The Gravity Equation in International Trade: An Explanation*

Thomas Chaney†
University of Chicago, NBER and CEPR

September, 2011
PRELIMINARY AND INCOMPLETE

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 their respective sizes, measured by their GDP, and inversely proportional to the geographic distance between them. While the role of economic size is well understood, the role played by distance remains a mystery. In this paper, I propose the first explanation for the gravity equation in international trade. This explanation is based on the emergence of a stable international network of importers and exporters. Firms can only export into markets in which they have a contact. They acquire contacts by gradually meeting the contacts of their contacts. I show that if, as observed empirically, (*i*) the distribution of the number of foreign countries accessed by exporters is fat tailed, (*ii*) there is a large turnover in exports, with firms often going in and out of individual foreign markets, and (*iii*) geographic distance hinders the initial acquisition of contacts in an arbitrary way, then trade is proportional to country size, and inversely proportional to distance. Data on firm level, sectoral, and aggregate trade support further predictions of the model.

*I want to thank Fernando Alvarez, Xavier Gabaix, Sam Kortum and Bob Lucas for helpful discussions. I am indebted to Jong Hyun Chung, Stefano Mosso and Adriaan Ten Kate for their research assistance.

†Contact: Department of Economics, The University of Chicago, Chicago, IL 60637. Tel: 773-702-5403. Email: tchaney@uchicago.edu.

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 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.[1] 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.

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

[1] 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.

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 inversely related to 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 (forthcoming) 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[2] 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 change 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 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

[2]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.

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.

The remainder of the paper is organized as follows. In section 1, I present a theoretical model of firm level and aggregate trade. In section 1.1, I spell out an economic model of trade subject to matching frictions. In section 1.2, I characterize the patterns of firm level trade. In section 1.3, I show that aggregate trade obeys the gravity equation. In section 2, I test empirically the main theoretical predictions of the model. I relegate to Appendix A all mathematical proofs, and to Appendix B the description of the data and robustness checks.

1 Theory

In this section, I develop a simple model of the formation of a stable network of importers and exporters. The model is a extension of the Krugman (1980) model of international trade in differentiated goods subject to matching frictions similar to the Chaney (2011) model of trade networks.

1.1 A model of trade subject to matching frictions

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. 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 1.2.

There are two types of goods: final goods and intermediate inputs. Final goods are produced by combining differentiated intermediate inputs and labor. 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. 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.

The static problem of the firm: Consider what happens within period t. For the moment, I drop the time t index. Firm i buys intermediate inputs from a continuum of suppliers $k \in K_i$ and sells its output to a continuum of customers $j \in J_i$. Both K_i and J_i will be endogenously determined dynamically below. Firm i combines labor with $q_i(k)$ units of differentiated intermediate inputs from each supplier k to produce Q_i units of output,

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

with $0 < \alpha < 1$ the share of intermediate inputs in production and $\sigma > 1$ the elasticity of substitution between any two intermediate inputs. Firm i faces the same iso-elastic demand from any customer $j \in J_i$,

$$p_j(i) q_j(i) = \frac{p_j(i)^{1-\sigma}}{\int_{k \in K_j} p_j(k)^{1-\sigma} dk} \alpha X_j \tag{2}$$

with $p_j(i)$ the price charged by i to customer j, and X_j the total spending on intermediate inputs and labor by j. Given these iso-elastic demands, firm i charges all its customers the same constant mark-up, $\frac{\sigma}{\sigma-1}$, over its marginal cost,

$$p_j(i) = p_i = \frac{\sigma}{\sigma-1} w^{1-\alpha} \left(\int_{k \in K_i} p_i(k)^{1-\sigma} dk \right)^{\frac{\alpha}{1-\sigma}} \tag{3}$$

with w the competitive wage rate. For simplicity, I will consider a symmetric equilibrium where all firms within a cohort have the same number of suppliers and customers[3] and therefore charge the same price: $\|K_j\| = \|K\|$ and $p_j(k) = p$ for any $j, k \neq i$. Given this symmetry assumption, the demand equation (2) and the pricing equation (3), the total sales of firm i only depend on the number of suppliers and of customers,

$$p_i Q_i = \int_{j \in J_i} p_i q_j(i) dj = \|K_i\|^{\alpha} \times \|J_i\| \times \left(\alpha \frac{X}{\|K\|^{1+\alpha}} \right) \tag{4}$$

[3]In such a symmetric equilibrium, all the complexity of the input-output structure of the economy is assumed away. See Carvalho (2010), Acemoglu, Ozdaglar and Tahbaz-Saleh (2010) or Atalay, Chaney, Hortacsu and Syverson (2011) for models with such a complex structure.

