Examined Assumptions: Three Essays on International Economics

by

Seth Kingery

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Public Policy and Economics)
in The University of Michigan
2017

Doctoral Committee:

Professor Alan Deardorff, Chair
Associate Professor Javier Cravino
Associate Professor Kyle Handley
Associate Professor Sebastian Sotelo

skingery@umich.edu

OCRID iD:0000-0002-7922-2479

For my mother who taught me to speak and my father who taught me to reason.

ACKNOWLEDGEMENTS

I would like to express my deepest thanks to Alan Deardorff, to whom I owe so much. I would also like to thank Sebastian Sotelo, Kyle Handley, Kathryn Dominguez, and Javier Cravino for their advice and patience. In addition, John Leahy and Chris House have helped my work on monetary economics enormously.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

Examined Assumptions: Three Essays on International Economics

by

Seth Kingery

Chair: Alan Deardorff

Chapter 1: Radiating Trade: Creating Gravity through Spatial Geometry (Job Market Paper)

This paper introduces a novel method of modeling the role of distance in trade models. Instead of creating gravity through model-specific and other implied costs, I create a simple abstraction of agents search process that produces gravity as a latent feature of spatial geometry. I then show how this simple model can be combined with more standard models, to produce gravity-fitting models more consistent with observed trade costs. Then I find a testable implication of this abstracted search process deriving from the geometry of a sphere. Finally, I show that this spherically adjusted model fits country-level trade-flow data extremely well; better, even, than the standard inverse distance relationship.

Chapter 2: Trade in Elasticity: Does Goods-Specific Wealth Elasticity Explain Trade Collapse during Recessions?

This paper seeks to examine the wealth and income elasticities of traded goods to see what explanatory power they might have in the sensitivity of trade to recessions. Using data from the consumer expenditure survey (CEX) I compute goods-specific income and wealth elasticities. Then, using the trade content of these sectors derived from the Input-Output Matrix, I find that elasticity very poorly predicts trade content. I also find that, without resorting to very strong assumptions about preferences, these elasticities can only explain a very small portion of the exaggerated responses of trade flows to recession. Ultimately, the paper examines the difference between income elasticity and deferrable expenditure, discusses why the latter cannot easily be identified in the CEX, and points to why deferability may be a more promising explanation of the trade-collapse phenomenon.

Chapter 3: Asymmetric Inflation: Consumer Credit Frictions and Stable Differences in Sector-Specific Inflation

This paper scrutinizes important structural features of consumer credit provision and possible consequences this may have for price changes between sectors. Because the underwriting of consumer loans is tied to wages and regulated more intrusively, I argue that credit provision to consumers is inherently less responsive than other types of lending to economic shocks or policy changes. In a simple model one that ignores the real economic implications of monetary policy I then examine how consumer good (CPI) inflation might be disproportionately affected by consumer credit provision. The paper then works through how increasing these consumer credit frictions might, ceteris paribus, in turn require larger policy interventions in order to achieve target inflation levels. Finally, I consider alternative strategies in this setting that could generate similar target inflation rates with less dramatic policy interventions.

CHAPTER I

Gravity and Geometry

1.1 Background

Most models of trade produce a "gravity" relationship for bilateral trade between countries. That is, bilateral trade flows are proportional to the product of the GDPs of the two countries divided by distance, as is well documented in the empirical evidence. In these models, this relationship is usually produced by including a cost for transporting goods between countries. Unfortunately, the costs needed to fit these models to the data are far larger than those observed in actual transportation costs. In addition, the gravity relationship has been extremely stable over time (such as in Disdier & Head 2008), despite falling transport costs, lengthening supply chains, and a host of other changes in the real economy. This paper proposes an alternative to trade costs: that trade falls with distance because trading partners far away are less likely to be found.

Put differently, consider an economic agent standing in the middle of a space uniformly packed with potential trading partners arranged in larger and larger concentric circles. Our agent must select a trading partner from among those at a given distance. As the circles get further out, there are more agents on each circle because the larger circumference will "fit" more potential trading partners. If the probabilities of trading with all the partners on a given circle are the equal (*i.e.*, there is no *ex ante* heterogeneity between trading partners at a given distance), then a larger circle implies a lower probability of our economic agent selecting any *specific* trading partner on a given circle. Therefore, due to geometry, as our agent chooses among partners at a greater distance, the probability of selecting any one partner at that distance decreases because there are more options. And it happens that the decrease in selection probability with distance implied by this process is exactly consistent with the gravity relationship.

In order to examine this phenomenon more deeply, this paper will construct a simple model in which agents using a uniform random radial search pattern try to locate buyers. (The search process was chosen for simplicity, to the point of being somewhat abstract, but the paper shows why this is an underlying feature of any probabilistic search going outward from some starting point.) This search process, combined with a very simple model of an economy, produces gravity exactly. And because this model derives gravity from the geometry of search alone, this result is relatively insensitive to other modeling decisions, such as the structure of preferences, production, or the size of trade costs. This makes the geometric process easy to combine with existing models of preferences and trade costs, as will be demonstrated below. But more important, it implies that this geometric process *could explain why the gravity relationship has proven so stable across time and region despite changes in the global economy and trade technology.*

This insensitivity to economic changes also has substantial implications for how trade counter-factuals and welfare gains are modeled. But before developing the geometric process and examining those features of the model, it is helpful to first discuss key findings in the literature and some of the shortcomings of the current prevailing technique for achieving gravity: "implied" trade costs.

1.1.1 Gravity and Costs

The gravity relationship in trade was first noted more than 50 years ago (Tinbergen 1962). In its simplest form, it states that trade (for simplicity we will simply say exports) from county i to country j is of the form

$$X_{ij} = \zeta \frac{Y_i Y_j}{\delta_{ij}^{\epsilon}}$$

where Y_i is the GDP of country i, δ_{ij} is the distance between the two countries, and ϵ is expected to be equal to one.[1] This basic relationship has been observed across sectors, in final and intermediate goods (Miroudot et al 2009), and even for services (Kimura & Lee 2006, Walsh 2008). The equation is so powerfully present in the data that it has taken on a privileged status in the literature. Now, when trade flows deviate from gravity it is usually taken as evidence of trade distortion. But, while every popular contemporary model of trade reduces to gravity, the observation of gravity in trade ante-dated a theoretical motivation for its existence (Anderson 2010).

While the fact of the gravity relationship is well understood, the explanation for its existence is more contentious. As was pointed out by Obstfeld & Rogoff (2001) it is hard to establish a justification for gravity in a frictionless world. The friction almost universally employed to create a gravity relationship has been trade costs. The story is simple; traded goods accrue some marginal cost per unit of distance that they are transported. For modeling convenience this cost is assumed to come in the form of an iceberg cost, in which more than one unit of a good ($\tau(\delta) > 1$) must be shipped in order to provide one unit at distance (δ). A quantity of the good ($\tau - 1$) "melts" in transit, so the cost of producing that extra good represents the "trade

[1] Recent surveys on this topic include Anderson & van Wincoup (2003) and Anderson (2010). Also, note that ϵ has been deeply studied, and in recent years there have been several papers estimating this value at something less than one, but it is presented as it is here because this was the original conception, and the paper shows a strong theoretical reason this should be the case.

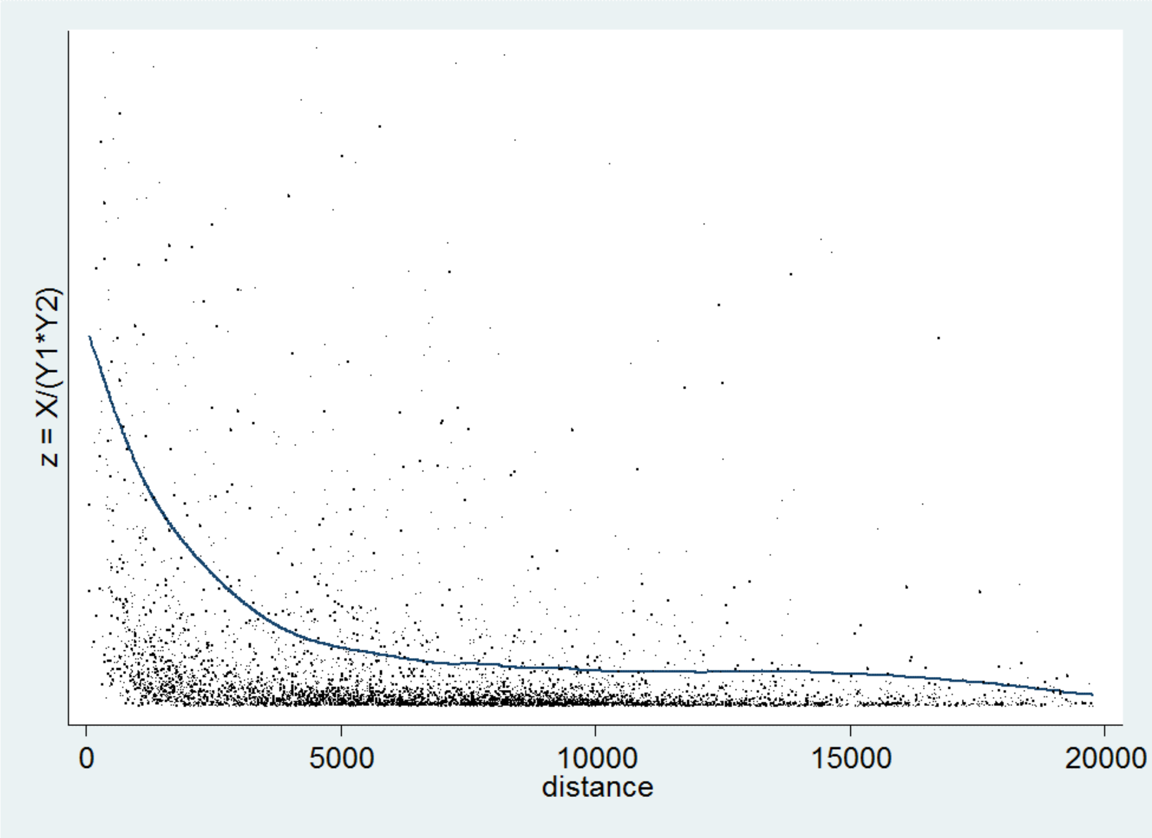


Figure 1.1: This is a local linear smoothing over the data of $\gamma_{ij} = \frac{X_{ij}}{Y_i Y_j} \approx \frac{\zeta}{\delta_{ij}^{\epsilon}}$. It is superimposed on a scatter plot of the raw data (some extreme values of which, on the left hand side below distance = 500, have been truncated). Note that the Earth is 40,075 km in circumference at its largest, so 20,000km is the approximately the maximum achievable distance.

cost". But the exact structure of how distance is transformed into trade costs often is elided over (Anderson & van Wincoop 2003 being a rare exception).

One reason could be because of the necessary constraint that faces the gravity modeling exercise. Three essential features form a trilemma: (i) the gravity relationship, (ii) the preference (or supply) structure in the model, and (iii) trade costs. These three interact such that choosing any two pins down the third.[2] For instance, Novy (2013) has elegantly demonstrated how changing the choice of utility transforms

[2] It is helpful to note how most models, even supply-driven ones, simplify to an Armington (1968) style gravity structure. For derivations of this, see Arkolakis et al (2012), Deardorff (1998), Dixon et al (2016).

the implied trade costs. This paper will not move beyond a standard CES-Armington setting, but the implication of this trillema holds. Using Anderson & van Wincoup's (2004) notation, define the price of a traded good as $\tau = \boldsymbol{\beta} * \delta^{\rho}$ where $\tau - 1$ is trade cost, δ is distance, ρ is the elasticity of cost with respect to distance, and $\tau = \boldsymbol{\beta} = (1 + \sum \beta_i)$ is some scalar that includes β_i cost effects (typically including borders, language, et cetera). An example of this trilemma is the well noted relationship in the Armington model between price growth (ρ), the trade elasticity (σ), and the gravity model distance exponent (ϵ): $\rho(1 - \sigma) = \epsilon \approx 1$.

1.1.2 Observed Trade Costs are Small and Linear

Before proceeding, it is important to pause here and discuss two things. First, it is helpful to note that observed trade costs in the data on freight rates, et cetera, are actually quite small relative to the values estimated from gravity equations. Second they are increasing at roughly a constant marginal rate per unit of distance when one examines shipments to non-bordering countries. In contrast, for purposes of gravity modeling, trade costs are usually modeled as increasing logrithmically at a rate consistent with a $\rho \approx 0.3$, as in Hummels (1999) and Anderson & van Wincoop (2003). However, when looking at countries that are not close, the data suggest that marginal cost of distance over the ocean is nearly constant (which justifies how costs will be modeled later in this paper). To show both the structure and the small size of observed trade costs, consider the following figures. Using sector-specific data of US imports that include a detailed summary of trade costs, it is possible to measure the observable physical trade costs experienced by US importers. The data cover 13 years, 1991 to 2003, and include imports from 188 countries divided into sectors by SIC code, altogether over 3,000,000 observations. The data also include transaction values, so it is easy to produce a percentage trade cost of transporting the physical good. This ad valorem trade cost ($\tau_D - 1$) derived from the data implies a data motivated traded-good price (τ_D). Figure 2 displays these plots for several of the most well represented sector categories.

