} x^2 \frac{f_K(x)}{K} dx = \Delta_0 \left(\frac{K}{K_0}\right)^{\frac{\beta}{\beta-\delta}} \tag{17}$$

where $\Delta_0 \equiv \int_{\mathbb{R}} x^2 g_0(x)\, dx$ is the average (squared) distance from a firm's initial contacts. While a firm's initial contacts are some distance away, each wave of new contacts come from firms who are themselves further away. As a consequence, each wave of new contacts is geographically more dispersed than the previous one. This effect is compounded by the fact that a firm's contacts are also acquiring more and more remote contacts. Since a firm acquires more contacts as it ages, the more contacts a firm has, the more dispersed these contacts are.

[10] Note that this simple model with a constant growth rate of the population and of the number of contacts corresponds to the Steindl (1965) model of firm growth. More elaborate stochastic models such as Gabaix (1999) or Luttmer (2007) deliver an invariant size distribution that is close to Zipf's law in a more general set-up. I choose to use the simpler Steindl model while adding substantial complexity on the geographic dimension of the model. I conjecture that including the stochastic elements of Gabaix (1999) or Luttmer (2007) would not change my results.

[11] f_K is defined as $f_K = f_{t(K)}$ where $t(K)$ s.t. $K_{t(K)} = K$ is the age a firm has to reach to have K contacts.

Note that the particular moment $\Delta(K)$ only depends on two parameters of the distribution g_0: Δ_0 and K_0. For any two economies with different g_0 and g_0' but with the same Δ_0 and K_0, the average (squared) distance from a firm's contact will evolve in the exact same way as a firm acquires more contacts, no matter how different g_0 and g_0' are otherwise. This result arises for the same reason that the n-th derivative of the composition of several functions only depends on their first n derivatives: $\Delta(K)$ is the second moment of the p.d.f. $g_K = f_K/K$, which is given by the second derivative of the moment generating function of g_K; this second derivative does not depend on any derivative of the moment generating function of g_0 above the second one.

I do not need to characterize the geographic distribution of contacts of firms any further to derive the main proposition of this paper regarding aggregate trade, presented in the next section. I will however show a particular asymptotic property of this general solution. It will prove useful for a variety of questions, and in particular to derive precise predictions for the aggregate welfare cost of trade disruptions which I study in Section 3. The following proposition[12] characterizes the geographic distribution of a firm's contacts, for a special case for the initial g_0 as well as asymptotically for a large t.

Proposition 1 *The distribution of a firm's contacts, $\frac{f_t}{K_t}$, converges when t grows large to a Laplace distribution (a 2-sided exponential),*

$$\frac{f_t(x)}{K_t} \underset{t\to\infty}{\sim} Laplace\left(0, e^{\beta t/2}\sqrt{\Delta_0/2}\right)$$

This property holds exactly for all t's if $g_0 \sim Laplace\left(0, \sqrt{\Delta_0/2}\right)$.

One can see from this proposition that as a firm grows bigger, the location of its contacts is less and less affected by distance. The distribution of a firm's contacts converges to what resembles a uniform distribution over the entire real line in a strong sense: for any two locations x and y, no matter how far x is from y, the fraction of contacts in x and in y become equal for t large. In other words, the world does become "flat" for *individual* firms as they grow large. But, as I will show in the next section, this does not mean that the world becomes "flat" in the *aggregate*.

Having characterized the distribution of contacts for all firms, I analyze next the aggregate distribution of contacts, and the structure of aggregate trade flows in this economy.

[12] I am grateful to Xavier Gabaix for suggesting this extension.

2.2 The geography of aggregate trade

Under the simplifying assumption that the share of intermediates in production is a half, all shipments sizes are the same, and a firm exports one shipment to each of its contacts. The volumes of trade between two locations, both at the firm and at the aggregate levels, are then simply proportional to the number of shipments between those locations. I have shown in the previous section that older firms have more numerous and dispersed contacts. Knowing the distribution of contacts of each firm, I can characterize the patterns of aggregate trade flows between firms in any set of locations. The following two propositions show that aggregate trade flows in this model obey the gravity equation in international trade.

Proposition 2 *For any distribution g_0 of initial contacts that is symmetric and admits a finite variance, aggregate trade flows between two countries A and B are approximately proportional to their respective sizes (GDP_A and GDP_B), and inversely related to the distance between them ($Dist_{A,B}$),*

$$T_{A,B} \propto \frac{GDP_A \times GDP_B}{(Dist_{A,B})^{1+\epsilon}}$$

with $\epsilon \equiv 2 \min\left(\frac{\frac{\gamma}{\beta-\delta}-1}{\frac{\beta}{\beta-\delta}}, 1\right)$, γ the population growth rate and β (resp. δ) the birth (resp. death) rate of contacts.

Proposition 3 *If the distribution of export sizes among individual firms is close to Zipf's law, then aggregate trade flows between two countries are approximately proportional to their respective sizes and inversely proportional to the distance between them. The canonical gravity equation holds,*

$$T_{A,B} \propto \frac{GDP_A \times GDP_B}{Dist_{A,B}}$$

The role of economic size in this model is relatively straightforward, and in essence similar to the role of size in existing trade models. If an exporting country doubles in size, it has twice as many firms (each with its own foreign contacts) and aggregate exports double. Symmetrically, if an importing country doubles in size, its aggregate imports double. Note also that as in traditional trade models, this argument is exact only for the case of small economies far from each other. If a country becomes a non-negligible fraction of the world, part of world trade will now take place within its borders, so that the elasticity of aggregate trade with respect to size eventually decreases below unity for large countries. In addition, if the size of two countries A and B becomes

large relative to the distance between them, then the distance between any to locations in those countries will no longer be approximately equal to the distance between the two countries, and one would get the illusion that a (trade enhancing) contiguity effect exists.

The role of distance on the other hand is novel compared to existing trade models.

While the exact intuition behind the precise functional forms in Lemma 2 is mathematically arduous, a simplified explanation would be as follows. Each cohort has a different distribution of contacts. From Equation (16), the distribution of the number of contacts in the population is a power law. From Equation (17), the variance of the distributions of contacts of each firm (the average squared distance from the firm's contacts) is a power function of the number of contacts of this firm. So the variances of the various distributions of contacts are themselves power law distributed. It turns out that the aggregation of a family of distributions with power law distributed variances is approximately a power law. This result is powerful and holds no matter what the exact shape of each distribution is. In particular, I do not need to impose any restriction on how distance affects the formation of contacts.[13] The result also holds under fairly more general conditions that the strict conditions I impose on the above model. For instance, I can relax the simplifying assumption that newborn firms only meet other newborn firms of the very same age. As long as newborn firms only meet existing firms (of any age) which themselves know firms a finite (squared) distance away, no matter how large, Equation (17) will hold, and the main proposition holds. I show formally in the next section 2.3 that it is robust to relaxing several simplifying assumptions.

The intuition for why a higher γ, lower β or higher δ increase the exponent on distance in the gravity equation is more straightforward. The contacts of younger firms are geographically less dispersed than those of older firms. The faster the population growth rate, i.e. the higher γ, the more younger firms there are relative to older ones: aggregate trade declines faster with distance. The less frequently firms acquire new contacts, i.e. the lower β, the fewer chances firms have to expand their network of contacts towards longer distances: firm level and aggregate trade declines faster with distance. δ plays the opposite role to β: the higher δ, the faster aggregate trade declines with distance.

[13]While I assume that distance affects the creation of initial contacts, I only impose that new contacts are symmetric (they are equally likely to be formed "eastward" or "westward"), and they occur on average at a finite (squared) distance. Beyond these two minimal regularity conditions, the relationship can take any arbitrary shape.

Proposition 3 shows that the -1 distance elasticity of aggregate trade is related to Zipf's law for the distribution of the size of firm level exports. Formally, it is the same assumption that generates Zipf's law for the distribution of firm level exports $\left(\frac{\gamma}{\beta-\delta} \approx 1^{+}\right)$ that also makes aggregate trade approximately inversely proportional to distance $\left(1+2\frac{\gamma/(\beta-\delta)-1}{\beta/(\beta-\delta)} \approx 1^{+}\right)$. In this model, firms that export a lot, i.e. firms with many contacts, are also firms that export far away. The same parameter condition that gives the highest share of total exports to large firms, Zipf's law, also gives the highest share in aggregate exports to firms that export far away. With exports a power function of distance, this corresponds to the gravity equation with a -1 distance elasticity of trade.

This result however is not tautological. Zipf's law describes the distribution of total sales of individual firms within the population, and the gravity equation describe how much a country exports at various distances. Zipf's law has nothing to say about where firms sell their output, and the gravity equation has nothing to say about which firm sells how much. While Zipf's law is a statement about *how much* different *firms* sell, the gravity equation is a statement about *where* different *countries* export.[14]

On a more conceptual level, this model departs from existing traditional models in its treatment of distance and trade barriers. In existing models, distance captures or proxies physical trade barriers. In this model, distance captures informational barriers and the network that transmits information. As in the Krugman (1980) model, the premise of this model is that if left unhindered, all firms would export to all countries. In the Krugman (1980) model, trade barriers are the only impediment to trade; they can be circumvented to the extent that firms can cover those trade costs. In this model on the other hand, while informational barriers can also be circumvented by paying some direct cost (the g_0 function is a very general reduced form for the direct cost of information acquisition), more importantly, they can be circumvented indirectly when people interact and share information. This feature implies that information about potential foreign contacts is transmitted along individual connections. Advances in transportation or communication technologies affect physical trade barriers, the direct cost of information (the function g_0), even the frequency of interactions, but it does not remove the need for direct interactions. A model that only features direct costs will fail to explain why distance plays the same role today as it did a century ago.