To save on notations, I will make from now on a slight abuse of language and use K (resp. J) to denote the mass of suppliers (resp. customers), instead of $\|K\|$ (resp. $\|J\|$).

It is clear from the previous Equation (4) that the number of suppliers and customers increases output, sales, and ultimately profits. The mass of suppliers, 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 new contacts. The notion that the diversity of intermediate inputs plays a role similar to capital has been presented since at least Romer (1990).[4] 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. I now turn to this dynamic decision.

The dynamic problem of the firm: A firm is born with a mass $K_0 = J_0$ of suppliers and customers. Those contacts are randomly distributed along the real line $\mathbb{R}$ according to the symmetric p.d.f. g_0. One can think that a potential entrant pays a fixed entry cost to create a new firm. This fixed cost buys both the blueprint for a new variety and this initial set of upstream and downstream contacts. This fixed cost will be reimbursed by a discounted stream of per period profit over the lifespan of the firm. Imposing a free entry condition would pin down the number of entrants each period.

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 instead 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 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. Facing the threat of this outside option, firm i sets a constant price p_I to reveal the name of one of its suppliers to its existing customers. The price p_I is set just low enough to prevent firm $j \in J_i$ to look for contact directly instead. Just as i sells information about its suppliers, it also buys at the same price p_I information about new suppliers from its existing suppliers.

[4] 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.

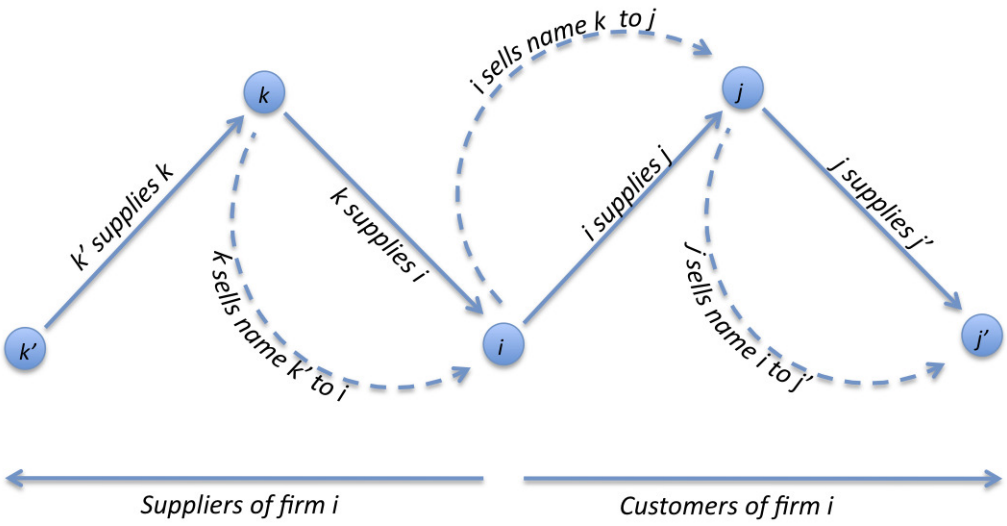

Figure 1: Firms buy and sell information about suppliers and customers.

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'.

In addition to the direct cost p_I of buying information, firm i faces a convex adjustment cost $G(I_i, K_i)$. The adjustment cost function is assumed increasing and convex in I, $G_I, G_{II} > 0$, decreasing in K, $G_K < 0$, and homogenous of degree one in I and K. This convex adjustment cost function is analogous to the adjustment cost assumed in the classical investment literature, 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 G 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.

Note that as in classical investment models, firm i has two reason 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, is set by an arbitrage condition, and the number of requests for names firm i receives depends on the decisions of its customers, $j \in J_i$. 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 an exogenous rate δ.