In order to first adjust for endogeneity in purchasing decisions, a semi-parametric local linear smoothing is estimated for each sector (*see* Figure 2). Because the data is from the US, there are important considerations with regard to distance. As demonstrated by Coughlin and Novy (2016), there is a notable spatial component to border effects, and being a large country between two oceans the US has few nearby trading partners. Once one looks further from the borders, the increase in marginal cost is very nearly linear. To tell a simple story, this is about the distance at which all trade is by ocean and the marginal cost of transporting goods one more kilometer on the ocean is more or less a very small fuel cost. But the most important insight from these is to note how remarkable small the costs are. At the most extreme distance, the estimated trade cost is about 10% of the value of the good. This is extremely small relative to that implied by the standard models.[3]

Finally, crude regressions are run on the data to provide rough values of the growth parameters of trade costs (*see* Table 1):

$$T_{tki} = \beta_\delta * d_i + \sum \beta_t + \varepsilon_{tki}$$

Here, k enumerates each good, i the country from which the US imports, and t the year where β_t is a fixed effect for each year. In the first column of Table 1, T_{tki} is the ad valorem trade cost (also referred to as $\tau_D - 1$), and d_i is the distance (δ) in kilometers of country i. In this setting, $\beta_\delta = 0.00000289$ represents the constant marginal cost as a percentage of a goods value for shipping it one kilometer further. In the second column of Table 1, T_{tki} is logged ad valorem trade cost, $ln(\tau_D - 1)$, and d_i is logged distance, $ln(\delta)$, implying this $\beta_\delta = 0.319$ is the growth elasticity of

[3] Note that there are many other trade costs in used in these models, such as tariffs, but those are not correlated with distance and therefore will not contribute to generating gravity.

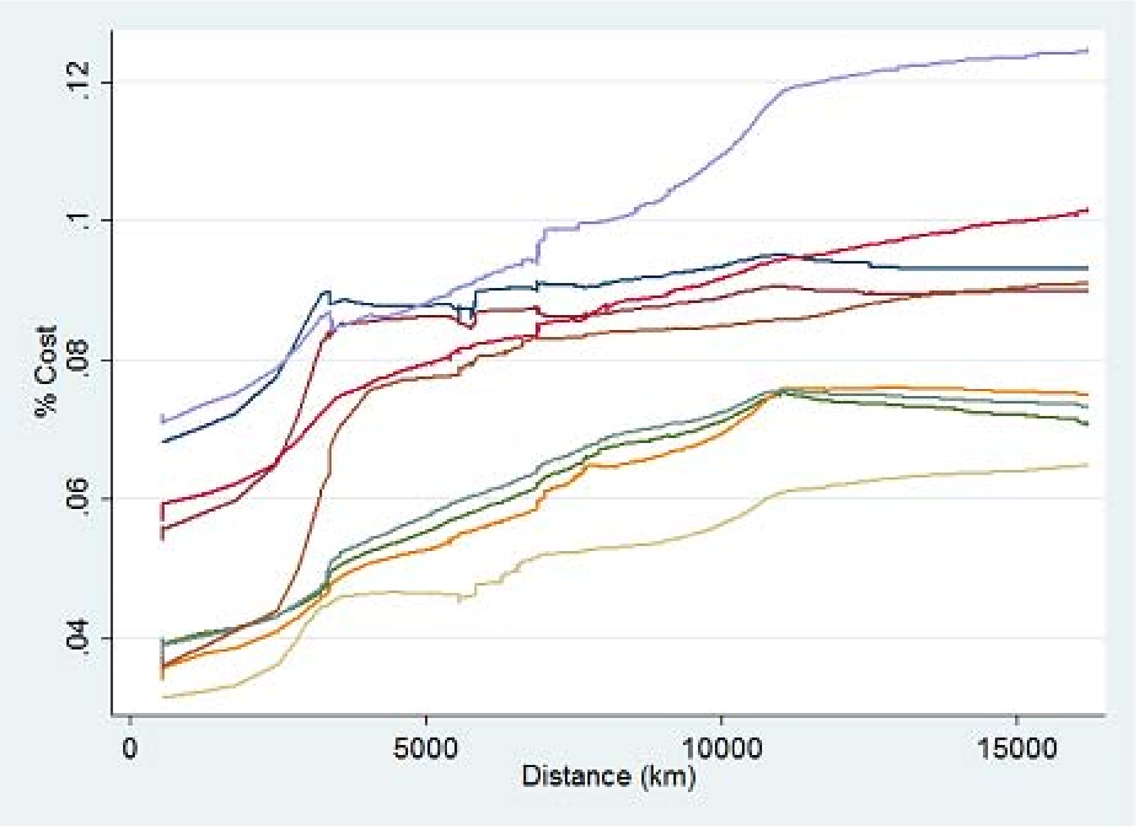


Figure 1.2: a local linear smoothing of observed trade costs in several sectors. The cost values are given as percentage of the total cost of the transported goods. Sectors represented were those most prevalent in the data, including fruit, sea food, textiles, children's toys, and auto-parts. Note that costs over short distances are harder to fit to a simple trend.

trade costs. This is consistent with what has been found in the literature (Hummels 1999, among others). Both coefficients are similarly and highly significant, though the second specification has a more favorable R^2.[4]

[4] Though not explored here, it is interesting to consider how running the model in logs changes the importance of different observations. In a variable like $\tau_D - 1$ which are usually less than one, values approaching zero are mapped to values of larger and larger absolute (in this case negative) size, so fitting observations near the origin is far more important. But observations near the origin are those least reliable in the data. So note how much the R^2 changes in the third specification when the same logged regression is run on the same logged regression *away* from the origin, $ln(\tau_D)$.

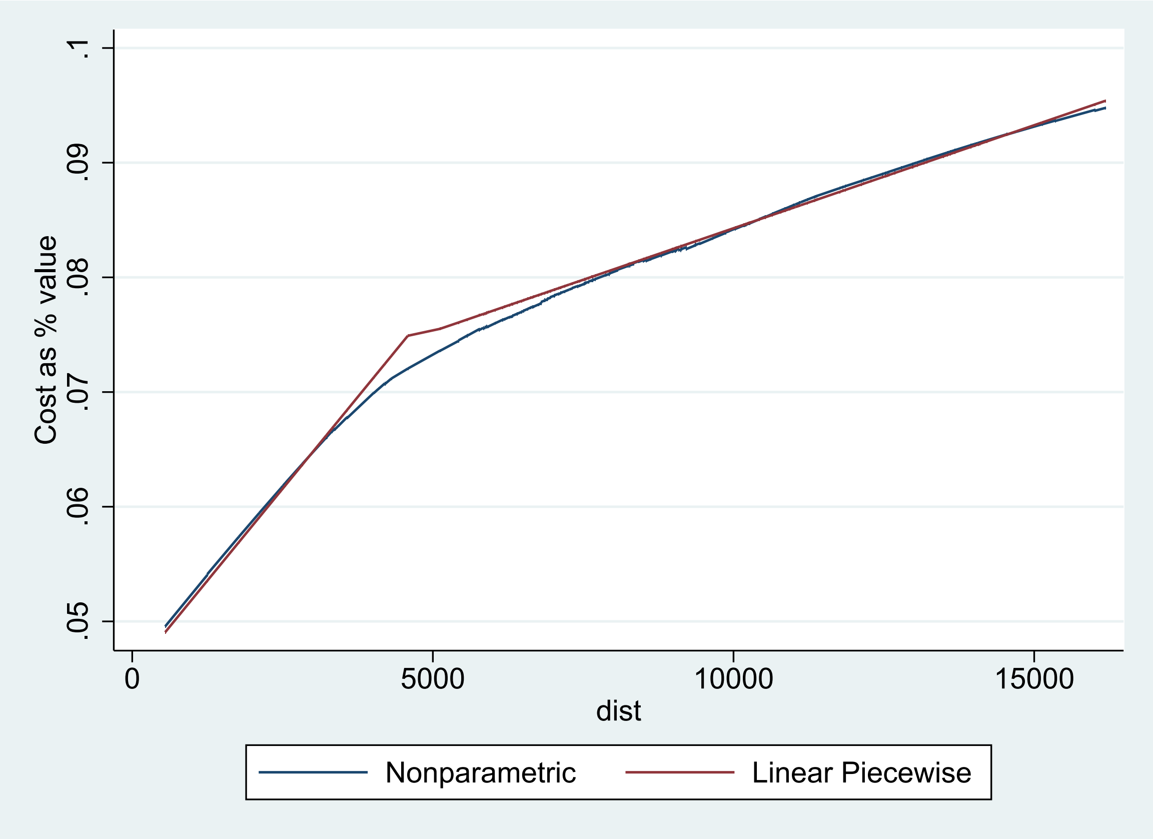


Figure 1.3: a local linear smoothing of the smoothed trends in all the sectors examined in the data. Again, the cost values are given as percentage of the total cost of the transported goods. The piecewise model imposes a "kink" at 5,000 km justified on the basis that the continental US 4,400 km wide, so only goods traveling from a foreign country to the US market over greater distances are likely to all be transported by the same method to the same port of entry. The fit is at least consistent with the notion that marginal costs for sea transport are lower and reasonably constant.

1.1.3 "Implied" Costs Grow with Distance

The gravity relationship is very well studied and easily observed and the choice of functional form for preferences or production is constrained by questions of tractability. Therefore, the structure of trade costs is usually the feature of the trilemma that has to give. So it is unsurprising that existing trade models tend to predict costs that are not consistent with observed measurements, but rather larger ones "implied" by the data. For instance Balistreri & Hillberry (2006) found that, for typical parameter estimates, this logic would imply that 50% of traded goods (or equivalent value) melt in transit.[5]

There are two ways in which these implied trade costs are troublesome. First is the fact, that multiplicative "implied" costs grow with distance in the Armington setting. Using the trends observed in the data, consider a form for trade costs $\tau = 1 + \beta_D * \delta^\rho$ where $\beta_D = 0.0044$ and $\rho = 0.319$ (from Specification 2 in Table 1). This trade cost $(\tau - 1)$ is plotted in Figure 4 with a **blue** line over the relevant distances $(0 - 20,000km)$. In most specifications of fitted models, a nontrivial "implied" cost (β_I) must be added into $\boldsymbol{\beta}$ in order to explain home-bias, among other things, because β_D is far too small to explain the effect. This discussion will use simulated values built on a small estimate of the cost leap for goods crossing any border: $\beta_I = 0.1$ (note that this value is actually smaller than typical border effect estimates).

There is no particularly intuitive story why this border effect implied cost should combine with our observed cost in a *multiplicative* way. The casual, intuitive explanation of this cost leap is that it is paid once, at the border, and should not vary beyond the border. This would suggest the implied cost is added to the observed cost, giving $\tau = 1 + \beta_D * \delta^\rho + \beta_I$ (this $\tau - 1$ is shown in Figure 1 with the **red** line). But of course, in the standard Anderson & van Wincoup setting, this value *is* included multiplicatively, meaning trade costs are $\tau = \boldsymbol{\beta} * \delta^\rho \approx 1 + (\beta_D + \beta_I) * \delta^\rho$ (this $\tau - 1$ is shown with the **green** line).[6]

Stated roughly, there is no obvious justification why the distance-growth elasticity $(\rho = 0.319)$[7] of observed costs in the data (β_D) should describe the distance-growth of implied costs (β_I). In fact, it's not clear that β_I should grow with distance at all. If we take the assumptions of the Armington model as given and the implied costs

[5] For a few papers that have estimated strikingly large trade costs, see Hummels (1999), Chen & Novy (2011), and Costinot & Clare (2013).

[6] Anderson & van Wincoup's specification is not of this form but I am, at this stage, using a different ρ as will be discussed below. The graphed equation motivates the same issue in the modeling while being easier to look at.

[7] This is the growth rate of costs, not τ_D as will be discussed below.