[14]The mathematical properties that generate Zipf's law and the gravity equation are also different. Zipf's law is derived as the solution to a differential equation, while the gravity equation is derived from the regularly-varying property of a sequence of functions. The only direct connection between both results is that the same stationarity condition is required to get a -1 coefficient for the power law distribution of firm exports and for the distance elasticity of trade.

In this model on the other hand, the shape of aggregate trade flows is immune to changes in the g_0 function. The patterns of trade at the *firm level* do change with changes in g_0. But if direct interactions between people play a role today as they did a century ago, this model predicts that the role of distance in *aggregate* trade flows will remain essentially unchanged.

2.3 Robustness

I now discuss how my main result are robust to relaxing some of the simplifying assumptions I made along the way.

I first show that as long as bigger firms export further away on average than smaller firms, and as long as Zipf's law for firm sizes holds, the gravity equation holds for aggregate trade. Second, I allow an active intensive margin of shipments, allowing for instance larger firms to sell more to each of their contacts. Third, I show how the simplifying assumption that information spreads only within cohorts can be substantially relaxed without altering the main results. Finally, I present a very simple example through which the reader can get an intuitive feel for the main results.

Minimal conditions for the gravity equation.— I have presented in sections 1 and 2 a stylized model of the dynamic formation of a network of input-output linkages. This explicit model sheds light on many aspects of both firm level and aggregate trade flows, as well as on the structure of production along complex vertical production chains. To derive the gravity equation in international trade, it is however sufficient to verify that Equations (16) and (17) hold. If some readers are not satisfied with the entire model I presented so far, or if they only care about the gravity equation, the following proposition offers minimal conditions under which the gravity equation of international trade obtains.

Proposition 4 *If the distribution of firm sizes is Pareto with shape parameter λ and if the average (squared) distance of exports for firms of size K is proportional to K^{μ}, then aggregate trade flows between two countries A and B are approximately proportional to their respective sizes (GDP_A and GDP_B), and inversely related to the distance between them ($Dist_{A,B}$),*

$$T_{A,B} \propto \frac{GDP_A \times GDP_B}{(Dist_{A,B})^{1+\epsilon}}$$

with $\epsilon \equiv 2\frac{\lambda-1}{\mu}$. Furthermore, if the distribution of firm sizes is close to Zipf's law ($\lambda \approx 1^+$) then aggregate trade is inversely proportional to distance ($\epsilon \approx 0^+$).

I prove this proposition in the online appendix[15] for the special case where the geographic distribution of firm level exports is approximately Gaussian, and I conjecture that it holds very broadly. This proposition says that as long as larger firms export further away on average, and as long as the distribution of firm sizes is close to Zipf's law, the gravity equation of aggregate trade flows holds. Given the the relative stability of the distribution of firm sizes over extended periods of time, this result explains the surprisingly stable role played by geographic distance in shaping aggregate trade flows. This proposition also shows that the gravity equation ought to hold in any model that matches the observed average distance of exports across firms of different sizes and the distribution of firm sizes.

Allowing an active intensive margin of shipments.— I have analyzed the trade patterns of the model of input-output linkages of section 1 under the simplifying assumption that the share α of intermediate inputs in production is a half. Under this simplifying assumption, the size of a firm in Equation (8) is simply proportional to the number of contacts it has, and all shipments in Equation (9) are of equal size irrespective of firm sizes. This is obviously a knife-edge simplifying assumption. I now prove that the main proposition holds when the share of intermediate inputs α takes arbitrary values in $(0, 1)$.

The dynamic acquisition of contacts remains exactly unchanged. The only change when $\alpha \neq 1/2$ is that the direct connection between a firm's contacts, its size and the value of its exports, is lost. When the share of intermediate inputs α differs from a half, the size of a firm's shipments to downstream firms and its total sales including those shipments as well as sales to the local final goods producer are no longer exactly proportional. Combining Equations (8) on sales, (9) on shipments and (16) on the distribution of contacts, it is obvious that both are Pareto distributed. But the Pareto exponent for the sales to downstream firms is $\frac{(1-\alpha)\gamma}{(2-3\alpha)(\beta-\delta)}$, while the Pareto exponent for total sales including those to the local final goods producer is $\frac{(1-\alpha)\gamma}{\alpha(\beta-\delta)}$. Only when $\alpha = 1/2$ are both exponents the same and equal to $\frac{\gamma}{\beta-\delta}$. Since I assume that only intermediate inputs are traded, while the final goods is only produced with locally sourced inputs, it is the former exponent $\frac{(1-\alpha)\gamma}{(2-3\alpha)(\beta-\delta)}$ which matters for aggregate exports. Combining Equations (9) on shipments and (17) on the distance of exports, a firm that sells X dollars to downstream firms sells to firms that are at an average (squared) distance proportional to $X^{\frac{(1-\alpha)\beta}{(2-3\alpha)(\beta-\delta)}}$.

Apart from those changes in exponents, Proposition 2 remains unchanged. The gravity

[15] See `https://sites.google.com/site/thomaschaney/Distance_Appendix.pdf`.

equation holds for aggregate exports even when the share α of intermediate inputs is not a half. Generally, the distance elasticity of aggregate trade will be equal to $1+\epsilon$ with $\epsilon = 2\min\left(\frac{\frac{(1-\alpha)\gamma}{(2-3\alpha)(\beta-\delta)}-1}{\frac{(1-\alpha)\beta}{(2-3\alpha)(\beta-\delta)}}\right)$. Once again, if the distribution of total sales to downstream firms obeys Zipf's law $\left(\frac{(1-\alpha)\gamma}{(2-3\alpha)(\beta-\delta)} \approx 1^{+}\right)$, then aggregate trade is approximately inversely proportional to geographic distance ($\epsilon \approx 0^{+}$).

Allowing between cohorts communication.— The strongest assumption I made is that newborn firms only connect to other newborn firms, so that information spreads only within cohorts and never between them. Relaxing this assumption adds a large amount of complexity to the model, and would prevent me from deriving most of the intermediate results above. It is however possible to relax this assumption substantially, allowing firms to communicate between cohorts, while still keeping essentially unchanged Proposition 2, namely the gravity equation for aggregate trade.

The key simplification that this assumption buys is that at age t, a firm communicates with other firms of the same age t. So if the contacts of this firm are distributed according to f_t, then all of its contacts have the same distribution f_t, where only the coordinates are shifted. If a firm communicates with firms of different ages, then the contacts of its contacts are no longer distributed according to the same f_t, but according to some distribution F_t which represents a weighted average of the distributions f_s of its contacts of various ages s. The weights of those various f_s's themselves evolve endogenously as a function of the state of the entire network.

I prove in the online appendix[16] that despite this complication, the average squared distance of exports of all firms still grows exponentially with age, a result equivalent to Equation (17). According to Proposition 4, this result, along with the very same distribution of firm sizes as in Equation (16) warrantees that the gravity equation for aggregate trade still holds. In other words, firms when they are born may form contacts with any subset of the existing set of plants. The dynamic network of input-output linkages that emerges from this relaxed assumption is vastly more complex than the one I have analyzed, but the gravity equation for aggregate trade still holds.

It should be noted however that I still require one very important condition: when a firm is born, it must form contacts with other firms that export at a *finite* (squared) distance on average. This assumption is not innocuous. For instance, a newborn firm cannot sample from the existing firms according to the actual population weights. If it did, and if the gravity equation were to

[16] See `https://sites.google.com/site/thomaschaney/Distance_Appendix.pdf`.

hold in the aggregate, that would mean that the average distance of exports would have to be *infinite*: the function $1/x$ does not have a finite second moment over any open set $[a, +\infty)$. This brings a contradiction. In this special case where a newborn firm meets existing firms according to the population weights, trade can only be inversely proportional to the square of distance, and only for the knife-edge combination of parameters $\gamma - (\beta - \delta) = \beta$.[17] But as long as newborn firms form contacts with existing firms that export at a finite (squared) distance on average, the essential elements of my model are preserved, and the gravity equation holds.

A very simple closed-form example.— To understand on a more intuitive level why the aggregation of a family of distributions with power distributed variances is approximately a power law itself, consider the following simplified set up: assume that each of these distributions can be approximated by a uniform distribution. A firm with K contacts with a variance $\Delta(K)$ has therefore a constant density $K/4\sqrt{\Delta(K)}$ over the interval $\left[-2\sqrt{\Delta(K)}, +2\sqrt{\Delta(K)}\right]$. Only those firms that have contacts distributed with a standard deviation higher than $x/2$ will export at a distance x. The aggregate amount exported at a distance x is then the sum (integral) of the number (density) of contacts of each of those firms. Since the K's are power law distributed, and the $\sqrt{\Delta(K)}$ are a power function of K, the amount exported is a power function of x (the integral of a power function is a power function). Formally, using Equation (16) for the distribution of firm sizes, $F(K)$, and Equation (17) for the link between size and distance of exports, $\Delta(K)$, the fraction (density) of firms that export at a distance x, which I denote $\varphi(x)$, is given by the following expression,

$$\varphi\left(x = 2\sqrt{\Delta(K)}\right) \propto \int_K^{+\infty} \frac{k}{4\sqrt{\Delta(k)}} dF(k) \propto \frac{1}{x^{1+2\frac{\gamma/(\beta-\delta)-1}{\beta/(\beta-\delta)}}}, \forall x \geq 2\sqrt{\Delta_0}$$

The algebra in this very simple example is straightforward. There is no need to use the Laplace transform of the f_t's, nor to use Karamata's Abelian and Tauberian theorem to characterize the asymptotic behavior of aggregate trade in φ. Everything instead is calculated in closed form. The gravity equation holds exactly for all distances above some minimum threshold, and not only asymptotically for long distances. I hope this example provides some intuition to the reader.

[17] I am grateful to Michal Fabinger for pointing out this case to me.

