Why So Little Trade?

Trade is tiny compared to potential. Big countries trade more (e.g. US and China) but small ones appear more open (e.g. Belgium). Distance and borders apparently kill a lot of trade, given relative country size. T. Friedman’s “the world is flat” is hugely wrong — world trade < 10% of its potential.

The gravity model organizes these visible regularities in a model that can infer the size of barriers to trade. The model applies within countries and within regions (and even within institutions such as BC). The model is also very useful for describing migration and foreign direct investment.

A big achievement of the model is to make sense of multi-country interactions. Intuitively, the more country A trades with country B, the less is left over for trade with country C. Frictions between B and C thus affect A’s trade with either B or C.
Given the estimated relationships of the model, useful comparative static experiments are conducted. For example, the effects of free trade agreements such as NAFTA can be measured. These effects include trade flow changes. They can also include wage and welfare effects when gravity is combined with supplemental data and the sort of models of the structure of production to be analyzed later in the course.

These notes develop the model of how trade patterns are determined by size (of countries, regions and sectors) and trade frictions. The model is applied to data to make sense of the world. Some further implications are drawn.
1 Size and Trade Patterns

Size is measured by national product instead of geographic area. Bring some notion of big economies to interpret the width of arrows in relation to economic size and the power of the relationship is obvious.
The size and trade relationship is much more obvious in this display of Japan’s trade with EU members. For clarity, each EU member’s GDP and trade with Japan are normalized (divided by) the GDP and trade of Greece respectively.

Notice the good fit of the line (75% to 85% of variation of trade is explained by size alone) and the slope of the best-fit line is $\approx 1$. 

Figure 1: Trade is proportional to size
(a) Japan’s exports to EU, 2006
(b) Japan’s imports from EU, 2006
Notice that the log of trade/GDP is powerfully decreasing in log of distance in the diagrams. And the slope is less than $-1$ but fairly close to it. If we interpret a slope of $-1$, it means that doubling distance cuts trade in half, all else equal.

The bottom panel suggests that other factors besides distance play into trade size, indicating that a richer set of factors need to be investigated.
### Trade Compared with GDP

**TABLE 1.2**

Countries with the highest ratios of trade to GDP tend to be small in economic size. Countries with the lowest ratios of trade to GDP tend to be very large in economic size.

<table>
<thead>
<tr>
<th>Country</th>
<th>Trade/GDP (%)</th>
<th>GDP ($ Billions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hong Kong</td>
<td>27.1%</td>
<td>3.77</td>
</tr>
<tr>
<td>Switzerland</td>
<td>22%</td>
<td>270</td>
</tr>
<tr>
<td>Singapore</td>
<td>15.5%</td>
<td>3.277</td>
</tr>
<tr>
<td>China</td>
<td>13.8%</td>
<td>6.575</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>6.0%</td>
<td>2.076</td>
</tr>
<tr>
<td>Japan</td>
<td>5.5%</td>
<td>1.13</td>
</tr>
<tr>
<td>Mexico</td>
<td>5.4%</td>
<td>1.05</td>
</tr>
<tr>
<td>Germany</td>
<td>5.0%</td>
<td>1.70</td>
</tr>
<tr>
<td>France</td>
<td>4.5%</td>
<td>0.850</td>
</tr>
</tbody>
</table>

Size and Trade Summary Observations

- bilateral trade rises with the size of either trading partner
- countries further apart trade less
- borders appear to impede trade a lot

The gravity model explains these patterns.

### 2 Frictionless Gravity

**Frictionless World Benchmark**

Size has a lot of explanatory power. Bigger incomes buy more from everywhere; bigger sales sell more to everywhere.

Developing a model where size alone determines trade patterns is thus useful in abstracting from complex frictions.

A particularly important use for such a model is to provide a benchmark for what a frictionless world would look like.
Frictionless Gravity Model

Assume:

- demand at each destination for goods from all origins
- market clearance
- perfect arbitrage with, for now, no trade costs

Expenditure by \( j \) is \( E_j \); sales of \( i \) is \( Y_i \); world sales is \( Y \).

In a completely smooth homogenized world, the exports flow from \( i \) to \( j \), \( X_{ij} \) is given by:

\[
\frac{X_{ij}}{E_j} = \frac{Y_i}{Y} \Rightarrow X_{ij} = \frac{Y_i E_j}{Y} \tag{1}
\]

Equation (1) \( \Rightarrow \) the share of \( j \)'s expenditure on goods from \( i \) is equal to the share of world expenditure (= sales) on goods from \( i \). Figure 1 says (1) fits well, \( X_{ij} \) proportional to \( E_j \). Natural frictionless benchmark.

Frictionless Gravity Equilibrium

Market clearance (material balance) implies that

\[
\sum_j X_{ij} = Y_i. \tag{2}
\]

World budget constraint \( \Rightarrow \sum_j E_j = \sum_i Y_i = Y \). Thus (1) is consistent with market clearance: check by summing right hand side of (1) over \( j \) and using the world budget constraint to set \( \sum_j E_j/Y = 1 \).

If we impose a no net trade budget constraint for each country then \( E_i = Y_i \), hence \( X_{ij} = Y_i Y_j/Y \), but the more general specification (1) is far more realistic.

Implications of Frictionless Gravity

Define \( s_i = Y_i/Y \), country \( i \)'s share of world sales, and \( b_i = E_i/Y \), country \( i \)'s share of world expenditure. Assume balanced trade for now, \( b_i = s_i \). Then:

\[
X_{ij} = s_i s_j Y.
\]

Implications

1. Any country trades more with bigger partners.
2. Smaller countries are more naturally open:

\[
\sum_{i \neq j} X_{ij}/Y_j = 1 - s_j
\]

which is decreasing in \( s_j \).
3. Faster growing country pairs have increasing share of world trade: \( X_{ij}/Y = s_i s_j \) is increasing in \( s_i, s_j \).