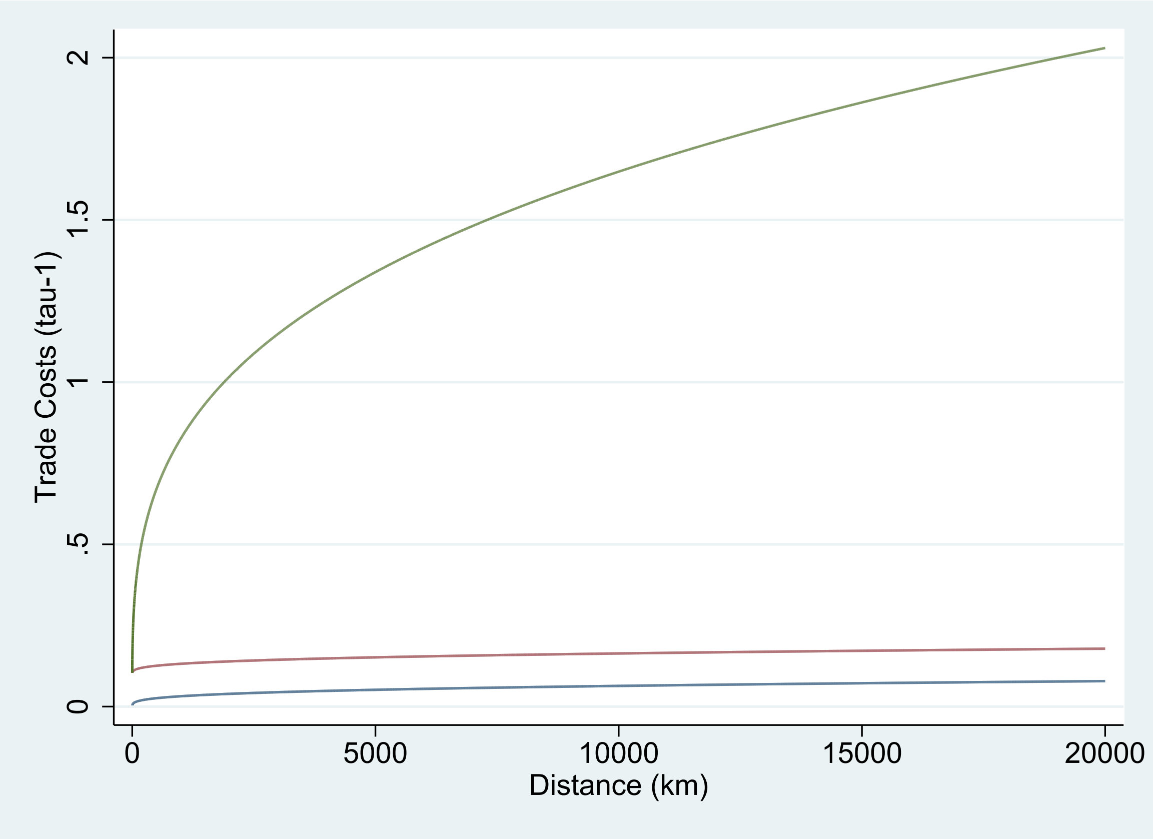


Figure 1.4: comparing multiplicative versus additive model-"implied" trade costs ($\tau-1$). The **blue** line gives the actual growth trend of shipping costs in the data. The **red** line gives the same costs with a small implied cost added as a level. The **green** line gives the same additional implied cost multiplied into $\tau - 1$. The area between the red and green lines represents model-implied costs that *increase* with distance, as typical in most Armington-related models.

as correctly estimated, the area between the red and green lines represents "implied" costs that are *growing* in distance that are *not* justified. At least not without a compelling story about why, for instance, the home-bias effect must increase with distance at the same rate as shipping costs.[8]

In order to explain these implied distance-variant costs that are not easily ob-

[8] If more intuitive, unvarying "implied" costs were instead used, there would still be tricky implications. If implied costs were included additively, then the elasticity of trade cost growth would have a different exponent (ρ'): $\tau - 1 = \beta_D * \delta^\rho + \beta_I = \boldsymbol{\beta'} * \delta^{\rho'}$. The larger implied costs β_I grow, the smaller ρ' would become. For instance, using the values of β_D, ρ, and β_I from our simulation, that would imply $\rho' = 0.09$. So even "invariant" model-generated costs would produce parameters that cannot be taken directly from cost data.

served, there have been several alternative strategies. One is to introduce information frictions like in Rauch (1998) and Allen (2014). And a search-style effect is corroborated by the observation of network effects, such as in Rauch & Trindade (2002) or Egger et al (2015). One striking result, which will be referred to again later, is that the distance coefficient falls enormously (65%) for eBay purchases, in a setting where search is not performed in physical space (Lendle et al (2015)).

1.1.4 "Implied" Costs are Very Large Compared to Observed Costs

The second way in which implied trade costs are troublesome is that they are strikingly enormous. Consider the equation relating cost growth to gravity: $\rho(1-\sigma) = \epsilon \approx 1$. This elegant relationship derives from how prices feed into the Armington CES demand function. But note that this is a statement about price, not trade costs. So when considering the parameter values derived from the data, the correct value of ρ is not that of observed trade costs ($\tau_D - 1$).[9] Instead it the relevant growth elasticity is that of total observed costs (τ_D), which is estimated in Specification 3 of Table 1 as $\rho = 0.015$. Substituting this value into the gravity-cost relation, if $\epsilon = 1$ and $\rho = 0.015$ this implies $\sigma \approx 68$, which is far outside the realm of estimated values of trade elasticity. In order to achieve even an upper-bound plausible estimate of trade elasticity from the literature, say $\sigma = 6$, ρ would need to be more than ten times larger than what is found in observed costs.

Stated in less theoretical terms, the fitted value from the data for the trade cost of shipping a good one km is 0.4% of its value. The fitted cost for shipping it 20,000 km is 8.2% of its value. That may be a 20-fold increase in cost over that distance, but it is only a 7.8% maximum increase in price due to shipping costs. The trade costs implied by almost any standard fitted model suggest maximum-distance trade

[9] There has been some ambiguity in the literature on this point. The growth elasticity of $\tau_D - 1$ is approximately $\rho = 0.3$ which, if incorrectly substituted into the cost-gravity equation is consistent with a trade elasticity of $\sigma \approx 4$ which looks tantalizingly plausible, though invalid.

costs many times the value of the good, which is an order of magnitude larger than observed shipping costs. Therefore, the trade cost story of gravity is nearly entirely dependent on unobserved costs implied by trade models. And because the costs are unobserved, it is very hard to produce a test of the validity of the framework at large, let alone the constituent parts of a given model.

1.2 Motivation

The gravity relationship in trade is simple and powerful, but any explanation of it is nearly perfectly theoretical. The intent of this paper is to produce a novel explanation and provide testable implications to validate or reject that model. A guiding inspiration for this alternate framework will be to look at gravity in another setting. The gravity model of trade is so called because it so resembles the equation describing physical gravity, which states that the force between two bodies is proportional to the product of their masses divided by the *square* of their distance (in contrast to the trade setting in which the divisor is distance to the first power). In subsequent (and unresolved) debates among physicists the inverse square portion of the equation has been noted as interesting because this is the same functional form as the rate at which a light source dims over distance. To see why light dims at this rate, simply consider the surface of a sphere ($A = 4\pi r^2$) around a light source. As the radius of the sphere increases, the same amount of light will be spread over an area increasing quadraticaly, so the amount of light at any point will *decrease* at an inverse square rate. This, however, is in three dimensional space. If we restate the problem in only two dimensions, something radiated in all directions (within a plane) from a central point would dissipate as it traveled outward on the circumference of an expanding circle ($C = 2\pi r$). So while a flashlight in 3-space would dim at an inverse square rate with distance, a flashlight in 2-space would only dim at the inverse of distance. Therefore, the simple observation at the center of this paper is that a gravity-style relationship between objects in a two-dimensional physical space can be generated by *any* model in which the "force"—or in this case trade—*radiates blindly from a source point.*[10]

1.2.1 Gravity and Distance

To provide intuition for the process to be used in the final model, let us first consider two example search processes. In the first case, consider an individual trying to pick from a finite set of objects of uniform shape (for simplicity, circles) and size (radius $\frac{\varepsilon}{2}$) from the center of a large, flat, dark room using only a laser pointer (see the left hand side Figure 5 below).[11] If the agent is blindly trying to find these objects by casting a laser beam into the dark, this can be thought of as the agent simply choosing a random azimuth (θ) for the beam to be cast in. To consider the likelihood of an object being found by this method, the probability of discovery is merely the likelihood of some azimuth being selected that shines on the object in question. If the azimuth is chosen randomly we can describe it as a uniform random variable ($\theta \sim \text{U}[0, 2\pi]$).[12] If a given object A, located at distance d_A, is not in the "shadow" of any other as seen from the origin, the likelihood of discovery is the portion of the two-dimensional "horizon" it occupies as seen by the searcher. The portion of the horizon occupied is (roughly) the circle's diameter (ε) divided by the measure of the horizon at the circle's distance ($2\pi d_A$).[13] Thus the general likelihood of discovery for

[10] It is important here to note an excellent 2016 paper by Ferdinand Rauche who discusses this insight in his recent publication "The Geometry of the Distance Coefficient in Gravity Equations in International Trade". We were unaware of each other's work until recently, but he explores this first geometric notion elegantly and it is recommended as a complimentary discussion of this foundational issue.

[11] At this point, the shape of the room is not important, but for the time being we will think of it as square, as we will later be examining how the model holds for arbitrary numbers of objects drawn randomly from a Cartesian uniform distribution.

[12] Subsequently there will be discussion of the ways in which the uniform distribution differs in polar versus Cartesian coordinates. Please note that this variable is being drawn from one dimension where no such distinction is necessary.

[13] It should be noted that the portion of the horizon occupied by the circle is slightly less than the diameter. But as distance increases and object size decreases this difference approaches zero. Since these are the exact circumstances for which the model will be considered in going forward, the issue

an "unobstructed" object in a single search is noted to be $\frac{1}{2\pi}\frac{\varepsilon}{d}$. This only considers the case that the object is unobstructed, but the size of the objects was not specified at the outset, or important to the functional form of the relative likelihood of finding objects. So if the objects were distributed in a uniform random fashion, there will exist some ε sufficiently small so as to make all objects unobstructed. To clarify, consider the case in which N objects are selected from a random uniform distribution in Cartesian space inside the room ($f_C(x, y) = 1/(2D)^2$) where D is the distance from the center to the edge of the room.[14] Because the probability of any two points falling on exactly the same ray from the middle of the room is zero in continuous space, there must exist *some* choice of ε so that all the objects are fully visible. In this setting, the probability of finding any given object continues to take the same form as above, but no assumption about the arrangement of the objects in the room has been necessary. This insight will be used later when we establish an analogous search process for searching over a continuum of points.

The second, ultimately equivalent example will be to consider the same unfortunate searching in the dark, but now they are lucky enough to be searching with a flashlight instead of a laser pointer (see the right hand side Figure 5). The searcher will again choose a random azimuth to shine the flashlight ($\theta \sim$ U[0,2π]), and for every possible azimuth choice the light cast in this direction has a fixed aperture (θ_F). Now it is possible to illuminate multiple objects at once, so the individual will select between the set of illuminated objects in proportion to how brightly they are illuminated. Consider a choice of azimuth so that only objects B and C located at distances d_B and d_C are fully illuminated at the same time. The brightness of the objects, as seen by the searcher, will be a function of the fraction of the flashlight beam that is striking them. This is a function of how much of the two-dimensional

is ignored for simplicity.

[14] Specifically $(x, y) \sim U[(-D, -D), (D, D)]$.

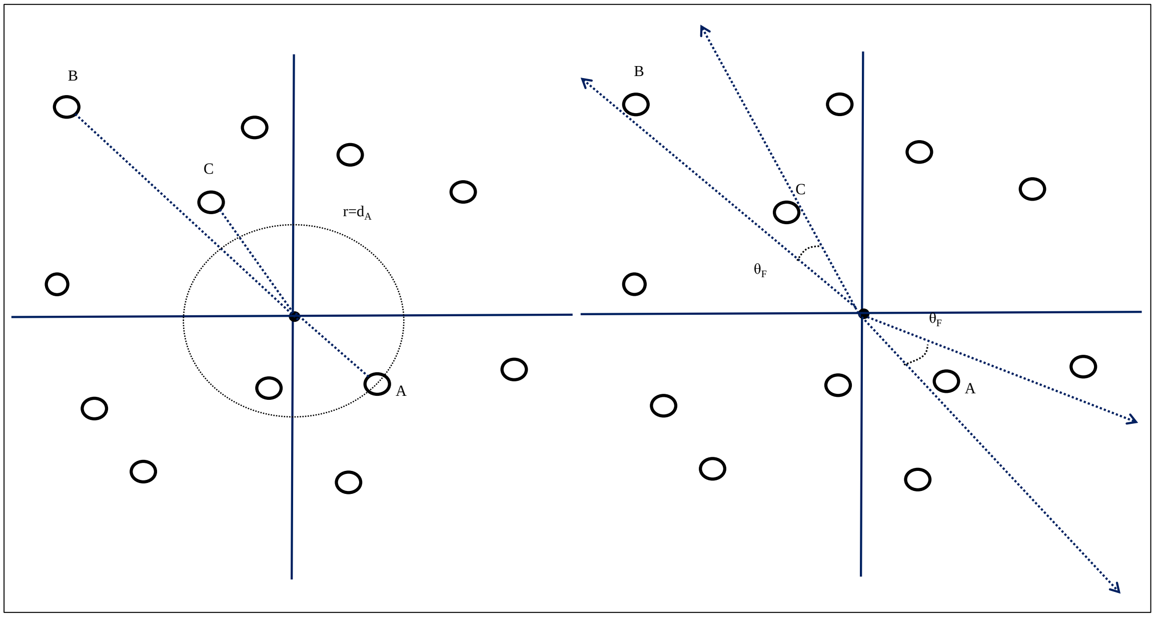


Figure 1.5: On the left hand side is illustrated the "laser pointer" version of the model in which an agent searches for objects in a dark room by firing it along random azimuths. The two dimensional "horizon" that A sits upon is illustrated. The right hand side illustrates the "flashlight" version of this same problem.