3 Static and dynamic costs of trade disruptions

In this model, strictly speaking, there are *in the long run* no welfare or efficiency gains from international trade integration, nor any welfare or efficiency losess from trade disruptions. There are however potentially large welfare costs from trade disruptions *in the short run*. I will first briefly discuss the long run properties of this model, and then analyze the short and medium run response to trade shocks.

3.1 The long run (non-existent) gains from trade

The prediction of no gains from trade in the long run is very stark, and comes from the specific assumptions of the model. It is however worth discussing briefly.

My model is a direct extension of Krugman (1980), but its predictions regarding the long run gains from trade differ starkly from the classical Krugman model. While in the Krugman model, in the absence of trade frictions, consumers would purchase all goods produced worldwide, in my model, even in the absence of any trade frictions, each firm would only purchase a (potentially small) subset of all intermediate goods available worldwide. This is due to the adjustment cost function that prevents firms from using too many intermediates at any point in time. The constraint on the use of new intermediate inputs is not external (trade friction prevent firms from using foreign intermediates), but rather internal (firms choose not to use all intermediates because of adjustment costs). In that sense, if a large enough country[18] were to move to complete autarky, the efficiency of any one firm, and hence aggregate efficiency, would not be affected.

This is of course an extreme prediction. It is likely that part of the adjustment cost of bringing new suppliers into a firm's production process corresponds to finding the right supplier. To the extent that a firm needs specific inputs, there may not be an infinite number of potential suppliers in every location of every country, as I assume. But more constructively, this prediction only holds in the long run. In the short and medium run, disrupting existing trade linkages can potentially entail large aggregate welfare and efficiency costs. I now turn to the more interesting short run prediction of my model.

[18] Formally, a country is "large" in my model if it contains a positive measure of locations. A "large" country can therefore be arbitrarily small compared to the size of the world.

3.2 The short and medium run cost from trade disruption

This model features complex production chains that cross national boundaries many times. A disruption to the international flow of goods would affect those production chains adversely. Moreover, since the process of acquiring upstream suppliers and downstream customers is a timely and costly process, rebuilding the affected segments of those production chains would take time. I consider now a simple example of such a disruption to the international flow of goods, and analyze its impact on aggregate welfare, both in the short and in the medium run.

The full model I build in sections 1 and 2 is complex, and ultimately, my ability to derive analytical solutions to the model rely on some strong simplifying assumptions. The most important assumption is that of symmetry: a firm of age t not only is connected only to other firms of age t, but those other firms have the same geographic distribution of contacts, where only coordinates are shifted. A generic disruption to the flow of goods would break this symmetry assumption: for instance, if say a large country (a long interval of the real line) were to move to complete autarky, then firms that are near the international border of that country (near the bounds of the interval) would be more severely affected than firms in the center of that country (in the middle of the interval). Since production chains are infinite, different segments of those chains would be affected differently by such a move to autarky. While this is an interesting example to consider, the added complexity would force me to make a series of ad hoc statements on all parameters of the model. I consider instead a simple example of trade disruption. Most of the intuition for a more complex type of disruption is contained in that example, but the analysis remains simple.

A simple example of trade disruption.— Consider the following unanticipated shock: at a given point in time, any contact between two firms at a distance from each other greater than $\bar{x}$ is severed. This shock is similar to a country of size $2\bar{x}$ moving to complete autarky, except that is preserves the geographic symmetry of the system. The lower $\bar{x}$ is, the more trade is disrupted.

Because older and larger firms are more likely to export over long distances, those firms are more likely to lose some of their contacts. This direct effect reduces trade. In addition, with the loss of contacts, the productivity of a firm goes down, so that it produces lower quantities, which further reduces trade. This affects not only the firm directly, but with each of the supplier of the firm losing suppliers themselves, this productivity effect is amplified along all the production chain. As for the direct effect, older and larger firms are connected to more complex production chains that span wider geographic ranges, so that production and trade along those more complex

production chains is more severely affected by trade disruption.

I now characterize all those effects explicitly.

Aggregate welfare before the shock.— Before the shock, in the symmetric equilibrium I consider in section 1, a firm with K suppliers only buys from other firms with K suppliers, so that the price each firm along those K-type production chains is explicitly solved for in Equation (6). Along a balanced growth path equilibrium, the distribution of the number of contacts across firms is invariant, and only the overall mass of firms grows. All workers worldwide have the same wage, and their per period utility U depends on this wage w and the local ideal price index P they face, $U = w/P$. Aggregate welfare before the shock is,

$$U_{\text{before}} = w/P = w\left(\int p_\omega^{1-\sigma} d\omega\right)^{\frac{1}{\sigma-1}} = w\left(M\int_0^\infty \gamma e^{-\gamma t} K_t^{\frac{\alpha}{1-\alpha}} dt\right)^{\frac{1}{\sigma-1}} \tag{18}$$

Note that larger firms (larger K's) contribute more to aggregate welfare, as in any model with heterogeneous firms such as Melitz (2003). Those firms also export further away on average, so they will be harder hit by any trade disruption.

Aggregate welfare after the shock.— After the shock, all trade linkages over a distance $\bar{x}$ are severed. Because of the presence of adjustment costs, firms that lose their distant contacts will not instantaneously rebuild those contacts. So upon impact, a fraction of the suppliers of all firms are simply lost. Losing those suppliers induces a productivity loss, and ultimately, welfare drops in every location. Aggregate welfare after the shock is,

$$U_{\text{after}} = w\left(M\int_0^\infty \gamma e^{-\gamma t}\left(K_t\int_0^{\bar{x}} g_t(x)\,dx\right)^{\frac{\alpha}{1-\alpha}} dt\right)^{\frac{1}{\sigma-1}} = w\left(M\int_0^\infty \gamma e^{-\gamma t}\left(K_t\left(1-e^{-\bar{x}/\sqrt{\Delta_t}}\right)\right)^{\frac{\alpha}{1-\alpha}} dt\right)^{\frac{1}{\sigma-1}} \tag{19}$$

Upon impact of the shock, the measure of firms M does not change, as firms have already paid the sunk entry cost. Because it is normalized, the nominal wage w does not change either. The only variable that adjusts is the efficiency of various firms, as they lose some suppliers from the disruption of trade linkages.

A firm with K_t suppliers has suppliers distributed over space according to the p.d.f. $g_t(x)$. After the shock hits, this geographic distribution of suppliers is truncated above $\bar{x}$ as can be seen in the first equality of Equation (19). Larger firms have more dispersed suppliers, so they will tend to have a larger fraction of their suppliers at a distance beyond $\bar{x}$, and they will tend to lose a larger fraction of their suppliers. This can be seen explicitly in the second equality of Equation (19),

where I use the special closed-form case $g_0 \sim Laplace$ from proposition 1. A firm with K_t suppliers has contacts that are at an average (squared) distance Δ_t away. The more suppliers a firm has, the larger K_t, the larger is Δ_t, and the smaller is the fraction $\left(1 - e^{-\bar{x}/\sqrt{\Delta_t}}\right)$ of suppliers that are at a distance lower than $\bar{x}$. Larger firms are therefore more affected by trade disruption than smaller ones. Obviously, because of the roundabout nature of production chains, the higher the share of intermediate inputs in production, α, the larger the welfare cost of trade disruptions, as the effect of the disruption of input-output linkages for one firm cascading down to its downstream suppliers gets magnified.

Transitional dynamics.— After the initial shock, new firms enter, and existing firms gradually rebuild linkages with new suppliers, under the constraint that no supplier can be at a distance larger than $\bar{x}$. Both margins, the entry of new firms, and the growth of existing firms, work together to bring back the economy to a balanced growth path equilibrium.

Analyzing the free entry condition in Equation (11), the drop in the ideal price index due to the disruption of trade linkages makes entry more profitable, and induces an influx of new firms at a rate above its steady state. Analyzing the optimal investment equation in Equations (23) or (24), the fall in the price index increases firm level profits for all firms (π), and the loss of existing contacts (K) increases further the level of profits per supplier (π/K) for all firms. Both forces increase the investment rate for all firms above its steady state level. Moreover, while profits increase by the same proportional amount for all firms, larger firms (higher K firms) lose disproportionately more suppliers, so that the profits per supplier increase more for larger firms, and those firms invest into acquiring new suppliers at a faster rate than smaller firms.

All those forces together, the higher entry rate of new firms, the higher investment rate into suppliers of existing firms, and all the more so for larger firms that have been harder hit by the trade disruption, bring back the economy towards its balanced growth path equilibrium. During the transition, and compared to the balanced growth path equilibrium, there are too many small firms, and too few large ones. Because of the convexity of the adjustment cost function, firm growth cannot restore the economy to its balanced growth path instantaneously, and initially, the entry margin is the most active.

On a more intuitive level, right after the shock hits the world economy, there are "too many" workers chasing "too few" firms. The entry of new firms and the growth of existing ones restores the balance between the size of the labor force and the diversity of firms.

This analysis of the dynamic response to a trade disruption is but a simple example. It shows how the model developed in this paper can be used to study several aspects of the dynamic response to various shocks of an economy characterized by complex vertical production chains. In the next section, I empirically test some of the main predictions of the model.

4 Empirics

The theoretical model above predicts that if the distribution of firm level total exports is close to Zipf's law, and if the average (squared) distance of a firm's exports is a power function of this firm's number of contacts, then aggregate exports follow the gravity equation. The model also predicts that any deviation away from Zipf's law should be associated with systematic deviations away from the -1 elasticity of aggregate trade to distance. I test those two predictions in turn.

4.1 Aggregate trade: Zipf's law and the canonical gravity equation

Proposition 3 predicts that if the distribution of firm sizes follows Zipf's law, then aggregate trade is inversely proportional to distance.

I test this prediction using data on firm level exports for France in 1992.