Firm i solves the following dynamic optimization problem,

$$\max_{I_i(t)} V(0) = \int_0^{+\infty} e^{-\rho t} \left(K_i(t)^{\alpha} \left(\frac{\alpha J_i(t) X(t)}{\sigma K(t)^{1+\alpha}} \right) - p_I(t) I_i(t) + p_I I(t) - G(I_i(t), K_i(t)) \right) dt \tag{5}$$

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

Firm i maximizes a discounted stream of profits, with an exogenous discount rate ρ. The first term represents per period profits, net of spending and receipts on information acquisition. It is a fraction $1/\sigma$ of the aggregate sales derived in Equation (4). In addition, firm i purchases information about I_i new suppliers at a price p_I each, and it sells information about I suppliers at a price p_I each. Finally, firm i pays to convex adjustment cost G to acquire new suppliers.

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,

$$I_i = \beta \left(\frac{p_I}{w}, \rho, \delta \right) \times K_i$$

where the function β sumarizes the contributions of the production function and the adjustment cost function that are relevant for the optimal investment decision. In general equilibrium, I could solve for the sequence of prices $p_I(t)/w(t)$. In this model as in any model where growth is driven by the accumulation of one factor of production (here K) combined with labor under constant returns to scale, growth ultimately is driven by population growth, as in Solow (1956). While I do not explicitly solve for such a steady state growth path, I will assume that the economy is along such a path, so that $I/K = \beta$ is constant.

Because all firms are charging the same price p_I per contact information, firm i has no reason to direct its search for new suppliers to any particular $k \in K_i$. To break this indeterminacy, I assume that the I_i new names come uniformly from all existing suppliers K_i. This means that any one of the existing suppliers $k \in K_i$ reveals one of the names of its suppliers, $k' \in K_k$, with a probability βdt over a small time interval dt. To break the indeterminacy of which name $k' \in K_k$ gets revealed by firm k, I simply assume that k draws k' at random among its existing K_k contacts.

Here is a recap of the conclusions of the model: A firm is born with an initial mass K_0 of suppliers, distributed geographically according to the p.d.f. g_0. 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. The next section characterizes explicitly the dynamic evolution of firm level trade flows, i.e. trade between the suppliers and customers of this model.

1.2 Firm level trade flows

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. With the exception of the population growth rate γ, all the parameters introduced in this section $(K_0, g_0.\beta, \delta)$ are the same as the ones in the economic model above. I treat the arrival rate of new contacts, β, as a parameter, knowing from the model above that it is the solution to the dynamic optimization problem of the firm.

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 newly born firms: firms within each cohort gradually get connected to each other.[5]

Death of a firm: I assume that firms are infinitely lived. This assumption is inocuous, 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 infinitly long tail of infinitesimally small firms and firm sizes shrinking on average.

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,

$$f_t : \mathbb{R} \to \mathbb{R}^+ \text{ and } K_t \equiv \int_{\mathbb{R}} f_t(x)\, dx \tag{6}$$

$f_t(x)$ is the density of contacts of a firm of age t in location x. In other words, the mass of contacts a firm of age t 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,

$$\frac{\partial f_t(x)}{\partial t} = \beta \int_{\mathbb{R}} \frac{f_t(x-y)}{K_t} \times f_t(y)\, dy - \delta f_t(x) \tag{7}$$

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 first term with the integral sign on the right hand side of Equation (7) corresponds to the creation of new contacts. It can be decomposed into four components, β, $\frac{f_t(x-y)}{K_t}dx$, $f_t(y)\, dy$ and the integral

[5] While this simplifying assumption is strong, relaxing it would force me to keep track of the entire system of firms simultaneously, and would render the model analytically intractable. Numerical simulations suggest that the main results of the paper are robust to relaxing this assumption.

sign $\int_{y\in\mathbb{R}}$. The first component, β, correspond 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 a contact in location y reveals the name of one of her contacts in a neighborhood dx of x.[6] 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. 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 (7) 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 (7) 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,

$$K_t = K_0 e^{(\beta-\delta)t} \tag{8}$$

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).