"horizon" they occupy $(\frac{1}{2\pi}\frac{\varepsilon}{d})$.[15] This measure will be referred to as an object's "arcwidth" ($A_x = \frac{1}{2\pi}\frac{\varepsilon_x}{d_x}$) going forward. Therefore, conditional on the two objects being illuminated, the selection probability of selecting a given object is simply the ratio of that object's arcwidth and the total arcwidth of the illuminated objects, in this case $p^I_{BC} = \frac{A_B}{A_B + A_C}$. (Empty portions of the searchers horizon are ignored, because the searcher is not seeking them.) An object's likelihood of being fully illuminated by the searcher's choice of azimuth is a function of its size and the aperture of the flashlight given by $\frac{\theta_F - A_B}{2\pi} = p^\theta_B$.[16] Generalizing these processes to an arbitrary number of objects, we can increase the number (N) of randomly distributed objects as in the previous example. In this case it is possible to show that the probability of selecting

[15] To clarify, this search is taking place in two dimensional space, so all the light "rays" exist only in the plane. As mentioned above this is not how a light source actually dissipates in three-space.

[16] In this example, the case of "partial" illuminations will be ignored, because as the number of objects increases and the size of the objects decreases this likelihood will approach zero.

random object x will converge towards $p_x = \frac{1}{2\pi g(N)} \frac{\varepsilon_x}{2\pi d_x}$ for some function $g(N)$. The key result is that the functional form follows gravity.

This second model is much more complex than the first, but it allows us to visualize how "rays" or "beams" of search expand and dissipate over distance. This will be useful when we extend these results to the surface of a sphere. But first we will take the intuition of these simple models, and extend it to searches over all the points of continuous space ($\mathbb{R}^2$) rather than discrete objects.

1.2.2 Gravity and GDP

The toy models of search only explain the denominator of the gravity equation. Just as important is explaining the relationship in the numerator: why should trade be driven by a product of GDPs? In a Sept. 1, 2015 column, Paul Krugman was discussed his general thoughts on the gravity equation. In describing his own intuition he said:

> Think about two cities with the same per capita GDP. They will trade if residents of city A find things being sold by residents of city B that they want, and vice versa.
>
> So whats the probability that an A resident will find a B resident with something he or she wants? Applying what one of my old teachers used to call the principle of insignificant reason, a good first guess would be that this probability is proportional to the number of potential sellers Bs population.
>
> And how many such desirous buyers will there be? Again applying insignificant reason, a good guess is that its proportional to the number of potential buyers As population.
>
> So other things equal we would expect exports from B to A to be proportional to the product of their populations.

This is, in fact, nearly identical to the rationale that will be used in this paper. The basic assumption will be that the number of buyers and sellers, populations, GDP, *and* area of a country will first-order be the same.

1.3 The Planar Model

1.3.1 Model Economy

We hope to construct a model that draws attention to the general insight that the gravity equation can be a characteristic of physical space perceived in radial coordinates, rather than something constructed ad hoc for the search model in this paper. In order to do this, all of the choices (production, utility, et cetera) going forward in this model will be out of a desire for simplicity. More complicated variations will be alluded to, but this paper attempts to prove the concept in only the most essential form.[17]

In this model, each point in the two-space continuum is an agent (x, y). These sellers search for buyers by making randomized search in polar coordinates. (It is not critical to the results if buyers search for sellers or the reverse, as long as only one side of the transaction searches. For the purposes of this paper, searchers are sellers.) Each point on the continuum is both a buyer and a seller, and each has an identical endowment of one unit of their good type, unique to each point on the continuum. The utility for each consumer is also extremely simple:

$$\mathbb{U}_{(x,y)} \left\{ c^{(x,y)}_{(x_i,y_j)} \right\} = max \left\{ c^{(x,y)}_{(x_i,y_j)} \right\}$$

where $c^{(x,y)}_{(x_i,y_j)}$ is the consumption by agent (x, y) of the good produced by agent (x_i, y_j)[18] and agents cannot consume their own good type. Transport costs are nil, and in equilibrium all points should be paired; therefore the model achieves equilibrium when all sellers part with their good at the identical global price (because all goods are perfect substitutes and there are no trade costs) and use their revenue to

[17] I intend to examine more complex variations in subsequent work.

[18] The utility and market clearing can be made much more complicated and the model will still hold. But such additions are not necessary for the key results here and will be examined in subsequent work.

buy one unit of the good sold by the seller who paired with them.

The search process itself is similarly simple. The seller searches in polar coordinates, and their pairing is created probabilistically based on an independently uniformly distributed azimuth ($\theta \sim$ U[0,2π]) and distance ($\delta \sim$U[0,D]). In other words, and in keeping with the notion of "radiating" trade, the agent picks a direction and distance at random. If the target agent is already paired then the seller searches again. It is important to note that all searchers search over identical spaces, and that the draws of azimuth and distance are not serially correlated.[19] Therefore the points removed from the search space on each iteration do not alter the shape or descriptive statistics of the probability space for subsequent iterations, and the search can be repeated however many times is needed to provide all sellers with a match.[20]

This process produces a network of sales between agents across the continuum. The structure of these pairings, as the search space is (in probability) identical for all agents from their location, is distributed the same for each agent. To intuit the pairing

[19] The timing of the search process can produce an interesting math problem, but one that is also ultimately uninformative. A simple variation to consider in lieu of simultaneous search is to have each agent make a separate random uniform draw at the start of each round of search to determine if they will be "active" (search as a seller only) or "passive" (wait to be found as a buyer only). Because search efficiency is not relevant to this paper, the extra stage in the search process is excluded. However, it is worthwhile to note how this could be useful for more complex utilities, sales, and market clearing.

[20] It should be noted here that, while it is obvious that this process should be 1-to-1 (all sellers find buyers), in an uncountable space it is not as clear why the process should be onto (all buyers are paired with sellers). In fact, it is true that an infinitesimally small (Lebesgue measure zero) set of buyers will go unpaired because they are sought stochastically. If we desired we could exclude these agents' sales from the model because they now lack the purchase necessary for market clearing, then their buyers' sales could be excluded and so on. But we can ignore this for three reasons: (i) after some number of "chasing market failure" iterations the set will close and a dense coverage of points with the desired characteristics would exist, (ii) the model could, with minor alteration, be infinitely repeated so that agents could "save" money across periods using fiat money, or (iii) if market clearing was defined in terms of integration over ε-neighborhoods the missing buyers would become irrelevant. In all three cases we will end up with the same structure for the pairing function and relevant characteristics to the equilibrium while imposing stronger assumptions and increasing mathematical complexity. Therefore, we will simply ignore the "dust" of unpaired buyers. Also note that while the distance variable has a finite bound, but the search plane as we have articulated it at this point does not. Therefore, in the planar case agents do not search over the entire space. This will change in the spherical case. In the meantime, note that even if every agent does not search the full space, every point in space is searched by the same number of agents. This means the assumptions about all points having equal likelihood of selection on each iteration hold.

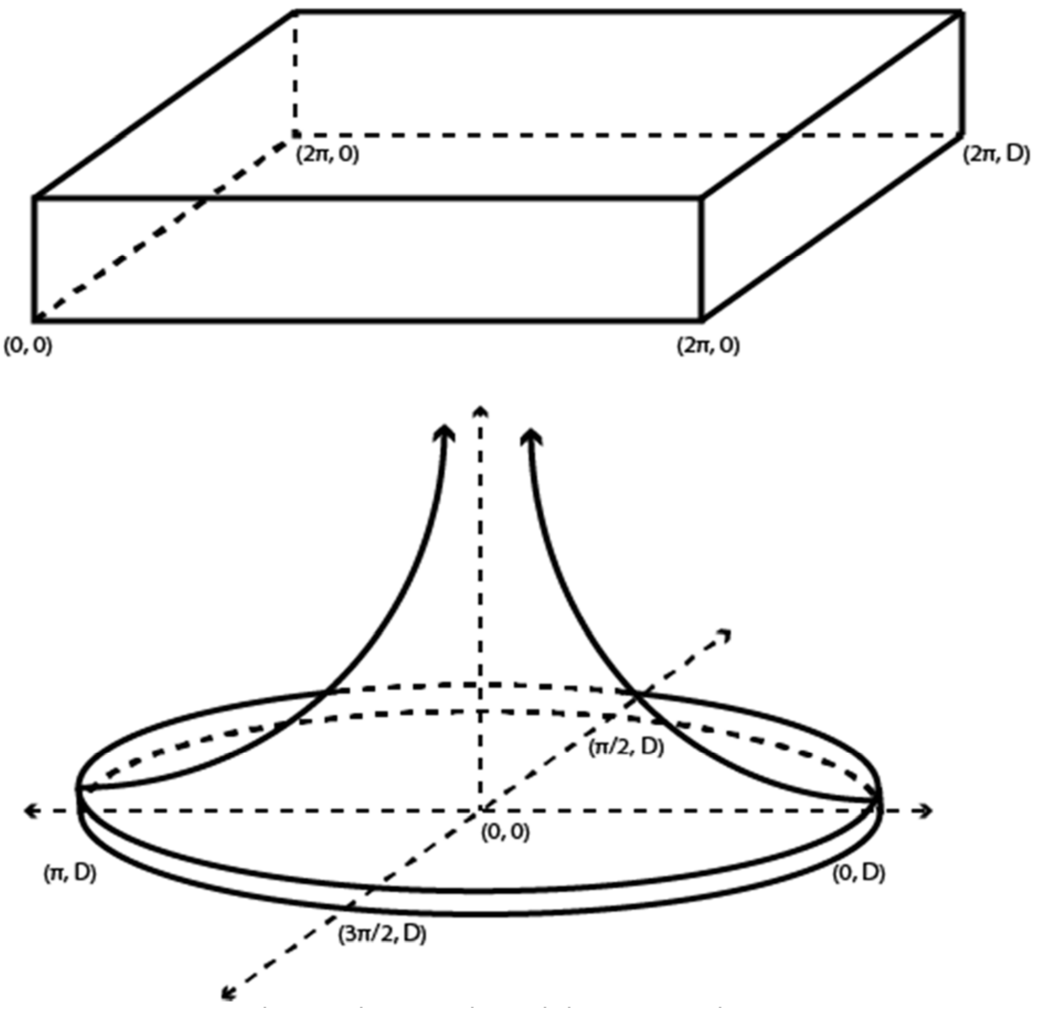


Figure 1.6: the visualization of the transformation of the pdf of a Cartesian uniform distribution into polar coordinates. Note that the z-axis is not to the same scale in both graphs.

probability that this implies, consider the visualization of a uniform two variable pdf in Cartesian space: a rectangular prism of uniform height (see Figure 6). When translated into polar coordinates, this prism is transformed by compressing all the points on the edge corresponding to distance zero into a single point at the origin. This produces a mass point, and an instantaneous pdf value of infinity at the origin.[21]

[21] While this is not intuitive, it does not change the well-defined nature of the transformed pdf for the same reason that the integral of $1/x$ is $ln(x)$. Also, none of the results in this paper examine the case that distance equals zero, because agents do not pair with themselves.

Similarly, the opposite edge of the prism corresponding to a distance of D is stretched around the origin reducing its height. From this simple picture, it is easy to see that a uniform Cartesian distribution is *not* uniform in polar coordinates.[22]

More formally, consider a random variable for polar coordinates ($R = (\delta, \theta)$) that is distributed random uniform when plotted in Cartesian space ($f_P(R) = \frac{1}{2\pi D}$, which is the pre-transformation rectangular prism in Figure 6). Now we consider what this variable would look like seen in polar space. But, since in basic definitions and concepts are not preserved in polar coordinates[23], we will instead plot this random variable in polar space and transform it into Cartesian coordinates ($X = (x, y)$) with a new pdf ($f_C(X)$). To do this, we need only make a standard transformation of random variable (the general form for this is $f_C(X) = \; f_P(R) \; |J_{g^{-1}}|$). Let $g : R \to X$ be the function that changes polar coordinates into Cartesian ones.[24] In order to perform a change of random variable, we will then need to consider $g^{-1} : X \to R$ where $g^{-1}(X) = (\sqrt{x^2 + y^2}, arctan(\frac{x}{y})) = R$. The determinant of the Jacobean (the matrix of partial derivatives) with respect to X is

$$|J_{g^{-1}}| = \left|\frac{\partial g^{-1}}{\partial X}\right| = \begin{vmatrix} \frac{x}{\sqrt{x^2+y^2}} & \frac{1}{1+(\frac{x}{y})^2}\left(\frac{-y}{x^2}\right) \\ \frac{y}{\sqrt{x^2+y^2}} & \frac{1}{1+(\frac{x}{y})^2}\left(\frac{1}{x}\right) \end{vmatrix} = \frac{1 + (\frac{x}{y})^2}{\sqrt{x^2+y^2}(1 + (\frac{x}{y})^2)} = \frac{1}{\sqrt{x^2+y^2}}$$

Note that this is simply $1/\delta$. This is a key observation of the paper. The fact that

[22] Probability in polar coordinates lacks several intuitive characteristics of probability on the plane, so it is necessary to define what a "uniform" distribution is in this setting. For our purposes, a distribution is "uniform" if the cumulative probability over any two regions of equal size in the domain of equal area are equal to each other. The underlying issue is that points in polar space are more "densely" packed around the origin (which is also the driving observation of this paper). So even if the function we are evaluating ($f_P(R)$) has a constant value across all coordinates, its integral over different regions of equal area could be dramatically different.