For more implications when unbalanced trade is modeled, see Anderson (2011).
Aside on Growth Rate Accounting

Point 3 on the previous slide uses a basic property of growth rate accounting.

Let \( \hat{x} \) denote the growth rate of \( x \). Then (the Hat Rule)
\[
x = y/z \Rightarrow \hat{x} = \hat{y} - \hat{z}
\]
and
\[
x = yz \Rightarrow \hat{x} = \hat{y} + \hat{z}.
\]
Thus the rate of growth of \( X_{ij}/Y \) is equal to \( \hat{s}_i + \hat{s}_j \). And \( \hat{s}_i = \hat{Y}_i - \hat{Y} \).

The Hat Rule follows from log differentiation: \( d \ln y/z = d \ln y - d \ln z = \hat{y} - \hat{z} \).

3 Gravity with Frictions

Evidence of Frictions
Trade is much smaller than indicated by (1). US has 25 percent of world GDP, exports should be 75% of GDP vs. 10-15% actual. \( X_{US,ROW}/Y_{US} = 0.75 \).

- Trade falls sharply with distance (with effect \( D_{ij} \) reflecting distance between \( i \) and \( j \)):
\[
X_{ij} = \frac{Y_i E_j}{Y} \frac{1}{D_{ij}} .
\]
(3) gives a pretty good fit with actual trade data (viz. Figure 2). Implication: doubled distance \( \Rightarrow \) halved trade.

- Crossing borders further reduces trade a lot. i.e., \( D_{ij} = d_{ij} B_{ij} \) where \( B_{ij} > 1 = B_{ii} \) for \( i \neq j \) in equation (3) is the effect of a border between \( i \) and \( j \) and \( d_{ij} \) is distance.

Economic Theory of Gravity
Equation (3) or its variants allowing for borders and other effects was inspired by Newton’s Law of Gravity (where \( D_{ij} = d_{ij}^2 \), the square of bilateral distance between \( i \) and \( j \), \( Y_i \) is mass at \( i \) and \( 1/Y \) is the gravitational constant). Thus it is called the gravity equation or model. It has no economic theory behind it.

Economic theory leads to an expenditure share called structural gravity. (It is derived from one of three well justified foundations.)
\[
\frac{X_{ij}}{E_j} = \frac{Y_i}{Y} \left( \frac{D_{ij}}{Y P_j} \right)^{1-\sigma},
\]
where \( \sigma > 1 \), hence \( 1 - \sigma < 0 \).
Behind the Share Equation

The intuitive meaning of (4) is that bilateral trade falls as the ‘economic distance’ between origin and destination rises. Equation (4) restricts the responsiveness to a constant elasticity \(1 - \sigma, \sigma > 1\). [For the derivation see Anderson (2011).]

Equation (4) is linear in logarithms — suggests fitting a straight line on data points relating log trade shares to log distance. The slope is \(1 - \sigma\) to be inferred.

Essentially this is the empirical procedure used. But complicated by needing multivariate inference. \(\Pi_i\) and \(P_j\) act on the bilateral trade flow data as common exporter and importer country shifters (fixed effects to be inferred).

Economic Theory of Gravity

Repeating the share equation (4)

\[
\frac{X_{ij}}{E_j} = \frac{Y_i}{Y} \left( \frac{D_{ij}}{\Pi_i P_j} \right)^{1-\sigma}.
\]

The right hand side of share equation is in two parts. \(Y_i/Y\) is the frictionless expenditure share prediction. \((D_{ij}/\Pi_i P_j)^{1-\sigma}\) is the effect of trade frictions.

\(\Pi_i, P_j\) are indexes of all outward and inward bilateral trade costs \(D_{ij}\), respectively, called outward and inward multilateral resistance. \(\Pi_i\) is the appropriate ‘average’ portion of trade costs borne by seller \(i\) to all destinations. \(P_j\) is the appropriate average portion of trade costs borne by buyer \(j\) from all sources. Behind the scenes, all 3rd party effects are incorporated into multilateral resistance. (The analogy with Newtonian gravity is completed by noting that multilateral resistance is the solution to the famous N body problem in the case of economic gravity.)

\(\Pi_i\) and \(P_j\) are not observable but can be inferred along with \(D_{ij}\).

\(\Pi_i\) and \(P_j\) are also interpreted as sellers’ and buyers’ incidence of trade costs. The incidence interpretation is seen clearly by recalling the standard incidence analysis of the first course in economics. Supply and demand schedules in the hypothetical world market intersect in a frictionless equilibrium at the price \(p^*\). With trade frictions driving a wedge between sellers price \(p\) and buyers price \(pt\) due to a trade cost factor \(t > 1\), the equilibrium volume of trade moves to the left of the frictionless volume as shown on the figure below. The full trade cost \(t\) is divided into sellers incidence \(\Pi\) and buyers incidence \(P\) using the frictionless price \(p^*\), the sellers price \(p\) and \(\Pi = p^*/p\). Then \(t/\Pi = P\), the buyers incidence.
Standard Incidence Analysis

Equilibrium buyers price \( p_t \), sellers price \( p \), trade cost factor \( t \). Reference: In this hypothetical equilibrium, it is as if all sellers pay a cost factor \( \Pi_i \) to get their goods to a hypothetical world market (or for each unit shipped only \( 1/\Pi_i \) units arrive on the world market) while all buyers pay a cost factor \( P_j \) (or for each unit purchased only \( 1/P_j \) units arrive at destination). For each country in the aggregate the incidence breakdown is just as in the figure above: sellers bear \( \Pi \) of the total trade cost to get their goods to the world market, while buyers bear \( P \) to bring their bundle of goods (one from each origin country) home from the world market.

Behind the Share Equation, 2

While share equation (4) is certainly quite special, it is more general than it at first appears.