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

[6] 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 $-y$: $f_{0,t}(x) = f_{y,t}(x-y)$.

is Pareto distributed. The fraction $F(K)$ of firms with K or fewer contacts is given by,

$$F(K) = 1 - \left(\frac{K}{K_0}\right)^{-\frac{\gamma}{\beta-\delta}} \text{ for } K \geq K_0 \tag{9}$$

From Equation (8), 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 (8) 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, $\frac{\gamma}{\beta-\delta} \approx 1^+$.[7] Beyond this empirical evidence, the assumption that $\frac{\gamma}{\beta-\delta} \approx 1^+$ seems to be a good candidate for a stationary system, where the number of contacts of existing firms grows approximately at the same rate as the population as a whole. While deviations from this stationary benchmark are to be expected in the data, these deviations ought not to be 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.[8] The average (squared) distance from the contact of a firm with K contacts, $\Delta(K)$, increases with its number of contacts,

$$\Delta(K) \equiv \int_{\mathbb{R}} x^2 \frac{f_K(x)}{K} dx = \Delta_0 \left(\frac{K}{K_0}\right)^{\frac{\beta}{\beta-\delta}} \tag{10}$$

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.

[7] 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.

[8] $f_K = f_{t(K)}$ with $t(K)$ s.t. $K_{t(K)} = K$.

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 distributions g_0 and g_0' that share 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.

Note the following interesting special case, and asymptotic result: if the initial distribution of contacts g_0 is a Laplace distribution (a symmetric two-sided exponential distribution), then each subsequent distribution $g_t = f_t/K_t$ is also Laplace, with the variance increasing at a constant rate β. Moreover, for any initial distribution g_0, the distribution g_t converges to a Laplace distribution as t grows large.

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.

1.3 Aggregate trade flows

Each firm trades one unit with each of its contacts. 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 lemma and proposition show that aggregate trade flows in this model obey the gravity equation in international trade.

Lemma 1 *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.

Proposition 1 *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 a small economy, i.e. one that has a negligible size relative to the rest of the world. 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 one for large countries.

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 1 is mathematically arduous, a simplified explanation would be as follows. Each cohort has a different distribution of contacts. From Equation (9), the distribution of the number of contacts in the population is a power law. From Equation (10), 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.[9]

To understand 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

[9]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.

$\left[-2\sqrt{\Delta(K)}, +2\sqrt{\Delta(K)}\right]$. 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 (9) for the distribution of K's and Equation (10) for $\Delta(K)$, the fraction (density) of firms that export at a distance x, which I denote $\varphi(x)$, is given by the following expression,

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

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.

Proposition 1 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 $\left(1 + 2\frac{\gamma-(\beta-\delta)}{\beta} \approx 1^+\right)$. 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 exports of individual firms within the population. It says nothing about where those exports go. While Zipf's law is a statement about *how much* different firms export, the gravity equation is a statement about *where* firms export.[10]

[10]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.

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 by the direct interaction between people. This feature implies that information about potential foreign contacts is transmitted along individual connections. Advances in transportation or communication technologies affect physical trade barriers as well as the direct cost of information (the function g_0), but not 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. In this model on the other hand, the shape of aggregate trade flows is immune to changes in the g_0 function. If direct interactions between people play a role today as it did a century ago, this model predicts that the role for distance will remain unchanged.

2 Empirics

THIS SECTION IS CURRENTLY VERY PRELIMINARY AND INCOMPLETE...

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.

To confirm this prediction, using firm level data on export for France in 1992, I show in Figure 2 the relation between the log of the rank of a firm, versus the log of its size. The relationship is very close to Zipf's law for large exporters. In Figure 3, I show 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. The relationship between $\Delta(K)$ and K is well approximated by a power function. The estimated slopes of these two relationships would predict according to my model that the number of firms that export to country c at a distance $Dist_c$ should be proportional to $1/Dist_c^{1.17}$. In the data, it is proportional to $1/Dist_c^{1.16}$.