[23] Among other things, integrating with respect to both variables and integrating over area are different operations in polar coordinates. This is because in two dimensional Cartesian integration, $\int f(x,y)\; dA \;\; = \;\; \int f(x,y) dxdy$. But in polar coordinates the area term is $dA \;\; = r \; dr \; d\theta$. As a consequence, in this setting a pdf and a cdf do not have the usual mathematical relationships to one another.

[24] The transformation of a random variable does not require us to use this function, only its inverse. But the function in question is $g\,(R) = (r \; cos\,(\theta)\,, r \; sin\,(\theta)) = (X)$ which itself has a Jacobean whose determinant is simply $|J_g| = r$.

pairing likelihood falls with distance in this case is not a clever or obscure artifact of the choice of $f_P(R)$, but is in fact derived from viewing the search space in polar coordinates. So any choice of pdf for R will have to be transformed by a gravity-style expression, and a many choices of $f_P(R)$ will produce a gravity-style pairing outcome.[25] In our specific example, the transformed pdf ($f_C(X)$) is of the form

$$f_C(X) = \ f_P(R)\ \left|J_{g^{-1}}\right|\ = \frac{1}{2\pi D}\frac{1}{\sqrt{x^2+y^2}} = \frac{1}{2\pi D}\frac{1}{\delta}$$

in Cartesian space. From this, we can see that, conditional on any choice of θ, the transformed pairing probability in the plane is of the form

$$p_P(\delta)\ \ =\ \ \frac{1}{2\pi D}\frac{1}{\delta}\ for\ 0 \leq \delta \leq D$$

which holds for *any* choice of θ.

At this juncture, it is necessary to address the issue of the leading coefficient for the pairing likelihood equation. It seems appropriate that the integral of this pdf, over the full surface of the planar disk, should equal one. Unfortunately, the actual globe (and the data) are not uniformly covered with potential agents. In fact, most of the globe is in fact "empty" (at least for trade purposes) space. Therefore the exact leading coefficient of the pairing likelihood has the problem of being (i) very convoluted to compute empirically and (ii) different for every country on the planet (consider Fiji versus Austria). To deal with this issue, rather than encumber the model with false precision, the exact coefficient will not be discussed going forward. As will be seen, none of the following analysis will depend on this value (aside from noting that it is in all cases positive) and the reference (gravity through trade cost)

[25] The realm of options for $f_P(R)$ that still produce an ultimate gravity-style pairing likelihood is an interesting question I have not yet answered. While it is obvious that any $f_P(\delta,\theta)$ that is constant with respect to δ will produce a gravity style equation, it seems that there are many more functional forms that will produce a probability in distance that is "first-order" equivalent to gravity.

model makes no predictions on this point for comparison.[26] This generality will be explored in more detail when the model is taken to the data.

This pairing probability function, in the planar case, produces trading behavior among countries nearly identical to the standard gravity model. As each point in the plane is an agent that produces and consumes perfectly substitutable, fungible goods, a country would simply be some contiguous shape drawn on the plane. The GDP of that country would be the integral of output across all agents, which is simply area in this context. For simplicity, in this discussion we will treat all countries as circular. Therefore, a country can be described as some epsilon neighborhood of area z (ε_z), in which GDP is proportional to area.

Now consider the case that a country is vanishingly small (ε_A), which is reasonable considering that countries are very small as a portion of the surface of the Earth.[27] In this small country setting the pairing likelihoods from all points within the country to all other points approach being identical. This allows us to treat the entirety of such a country as being a continuum of agents located at an individual point on the globe. In this case, the expected number of a country's pairings in terms of distance becomes $f_A(\delta) = \beta_A \frac{\mathrm{A}}{\delta}$ or simply a mass of size A pairing identically and independently with probability $1/\delta$.[28] Therefore, given two small countries, ε_A and ε_B, which lie some distance apart d, the trade flow from ε_A to ε_B should be the pairing likelihood of a mass of size A pairing at distance d with a mass of size B, which is simply $f_{AB}\left(\delta\right) = Bf_A\left(\delta\right) = \beta_A \frac{\mathrm{AB}}{\delta} \; = \; \beta_A \; X_{AB}$, where X_{AB} is exports from A to B. This demonstrates that, in the context of this model, as country size approaches zero

[26] To be more precise, the way in which an agent searches over a subset of space does not affect the gravity result as long as underlying search space of δ, θ is uniform when plotted in Cartesian coordinates, or "piecewise uniform" in that it could permissibly be several different rectangular prisms instead of one.

[27] The (by far) geographically largest modern country, Russia, occupies a mere 3.2% of the planet's surface. For the 194 countries currently recognized by the United States the average country size is less than 0.15% of the planet's surface.

[28] In a plane that has no vacancies or voids (i.e., a planet with no oceans), $\beta_A = \frac{1}{2\pi D}$, but for the reasons mentioned above we will now start to transition to a more general statement of β.

and distance increases away from zero, the trade flows are described by the standard gravity equation.

1.3.2 Planar Model with Intensive Margin

Having created a simple search model that achieves gravity that is driven solely by the search process, trade costs are conspicuous in their absence. Trade costs that increase in distance do exist whatever their magnitude, and must have some impact on trade flows under any set of preferences. So the next step is to create a model that allows for trade costs to affect demand conditional on a successful search pairing. In order to do this, consumers will need to pick multiple goods and allocate their expenditure among them. Each consumer will now consume a home good (x_H) and a foreign good (x_F). A countrys income will be distributed equally among all citizens to simplify the model. This simplification is reasonable because all of the good specific distance effect is created through the demand function, whereas income effects of distance are spread across all goods. (It is important to note here that in any model that is isomorphic to Armington this is the case, so very little of the result relies on the simplification.) The price of the home good is set as the numeraire and trade costs are modeled as being linear in distance subject to a constant marginal cost (α) and a fixed cost of trade (T). So measured in home price, the price of x_F is $C_F = 1+\alpha\delta+T$.

For simplicity in this specific case we will impose CES preferences, though for reasons that will be apparent most any selection could be made. So agents maximize

$$\mathbb{U} = \left(x_F^{\frac{\sigma-1}{\sigma}} + x_H^{\frac{\sigma-1}{\sigma}}\right)^{\frac{\sigma}{\sigma-1}}$$

The first order conditions yield the requirement that $x_F = C_F^{-\sigma} x_H$, so we are left with a partial equilibrium demand curve $x_F = C_H^{-\sigma} Y_H$ where Y_H is the GDP of the home country.[29] But one infinitesimal agents consumption of a one good type has

[29] The standard general equilibrium solution of this system usually produces a demand curve with

no discernible impact on shared GDP of the country, so this will ultimately be the optimization for all agents in the country. So the demand function for all agents, conditional on a pairing, is $x_F = (C_F)^{-\sigma} Y$. So now it is possible to create a combined aggregate demand function for the country as a function of distance that combines both the pairing likelihood for a given distance (p_δ) and trade cost effect, scaled by the number of eligible partners in the foreign country (Y_F). This function is:

$$D(\delta, Y_H, Y_F) = p_\delta Y_F (C_F)^{-\sigma} Y_H = \underbrace{\frac{\beta_H}{\delta} Y_F}_{\text{Extensive}} \underbrace{(1 + \alpha\delta + T)^{-\sigma} Y_H}_{\text{Intensive}}$$

It is interesting to pause here and note that the two parts of this function actually have an interpretation that can be related to the existing literature. The pairing likelihood component describes the decline in varieties consumed with respect to distance, which would be empirically but not theoretically similar to the extensive margin as it is observed in the data. The demand-driven component would describe the decline in intensity of trade flows conditional on pairing, akin to the intensive margin. Usually when these terms are discussed in the literature they refer to different characteristics of the distribution of firms or consumers, which is not at all how they are being mentioned here. But in data analysis, they would look more or less the same; the number of firms participating in markets would decline with distance at one rate, and the participating firms would export less to markets they are present in at another rate.

One interesting implication of this model is that it provides a totally different motive for the existence of this margin that could be studied in greater detail. If this new geometrically motivated extensive margin exists, it would be unaffected by the size of trade costs. Using almost any preferences, this new hybrid framework would,

an exponent is $1 - \sigma$, as discussed by Head & Mayer (2015). The partial equilibrium setting closes off all wealth effects leading to a simpler solution. However, when estimated the specific form will prove irrelevant, so $1 - \sigma$ could just as easily be used.

when taken to the data, provide smaller estimates of trade costs in a way more consistent with observed data. But most intriguingly, the observed importance of the extensive margin in this model will in most cases be larger than that of the intensive margin (because the extensive margin is already gravity shaped). This would be consistent with trends that have been found in the data (Hummels & Klenow 2005).[30]

1.4 The Spherical Model

1.4.1 Motivation: Planar vs. Spherical

This search behavior produces a model that looks very similar to the iceberg cost model, without any iceberg cost. However, there is a pointed difference that occurs when one considers placing the model on a globe rather than a plane: the radii converge on the opposite side of the world. To refer back to the two intuitive, discrete models, consider the situation that the dark room has in fact become the surface of a sphere, and that the light from the laser pen or flashlight must now follow the two dimensional space along the surface of the sphere. Consider the laser pointer example first. If we assume the searcher is standing at the north pole of the globe, then any light rays from their position would follow lines of longitude over the surface and converge at the South Pole (see Figure 8). This would produce an increase in finding probability for objects located at the South Pole analogous to the increase near the North Pole. The same would occur in the flashlight example if flashlight rays also followed lines of longitude (the beam would get brighter again at the South Pole). In both cases we would be left with the absurd result that the “easiest” object to find would be one located at the opposite pole of the planet. We could instead make flashlight rays that were not bound to lines of longitude and simply spread away from the aperture as if on a plane, so that the arcwidth of the beam was the

[30] This role for the distance component of the extensive margin in also deep in most relevant structural models, as discussed by Chaney (2008), Helpman et al (2008), and Bernard et al (2011).

same as a function of distance.[31] This would allow us to describe a model in which things far away were more "dimly lit", but it would create new problems because the flashlight "beams" from different azimuths would cover different points until they approached the South Pole and where they could begin to cover some of the same points. So even in the expanding and diming flashlight beam case, the issue again arises near the South Pole due to duplicate coverage across choices of azimuth. In any case, transferring the search processes of the motivating models directly onto a sphere violates their underlying intuition which is that objects "shrink" and are harder to "find" as they are further away.

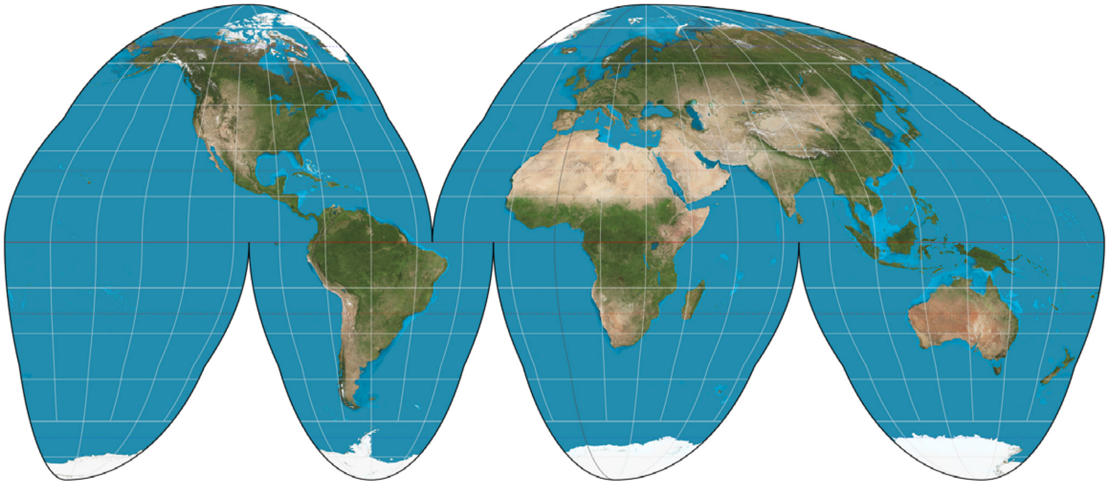


Figure 1.7: an example of a "Goode Homolosine" projection map. This projection preserves landmass areas better than most, and shows the portion of plane that must be removed in order to represent the globe in two dimensions. *Source*: Wikimedia Commons.

The basic problem with this conversion is the paucity of points to locate an object on a sphere for long distances when compared to a plane. It is the same reason why all map projections of a globe must be achieved by stretching objects near the edges of the

[31] This is not intuitive, as "rays" no longer travel in straight lines. One way to formalize this process would be to measure the arcwidth of the flashlight beam in the plane for every distance, and then create a beam on the surface of the globe so that the width is the same at every distance. This would mean that the "brightness" would fall in distance, but only the ray in the middle of the beam would travel in a straight line. Those at the edges would need to curve outward.