First, I confirm that Zipf's law for the distribution of firm level aggregate exports holds, at least for the larger firms. Figure 2 shows the relation between the log of the percentile of a firm, versus the log of its size. The relationship is very close to Zipf's law for large exporters, with a power law exponent of -1.0386.[19]

Second, I confirm that larger firms export over longer distances on average. Following the guidance of Chaney (2013), I use the number of countries a firm exports to as a proxy for its number of contacts, and then compute $\Delta(K)$ as the average (squared) distance of export among firms that export to K foreign markets. Figure 3 shows the relation between the log of the average (squared) distance from a firm's exports, $\Delta(K)$, versus the log of the number of foreign countries it exports to, K. The relationship between $\Delta(K)$ and K is well approximated by a power function, with a slope equal to .464 (s.e. 0.0108, adj-R^2=95.02%).[20]

[19] I follow Gabaix and Ibragimov (2011) to estimate the power law exponent for the distribution of aggregate export sizes among the 5% largest French exporters in 1992. Firms are ranked in decreasing size. I estimate by OLS,

$$\ln\left(Rank_i - \frac{1}{2}\right) = \text{constant} - \underset{(0.00185)}{1.0386}\ln\left(Size_i\right) + \epsilon_i,\ Adj - R^2 = 99.24\%$$

[20] The fact that the relationship is flatter than a straight line in a log-log scale for small K's can easily be explained

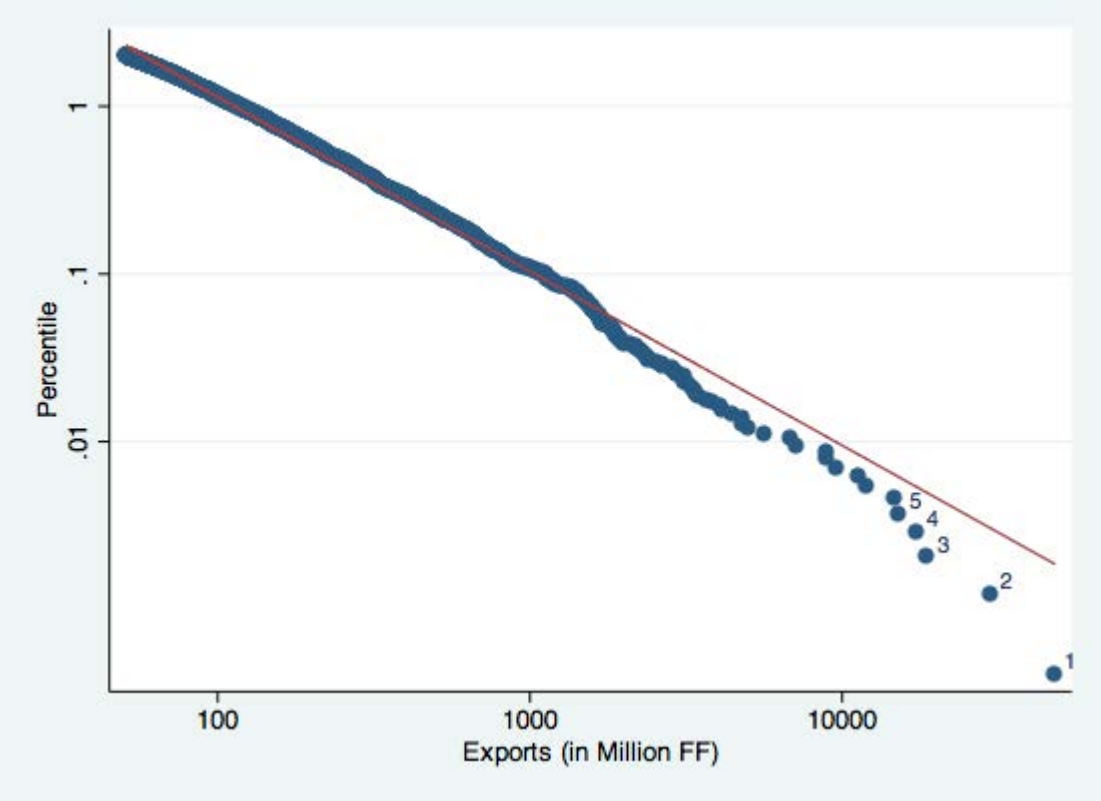

Figure 2: The distribution of firm level total exports is Zipf.

Notes: This graph shows on a log-log scale the fraction (in percentiles of the population) of firms that export more than x as a function of the value x of a firm's total exports (in million French Francs). This distribution is well approximated by Zipf's law for the largest firms, as shown by the straight fitted line in this log-log scale. The estimated slope is -1.0386 (s.e. 0.00185, adj-R^2=99.24%). Data: French exporters, 1992.

Third, I confirm that aggregate trade for French firms is approximately inversely proportional to distance. Here, the theory suggests to estimate the impact of distance on aggregate trade in a somewhat different way than what the empirical trade literature typically does. Specifically, the model predicts that the fraction of number of French firms that export at a distance D should be approximately proportional to $1/D$. To estimate this prediction, I draw concentric circles around France that are 500km's apart. Each country in the world falls into one 500km thick band. I then calculate the total number of French exporters that export to at least one country in each band. I also calculate the sum of the GDP of all countries in each band. I then estimate by OLS,

$$\ln\left(Number_b\right) = \alpha + \beta \ln\left(GDP_b\right) - \zeta \ln\left(Distance_b\right) + \epsilon_b$$

if in addition to meeting the contacts of their contacts, all firms also meet some contacts at random. See Chaney (2013) for a model that incorporates this additional source of contact formation. For larger firms, the contribution of this additional source of contacts becomes negligible.

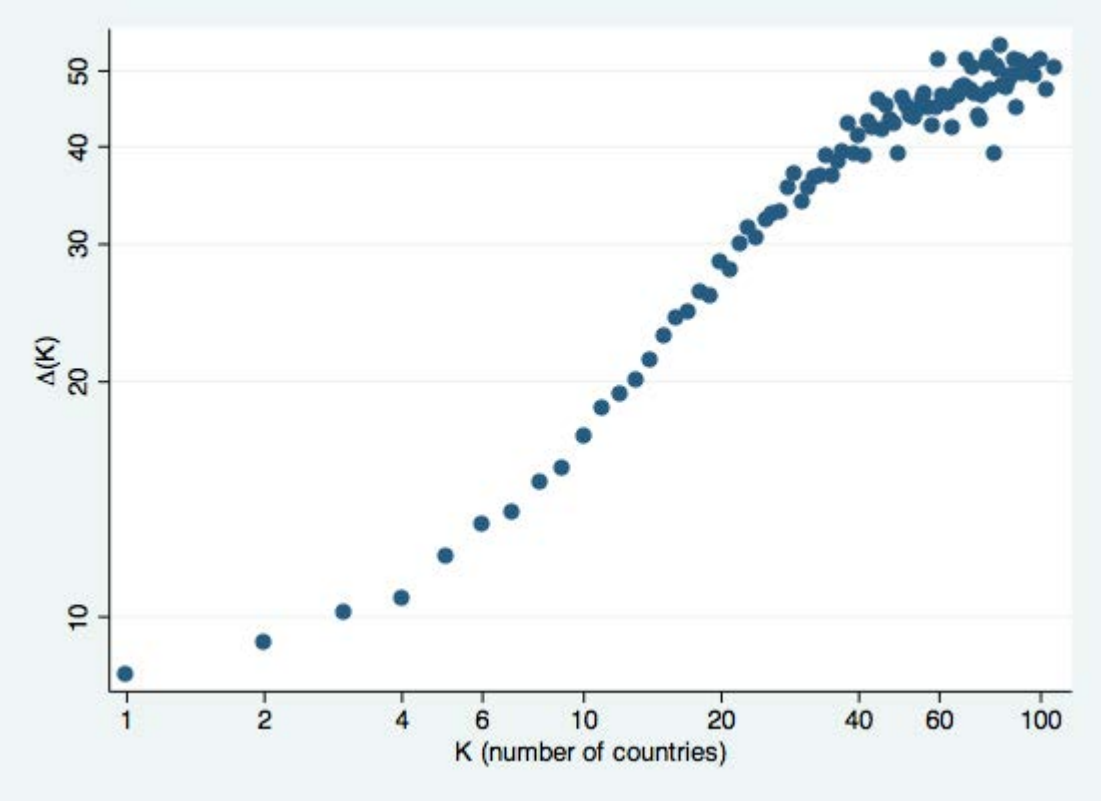

Figure 3: Average (squared) distance of exports versus number of export destinations.
Notes: This graph shows on a log-log scale the average (squared) distance of exports, $\Delta(K)$, among firms that export to K foreign countries, as a function of K. The relationship is close to a straight line in this log-log scale, suggesting that $\Delta(K)$ is well approximated by a power function of K. Distances are measured in thousand km's. Data: all French exporters, 1992.

where each $b = 1, \cdots, 40$ corresponds to a 500km's wide band centered around France, $Number_b$ is the number of French exporters that export to at least one country in band b, and GDP_b is the sum of the GDP of countries in band b. I find that $\zeta = 1.16$ (s.e. 0.15, adj-R^2=37%). The results are not sensitive to changing the thickness of each band, nor to measuring the nominal value of exports instead of the number of exporters. Note also that the distance coefficient ζ is not materially different from the estimate of a more conventional specification for the gravity equation, where the log of exports between two countries A and B is regressed on the log of their respective GDP's and on the log of the distance between them. This is due to the fact that there is approximately as much GDP mass for every distances away from France. In other words, from the point of view of France, there is not more income at a distance d than at any other distance d'. Interestingly, this fact happens to be approximately true for most countries in the world.