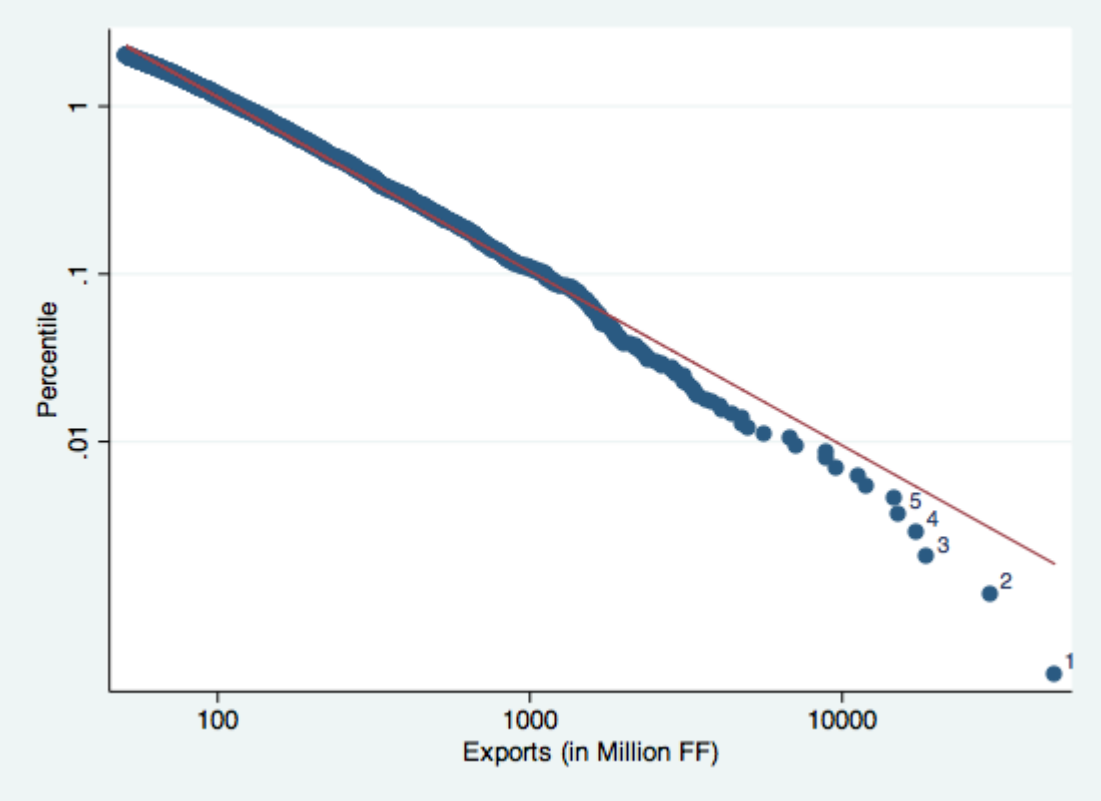

Figure 2: The distribution of firm level total exports is Zipf.

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: all French exporters, 1992.

Conclusion

This paper offers the first theoretical explanation for the gravity equation in international trade in the sense that it explains not only why trade is proportional to size, but also the mysterious -1 distance elasticity of trade. This explanation is 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.