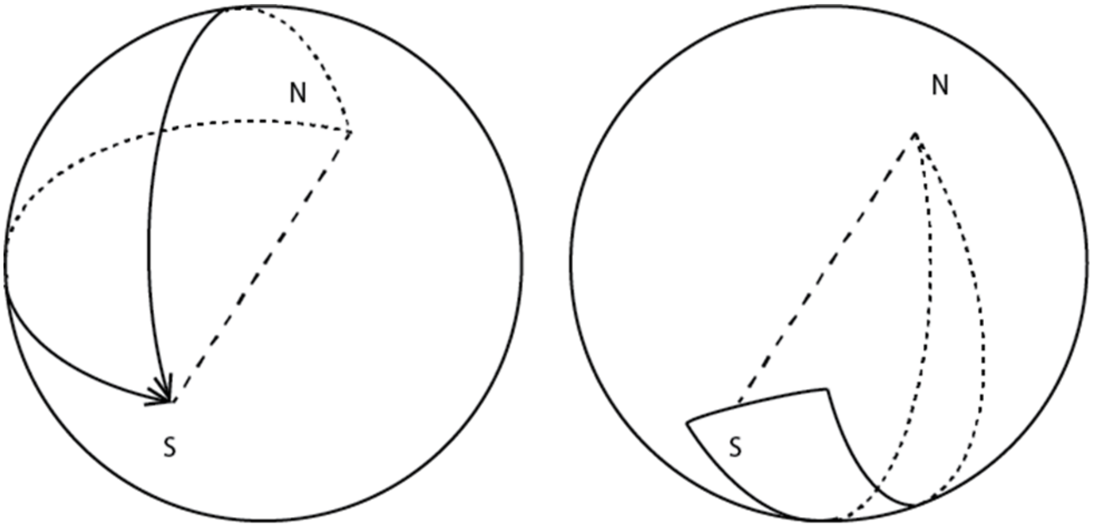


Figure 1.8: a picture illustrating the two example models transformed into spherical space without adjustment. The "laser pen" example is given at left, the "flashlight" example in which the light dims and disperses over distance as it would on a plane is given at right.

projection rather than shrinking them. To consider why this is, plot the circumference of circles (centered at the North Pole) on a globe as a function of distance from the Pole, versus the circumference of concentric circles drawn at the same distance on a plane (see Figure 9).

Going back to the motivating models, it is obvious that our problems arise from the way in which lines of longitude return to meet one another on a globe. If we could perform the search in the planar, light dispersing setting, and then translate that result in way isomorphic to searching on a globe, it would be possible to preserve the general intuition of the models. In order to make a plane isomorphic to a globe, objects near the edge of a planar disk would need to grow smaller, in the intuitive inverse of the distortion created by globe to plane projections. So, if we made the objects in the dark room shrink with distance, we could preserve the desired characteristics of the models.[32]

[32] As this process is fairly convoluted to do separately for the discrete-object example models, it will not be derived in detail.

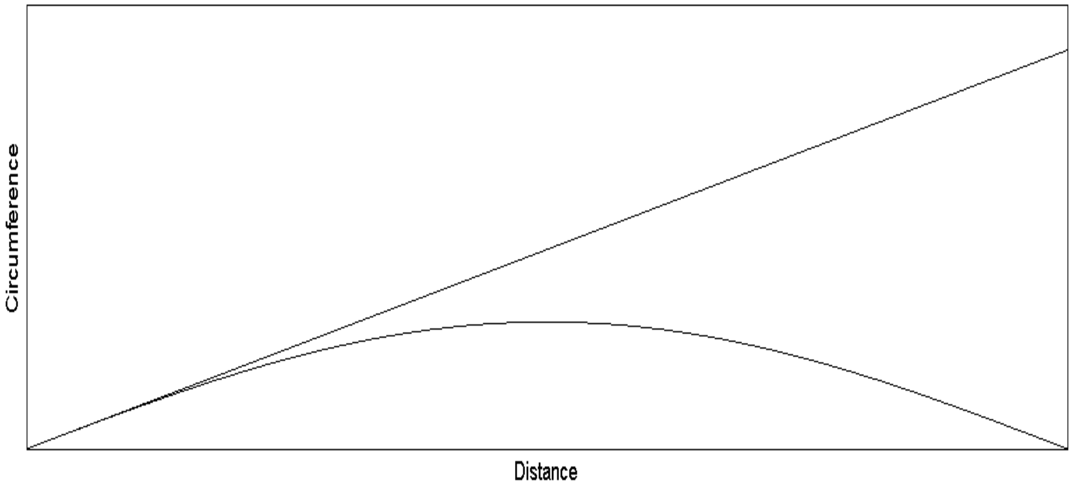


Figure 1.9: the circumference of a circle in a planar disk and on a sphere as a function of distance from the North Pole. The upper line is for the disk ($C = 2\pi d$) and the lower line is a sine function that will be derived in detail below.

1.4.2 The Spherical Adjustment

In order to make sense of the continuous model as applied to the sphere, we will make an analogous alteration. As all objects in the continuous case are of the same infinitesimal size, instead of altering their size we will remove surplus points from an agent's search plane so that the search field will be isomorphic to a sphere (see Figure 10). This removal will be made random with respect to azimuth (see the right-hand diagram in Figure 10). Specifically, the integrated "width" of all the candidate search points in the plane for the agent at a given distance δ_0 should be the same as the width of the points falling at δ_0 on the surface of a sphere (the proportion of the circumference that is not black in the left-hand diagram in Figure 10).[33] This

[33] There is a minor technical slight-of-hand taking place here. If an agent draws truly random points to be in or out of the search space at every possible distance based on the probability we will establish below, the Lebesgue measure of both the included and excluded sets for a given distance will be equal and equal to the entire 2-dimensional horizon at that distance. This is because the selected space of uncountably many points allows for no countable subcover of less than the total search space. A technically precise alternative would be to divide the horizon into countably many sets of vanishing size, and assign them randomly to be in or out of the search space based on the probability function. As the "integration" definition of the modified search space is only presented for intuitive understanding and we never ultimately are called upon to integrate over this space, the issue is ignored because it only adds complexity.Also note that this random removal of points

reduction in points requires a normalization of the pairing probability to adjust to the reduction in candidates. We will achieve this simply by normalizing for surface area on a globe as a function of distance from a point.

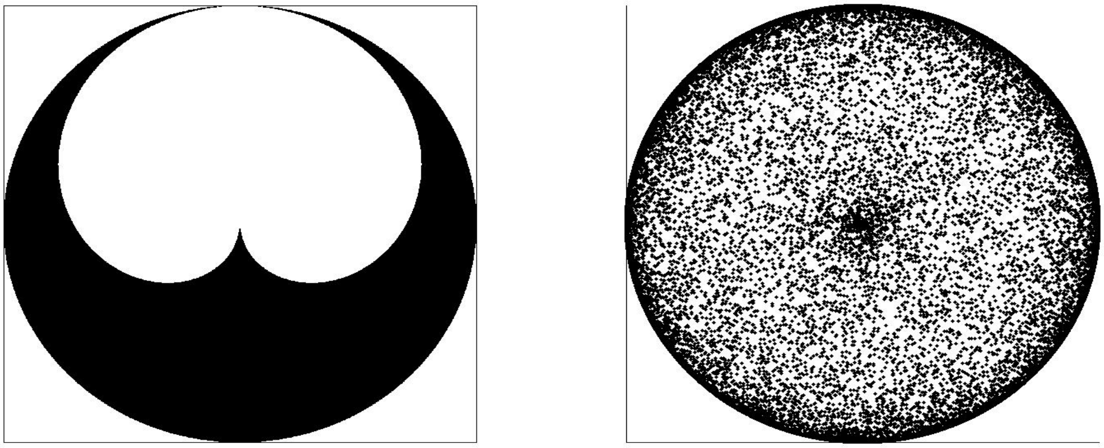


Figure 1.10: The image at left shows the surface of a sphere with a circumference of 2D laid flat on a disk with diameter 2D. The area in black represents all the points that exist only in the planar disk. The image at right displays approximately the same coverage (subject to graphing software limitations, where the non-zero size of "points" seems to indicate unneeded "darkening" near the origin). In this case the "missing" points on the sphere occur at the same frequency with respect to distance, but are randomized with respect to azimuth.

Let $A_P(\delta)$ be the instantaneous change in surface area of a circle drawn on a plane at distance δ from the point O. Let $A_G(\delta)$ be the change in area of a circle on a sphere (or globe). Define the function $p^G\left(\delta\right) = A_G(\delta)/A_P(\delta)$ to be the probability of a random point lying at distance δ being included in the search space after being adjusted to match the characteristics of the sphere. The function $p^G\left(\delta\right)$ is equal to the ratio of the circumferences of these two circles at distance (δ), which we will define as $C_G(\delta)/C_P(\delta)$. The denominator is obvious enough to compute, but $C_G(\delta)$ requires

does not affect the properties that made solving the search model so simple in the planar case. The resulting search space differs for every agent because it is stochastic, but the ex ante search probability functions are identical for every agent. The random removal does not alter that the search process (i) is not serially correlated, (ii) is 1-1, and (iii) is close enough to onto to make unpaired buyers not an issue.

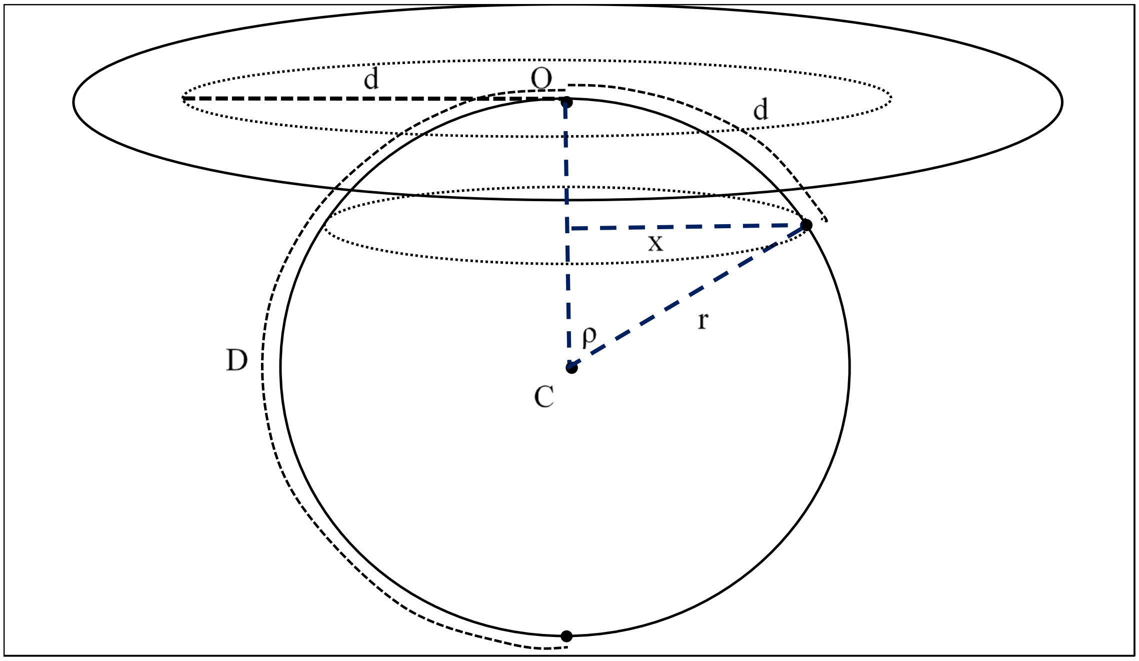


Figure 1.11: this diagrams the problem of determining circumference on a sphere versus on a disk. The labels and definitions used in the derivations are provided.

some simple trigonometry. As the circumference of the sphere we are examining is 2D (because D is defined to be the maximum distance achievable on the globe), we can deduce that the radius of this sphere is $r = \frac{D}{\pi}$. Define $\rho(d)$ to be the angle between the origin and the edge of the circle of size d as seen from the middle of the sphere, which we can see (in radians) is $\rho(\delta) = \frac{\pi\delta}{D}$. We can construct a right triangle (seen in Figure 11 as having edges x and r and angle ρ) from the middle of the sphere to the edge of the circle at distance δ to the line between the origin and the middle of the sphere. By the definition of the sine function, we can determine that the opposite edge (between the ray and circle) is of length $x = r\ sin(\rho) = \frac{D}{\pi} sin(\frac{\pi\delta}{D})$. Therefore

$$C_G(\delta)\ = 2\pi x\ =\ 2D\ sin(\pi \frac{\delta}{D})$$

and

$$p^G\left(\delta\right) = A_G(\delta)/A_P(\delta) \;=\; 2D\ sin(\frac{\delta}{2D})/2\pi\delta = \frac{D}{\pi}\frac{sin(\frac{\pi}{D}\delta)}{\delta}$$

The new pairing probability of the model is the joint likelihood of (i) a point being chosen using the planar search model and (ii) the likelihood that the point in question is included in the sphere-adjusted search space. Because these two probabilities are independent, we see that the pairing probability in the sphere adjusted space as a function of distance, $p_S(\delta)$, becomes

$$p_S\left(\delta\right) = \; p_P\left(\delta\right)p^G\left(\delta\right) = \frac{\beta}{\delta}\frac{D}{\pi}\frac{\sin\left(\pi\frac{\delta}{D}\right)}{\delta} = \beta\frac{\sin\left(\pi\frac{\delta}{D}\right)}{\delta^2}$$

(Note that the leading coefficient has again been generalized for the same reasons as in the planar case.) This pairing likelihood will produce matching behavior very much like that of the behavior found in the planar model near the origin, which can be intuited in two ways: (i) nearer the origin the sphere becomes approximately flat, and (ii) the Maclaurin first order approximation of the function sin(x) is simply x, yielding a first order approximation of the full function of $1/x$ in the neighborhood of zero.