The three theoretical justifications for (4) are:

1. consumers (producers) gain from variety of goods consumed (used in production)
2. consumers (producers) differ in their ideal varieties, characterized by a probability distribution, and equation (4) gives the proportion of buyers who prefer the variety of region $i$.

3. consumers want only one variety of any good but there are many goods and producers differ in their productivities, characterized by a probability distribution. The shares in equation (4) refer to proportions of all goods produced by $i$ and sold to $j$.

With fairly plausible restrictions on the way gains from variety work in item 1 and on the probability distributions in items 2 and 3, equation (4) results. In the gains from variety case the parameter $\sigma$ is the elasticity of substitution between varieties. In the probability distribution cases 2 and 3 the parameter $1 - \sigma$ is a dispersion parameter (like the variance parameter of the normal distribution) of the probability distribution. For relatively homogeneous types of products, variant 3 is the natural interpretation while for differentiated products variants 1 or 2 are the natural interpretation.

**Gravity, Migration and Investment**

The gravity model can also be applied to migration of labor and to foreign direct investment. The same structure yields similarly good fit with data. The theoretical underpinnings change a bit in details that do not matter for many purposes.

Gravity also tends to explain portfolio investment (cross country ownership of stocks and bonds) but here the theory base is lacking.
4 Empirical Gravity

Empirical Gravity

Example: econometric inference of coefficients $\delta, b$ in assumed function:

$$D_{ij} = d_{ij}^\delta b_{ij}$$

(5)

where $d_{ij}$ is the distance between $i$ and $j$, $b_{ij} = 1$ if there is a border between $i$ and $j$ and $b_{ij} = 0$ if there is no border (inter- or intra-regional trade), $b > 0$ is the border resistance. It can be identified when shipments data includes internal as well as external trade.

Specification (5) makes the border resistance the same for all countries, but this can be relaxed to allow variation by country and also by direction of trade.

More elaborate versions of (5) allow for language differences, contiguity, former colonial ties, etc.

Inference

Infer best fitting coefficients $\delta, b$ along with $\Pi_i$ and $P_j$ from

$$\frac{X_{ij}}{Y_iE_j/Y} = \left(\frac{\delta^\delta b_{ij}}{\Pi_iP_j}\right)^{1-\sigma} \epsilon_{ij}$$

(6)

where $\epsilon_{ij}$ is the random error term representing the forces not explained by the model.

Estimated equations like (6) usually “explain” 90% of the variation in trade. Coefficients are precisely estimated.

Coefficient $1 - \sigma$ can be inferred if some trade cost is directly observed. Otherwise, distance elasticity $\delta(1 - \sigma)$ is inferred, for example. $\sigma$ itself is not needed for many purposes.

$\delta(1 - \sigma)$ et al. coefficients are inferred. Distance elasticities $\delta(1 - \sigma) \approx -1$ typically.

Border effects $-3 \leq (1 - \sigma) \ln b \leq -1$ typically.

$1 - \sigma$ can be inferred if some trade cost element of $D_{ij}$ is directly observed. Typically $10 > \sigma > 6$.

Gravity fits well with either aggregate or disaggregated trade flow data. But estimates on aggregated data (i) bias downward the average effect of trade costs and (ii) mask large variation across sectors in the effect of distance, for example, on trade flows.

Gravity fits well to explain services trade flows. See for example Anderson, Milot and Yotov (2013).
Structural Gravity Implications

Use the estimated coefficients to form the ratio of predicted (indicated by a tilde) to predicted frictionless trade:

\[
\frac{\tilde{X}_{ij}}{Y_i E_j / Y} = \left( \frac{\tilde{D}_{ij}}{\tilde{\Pi}_i \tilde{P}_j} \right)^{1-\sigma}
\]

(7)

A particularly interesting instance of (7) is for internal trade, \( i = j \). This is Constructed Home Bias, the predicted excess (relative to frictionless) amount of local trade.

\[
CHB_i = \left( \frac{\tilde{D}_{ii}}{\tilde{\Pi}_i} \right)^{1-\sigma}
\]

(8)

Home Bias

Most all countries’ international trade is far less than predicted by frictionless benchmark. For example, US has 25% of world GDP, frictionless export share should be 75% vs. actual share \( \approx 10\% \).

Openness to trade is associated with gains from trade (to be demonstrated subsequently). CHB is a measure of foregone potential gains.

Conversion of CHB to a welfare measure requires estimate of trade elasticity \( 1 - \sigma \), emphasized in Arkolakis, Costinot & Rodriguez-Clare (AER, 2012).

CHB in Manufacturing


The results below add another major implication about the relationship of size and trade: While bigger producers trade less in the frictionless benchmark, they tend to trade more (a lot more) relative to the frictionless benchmark.

- CHB is very large
- CHB is (much) lower for bigger producers
- the lower CHB is due almost entirely to lower \( \Pi \), interpretable as sellers’ incidence of trade costs.
- falling (rising) CHB over time is due to falling (rising) \( \Pi \), associated with increasing (decreasing) sales shares \( Y_i / Y \).
- lower \( \Pi \) is equivalent to productivity improvement
Table 1: Constructed Home Bias Indexes by Country