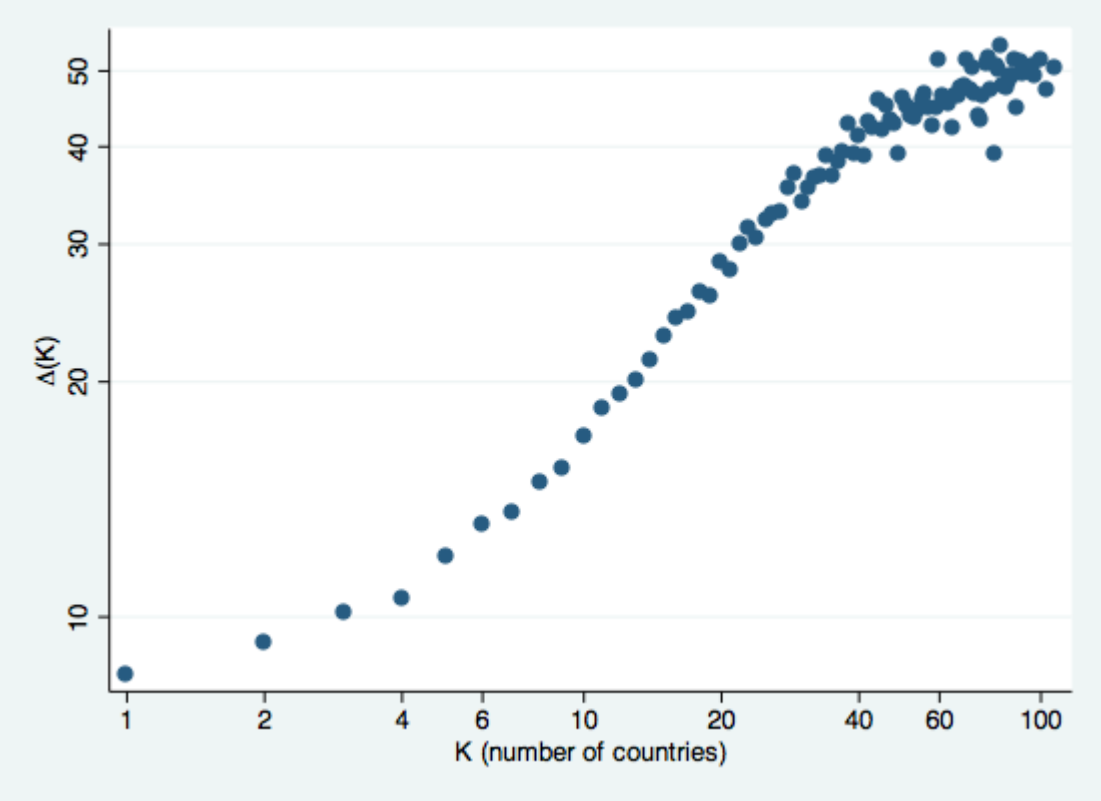

Figure 3: Average (squared) distance of exports versus number of export destinations.

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.

References

Acemoglu, Daron, Asuman Ozdaglar and Alireza Tahbaz-Saleh. 2010. "Cascades in Networks and Aggregate Volatility," MIT, *mimeo*.

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.

Antràs, Pol and Arnaud Costinot. Forthcoming. "Intermediated Trade," *Quarterly Journal of Economics*.

Arkolakis, Costas, Arnaud Costinot and Andres RodrÌguez-Clare. Forthcoming. "New Trade Model, Same Old Gains?" *American Economic Review*.

Atalay, Enghin, Ali Hortaçsu, James W. Roberts, and Chad Syverson. 2010. "On the Network Structure of Production," University of Chicago, *mimeo.*

Atalay, Enghin. 2011. "An Exact Expression for the In-degree Distribution of the Jackson and Rogers (2007) Model," University of Chicago, *mimeo.*

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

Bramoullé, Yann and Brian W. Rogers. 2009. "Diversity and Popularity in Social Networks," CIRPEE Working Paper No. 09-03.

Burchardi, Konrad B. and Tarek A. Hassan. 2010. "The Economic Impact of Social Ties: Evidence from German Reunification." Chicago Booth, *mimeo.*

Carvalho, Vasco M. 2010. "Aggregate Fluctuations and the Network Structure of Intersectoral Trade," CREI *mimeo.*

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

Chaney, Thomas. 2011. "The Network Structure of International Trade," NBER WP 16753.

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): 537-542.

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," University of Michigan *mimeo.*

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 Francis Kramarz. Forthcoming. "An Anatomy of International Trade: Evidence from French Firms," *Econometrica.*

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. 2008. "Power Laws," *The New Palgrave Dictionary of Economics*. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2008.

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*, 50:213-224.

HELPMAN, Elhanan, Marc J. MELITZ, and Yona RUBINSTEIN. 2008. "Estimating Trade Flows: Trading Partners and Trading Volumes," *Quarterly Journal of Economics*, 123: 441-487.

HIDALGO, César A., B. KLINGER, Albert-László BARABÁSI, and Ricardo HAUSMANN. 2007. "The Product Space Conditions the Development of Nations" *Science*, 317(5837): 482-487.

JACKSON, Matthew O. 2010. "An Overview of Social Networks and Economic Applications," *The Handbook of Social Economics*, edited by J. Benhabib, A. Bisin, and M.O. Jackson, North Holland Press.

JACKSON, Matthew O., and Brian W. ROGERS. 2007. "Meeting Strangers and Friends of Friends: How Random Are Social Networks?" *American Economic Review*, 97(3): 890-915.

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.

RAUCH, James E. and Vitor TRINDADE. 2002. "Ethnic Chinese Networks in International Trade," *Journal of International Economics*, 84(1): 116-30.