According to Proposition 3, the distance elasticity of aggregate trade, ζ, should be directly

related to the power law exponent of the distribution of firms' exports, and to the elasticity of the average (squared) distance of exports among firms that export to K countries with respect to K. The estimated elasticities of these two relationships would predict according to my model that the number of firms that export at a distance $Dist_b$ should be proportional to $1/Dist_b^{1.17}$. In the data, it is proportional to $1/Dist_b^{1.16}$. The elasticity coming from the theory (1.16) is very close to the elasticity in the data (1.17), so that the prediction of the theory regarding aggregate trade holds tightly in the data.[21]

The reader may be concerned that there is something circular in this statement, as I am using the same data on French exporters exporting to various distances for both the predicted distance elasticity and the direct empirical measure of this distance elasticity. In the next section, I test a stricter prediction from the theory, and I use data from different sources, so that any suspicion of circularity ought to be dispelled.

4.2 Sectoral trade: departures from Zipf and the canonical gravity equation

The prediction from the theory is actually more subtle. Not only does the theory predict that *if* Zipf's law for the distribution of firm sizes holds, *then* aggregate trade is inversely proportional to distance, but it also predicts that *if* Zipf's law for the distribution of firm sizes *does not* hold, *then* one should expect specific departures from the -1 distance elasticity of aggregate trade. This more nuanced prediction corresponds to Proposition 2.

To test this stronger prediction, I disaggregate my dataset into industrial sectors. While in the aggregate, Zipf's law is a good approximation for the distribution of firm sizes, at least among the largest firms, there are systematic departures away from Zipf's law across different sectors, as documented for instance in Helpman, Melitz and Yeaple (2004). Formally, Proposition 2 links the power law exponent for the distribution of firm sizes in sector s and the distance elasticity of aggregate trade in sector s as follows,

$$\ln\left(\text{Rank}_i^s - \frac{1}{2}\right) \approx c - \alpha_s \ln\left(\text{Size}_i^s\right) + \epsilon_i^s \tag{20}$$

$$\ln \frac{X_{A,B}^s}{GDP_A GDP_B} \approx c - \zeta_s \ln Dist_{AB} + \epsilon_{AB}^s \tag{21}$$

$$\Rightarrow \zeta_s \approx (1-\beta) + \beta\alpha_s + \epsilon_s \tag{22}$$

where I disaggregate the data into 42 manufacturing sectors in the year 1992. I estimate those three equations in turn.

[21] The theory predicts that $\zeta = 1 + 2\frac{1.0386-1}{.464} \approx 1.17$, compared to the estimated $\zeta \approx 1.16$ in the data.

First, to estimate the power law coefficient for the distribution of firm sizes in Equation (20), I only use data on French firms. Following Gabaix and Ibragimov (2011), I estimate the probability that a firm is larger than a given size by the rank of a firm of such size minus $\frac{1}{2}$. I estimate Equation (20) using OLS separately for each 42 sectors. I run three different specifications: first, I use data on all firms in sector s; second I only use data on the largest 30% of firms within sector s (I have only 41 sectors with this restricted sample); and third I use only data on the largest 10% of firms in sector s. These sets of regressions provide me with three alternative series of estimated $\hat{\alpha}_s$.

Second, to estimate the distance elasticity of trade in Equation (21), I use only data on bilateral aggregate trade flows between pairs of countries. The data on nominal trade flows comes from the NBER,[22] the data on nominal *GDP* comes from the Penn World Tables,[23] and the data on bilateral distance between countries comes from the CEPII.[24] In order to avoid any endogeneity problem, I remove France from my sample of countries. I estimate Equation (21) using OLS separately for each 42 sectors. These regressions provides me with a series of estimated $\hat{\zeta}_s$.

Third, with the estimates $\hat{\alpha}_s$ and $\hat{\zeta}_s$ in hand, I can test the main prediction from the theory in Equation (22). Before doing so, several observations are in order. First, to avoid any endogeneity problem, I do not use any data on the elasticity of the average (squared) distance of exports with respect to the number of contacts of a firm. In other words, the power law exponents $\hat{\alpha}_s$ are estimated using only data on the total export sales of French firms, and no geographic information is used. The distance elasticities $\hat{\zeta}_s$ on the other hand are estimated using only data on aggregate trade flows between countries other than France, and no data on French firms is used. Second, my identifying assumption is that variations in the elasticity of the average (squared) distance of exports with respect to the number of export destinations are orthogonal to variations in the firm size distribution parameter α_s. Finally, while any measurement error in the LHS variable $\hat{\zeta}_s$ will be captured by the error term ϵ_s in Equation (22), I have to account for the fact that the RHS variable $\hat{\alpha}_s$ are estimated, and therefore subject to systematic measurement error. I use a very reduced form approach to address this measurement error issue, and I weight each observation $\hat{\alpha}_s$ by its estimated precision, calculated as the inverse of the standard error of $\hat{\alpha}_s$ estimated in Equation (20). I conjecture that a more systematic GMM estimation procedure would produce similar results.

[22] See Feenstra et al. (2004).
[23] See `http://pwt.econ.upenn.edu/`.
[24] See Mayer and Zignago (2006).

Table 1: Firm size distribution and the geography of aggregate trade across sectors.

Dep. var.: $\hat{\zeta}_s$, distance elasticity of aggregate trade			
	all firms	top 30%	top 10%
	(1)	(2)	(3)
$\hat{\alpha}_s$ (rank-size)	.546*	.224*	.204**
	(0.297)	(0.123)	(0.102)
Constant	.386**	.465***	.432***
	(0.161)	(0.127)	(0.119)
N. obs.	42	41	38
R^2	0.08	0.08	0.10

Notes: This table shows the result of the estimation of Equation (22). The dependent variable $\hat{\zeta}_s$ is the estimated distance elasticity of trade in sector s in 1992, from Equation (21). The explanatory variable $\hat{\alpha}_s$ is the estimated power law exponent for the distribution of the size of aggregate exports for French firms in sector s in 1992, from Equation (20). Each observation is weighted by the estimated precision (inverse of the estimated s.d.) from Equation (20). Column (1) uses all French firms in sector s to estimate $\hat{\alpha}_s$. Column (2) uses only the largest 30% firms in sector s. Column (3) uses only the largest 10% firms in sector s. *, ** and *** respectively mean statistically different from zero at the 10%, 5% and 1% level of significance.

The estimation results for Equation (22) are presented in Table 1. In all 3 specifications, a thicker upper tail for the distribution of firm sizes (a smaller $\hat{\alpha}_s$) tends to be systematically associated with a milder negative impact of geographic distance on aggregate trade flows (a smaller $\hat{\zeta}_s$). This is directly in line with the theory which predicts that when large firms account for a larger share of sales in a sector, aggregate trade will fall off with distance at a slower pace. Intuitively, this is simply due to the fact that large firms tend to export over longer distance, so that when there are more large firms, there is more trade going to distance places, and distance seems to be less of an impediment to trade in the aggregate.

In all three specifications, the impact of the power law exponent of the firm size distribution in a sector on the distance elasticity of aggregate trade in that sector has the expected sign, and is significant at the 10% level when data on all firms are used, and at the 5% level when on the larger 30 or 10% firms are used. The distribution of firm sizes is typically well approximated by a power law in the upper tail, but less so in the lower tail. This explains why the results when all firms are used are less significant. Note also that the R^2 for these regressions are low (around 8-10%), as I abstract from any other sources of variation in the role that distance plays in shaping aggregate trade flows.

Conclusion

This paper offers the first theoretical explanation for the gravity equation in international trade. I explain not only why trade is proportional to size, but also the mysterious -1 distance elasticity of trade. In my model, larger firms endogenously export over longer distances. The impact of distance on aggregate trade therefore depends on the shape of the distribution of firm sizes. If firm sizes are well approximated by Zipf's law, as the data suggests, then the distance elasticity of trade ought to be close to -1. This result holds irrespectively of the precise impact of geographic distance on firm level trade. Unlike in existing models, my explanation is therefore immune to the critique that the impact of distance on trade ought to change with changes in the technology for trading goods, in the types of goods traded, in the political barriers to trade, in the set of countries involved in trade, etc. As long as the individuals that make up firms engage in direct communication with their clients and suppliers, and as long as information permeates through these direct interactions, one ought to expect that aggregate trade is close to proportional to country size and inversely proportional to distance.

References

Acemoglu, Daron, Vasco M. Carvalho, Asuman Ozdaglar and Alireza Tahbaz-Salehi. 2012. "The Network Origins of Aggregate Fluctuations." *Econometrica*, 80(5): 1977-2016.

Ahn, JaeBin, Amit K. Khandelwal and Shang-Jin Wei. Forthcoming. "The Role of Intermediaries in Facilitating Trade," *Journal of International Economics*.

Anderson, James E. 1979. "A Theoretical Foundation for the Gravity Equation." *American Economic Review*, 69(1): 106–16.

Anderson, James E., and Eric van Wincoop. 2003. "Gravity with Gravitas: A Solution to the Border Puzzle," *American Economic Review*, 93(1): 170–92.

Arkolakis, Costas, Arnaud Costinot and Andres RodrÌguez-Clare. 2012. "New Trade Model, Same Old Gains?" *American Economic Review*, 102(1): 94-130.

Atalay, Enghin, Ali Hortaçsu, James W. Roberts, and Chad Syverson. 2011. "Network Structure of Production." *Proceedings of the National Academy of Sciences*, 108(3): 5199-202.

Barabási, Albert-László and Réka Albert. 1999. "Emergence of Scaling in Random Networks," *Science*, 286: 509-12.

Chaney, Thomas. 2008. "Distorted Gravity: The Intensive and Extensive Margins of International Trade," *American Economic Review*, 98(4): 1707-21.

Chaney, Thomas. 2013. "The Network Structure of International Trade," Toulouse School of Economics, *mimeo*.

Combes, Pierre-Philippe, Miren Lafourcade and Thierry Mayer. 2005. "The Trade-Creating Effects of Business and Social Networks: Evidence from France," *Journal of International Economics*, 66(1):1-29.

de Haan, Laurens. 1976. "An Abel-Tauber Theorem for Laplace Transforms," *Journal of the London Mathematical Society*, s2-13(3): 537-42.