<table>
<thead>
<tr>
<th>Countries with Lowest CHB</th>
<th>Countries with Highest CHB</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISO 1996 %ΔCHB</td>
<td>ISO 1996 %ΔCHB</td>
</tr>
<tr>
<td>USA 3 -11</td>
<td>MUS 658 -37</td>
</tr>
<tr>
<td>JPN 5 41</td>
<td>LVA 679 -43</td>
</tr>
<tr>
<td>DEU 8 0</td>
<td>CRI 694 -33</td>
</tr>
<tr>
<td>FRA 11 2</td>
<td>EST 704 -54</td>
</tr>
<tr>
<td>CHN 12 -51</td>
<td>AZE 737 154</td>
</tr>
<tr>
<td>ITA 13 -14</td>
<td>BOL 778 -5</td>
</tr>
<tr>
<td>GBR 14 15</td>
<td>JOR 866 -28</td>
</tr>
<tr>
<td>CAN 19 0</td>
<td>PAN 872 -15</td>
</tr>
<tr>
<td>HKG 20 -22</td>
<td>OMN 948 -49</td>
</tr>
<tr>
<td>KOR 20 -28</td>
<td>SLV 1186 -54</td>
</tr>
<tr>
<td>ESP 25 -2</td>
<td>MDA 1224 24</td>
</tr>
<tr>
<td>BRA 32 56</td>
<td>TZA 1254 -28</td>
</tr>
<tr>
<td>MYS 36 -52</td>
<td>MNG 1328 139</td>
</tr>
<tr>
<td>BLX 37 30</td>
<td>SEN 1336 -5</td>
</tr>
<tr>
<td>NLD 39 24</td>
<td>TTO 1577 -58</td>
</tr>
<tr>
<td>RUS 39 33</td>
<td>ARM 2423 -39</td>
</tr>
<tr>
<td>AUT 41 13</td>
<td>KGZ 2554 276</td>
</tr>
<tr>
<td>IND 44 -13</td>
<td>MOZ 2630 -44</td>
</tr>
</tbody>
</table>

Size and CHB in Services Trade
\[ \ln CHB_i \text{ vs. } \frac{Y_i}{Y} \text{, Canada’s provinces (Anderson, Milot and Yotov, 2011).} \]

**Size and Sellers’ Incidence**
Multilateral resistance is interpreted as average incidence: outward for sellers $\Pi_i$, and inward for buyers $P_j$ for each $i$ and $j$.

Why does $\Pi_i$ get smaller as size gets larger?

- A big country has less of its total shipments forced into crossing borders (as in the frictionless model), so it incurs less trade cost on average on its shipments $\Rightarrow$ lower sellers’ incidence of trade costs.

- This is a tendency only, not a one-to-one relationship. Statistically close relationship in the sense of high negative correlation.

An important force acting as if economies of scale, even though no economies of scale in model.

**Interpreting the Size-Incidence Effect**

The base for interpreting results comes from looking into the structure of sellers and buyers incidence. There are $N$ countries.

Multiply (4) through by $E_j$, sum over $j$ and use (2) to cancel $Y_i$ from both sides of the resulting equation:

$$\Pi_i^{1-\sigma} = \sum_j \left( \frac{D_{ij}}{P_j} \right)^{1-\sigma} \frac{E_j}{Y_j}, i = 1, ..., N.$$  \hspace{1cm} (9)

Next, multiply (4) through by $E_j$, sum over $i$ and use the budget constraint $E_j = \sum_i X_{ij}$ to cancel $E_j$ from both sides:

$$P_j^{1-\sigma} = \sum_i \left( \frac{D_{ij}}{\Pi_i} \right)^{1-\sigma} \frac{Y_i}{Y_j}, j = 1, ..., N.$$  \hspace{1cm} (10)

**Interpreting Results**

Notice that in (9), the larger is $E_i/Y$, all else equal, the smaller will be $\Pi_i$. This follows because:

- For local sales, no border crossing cost is paid so $D_{ii} < D_{ij}, i \neq j$.

- The bigger country sells more of its output at home naturally (as in the frictionless world), hence less of it must incur the border crossing cost than for a smaller country, all else equal.

- Formally, $(D_{ii}/P_i)^{1-\sigma}$, which is bigger than $(D_{ij}/P_j)^{1-\sigma}$ for $i \neq j$ gets a bigger weight for bigger countries.

- Thus the right hand side of (9) is bigger for bigger countries.

- The larger right hand side of (9) must, for equality to be preserved, imply a larger $\Pi_i^{1-\sigma}$ on the left hand side $\Rightarrow$ smaller $\Pi_i$. 
All else is not equal, but this simple insight turns out to predominate.

The full details of how size relates to the $\Pi_i$s and $P_j$s comes from solving the system of $2N$ equations (9)-(10) with given $D_{ij}$s, $E_j$s and $Y_i$s.

The $P_j$s turn out not to vary nearly as much as the $\Pi_i$s over countries and across time. Just why this is so is a complex matter still not fully understood. See Anderson and Yotov (2010) for more discussion. Another way to look at size-incidence effect is that smaller producers tend to be (greatly) disadvantaged by the global incidence of resistance to trade. It is important to note that this disadvantage operates independently of any economies of scale, though indeed economies of scale may add further to the disadvantage. Another implication is that inferences of disadvantage due to economies of scale may falsely be drawn from observing the effects of what is actually differential incidence of resistance to trade.

**Size and Trade Again**

Summarizing the key insights from the gravity literature:

1. Any country naturally trades more with bigger partners.
2. Smaller countries are more naturally open.
3. Faster growing country pairs have increasing share of world trade
4. Given natural openness, bigger countries tend to trade much more because they tend to have much lower Constructed Home Bias.
5. Bigger countries lower CHB due mostly to lower outward multilateral resistance = sellers’ incidence.

**Projections**

A valuable use of the empirical gravity model is to project missing data.

- Use estimated gravity coefficients (estimated using other data and (7)) to calculate

$$\frac{Y_iE_j}{Y} \left(\frac{d_{ij}^p \delta_{ij}}{\Pi cP_j}\right)^{1-\sigma} = \tilde{X}_{ij}.$$ 

- Can be used to check or replace bad or suspicious data (e.g. smuggling)
- Used to forecast effects of big changes — e.g., fall of Iron Curtain on E. European countries.
- Can be used to forecast effects of building a canal, bridge, etc.
5 Measuring Globalization and Other Open Questions

5.1 Missing globalization

Estimation of gravity coefficients on various datasets yields no recent evidence of decreasing coefficients on distance (or other frictions such as borders). This appears puzzling — the Mystery of the Missing Globalization. One answer to the puzzle is that if all trade costs (distance) shrink(s) uniformly, all relative trade costs (distances) $t_{ij}/t_{kl}$ ($d_{ij}/d_{kl}$) remain the same and so would all relative trade $X_{ij}/X_{kl}$. Thus no decrease in the distance coefficient would be found in estimated gravity equations across time, even though the world was ‘getting smaller’. (This point can be formally demonstrated by analyzing the effect of a uniform proportional change in all $D$s on the system of equations (9)-(10). Multiply all $D$s by factor $\tau$ and all initial $P$s and $\Pi$s by $\tau^{1/2}$. All equations continue to hold.)