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

Proposition 2 *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,[11] *$\mathcal{B}^{-1}$ 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[12] in Equation (7), I can rewrite it in a compact form,

$$\frac{\partial f_t}{\partial t} = \beta\frac{f_t * f_t}{K_t} - \delta f_t \tag{11}$$

with initial condition $f_0 = K_0 g_0$. I will first solve for K_t, and then solve for f_t. Integrating Equation (11) 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 (11), 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 transform of the convolution of two function is the product of their Laplace transforms, I get the following ordinary differential equation,

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

[11] 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.

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

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 (8), (9) and (10) 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 (8). 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 (9). 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 (10). 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 3 *If the initial distribution of contacts is Laplace, $g_0 \sim$ Laplace $\left(0, \sqrt{\Delta_0/2}\right)$, then the distribution of contacts remains Laplace at all subsequent period, $g_t = f_t/M_t \sim$ Laplace $\left(0, e^{\beta t/2}\sqrt{\Delta_0/2}\right)$.*

Moreover, for any initial distribution g_0, asymptotically as $t \to \infty$, the p.d.f. of contacts, $g_t = f_t/K_t$, converges to a Laplace $\left(0, e^{\beta t/2}\right)$ distribution,

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

Proof. For simplicity, consider a normalized Laplace distribution $g_0 \sim$ Laplace $(0,1)$. The proof can trivially be extended to $\sqrt{\Delta_0/2} \neq 1$. The Laplace transform of g_0 is $\hat{g}_0(s) = \frac{1}{1+s^2}$. From Proposition 2, 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.

From Equation (11), 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 without proof that $\lim_{t\to\infty}\frac{\partial}{\partial t}h=0$, h must asymptotically satisfy the following ordinary differential equation,

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

which admits the solution,

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

This completes the proof. ■

Lemma 1 (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 ($Dist_{A,B}$),*

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

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 $\varphi\left(x\right)$ 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,

$$\varphi\left(x\right)\equiv\frac{\gamma-\beta+\delta}{M_0}\int_0^\infty e^{-\gamma t}f_t\left(x\right)dt$$

I will prove that $\varphi\left(x\right)$ is equal to $1/x^{1+\epsilon}$ for $x\to+\infty$, up to a slowly varying function L,[13]

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

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

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

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 2 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' hypergeometric function ${}_2F_1$, and invoking one among the hundreds of useful properties of the hypergeometric function,[14] I get,

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

[14] 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 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.

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,

$$\begin{aligned}
&\lim_{s \downarrow 0} \frac{1-\hat{\varphi}(\lambda s)}{1-\hat{\varphi}(s)} = \lim_{s \downarrow 0} \frac{\frac{\partial}{\partial s}\hat{\varphi}(\lambda s)}{\frac{\partial}{\partial s}\hat{\varphi}(s)} \\
&= \lim_{s \downarrow 0} \frac{\frac{\partial}{\partial s}h(\lambda s)\left[\frac{\alpha}{\alpha-1}{}_2F_1(1,1-\alpha,2-\alpha,-h(\lambda s)) + \frac{\alpha}{2-\alpha}h(\lambda s){}_2F_1(2,2-\alpha,3-\alpha,-h(\lambda s)) + \Gamma(1-\alpha)\Gamma(1+\alpha)\alpha(h(\lambda s))^{\alpha-1}\right]}{\frac{\partial}{\partial s}h(s)\left[\frac{\alpha}{\alpha-1}{}_2F_1(1,1-\alpha,2-\alpha,-h(s)) + \frac{\alpha}{2-\alpha}h(s){}_2F_1(2,2-\alpha,3-\alpha,-h(s)) + \Gamma(1-\alpha)\Gamma(1+\alpha)\alpha(h(s))^{\alpha-1}\right]} \\
&= \begin{cases} \lambda^2 & \text{when } \alpha > 1 \text{ so that the second and third terms vanish} \\ \lambda^{2\alpha} & \text{when } \alpha < 1 \text{ so that the first and second terms vanish} \end{cases} \\
&= \lambda^{\epsilon}
\end{aligned}$$

This completes the proof. ■

Proposition 1 (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 (9), 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 Lemma 1, one gets $\epsilon \approx 0^+$, so that the canonical gravity equation holds for aggregate trade flows. ■

B Data