Disdier, Anne-Célia and Keith Head. 2008. "The Puzzling Persistence of the Distance Effect on Bilateral Trade," *Review of Economics and Statistics*, 90(1): 37-48.

di Giovanni, Julian and Andrei A. Levchenko. 2010. "Firm Entry, Trade, and Welfare in Zipf's World," *Journal of International Economics*, 89(2): 283-96.

Eaton, Jonathan, Marcela Eslava, C.J. Krizan, Maurice Kugler, and James Tybout. 2010. "A Search and Learning Model of Export Dynamics," Penn State University, *mimeo*.

Eaton, Jonathan and Samuel Kortum. 2002. "Technology, Geography, and Trade," *Econometrica*, 70(5): 1741–79.

Eaton, Jonathan, Samuel Kortum, and Sebastian Sotelo. 2010. "International Trade: Linking Micro and Macro," University of Chicago, *mimeo*.

Erdös, Paul and Alfréd Rényi. 1959. "On Random Graphs," *Publicationes Mathematicae*, 6:290-7.

Feenstra, Robert C., Robert E. Lipsey, Haiyan Deng, Alyson C. Ma, and Hengyong Mo. 2004. "World Trade Flows: 1962-2000," NBER WP 11040.

Gabaix, Xavier. 1999. "Zipf Law for Cities: an Explanation," *Quarterly Journal of Economics*, 114(3): 739-67.

Gabaix, Xavier, and Rustam Ibragimov. 2011. "Rank-1/2: A Simple Way to Improve the OLS Estimation of Tail Exponents," *Journal of Business Economics and Statistics*, 29(1): 24-39.

Halpern, Lázló, Miklós Koren and Adam Szeidl. 2009. "Imported Inputs and Productivity," Central European University *mimeo*.

Hayashi, Fumio. 1982. "Tobin's Marginal q and Average q: A Neoclassical Interpretation," *Econometrica*, 50:213-24.

Helpman, Elhanan, Marc J. Melitz, and Yona Rubinstein. 2008. "Estimating Trade Flows:

Trading Partners and Trading Volumes," *Quarterly Journal of Economics*, 123: 441-87.

Helpman, Elhanan, Marc J. Melitz, and Stephen Yeaple. 2004. Export Versus FDI with Heterogeneous Firms," *American Economic Review*, 94(1): 300-16.

Melitz, Marc, Elhanan Helpman, and Stephen Yeaple. 2004. Export Versus FDI with Heterogeneous Firms. *American Economic Review* 94: 300-316.

Krugman, Paul. 1980. "Scale Economies, Product Differentiation, and the Patterns of Trade," *American Economic Review*, 70(5): 950-59.

Kummer, Ernst Eduard. 1836. "Über die hypergeometrische Reihe $1 + \frac{\alpha\beta}{1.\gamma}x + \frac{\alpha(\alpha+1)\beta(\beta+1)}{1.2.\gamma(\gamma+1)}x^2 + \frac{\alpha(\alpha+1)(\alpha+2)\beta(\beta+1)(\beta+2)}{1.2.3.\gamma(\gamma+1)(\gamma+2)}x^3 + \text{etc.}$" (in German). *Journal für die reine und angewandte Mathematik*, 15: 39–83.

Lucas, Robert E. Jr. 1967. "Adjustment Costs and the Theory of Supply," *Journal of Political Economy*, 75:321-334.

Luttmer, Erzo G. J. 2007. "Selection, Growth, and the Size Distribution of Firms," *Quarterly Journal of Economics*, 122(3): 1103-44.

Mayer, Thierry, and Soledad Zignago. 2006. "Notes on CEPII's Distances Measures," *mimeo.*

McCallum, John. 1995. "National Borders Matter: Canada-U.S. Regional Trade Patterns," *American Economic Review*, 85(3): 615-623.

McPherson, Miller, Lynn Smith-Lovin, and James M. Cook. 2001. "Birds of a Feather: Homophily in Social Networks," *Annual Review of Sociology*, 27: 414-44

Melitz, Marc J. 2003. "The Impact of Trade on Intra-Industry Reallocation and Aggregate Industry Productivity," *Econometrica*, 71(6): 1695-1725.

Romer, Paul M. 1990. "Endogenous Technological Change," *Journal of Political Economy*, 98(5): S71-S102.

Santos Silva, J.M.C. and Silvana Tenreyro. 2006. "The Log of Gravity," *Review of Economics and Statistics*, 88:641-658.

Slater, Lucy Joan. 1966. *Generalized Hypergeometric Functions.* Cambridge, UK: Cambridge University Press.

Solow, Robert M. 1956. *"*A Contribution to the Theory of Economic Growth*," Quarterly Journal of Economics.* 70 (1): 65–94.

Steindl, Josef. 1965. "Random Processes and the Growth of Firms," Charles Griffin, London.

Tinbergen, Jan. 1962. "An Analysis of World Trade Flows," in *Shaping the World Economy*, edited by Jan Tinbergen. New York, NY: Twentieth Century Fund.

Appendix

A Mathematical proofs

A.1 Optimal investment

The firm choses an optimal investment policy so as to maximize a discounted sum of profits as in (10), taking as given the behavior of other firms. For a generic production function and adjustment cost function, the Lagrangian associated with the firm's program is,

$$\max_{I_t} \mathcal{L} = \int_0^{+\infty} e^{-rt} \left(\pi \left(K_t \right) - p_I w I_t - wC \left(I_t, K_t \right) + \lambda_t \left(\dot{K}_t - I_t - \delta K_t \right) \right) dt$$

where π is the profit function once the optimal choice of spending on labor and intermediate input has been made, and where I have omitted the proceeds from selling information, which are not affected by the firm's choices. The first order conditions for this program are,

$$\text{for all } t\text{'s, } \begin{cases} \frac{\partial \mathcal{L}}{\partial I_t} & = 0 \\ \frac{\partial \mathcal{L}}{\partial K_t} & = \frac{d}{dt} \frac{\partial \mathcal{L}}{\partial \dot{K}_t} \end{cases}$$

Those conditions along with $\dot{\lambda} = 0$ give the following characterization of the optimal investment choice,

$$\left(p_I w + wC_I \right) \left(\delta + r \right) = \pi_K - wC_K \tag{23}$$

When both the production function and the adjustment cost functions are homogenous, the solution to Equation (23) is such that the investment rate (I/K) is constant and does not depend on the level of capital K. It is easy to see this result for the special case of a quadratic adjustment cost function, and the Cobb-Douglas production function I use in Section 1, where closed form solutions can be derived. Assume the cost of bringing new suppliers into production is,

$$C \left(I, K \right) = \frac{\kappa}{2} \frac{I^2}{K}$$

Then the investment rate (I/K) simply solves,

$$w \left(p_I + \kappa \frac{I}{K} \right) \left(\delta + r \right) = \alpha \frac{\pi}{K} - w \frac{\kappa}{2} \left(\frac{I}{K} \right)^2 \tag{24}$$

For a given technology, the investment rate I/K is independent of a firm's size, scaled by K. Furthermore, if the share of intermediate inputs is $\frac{1}{2}$, then profits π are proportional to the

number of suppliers, K, so that π/K is constant across all firms, and all firms invest at exact the same rate.[25] If one firm finds it optimal to invest at this rate, in a symmetric equilibrium, all other firms will also invest at the same rate, and the conditions for π/K to be constant are indeed satisfied. I call the equilibrium investment rate β. The measure of supplier of any firm grows at a constant rate $\beta - \delta$.

A.2 Formal proofs of all the propositions in the paper

Proposition 5 *The geographic distribution of the contacts of a firm of age t is given by,*

$$f_t(x) = \mathcal{B}^{-1}\left[\frac{K_0 e^{(\beta-\delta)t}\mathcal{B}[g_0(x)]}{(1-e^{\beta t})\mathcal{B}[g_0(x)] + e^{\beta t}}\right]$$

where $\mathcal{B}$ is the two-sided bilateral Laplace transform,[26] *$\mathcal{B}^{-1}$ its inverse, K_0 is the mass of initial contacts of a newly born firm, g_0 is the p.d.f. of these initial contacts, β (resp. δ) is the Poisson birth (resp. death) rate of new (resp. old) contacts.*

Proof. Recognizing a convolution product[27] in Equation (14), I can rewrite it in a compact form,

$$\frac{\partial f_t}{\partial t} = \beta \frac{f_t * f_t}{K_t} - \delta f_t \tag{25}$$

with initial condition $f_0 = K_0 g_0$. I will first solve for K_t, and then solve for f_t. Integrating Equation (25) over $\mathbb{R}$, and using the fact that the integral of the convolution of two functions is the product of their integrals, I derive an ordinary differential equation for K_t,

$$\frac{\partial K_t}{\partial t} = \beta \frac{K_t \times K_t}{K_t} - \delta K_t = (\beta - \delta) K_t$$

with initial condition K_0. This ODE admits the simple solution,

$$K_t = K_0 e^{(\beta-\delta)t}$$

Plugging this result into Equation (25), taking the two-sided Laplace transform of this equation (I denote by $\hat{f}$ the transform of f), and using the convolution theorem which states that the Laplace

[25] Generally, when the share of intermediates differs from $\frac{1}{2}$, firms with a higher π/K invest at a higher rate. This corresponds to firms with either more customers, or fewer suppliers.

[26] The two-sided Laplace transform is closely related to the moment-generating function. For a random variable X with a p.d.f. f, the moment generating function μ_X is defined as $\mu_X(s) = \mathbb{E}\left[e^{sX}\right]$, while the Laplace transform $\mathcal{B}[f]$ is defined as $\mathcal{B}[f](s) = \mathbb{E}\left[e^{-sX}\right] = \int_{\mathbb{R}} e^{-sx} f(x)\, dx$, so that $\mu_X(-s) = \mathcal{B}[f](s)$. This definition extends to positive functions which are not probability densities.