The uniformity explanation does seem roughly plausible: better shipping and communications stimulate short distance trade within countries as well as long distance trade between countries.

Recent data on changes over time in trade costs:

- sea freight rates have risen (container rates) but quality is much better (containerization). See M. Levinson The Box.
- Some freight has shifted from surface to air; much more expensive but presumably time saved is worth the expense.
- Observed willingness-to-pay difference is lower bound estimate of quality improvement. Hard to sort out further effects of quality change, but may be large.
- Tariffs fell most among big countries in the 50s through 70s, so not too much action there.
- end of Multi-Fibre Arrangement is much more significant for textiles and apparel.
- some regulatory agreements have fostered trade, especially in services.
- Free Trade Agreements seem to foster trade much more than the implied tariff cuts. The inference is that the FTA fosters investment by firms in foreign marketing and commercial networks, etc.

World trade $T$ is rising (except for the 2008 sharp recession) relative to value of world shipments $Y$, good by good and overall. What is the explanation, if gravity coefficients are constant?

- measurement of constant gravity coefficients may be wrong.
- production/expenditure patterns shifting may induce trade-increasing changes. Specialization may explain the changing location of production.
Work is proceeding on both points. On the first bullet point, better methods may turn up movement of coefficients. On the second bullet point, a starting point is the following decomposition:

\[
\frac{T}{Y} = \sum_{i,j: i \neq j} \tilde{X}_{ij} / Y = 1 - \sum_i \tilde{X}_{ii} / Y
\]

\[
\tilde{X}_{ii} = CHB_i E_i / Y
\]

using the definition (8), so

\[
\frac{T}{Y} = 1 - \sum_i CHB_i s_i b_i,
\]

Here \(b_i = E_i / Y\), the expenditure share of country \(i\). (i) if CHB falls for big (in sales and purchases) countries, then \(T/Y\) falls, even if CHB rises for small countries. (ii) \(s_i\) and \(b_i\) tend to track closely across countries. Globalization means they tend to be less closely associated.\(^1\) If \(s_i\) and \(b_i\) diverge more across countries, their correlation is lower and this raises \(1 - \sum_i CHB_i s_i b_i\), tending to raise \(T/Y\) all else equal.

5.2 What Are “Trade Costs”?

The measured trade costs inferred from gravity have a number of potential explanations:

- information costs
- non-monetary barriers — regulation, licensing,...
- taste differences
- extortion, insecure contracts

See Anderson & van Wincoop (2004) for a survey that attempts to quantify some of these components and relate them to the costs inferred from gravity.

5.3 Linking Gravity to Other Trade Models

Gravity can be nested inside the large family of general equilibrium trade models under assumptions that preserve modular structure. It is understood as a model describing the distribution of goods (or service, or people or foreign investment) within a sector for given total sales and total expenditure at each location. The assumptions permitting modularity are technical, but prominently include a simple restriction on the resource use in distribution itself: resources are used in the same proportion in both production and distribution. This is called the iceberg melting assumption: when goods leave the factory, it is as if some melts away en route to the buyer: of \(y\) units shipped, only \(y/t\) arrive at the destination.

---

\(^1\)For aggregate trade, \(b_i - s_i > 0\) implies international borrowing. For sectoral goods trade \(b_i - s_i \neq 0\) implies net trade, associated with bigger international trade.
Since the resources lost in distribution must be paid for, the seller receives $pty$ from the buyer, of which $py$ pays for the cost of production ($p$ is the “sellers’ price”) and $p(t - 1)y$ pays the distribution cost. We can equivalently think that the buyer pays the seller $py$ to obtain at destination the number $y/t = x$ of goods to consume due to iceberg melting. His full price per unit consumed is $\pi = pt$ where the buyers’ full price is solved from the following logic: $py = \pi x = \pi y/t \Rightarrow \pi = pt$. The iceberg melting metaphor is extremely useful for understanding and the simplification brings tremendous economy of thought to describing economic interaction.

It is possible to relax slightly the iceberg melting assumption and still preserve modularity. But more general (and realistic) distribution cost models require sacrificing modularity: all the allocation between sectors must be analyzed simultaneously with the distribution of goods and services within sectors.

How misleading the modular structure is, particularly the iceberg melting assumption, is an important open question.

6 Note on Feenstra-Taylor’s Gravity Treatment

Feenstra and Taylor treat gravity as a manifestation of monopolistic competition only in Chapter 6. Instead, it should be understood as nesting inside many possible models (such as the Ricardian, Specific Factors and Heckscher-Ohlin models of Chapters 2-4) that describe the allocation of resources across sectors and of the determination of “who produces what” within sectors. Gravity at the empirical level is really about measuring “trade costs” where the quotation marks emphasize that the measures may include differences in tastes or technology that effectively impede trade relative to a homogeneous smooth world. Gravity is also about the general equilibrium problem of allocating goods (from sellers’ perspective) or purchases (from buyers’ perspective) across markets, taking the levels of total sales and total expenditure in each location as given.