Returning to the reasoning set forth above, consider two small countries, ε_A and ε_B, which lie separated by some distance δ. As distance increases and country-size decreases, the trade flow between these two countries will approach $f_{AB}\left(\delta\right) = \alpha\ \mathrm{AB}\frac{sin(\pi\frac{\delta}{D})}{\delta^2} = X_{AB}$. This equation is the spherical-search modified statement of the gravity equation, which is nearly identical over short distances, but deviates in its predictions at extreme distance (see Figure 12). This is the other startling result of the paper: an alternative statement of the gravity equation derived from the most basic model of search and a simple adjustment for the differences between a plane and a sphere. The way in which gravity derives from search in polar space was shown above taking care to seek the weakest assumptions and greatest generality possible.

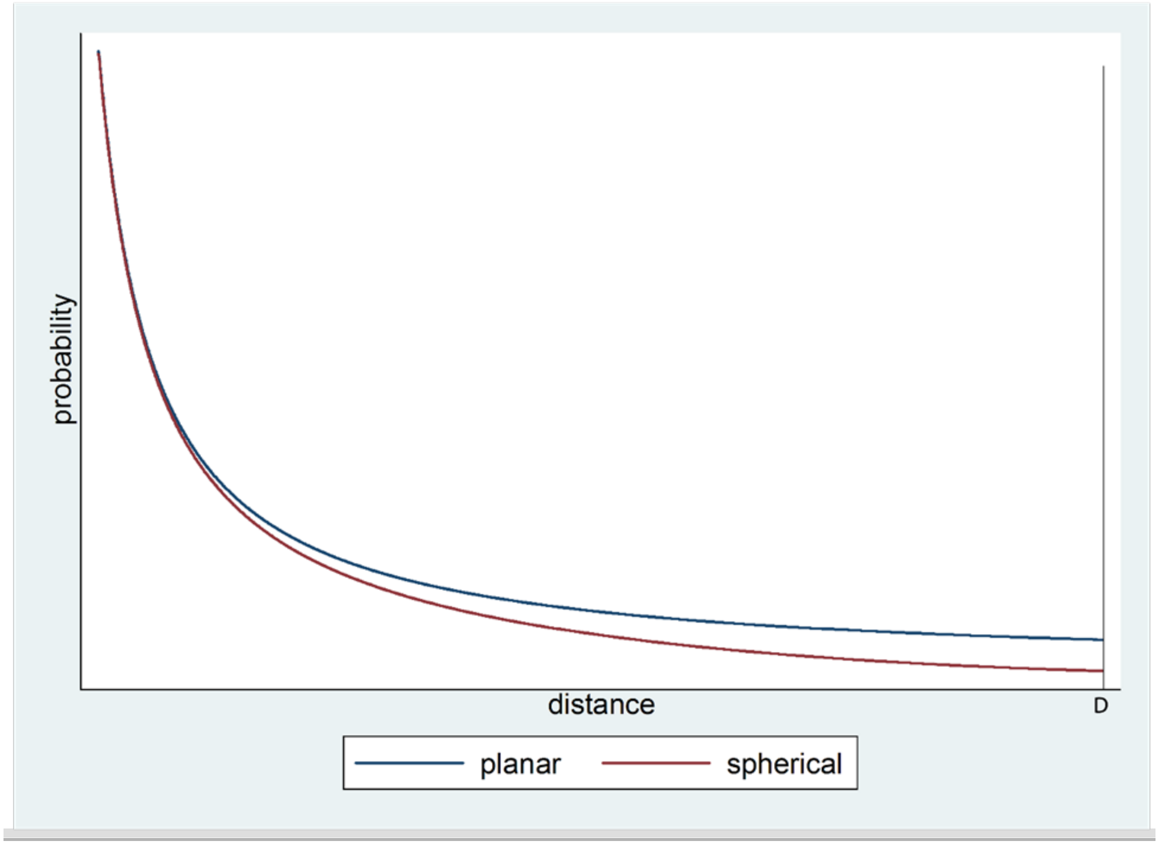


Figure 1.12: this is a comparison of the pairing pdfs for the original planar version of the model (blue) versus the spherical model (red). Note that they are identical near the origin, but that the pairing probability of the spherical model goes to zero as distance approaches the maximum on the globe.

The adjustment and result in the spherical case does not have a claim to generality as strong as the determinant of the Jacobean for converting from polar to Cartesian coordinates; there are other ways of distorting the planar pdf to match with a sphere and the one selected here might be the simplest but does not necessarily encompass or relate to the alternatives. However, our spherically-adjusted form for the gravity model differs from the standard model, so now it is possible to test their diverging predictions (and see whether the alternative structure is born out in the data).

1.4.3 Implications of Spherical Adjustment

This radiant model produces two noteworthy predictions, one minor and one significant. The first is that a continuum of agents searching in two dimensional space provides an additional reason why we might expect small countries to export a larger portion of GDP.[34] The gravity equation above was constructed by generalizing the results of the model to shrinking countries at increasing distances, in which case the search probability of all agents in a given country approach being identical. As distances shrink (or country shapes become irregular) this is less and less the case. The shortest distances describe countries' trade with themselves. The effect of country size on trade derives from the simple observation that if all agents search independently, an agent at the center of a large country is more likely to have their successful pairing "caught" by another agent in their own country when compared to an agent in a small country. So this gives a motivation for the fact that small countries export more without invoking more complex arguments about comparative advantage (factor endowments, increasing returns, et cetera). Unfortunately, the math underlying this discussion grows intractable very quickly, so will not be explored in depth here.

The second prediction is the deviation from standard gravity due to the spherical adjustment. Generally, any claim of deviation from standard gravity should met with skepticism due to the depth of the validation of the relationship. However, in this case the adjustment is so small it is extremely hard to observe in the data. To illustrate this, consider the simple regressions given in Table 2. The data are bilateral trade flows from the year 2000. The data cover 165 countries, and each observation is a bilateral trade link. Observations are only used for the purpose of this analysis if there are positive trade flows in both directions (exporter and importer) and GDP information is available for both countries, leaving us with 5887 useful observations.

In order to allow us to examine the functional form implied by the data, we will

[34] This is not clearly an implication unique to this model, as discussed in Lashkaripour (2006)

define a simple statistic (γ_{AB}) for each bilateral trade relationship

$$\gamma_{AB} = \frac{X_{AB} + X_{BA}}{Y_A Y_B} \approx \beta \frac{1}{\delta} \; or \; \beta \frac{sin(\frac{\pi}{D} d)}{d^2} \; for \; some \; \gamma \; > \; 0$$

This analysis uses the sum of bilateral exports in order to (i) remove trade imbalance issues from the analysis and (ii) reduce the impact of observation error and prevent selection bias[35].

Both the standard and modified gravity equations suggest a functional form but not a coefficient (β) for the relationship between γ and distance. Therefore, a simple method of examining how well the spherical model fits the data and comparing the performance of the models is linear regression of the data on generated values of the planar model ($x_p = \frac{1}{\delta}$) and the spherical model ($x_s = \frac{sin(\frac{\pi}{D}\delta)}{\delta^2}$). In both cases the coefficient values are uninformative, but the coefficients' significance and the model fit (adjusted R^2) can tell us how well the models are performing. Looking at the regression results (Table 2), both models fit nearly identically well both in size of the coefficient and the adjusted R^2. This similarity, and the powerful multicollinearity it implies, means that it will be extremely hard to run a comparison of these two models without more theoretical apparatus. But it also explains how the difference in functional form might have gone unnoticed before.

1.4.4 Spherical Model with an Intensive Margin

Above, a model was developed combining the search process with a cost-driven intensive margin was developed. It is now possible to move it into a spherical search context. If we combine the same constant-marginal cost intensive margin ($(C_\delta)^{-\sigma} Y_H$) motivated by the same home-versus-foreign CES choice with the new spherically

[35] Small countries tend to have less and worse data. Therefore, when looking at long distances, unidirectional data tend to be dominated by large rich countries as the point of origin. By requiring the availability of data in both directions it lessens the possibility of systematic bias arising from this.

adjusted extensive margin ($p_S(\delta)Y_F$), we get:

$$D_S(\delta, Y_H, Y_F) = p_S(\delta)Y_F(C_\delta)^{-\sigma}Y_H = \underbrace{\beta_H \frac{\sin\left(\pi\frac{\delta}{D}\right)}{\delta^2} Y_F}_{\text{Extensive}} \underbrace{\left(1+\alpha\delta+T\right)^{-\sigma} Y_H}_{\text{Intensive}}$$

This new demand function is a modest transformation of the extensive-margin-only spherically adjusted demand curve, owing to the fact that the effects of linearly increasing costs diminish quickly in a CES setting. But regardless, by separating out intensive and extensive margins there is now more hope of reducing the multicollinearity problem for comparing the spherical and planar versions of the extensive margin. Consider the new spherical demand equation in logs.

$$\begin{aligned} ln(\gamma_{FH}) &= ln\left(2\frac{D_S(\delta,Y_H,Y_F)}{Y_H Y_F}\right) \\ &= \ln\left(\sin\left(\pi\frac{\delta}{D}\right)\ \right)\ - 2\ln\left(\delta\right)\ - \sigma ln\left(1+\ \alpha\delta+T\right) + ln(\beta_H) + ln(\beta_F) + C + \varepsilon \end{aligned}$$

The proposed model actually makes very strong assertions about the estimated parameter values of two of the coefficients in any linear regression in logs: the coefficients of the log-sin term should be one and of the log-distance term should be negative two. However, the linear cost term is unknown, and so it would be difficult to examine this term using linear regression techniques. Fortunately, the Maclaurin-series expansion of this term is fairly straightforward: [36]

$$\begin{aligned} f\left(\delta\right) &= ln\left(1+\ \alpha\delta+T\right) \\ &= ln(1+T) + \left(\tfrac{\alpha}{1+T}\right)\delta - \left(\tfrac{\alpha}{1+T}\right)^2\delta^2 + \left(\tfrac{\alpha}{1+T}\right)^3\delta^3 - \left(\tfrac{\alpha}{1+T}\right)^4\delta^4 + \ldots \end{aligned}$$

Therefore, we can examine the relationship between this model and the data by

[36] For an example of using Taylor series analysis, see Baier & Bergstrand (2009)

running the regression

$$ln(\gamma_{FH}) = \rho_1 \text{ln}\left(\text{sin}\left(\pi\frac{\delta}{D}\right)\right) \; +\rho_2 \text{ln}\left(\delta\right) \; +\eta_1\delta+\eta_2\delta^2+\eta_3\delta^3+\eta_4\delta^4+D_H+D_F+C+\varepsilon$$

where the model predicts $\rho_1 = 1$, $\rho_2 = -2$, and $\eta_1{=}{-}\sigma\frac{\alpha}{1+T}$, $\eta_2{=}\sigma\left(\frac{\alpha}{1+T}\right)^2$, $\eta_3{=}{-}\sigma\left(\frac{\alpha}{1+T}\right)^3$ and so on with alternating sine. It is important to here note that, because log of costs is multiplied by *negative* σ the coefficients should alternate sine in the opposite pattern to the Maclaurin Series expansion.[37] The strongest rejections of the model would be $\rho_1 = 0$, $\rho_2 = -1$, and $\eta_i = 0$.

But before discussing the results of this regression, there is also a simpler variation that will be informative. Consider the model without *without* the extensive margin, wherein:

$$\begin{aligned} ln(\gamma_{FH}) &= -\sigma ln\left(1 + \; \alpha\delta + T\right) + ln(\beta_H) + ln(\beta_F) + C + \varepsilon \\ &\approx \eta_1\delta + \eta_2\delta^2 + \eta_3\delta^3 + \eta_4\delta^4 + D_H + D_F + C + \varepsilon \\ &\approx (1-\sigma)ln(1+\alpha\delta+T) + ln(\mathbb{P}_H) + ln(\mathbb{P}_F) + C + \varepsilon \end{aligned}$$

This regression examines only how constant marginal cost would effect demand in our crude partial-equilibrium setting. With no more search parameters to cause country-level idiosyncrasies it is not clear why country-level fixed effects should be present. However, running the regression with country-level fixed effects would be isomorphic to a standard Armington model with linear costs, wherein the fixed effects would now be interpreted as price level effects ($\mathbb{P}_i$) or multilateral resistance terms. This is helpful for our discussion, because it gives us an "implied" constant-marginal cost trade cost (which is not often estimated) for the data. It shows us how well the Taylor expansion fits the data. And most importantly, it allows us to examine if and how much implied trade costs fall by adding the extensive margin to the model.

[37] Estimates of the elasticity of substitution vary widely, but a common Armington estimate is an elasticity of substation of about 3.5.