[27] Remember that the p.d.f. of the sum of two random variables is the convolution of their p.d.f.'s.

transform of the convolution of two function is the product of their Laplace transforms, I get the following ordinary differential equation,

$$\frac{\partial \hat{f}_t}{\partial t} = \beta \frac{\hat{f}_t^2}{K_0 e^{(\beta-\delta)t}} - \delta \hat{f}_t$$

with initial condition $\hat{f}_0 = K_0 \hat{g}_0$. This ODE admits the solution,

$$\hat{f}_t = \frac{K_0 e^{(\beta-\delta)t} \hat{g}_0}{(1 - e^{\beta t})\, \hat{g}_0 + e^{\beta t}}$$

Taking the inverse Laplace transform, I recover the proposed solution for f_t. ■

Corollary *Equations (15), (16) and (17) are satisfied,*

$$\begin{aligned} K_t &= K_0 e^{(\beta-\delta)t} \\ F(K) &= 1 - \left(\frac{K}{K_0}\right)^{-\frac{\gamma}{\beta-\delta}} \text{ for } K \geq K_0 \\ \Delta(K) &= \Delta_0 \left(\frac{K}{K_0}\right)^{\frac{\beta}{\beta-\delta}} \text{ for } K \geq K_0 \end{aligned}$$

Proof.

Equation (15). Using the property of the Laplace transform, the total mass of contacts of a firm of age t, K_t, is the Laplace transform $\hat{f}_t(s)$ evaluated at zero,

$$K_t = \hat{f}_t(0) = K_0 e^{(\beta-\delta)t}$$

where I used the fact that since g_0 is a well defined p.d.f. that sums up to 1, $\hat{g}_0(0) = 1$.

Equation (16). The formula for K_t provides the following relation between a firm's number of contacts and its age,

$$e^t = \left(\frac{K_t}{K_0}\right)^{\frac{1}{\beta-\delta}}$$

The population grows at an exponential rate γ so that the fraction of firms younger than t is $\left(1 - e^{-\gamma t}\right)$. Since a firm of age t has a total number of contacts K_t, using the above expression for e^t, I get the proposed formula for the fraction of firms with fewer than K contacts,

$$F(K) = 1 - \left(\frac{K}{K_0}\right)^{-\frac{\gamma}{\beta-\delta}}$$

Equation (17). The average (squared) distance between a firm of age t and its contacts, Δ_t, is the variance of the p.d.f. f_t/K_t of the distribution of this firm's contacts. Again using

the property of the Laplace transform, this variance is simply the second derivative of $\widehat{f_t/K_t}(s)$ evaluated at zero. Simple algebra gives this second derivative,

$$\widehat{f_t/K_t}"(s) = \frac{e^{\beta t}\left(\hat{g}_0"(s)\left(\left(e^{\beta t}-1\right)\hat{g}_0 - e^{\beta t}\right) - 2\hat{g}_0'(s)^2\left(e^{\beta t}-1\right)\right)}{\left(\left(e^{\beta t}-1\right)\hat{g}_0(s) - e^{\beta t}\right)^3}$$

Since g_0 is a well defined symmetric p.d.f. with finite variance, I can use the following properties of its Laplace transform: $\hat{g}_0(0) = 1$ (a p.d.f. sums up to 1), $\hat{g}_0'(0) = 0$ (g_0 is symmetric) and $\hat{g}_0"(0) = \Delta_0$ (g_0 has a finite variance Δ_0). The previous expression evaluated at zero simplifies into the proposed formula,

$$\Delta_t = \widehat{f_t/K_t}"(0) = \Delta_0 e^{\beta t}$$

Plugging the expression $e^t = (K_t/K_0)^{\frac{1}{\beta-\delta}}$ into the above formula for Δ_t, I derive the proposed relationship between a firm's total number of contacts K and the average (squared) distance from its contacts, $\Delta(K)$,

$$\Delta(K) = \Delta_0\left(\frac{K}{K_0}\right)^{\frac{\beta}{\beta-\delta}}$$

■

Proposition 1 (reminded) *The distribution of a firm's contacts, $\frac{f_t}{K_t}$, converges when t grows large to a Laplace distribution (a 2-sided exponential),*

$$\frac{f_t(x)}{K_t} \underset{t\to\infty}{\sim} Laplace\left(0, e^{\beta t/2}\sqrt{\Delta_0/2}\right)$$

This property holds exactly for all t's if $g_0 \sim Laplace\left(0, \sqrt{\Delta_0/2}\right)$.

Proof. For simplicity, I normalize $\sqrt{\Delta_0/2} = 1$. The proof can trivially be extended to $\sqrt{\Delta_0/2} \neq 1$.

First consider the special case where g_0 is a normalized Laplace distribution, $g_0 \sim \text{Laplace}(0,1)$. The Laplace transform of g_0 is $\hat{g}_0(s) = \frac{1}{1+s^2}$. From Proposition 5, the Laplace transform of g is then,

$$\hat{g}_t(s) = \frac{\hat{f}_t(s)}{K_t} = \frac{\hat{g}_0(s)}{(1-e^{\beta t})\hat{g}_0(s) + e^{\beta t}} = \frac{1}{1+\left(e^{\beta t/2}s\right)^2}$$

where one recognizes the Laplace transform of a Laplace $\left(0, e^{\beta t/2}\right)$ distribution.

Consider now the general case where g_0 is not Laplace. I need to prove that,

$$\begin{cases} \hat{g}_t(s) & \underset{t\to\infty}{\propto} \frac{1}{1+\left(e^{\beta t/2}s\right)^2} \\ g_t(x) & \underset{t\to\infty}{\propto} \frac{1}{2e^{\beta t/2}}\exp\left(-|x|/e^{\beta t/2}\right) \end{cases}$$

From Equation (25), I derive an ordinary differential equation for $\hat{g}_t$,

$$\frac{\partial}{\partial t}\hat{g}_t = \beta\left(\hat{g}_t^2 - \hat{g}_t\right)$$

Now postulate that $\hat{g}_t$ is of the form $\hat{g}_t(s) = h\left(e^{\beta t/2}s, t\right)$. Then from the previous equation I derive a partial differential equation for $h(y,t)$,

$$\frac{\partial}{\partial t}h = \beta\left(h^2 - h + \frac{1}{2}y\frac{\partial}{\partial y}h\right)$$

Accepting that $\lim_{t\to\infty}\frac{\partial}{\partial t}h = 0$, h must asymptotically satisfy the following ODE,

$$\frac{1}{2}y\frac{\partial}{\partial y}h = h - h^2$$

which admits the solution,

$$h(y) = \frac{1}{1+y^2}$$

This completes the proof. ■

Proposition 2 (reminded) *For any distribution g_0 of initial contacts that is symmetric and admits a finite variance, aggregate trade flows between two countries A and B are approximately proportional to their respective sizes (GDP_A and GDP_B), and inversely related to the distance between them ($Dist_{A,B}$),*

$$T_{A,B} \propto \frac{GDP_A \times GDP_B}{(Dist_{A,B})^{1+\epsilon}}$$

with $\epsilon \equiv 2\min\left(\frac{\gamma-(\beta-\delta)}{\beta}, 1\right)$, γ the population growth rate and β (resp. δ) the birth (resp. death) rate of contacts.

Proof. I will prove first that aggregate trade is proportional to economic size, and second that it is inversely proportional to distance raised to the power $(1+\epsilon)$.

Size: In any location x, all firms of the same age t have the same volume of exports towards and the same volume of import from any other location. For any $\lambda > 0$, if a location, or any set of locations (any country) produces λ times as much in the aggregate, it will export and import λ times as much in the aggregate. Aggregate trade flows between any arbitrary set of locations (countries) are therefore proportional to the size of the importing and exporting countries.

Distance: Denote by $\varphi(x)$ the p.d.f. of aggregate exports from the origin towards any location $x \in \mathbb{R}$. It is the weighted average of the exports of firms in the origin of all ages towards location x, normalized to sum up to 1,

$$\varphi(x) \equiv \frac{\gamma - \beta + \delta}{K_0}\int_0^\infty e^{-\gamma t} f_t(x)\, dt$$

I will prove that $\varphi(x)$ is equal to $1/x^{1+\epsilon}$ for $x \to +\infty$, up to a slowly varying function L,[28]

$$\varphi(x) = L(x) \times \frac{1}{x^{1+\epsilon}}$$

Step 1: By virtue of Karamata's abelian and tauberian theorem, the p.d.f. $\varphi(x)$ is equal to $1/x^{1+\epsilon}$ for $x \to +\infty$, up to a slowly varying function i.i.f. its Laplace transform $\hat{\varphi}$ is such that $1-\hat{\varphi}(s)$ is equal to s^{ϵ} for $s \to 0$, up to a slowly varying function. See for instance de Haan (1976) for an application of Karamata's theorem to p.d.f.'s. Formally, this means that I need to prove,

$$\lim_{s \downarrow 0} \frac{1-\hat{\varphi}(\lambda s)}{1-\hat{\varphi}(s)} = \lambda^{\epsilon}, \forall \lambda > 0$$

Step 2: Taking the two-sided Laplace transform of φ which I denote by $\hat{\varphi}$, and using the properties of the Laplace transform, the formula for f_t in Proposition 5 and simple algebra, I get,

$$\begin{aligned}
\hat{\varphi} &= \frac{\gamma-\beta+\delta}{K_0} \int_0^{\infty} e^{-\gamma t} \hat{f}_t dt \\
&= \frac{\gamma-\beta+\delta}{K_0} \int_0^{\infty} e^{-\gamma t} \frac{K_0 e^{(\beta-\delta)t} \hat{g}_0}{(1-e^{\beta t})\hat{g}_0 + e^{\beta t}} dt \\
&= (\gamma-\beta+\delta) \int_0^{\infty} e^{-(\gamma-\pi+\delta)t} \frac{\hat{g}_0}{e^{\beta t}(1-\hat{g}_0)+\hat{g}_0} dt \\
&= -\frac{\gamma-\beta+\delta}{\beta} \sum_{n=1}^{\infty} \frac{\left(\frac{\hat{g}_0}{\hat{g}_0-1}\right)^n}{n+(\gamma-\beta+\delta)/\beta}
\end{aligned}$$

where I iteratively integrate by part to get the last expression.