7 References

8 Appendix: Technical Notes on Gravity

Newton’s physical law of gravity inspired the original gravity model of economic interaction over space. The intuition was that economic flows might plausibly vary with the masses of economic activity at origin and destination and inversely with the distance between origin and destination. Newton’s Law applied strictly predicts that the economic flow \(X_{ij}\) from origin \(i\) to destination \(j\) is

\[
X_{ij} = G \frac{Y_i E_j}{D_{ij}^2}
\]

(11)

where \(G\) is the gravitational constant, \(Y_i\) is the relevant economic activity mass at origin \(i\), \(E_j\) is the relevant economic activity mass at destination \(j\) and \(D_{ij}\) is the distance between \(i\) and \(j\). Matching the predicted value on the right hand side of equation (11) to observed economic flows on the left hand side, a first step replaces the gravitational constant with a constant relevant to the flows being studied. Even with a more relevant constant the prediction does not fit the data well, suggesting that Newton’s value of 2 for the exponent of distance, based on physical principles, should be replaced by a value appropriate to the data and the exponents equal to 1 for the mass variables should be replaced by exponents that improve the fit of the prediction to the data.

Application of (11) with exponents and the constant term altered to fit economic data yields \(\hat{X}_{ij} = a Y_i^b E_j^c D_{ij}^\delta\) where \(\hat{X}\) is the fitted value of \(X\) and parameters \(a, b, c, \delta\) are estimated by best fit methods. The first application was to migration flows within the UK (Ravenstein, 1889) and the first application to trade flows was by Tinbergen (1962). The original form of the gravity model gives a close fit to observed flows: a scatter plot of \(\hat{X}\) on observed \(X\) shows most points clustered close to a 45 degree line, much closer than most estimated economic relationships. Estimated \(b\) and \(c\) tend to be close to 1 and estimated \(\delta\) close to \(-1\) in many different applications. Despite this remarkable performance (most econometric relationships perform nowhere near as well), mainstream international economists were averse to a gravity model with no economic foundation.

Recent development of economic foundations for gravity modify the original form to be consistent with plausible economic structure. Structural gravity is now firmly embedded in the economic mainstream. An explosion of empirical research is the result, surveyed in Head and Mayer (2014). Structural gravity embedded in models of resource allocation across economic sectors has improved quantification of the consequences of globalization for trade patterns, the location of economic activity and the development of economies over time. The structural gravity model of economic interaction is useful due to a remarkably simple characterization of the distribution of economic activity across many origin and destination pairs, allowing a tractable model of the global interaction of many relatively large regions. The exposition below focuses on goods trade, but structural gravity models of migration and foreign direct investment are essentially the same. Other gravity models of economic interactions (portfolio investment, dissemination of ideas and culture) fit data well but still lack an economic foundation (Anderson, 2011).

The first bricks of the economic foundation of gravity are adding up constraints. Adding up constraints are not satisfied by the original form of gravity (11), seen as
follows. The total of sales by origin \( i \), \( Y_i \), must equal the sum of sales to each destination \( X_{ij} \). Similarly, the total expenditure by destination \( j \), \( E_j \), must equal the sum of purchases from each origin \( i \), \( X_{ij} \). Formally, \( Y_i = \sum_j X_{ij} \) and \( E_j = \sum_i X_{ij} \). Finally, world sales must equal world expenditure, \( Y = \sum_i Y_i = \sum_j E_j \). Equation (11) does not satisfy these elementary economic requirements without further restrictions. For example, remove the effect of distance in (11) by replacing its exponent 2 with the exponent 0, interpreted as moving goods in a frictionless world. Then the adding up constraints are satisfied if \( G = 1/Y \). Using this new constraint in a frictionless world (11) becomes

\[
X_{ij} = \frac{Y_i}{Y} E_j.
\] (12)

Equation (12) means that purchases in each destination \( j \) from each origin \( i \) are equal to total expenditure in \( j \), \( E_j \), times \( i \)'s global share of expenditure \( Y_i/Y \), a share that is common to all destinations \( j \) and equal to \( i \)'s global sales share. In a frictionless world with perfect arbitrage of prices, each destination \( j \) would face the same price \( P_i \) for shipments from \( i \), and this would be true for goods from every origin \( i \). Then (12) is consistent with an economic model where expenditure shares are identical across destinations that face the same set of prices. This line of reasoning points to a theory of expenditure shares as the next layer of bricks in the foundation.

Expenditure share structure predicts how shares at each destination \( j \) vary with trade frictions from each origin \( i \). A full development is needed because observed economic interactions are far from the frictionless benchmark (12). For example, if US national income is 25% of world income (approximately right) then the US should be spending about 75% of its national income on imported goods. This follows from using \( j \) as the US and \( i \) as the rest of the world, implying \( Y_{ROW}/Y = 0.75 = X_{ROW,US}/E_{US} \). The US actually spends around 15% of income on imports. Frictions can explain why international trade is so low relative to the benchmark. A useful structural economic model of the effect of frictions follows from a suitably simple specification of the expenditure share structure.

The identical shares requirement is consistent with standard models of expenditure if (i) goods are different according to place of origin, (ii) consumer preferences or technologies in the case of intermediate goods are identical across destinations, and (iii) preferences or technologies are invariant to income and size of output respectively. The first economic foundation for the gravity model of trade (Anderson, 1979) assumed Constant Elasticity of Substitution (CES) expenditure structure. Expenditure shares \( X_{ij}/E_j \) are given by

\[
\frac{X_{ij}}{E_j} = \beta_i \left( \frac{p_{ij}}{P_j} \right)^{1-\sigma}
\] (13)

where \( p_{ij}/P_j \) is the price of goods from \( i \) delivered to \( j \) relative to a price index of goods at \( j \), \( \beta_i > 0 \) is a ‘distribution’ parameter (one for goods from each origin \( i \) and \( \sum_i \beta_i = 1 \) to ensure that shares sum to 1) and \( \sigma \) is the elasticity of substitution parameter. To accord with observed behavior, \( \sigma > 1 \), meaning that a rise in the relative price of good \( i \) in destination \( j \) will reduce \( i \)'s expenditure share in \( j \). Since expenditure on each origin’s goods adds up to total expenditure, the sum over origins \( i \) of expenditure
shares (13) is equal to 1, an equation solved for the CES price index

\[ P_j = \left( \sum_i \beta_i p_i^{1-\sigma} \right)^{1/(1-\sigma)}. \]  

(14)

Trade frictions are assumed to raise the delivered price of good \( i \) in destination \( j \) by a constant ‘iceberg melting’ factor \( t_{ij} > 1 \), as if 1 unit departing the origin factory \( i \) yields \( 1/t_{ij} < 1 \) units at destination \( j \). Then the assumption of perfect arbitrage implies that \( p_{ij} = p_i t_{ij} \), destination prices are raised by exactly enough to cover ‘melting’.

The adding up condition for each seller \( i \) is \( Y_i = \sum_j X_{ij} = \sum_j \beta_i (p_i t_{ij} / P_j)^{1-\sigma} E_j \). Solve the adding up condition for \( \beta_i p_i^{1-\sigma} = Y_i / \sum_j (t_{ij} / P_j)^{1-\sigma} E_j \). Next, use the preceding equation to replace \( \beta_i p_i^{1-\sigma} \) in (13). The result is the structural gravity model (15)-(17).

\[ X_{ij} = Y_i E_j \left( \frac{t_{ij}}{\Pi_i P_j} \right)^{1-\sigma} \]  

(15)

where

\[ \Pi_i^{1-\sigma} = \sum_j \left( \frac{t_{ij}}{P_j} \right)^{1-\sigma} E_j \frac{1}{Y}. \]  

(16)

Replace \( \beta_i p_i^{1-\sigma} \) in the CES price index using (16) to yield.

\[ P_j^{1-\sigma} = \sum_i \left( \frac{t_{ij}}{\Pi_i} \right)^{1-\sigma} Y_i \frac{1}{Y}. \]  

(17)

The new variable \( \Pi_i \) is an index of the outward trade frictions facing shippers from \( i \). The price index \( P_j \) is rewritten using \( \Pi_i \) as expression (17), an index of inward trade frictions facing shipments to destination \( j \). Anderson and van Wincoop (2003) coined the term multilateral resistance for these indexes of bilateral resistance. The dependence of multilateral resistances on frictions from \( i \) and to \( j \) on all links and all masses in the world economy is implicit in the system of equations (16)-(17).

The importance of multilateral resistance is illustrated by the large role multilateral resistance plays in driving Canadian provinces to trade so much more with each other than do US states (the border puzzle posed by McCallum, 1995). Take a pair of provinces and a pair of states chosen so that origin and destination size is the same (\( Y_i^{US} / Y_i^{CA} = 1, E_j^{US} / E_j^{CA} = 1 \)) and they are the same distance apart. Then using (15) and canceling equal terms in numerator and denominator

\[ \frac{X_i^{CA}}{X_i^{US}} = \left( \frac{\Pi_i^{CA} P_j^{CA}}{\Pi_i^{US} P_j^{US}} \right)^{\sigma-1} \]

Since the Canadian economy is 10% the size of the US economy, far more of the trade of Canadian provinces must cross the border and incur border frictions than is the case for US states. This increases \( \Pi_i^{CA} \) and \( P_j^{CA} \) for province pairs in Canada relative to
Π^{US} and P_j^{US} for province pairs in the US (the solution to McCallum’s border puzzle proposed by Anderson and van Wincoop, 2003).

The bilateral trade flow in equation (15) comprises two elements. \( Y_iE_j/Y \) is the frictionless flow. \( \left( \frac{t_{ij}/\Pi_{i}}{P_{j}} \right)^{1-\sigma} \) is the systemic effect of trade frictions, the factor by which the bilateral flow differs from its hypothetical frictionless value. The bilateral friction of the original trade gravity model is replaced with a relative bilateral friction, where the denominator is the product of multilateral resistances. Intuition had suggested to some investigators that third party interactions must modify the simple bilateral form in (11) but no theory before Anderson (1979) was available to formalize this intuition. The initial borrowing from Newton offered no useful guidance since physics formula (11) is for the two body case (other masses are too distant to matter) and the knotty N-body problem yields nothing simple as a solution. In this context (15) is an elegantly simple economic theory of the equilibrium distribution of given supplies of goods from many origins to buyers at many destinations who spread given expenditures across goods from many origins.

The assumption that products are differentiated by place of origin in the gravity model is itself given an economic foundation in the modern theory of monopolistic competition where firms offer unique varieties to buyers with CES demand structure for all varieties (the love of variety model). This point is developed by Bergstrand (1989). System (15)-(17) still describes the model but the explanation of the set of \( Y_i \)'s is enriched.

Two subsequent model building blocks generate CES-type expenditure shares and thus the same structure as (15)-(17). Heterogeneity of buyers or producers accounts for why potentially all origins serve all destinations. In the buyer heterogeneity model (Anderson, de Palma and Thisse, 1992) each individual has a single preferred variety with individuals differing according to a probability distribution with dispersion characterized by a parameter playing the role of \( 1 - \sigma \). In the seller heterogeneity model (Eaton and Kortum, 2002) it is the dispersion parameter of the probability distribution of the labor productivity of sellers that plays the role of \( 1 - \sigma \). An origin-specific location parameter of the productivity distribution plays the role of \( \beta_i \). (The Eaton-Kortum model has seller heterogeneity at the level of countries, with identical atomistic competitive firms, so monopolistic competition is suppressed.) More general models still manage to approach the simplicity of (15)-(17) and extensions remain an area of active research.