Step 3: To save on notations, I introduce $\alpha = \frac{\gamma-(\beta-\delta)}{\beta}$ so that $\epsilon = 2\min[\alpha, 1]$. Manipulating the previous expression $\hat{\varphi}$, recognizing Gauss's hypergeometric function ${}_2F_1$, and invoking one

[28] A function L is said to be slowly varying around $+\infty$ i.i.f.

$$\lim_{x \to +\infty} \frac{L(\lambda x)}{L(x)} = 1, \forall \lambda > 0$$

among the hundreds of useful properties of this function,[29] I get,

$$\begin{aligned}
1-\hat{\varphi}(s) &= 1+\alpha\sum_{n=1}^{\infty}\frac{\left(\frac{\hat{g}_0(s)}{\hat{g}_0(s)-1}\right)^n}{n+\alpha} \\
&= \sum_{n=0}^{\infty}\frac{(1)_n(\alpha)_n}{(1+\alpha)_n\, n!}\left(\frac{\hat{g}_0(s)}{\hat{g}_0(s)-1}\right)^n \\
&= {}_2F_1\left(1,\alpha,1+\alpha,\frac{\hat{g}_0(s)}{\hat{g}_0(s)-1}\right) \\
&= \frac{\Gamma(\alpha-1)\Gamma(1+\alpha)}{\Gamma(\alpha)\Gamma(\alpha)}\left(-\frac{\hat{g}_0(s)}{\hat{g}_0(s)-1}\right)^{-1}\sum_{k=0}^{\infty}\frac{(1)_k(1-\alpha)_k}{k!(2-\alpha)_k}\left(\frac{\hat{g}_0(s)}{\hat{g}_0(s)-1}\right)^{-k} \\
&+ \frac{\Gamma(1-\alpha)\Gamma(1+\alpha)}{\Gamma(1)\Gamma(1)}\left(-\frac{\hat{g}_0(s)}{\hat{g}_0(s)-1}\right)^{-\alpha}\sum_{k=0}^{\infty}\frac{(\alpha)_k(0)_k}{k!(\alpha)_k}\left(\frac{\hat{g}_0(s)}{\hat{g}_0(s)-1}\right)^{-k} \\
&\text{for a sufficiently small } s \text{ such that } \left|\frac{\hat{g}_0(s)}{\hat{g}_0(s)-1}\right|>1 \text{ and a non-integer } \alpha \\
&= \frac{\alpha}{\alpha-1}\left(-\frac{\hat{g}_0(s)}{\hat{g}_0(s)-1}\right)^{-1}{}_2F_1\left(1,1-\alpha,2-\alpha,\left(\frac{\hat{g}_0(s)}{\hat{g}_0(s)-1}\right)^{-1}\right)+\Gamma(1-\alpha)\Gamma(1+\alpha)\left(-\frac{\hat{g}_0(s)}{\hat{g}_0(s)-1}\right)^{-\alpha}
\end{aligned}$$

Step 4: The following lemma will prove useful. If g is the p.d.f. of a random variable X symmetric around the origin and with a finite variance $0<Var(X)<+\infty$, then its Laplace transform $\hat{g}$ is such that for any $\lambda>0$,

$$\lim_{s\downarrow 0}\frac{\frac{1-\hat{g}(\lambda s)}{\hat{g}(\lambda s)}}{\frac{1-\hat{g}(s)}{\hat{g}(s)}}=\lambda^2$$

To prove this lemma, note that g being a well defined p.d.f., $\hat{g}(0)=1$. Using l'Hôpital's rule,

$$\lim_{s\downarrow 0}\frac{\frac{1-\hat{g}(\lambda s)}{\hat{g}(\lambda s)}}{\frac{1-\hat{g}(s)}{\hat{g}(s)}}=\lim_{s\downarrow 0}\frac{1-\hat{g}(\lambda s)}{1-\hat{g}(s)}\frac{\hat{g}(s)}{\hat{g}(\lambda s)}=\lambda\lim_{s\downarrow 0}\frac{\frac{\partial}{\partial s}\hat{g}(\lambda s)}{\frac{\partial}{\partial s}\hat{g}(s)}$$

I use the known result that $\hat{g}^{(k)}(0)=(-1)^k\mu_k$ where μ_k is X's k-th moment. Since X is symmetric, its first moment is zero, and $\hat{g}'(0)=\mu_1=0$. The limit is again indeterminate. Applying l'Hôpital's rule a second time, and by the assumption $0<Var(X)=\mu_2-\mu_1^2=\mu_2<+\infty$, I prove the proposed lemma,

$$\lambda\lim_{s\downarrow 0}\frac{\frac{\partial}{\partial s}\hat{g}(\lambda s)}{\frac{\partial}{\partial s}\hat{g}(s)}=\lambda^2\lim_{s\downarrow 0}\frac{\frac{\partial^2}{\partial s^2}\hat{g}(\lambda s)}{\frac{\partial^2}{\partial s^2}\hat{g}(s)}=\lambda^2\frac{\mu_2}{\mu_2}=\lambda^2$$

Note that the assumption of finite variance is a sufficient but not a necessary condition. For example, Student's t-distribution with 2 degrees of freedom satisfies the desired property although its variance is infinite.

[29] See `http://functions.wolfram.com/HypergeometricFunctions/Hypergeometric2F1/02/02/` or a modified version of Kummer's Theorem (1836), as presented by Slater (1966) in Equation 1.7.1.3 on page 31.

Step 5: Let $h(s) = \left(-\frac{\hat{g}_0(s)}{\hat{g}_0(s)-1}\right)^{-1} = \frac{\hat{g}_0(s)}{\hat{g}_0(s)-1}$ and note that $h(0) = 0$ and $1 - \hat{\varphi}(0) = 0$. Using l'Hôpital's rule and the above lemma for the penultimate equality, I can now characterize the limit of interest,

$$\lim_{s\downarrow 0} \frac{1-\hat{\varphi}(\lambda s)}{1-\hat{\varphi}(s)} = \lim_{s\downarrow 0} \frac{\frac{\partial}{\partial s}\hat{\varphi}(\lambda s)}{\frac{\partial}{\partial s}\hat{\varphi}(s)}$$

$$= \lim_{s\downarrow 0} \frac{\frac{\partial}{\partial s}h(\lambda s)\left[\frac{\alpha}{\alpha-1}{}_2F_1(1,1-\alpha,2-\alpha,-h(\lambda s)) + \frac{\alpha}{2-\alpha}h(\lambda s){}_2F_1(2,2-\alpha,3-\alpha,-h(\lambda s)) + \Gamma(1-\alpha)\Gamma(1+\alpha)\alpha(h(\lambda s))^{\alpha-1}\right]}{\frac{\partial}{\partial s}h(s)\left[\frac{\alpha}{\alpha-1}{}_2F_1(1,1-\alpha,2-\alpha,-h(s)) + \frac{\alpha}{2-\alpha}h(s){}_2F_1(2,2-\alpha,3-\alpha,-h(s)) + \Gamma(1-\alpha)\Gamma(1+\alpha)\alpha(h(s))^{\alpha-1}\right]}$$

$$= \begin{cases} \lambda^2 & \text{when } \alpha > 1 \text{ so that the second and third terms vanish} \\ \lambda^{2\alpha} & \text{when } \alpha < 1 \text{ so that the first and second terms vanish} \end{cases}$$

$$= \lambda^{\epsilon}$$

This completes the proof. ■

Proposition 3 (reminded) *If the distribution of export sizes among individual firms is close to Zipf's law, then aggregate trade flows between two countries are approximately proportional to their respective sizes and inversely proportional to the distance between them. The canonical gravity equation holds,*

$$T_{A,B} \propto \frac{GDP_A \times GDP_B}{Dist_{A,B}}$$

Proof. From Equation (16), the distribution of export volumes among individual firms is close to Zipf's law if $\frac{\gamma}{\beta-\delta} \approx 1^+$. Plugging this condition into Proposition 2, one gets $\epsilon \approx 0^+$, so that the canonical gravity equation holds for aggregate trade flows. ■

B Data

Firm level export data: The data on firm level exports come from the French customs, and are described in greater detail in Eaton, Kortum and Kramarz (2011). Until 1992, all shipments crossing the French border are reported, either by the owner of the (exporting) firm, or by authorized customs commissioners. Information about the identity of the exporting firm, the value of the shipment, the industrial sector, and the destination country is recorded. This information is then aggregated over a year. I use data on all French exporters. A data point is therefore a firm, year, destination country and value of exports (in French Francs) vector.

Distance data: I use data on bilateral distances between countries collected and constructed by the CEPII. The distance between two countries is calculated as a weighted arithmetic average

of the geodesic distances between the main cities in these countries, where population weights are used. Data on the location of the main cities in each country (latitude and longitude), as well as the population of those main cities are used to compute those distances. The construction of the data is described in further detail by Mayer and Zignago (2006).

Country size data: I use as a measure of a country's size its nominal *GDP* (in million US$) in the current year. The data are collected from the Penn World Tables and are described in further detail at `http://pwt.econ.upenn.edu/`.

Bilateral trade flows: To proxy for the intensity of firm level contacts between countries other than France, I use data on bilateral trade flows between countries. The data correspond to the nominal value (in US$) of aggregate trade flows between country pairs. The data are collected from the NBER, and are described in further detail in Feenstra et al. (2004).