Fitting the structural gravity model (15) to data is almost as straightforward as in the original gravity model. The multilateral resistance terms are commonly estimated using origin and destination fixed effects in standard regressions. [A full information alternative estimator using the full system (15)-(17) was applied by Anderson and van Wincoop (2003).] Two equivalent procedures are used. One is to divide both sides of (15) by \( Y_iE_j/Y \) and estimate. The other is to estimate (15) as it is, and to recover the multilateral resistances (up to a normalization) from the estimated fixed effects \( \chi_i \) and \( \mu_j \) using \( \Pi_{i}^{1-\sigma} = \chi_i/Y_i \) and \( P_{j}^{1-\sigma} = \mu_j/E_j \). In the absence of data on \( Y_i \) and \( E_j \) only the combined \( \chi_i \) and \( \mu_j \) can be identified. The heart of the analysis is estimating bilateral frictions \( t_{ij} \), done with loglinear functions of proxies. Proxies for friction such as bilateral distance are used along with measures for free trade agreements, common language, past colonial relationships, etc. More recent approaches have emphasized
the importance of network structure variables [Rauch and Trindade (2002), Chaney (2014)]. Estimated gravity equations fit very well and yield precisely estimated parameters.

While structural gravity is a static model with parameters identified by cross section variation of bilateral trade flows, it is also applied on panel data. On the time dimension, theory implies using origin and destination fixed effects that vary independently across time and these indeed have a lot of time variation, especially in the multilateral resistances (Anderson and Yotov, 2010). In contrast, there is little evidence for time variation in distance elasticities. Development of dynamic models of gravity is a challenging frontier of research.

A problem with the use of the model (15) is zeroes. All theoretical rationales discussed above imply that trade should be positive, even if very small. One explanation is that zeroes are due to observation error. When the flow is small, the observer may not see what ‘should’ be there. There is no doubt some truth to this explanation. It suggests econometric procedures to deal appropriately with the nature of the random error term (Santos-Silva and Tenreyro, 2006).

Two extensions of the gravity model give economic explanations for zeroes. One extension is fixed costs of serving a market (a portion of the trade iceberg breaks off and is lost before the berg begins melting on its trip to destination). A zero results when an origin is not productive enough to serve a destination. With heterogeneous firms in each sector and country, this model implies a selection effect when positive trade is observed: the most productive firms are able to pay the fixed export cost. A fall in bilateral trade costs acts on the volume of trade on the intensive margin as in (15) but also on a bilateral extensive margin through the entry of more origin firms into trade with the destination. Bilateral frictions inferred when controlling for selection appropriately are a combination of intensive and extensive margin changes. For a successful approach to estimating such models see Helpman, Melitz and Rubinstein (2008). A problem with application is the difficulty of finding proxies for fixed costs that are not also proxies for the $t_{ij}$s.

The alternative explanation for zeroes arises when the buyers at $j$ have willingness to pay for even a minimal amount (a choke price) that is less than the potential delivered price $p_{ij} = p_i t_{ij}$. Demand would be zero whether fixed costs are negligible or not. Non-CES demand structure is required for finite choke prices because CES share (13) with $\sigma > 1$ is positive for finite prices. For a (translog) example of expenditure shares with finite choke prices that yields a very tractable gravity model see Novy (2013).

The model (15)-(17) refers to goods in a single sector, which could be at any level of aggregation that is to be treated. The gravity model of distribution can be nested in any general equilibrium model of the allocation of resources between sectors in each location. Thus it is consistent with most such models developed in mainstream economics.

Gravity nested within general equilibrium models permits a number of applications measuring the effects of policy changes, actual or proposed. Head and Mayer (2014) classify such policy change analyses in order of the degree of interaction considered. Partial Trade Impact evaluates the effect of a change in a $t_{ij}$ on $X_{ij}$ holding all else equal in (15). Modular Trade Impact evaluates the effect of a change in $t_{ij}$ on all the trade flows and multilateral resistances using (15)-(17) with all $Y_i$ and $E_j$ held
constant. General Equilibrium Trade Impact includes the response of $Y_i$ and $E_j$ to changes flowing from MTI induced by the original change in $t_{ij}$. The last step requires a model to determine $Y_i$ and $E_j$ for all countries and sectors. Structural gravity is the trade distribution component of the full model.

The simplest full general equilibrium model (Anderson and van Wincoop, 2003) determines for each origin $i$ the factory gate price $p_i$ of $i$’s physical endowment $y_i$, hence $Y_i = p_i y_i$. The model is closed using $Y_i = E_i$, the balance of payments constraint of each origin $i$ with the rest of the world. Shifts in trade costs alter the factory gate prices $p_i$ of all origins $i$ and the full model is solved for the new $p_i$s and multilateral resistances. (A multi-sector endowments model is developed and applied in Anderson and Yotov, 2016.)

An attractive and simple alternative full general equilibrium model is due to Eaton and Kortum (2002). Production is Ricardian (labor is the only factor of production) and countries draw labor productivities from a Frechet probability distribution. In equilibrium each country specializes in a range of sectors and earns and spends income equal to the wage bill $w_i L_i = Y_i = E_i$. Full general equilibrium analysis includes the effect of changes in $t_{ij}$ on the equilibrium wage $w_i$, shifting $Y_i = E_i$. The shifts are the same as in the endowments model applied in Anderson and van Wincoop (2003), but the interpretation differs. The Eaton-Kortum model features a richer model of $Y_i$ in the sense that the composition of output of $i$’s good $Y_i$ is explained and all substitution is on the extensive rather than intensive margin. Costinot, Donaldson and Komunjer (2012) extend the Eaton-Kortum model to multiple sectors in contrast to the effectively one good composite of the original model. The extended model features inter-sectoral resource allocation in a particularly simple form. Equilibrium specialization occurs in a range of sub-sectors within each sector. The extended version implies that change in trade costs induces changes in the proportions of labor devoted to each sector as well as changes in the range of sub-sectors within each sector. Much recent research has used the Ricardian full general equilibrium model to explore such important policy changes as NAFTA (Caliendo and Parro, 2015).

References