The results of these regressions are given in Table 3. In the first two columns the models are estimated without fixed effects. The first column is the intensive-margin-only specification. Working under the assumptions that T is fairly small and that the first term in an alternative Taylor series is the most important and best estimated, we can hold up the estimate of $\eta_1 = -9.62e - 04$ as a rough estimate for marginal trade cost (α). Using an estimated trade elasticity of $\sigma = 3.5$ implies $\alpha = 2.75e - 04$ which is two orders of magnitude larger than our estimate from trade cost data ($\alpha = 2.89e - 06$). So this is consistent with the assertion that *marginal* "implied" costs are very large in the trade-cost framework. In the second column, the model is run with the intensive margin included. As we can see, the estimated values of ρ_1 and ρ_2 are near what the model predicts. The Taylor series coefficients are no longer significant, but their sine and relative amplitude is consistent with what the model require. Note that the amplitude of every coefficient is smaller than that of the cost-only regression, suggesting implied costs have indeed fallen. Also, the point estimate for $\eta_1 = -2.49e - 06$, though not significant, would be consistent with trade costs the correct order of magnitude when compared to what we see in the data.

Adding country level fixed effects only makes the success of the model grow more clear. First, the cost-only model is run and the implied marginal trade cost grows even larger. However it is worth noting that, if country shipping costs are idiosyncratic ($\tau_i = 1 + \alpha_i \delta + T_i$) then the inclusion of fixed effects might capture some of this heterogeneity. Unfortunately, the presence of such heterogeneity makes the Taylor expansion somewhat misspecified. This should not effect the sign or amplitude of the η_i estimates, but it does make the coefficient values hard to interpret independent of the fixed effect values, which would explain why adding fixed effects causes η's to grow larger. In either case, the addition of fixed effects causes the errors around the η estimates to fall in every specification, with or without an extensive margin.

Looking at Specification 2 and Specification 3, we again see the estimates of ρ_1

and ρ_2 are consistent with the spherically adjusted model. They again reduce the size of the marginal cost estimates (η_i) in a way consistent with premise of the model. Specifications 1, 2, and 3 together provide strong evidence of a spherically adjusted extensive margin in the data. But this evidence grows still more compelling when considering Specification 4: a "planar" extensive margin that does not include the sine function. When the model is estimated without the spherical adjustment, there are a few striking implications. First, the η_i estimates do not change substantially. This suggests that (i) the intensive margin is being identified, and (ii) switching between spherical and planar extensive margins is not changing how much of the variation is explained by the extensive margin. Second, the planar model (δ^{-1}) of the extensive margin is very nearly rejected and fits far worse than the other three specifications. And third, if we presumed that spherically adjusted model was correct, ρ_2 in Specification 4 would be akin to estimating a constant elasticity approximation of the extensive margin. Intriguingly, this value is consistent with extensive margin distance elasticities observed in the data for the decline of varieties over distance (Hummels 2005).

1.5 Conclusion

This paper has shown that a very simple model that generates gravity in a fundamentally different way than through trade costs. It utilizes a very basic search process, and creates gravity as a latent feature of search. This effect is independent of the structure of preferences or production. It then shows how to easily integrate this framework with more familiar models to generate a trade-cost driven intensive margin and a search-driven extensive margin. Before proceeding to greater refinement, it is helpful to pause here and note that this "planar" model is nearly as consistent with the data as the trade cost model in that it relies on a simplification of some true and obscure search process whereas the trade cost model relies on obscure, unmeasured trade costs of a specific structure.

But by examining the deeper characteristics of the search process, it is possible to find an actual falsifiable claim for the search model: the spherical adjustment. So next the paper develops the inevitable spherical search correction for the model. It also shows how to integrate this with more conventional models and takes the result to the data. The spherical adjustment is well validated in the data. It reduces the size of implied trade costs. And it produces a motive for the extensive margin that is independent of the model, and at a magnitude consistent with estimates in the literature.

This draws attention to a major open question: there has been no deep justification for the trade cost motivation for gravity. This is a very simple alternative architecture that achieves the same ends. Both have advantages and disadvantages, but the effort to distinguish them will be complex and deep.

Table 1.1: Shipping Cost Regressions

	Specification 1 $\tau_D - 1$	Specification 2 $ln(\tau_D - 1)$	Specification 3 $ln(\tau_D)$
Distance (α)	**2.89e-06** **(2.56e-09)**		
Constant (T)	**.0464** **(2.24e-05)**		
Log Distance		**0.319** **(1.61e-04)**	**0.015** **(4.63e-5)**
Log Constant		**-5.515** **(0.001)**	**-0.067** **(4.07e-4)**
R^2	0.0296	0.5654	0.033
N	3,029,547	3,029,547	3,029,547

Coefficients reported with (standard errors). All variables significant at 99% confidence or better are given in **bold**.

Table 1.2: Similarity of Planar & Spherical Models

	Specifcation 1 γ_{ij}	Specifcation 2 γ_{ij}
Planar Model (x_p)	1.01e-13 (16.50)	No
Spherical Model (x_s)	No	1.00e-13 (16.51)
Adj-R^2	0.0477	0.0477

Regressions of the implied functional form of $\gamma_{ij}(\delta)$ on the actual data. The t-statistics are given in parenthesis due to extreme small size of standard errors. All coefficients are highly significant. The values of γ in the data are very small, which is the reason for the small size of the coefficients.

Table 1.3: The Intensive and Extensive Model

Distance Regressor	Cost Only Without FE	Specification 1	Cost Only With FE	Specification 2	Specification 3	Specification 4 Planar Model
$\ln\left(\sin\left(\pi\frac{\delta}{D}\right)\ \right)$	No	**1.285** **(0.653)**	No	**1.200** **(0.528)**	**1.182** **(0.527)**	No
$\ln(\delta)$	No	**-2.377** **(0.606)**	No	**-2.016** **(0.503)**	**-1.680** **(0.506)**	**-0.643** **(0.207)**
δ	**-9.62e-04** **(9.05e-05)**	-2.49e-06 (0.00027)	**-1.33e-03** **(9.01e-05)**	**-7.97e-04** **(2.42e-04)**	**-9.57e-04** **(2.44e-04)**	**-7.12e-04** **(2.18e-04)**
δ^2	**1.36e-07** **(1.93e-08)**	3.70e-08 (4.32e-08)	**1.82e-07** **(1.87e-08)**	**1.41e-07** **(3.92e-08)**	**1.55e-07** **(3.92e-08)**	**1.03e-07** **(3.17e-08)**
δ^3	**-9.30e-12** **(1.56e-12)**	-4.10e-12 (3.09e-12)	**-1.16e-11** **(1.49e-12)**	**-1.02e-11** **(2.81e-12)**	**-1.08e-11** **(2.80e-12)**	**-6.79e-12** **(2.16e-12)**
δ^4	**2.30e-16** **(4.18e-17)**	1.52e-16 (8.53e-17)	**2.62e-16** **(3.92e-17)**	**2.70e-16** **(7.66e-17)**	**2.79e-16** **(7.65e-17)**	**1.54e-16** **(5.24e-17)**
Country FE	No	No	Yes	Yes	Yes	Yes
Border Effects	Yes	Yes	Yes	No	Yes	Yes
R^2	0.229	0.229	0.545	0.545	0.547	0.546

Coefficients reported with (standard errors). All variables at 95% confidence or better are given in **bold**.

CHAPTER II

Trade in Elasticity

2.1 Background

It has been widely observed that small contractions in output correspond with much larger reductions in trade volumes. The recent, 2008-2009, global downturn was no exception. US monthly exports and imports experienced year-on-year falls of 20% during 2009 while contraction of output was less than 5%, and this was small by global standards. Some East Asian economies experienced drops of as much as 40% in trade while output contracted far less.

Economists have explored several hypotheses to explain this: (i) recessions spur protectionism which suppresses trade, (ii) credit constriction during crisis disproportionately affects trade (Amiti and Weinstein 2009), (iii) the large share of intermediate goods in the trade bundle amplifies the effect of shocks (Levchenko, Lewis, and Tesar 2010), (iv) trade is disproportionately in durable/capital goods, which are also disproportionately affected by recessions (Eaton et al 2010), (v) in recessions firms reduce inventories leading to larger contractions in non-perishable goods, which are more often traded than perishable ones, and (vi) traded goods are disproportionately of higher quality and economic shocks shift consumers towards lower quality consumption (Berthou and Emlinger 2010). Efforts to test these hypotheses have met with mixed results, and the question is still generally considered open.

This paper examines a seventh alternative, that trade is dominated by income-/wealth-elastic goods. That is, rather than the trade fluctuations being driven at the firm level, through inventories, they may be driven by shifts in consumer behavior in response to the economic shock. It seeks to test whether household consumption responses are predictive of trade share.

This paper seeks to examine if goods that are traded are relatively income elastic. In order to test this hypothesis, it is necessary to determine some way of estimating the consumption responses caused by changes in income on a good-specific basis.[1] There is a great deal of work on estimating these characteristics for individual's consumption versus savings behavior. Unfortunately, using only two categories of allocation, all consumption and all savings, will be unworkably coarse in regard to the hypothesis of this paper.[2] This paper seeks to examine domestically produced goods versus imported ones; which requires more detailed information about consumption.

What is needed to answer this question are data that in some way disaggregate consumption into categories that are correlated with imported versus domestically-produced goods.

2.2 Methods

2.2.1 Wealth versus Income

This paper explores two categories of elasticity: wealth and income. In the textbook setting, income elasticity is the most discussed concept. In recent years, however, much more examination has been invested in looking at how disruptions in asset markets affect consumer behavior (*i.e.*, increases in housing prices cause homeown-

[1] There is a substantial literature computing aggregate elasticities for trade (e.g., Broda and Weinstein 2006). However, this hypothesis examines the issue at the household level distinct from firm behavior, meaning economy-wide aggregates are not useful.

[2] It seems appropriate to note a previous use of consumer-level data to examine trade issues. Broda and Romalis 2009 used consumer data to examine responses to price fluctuations based on trade. This paper has since been retracted, but the methods it sets out are reasonable and innovative.

ers' expenditure to increase without altering their real income) (e.g., Case, Shiller, and Quigley 2005, among many others). There is further value in this, because it allows the empirical tests to search for different trade-shift implications due to the *type* of macroeconomic shock. For instance, an ordinary cyclical recession will have proportionately similar impacts on household wealth and income, whereas financial crisis-induced recessions will disproportionately affect wealth-levels. As will become clear, the estimation strategy adopted ultimately does not require different treatment of the two parameters. So this paper will examine *both* measures, and report the results for both in the final tables.

2.2.2 Data

This paper utilizes two data sources: (i) the Consumer Expenditure Survey ("CEX") provided by the BLS and (ii) the Input Output Account ("IO Matrix") provided by the BEA. The CEX is quarterly data about expenditure, income, and assets from a sample of between nine and twenty-six thousand consumers in the US.[3] This paper looks at data from 1980 to 2003 as compiled by the NBER. In interviews participants report their expenditures divided into several sub-categories.[4] Participants remain in the sample for four months, and entry / exit is staggered.[5]

The IO Matrix divides the dollar cost of inputs for an industry across all possible source industries. The detailed form of the matrix provides six-digit category fineness, but this paper only utilizes the coarser summary version, which uses 128 industries. The BEA releases an IO Matrix once every five years but, due to the technical difficulties constructing concordances with the consumption data-set, I will only utilize that of 2002.

[3] For a complete list of these categories, see Appendix I.

[4] This paper uses the 36 sub-categories provided in the NBER version of the data. Finer data would be possible using a more raw form of the data from the BLS.

[5] For this reason, the dataset could have limited use as a panel. But, after seasonal adjustment, it is not clear that enough wealth variation could be observed for meaningful estimates.

In the analysis that follows, data from both sources will be aggregated into 28 "bins" that accord with one another.[6] This number is chosen because eight items in the CEX accounted for trivial or zero shares of trade at the industry level.[7] Luckily, these items also account for trivial shares of consumer expenditure. Also note that while every other entry in the CEX is matched to a classification in the IO Matrix, the reverse is not true. Many of the items in the IO table are rarely or never purchased directly by consumers.

2.2.3 Trade Share

This paper computes trade share from the IO Matrix in two different ways. The simplest is to divide the economy's imports of a good M_t^j, by total output of that good, Y_t^j.

$$\gamma_t^j = M_t^j / Y_t^j$$

But using this specification, certain consumer goods that are services have import shares of zero. This method fails to capture any information about the role of intermediate goods, which is vital when dealing with consumers who purchase heavily from service industries. So instead, I compute a weighted sum of imports as intermediates for each sector.

$$\varphi_t^j = \left[\sum_k \gamma_t^k C_{kt}^j \right] / Y_t^j$$

Here C_{kt}^j is the utilization of good type-k by the industry type-j (*i.e.*, the quantity of output from industry k that goes into industry j) at time t. So φ_t^j represents the sensitivity of *aggregate* imports to shocks in the output of good type-j, whereas

[6] The CEX does include "diary data" which is the raw itemized survey information from which the aggregates used in this paper are constructed. It would be possible to construct finer data using the diary data, but the validity of these results that this would produce is unclear. While the data are far finer than the aggregated bins, the CEX is notoriously imprecise at fine detail (individuals surveyed are asked quarterly about all of their consumption over the prior three months). Therefore some reasonable aggregation is necessary. Rather than construct novel categories from the raw, this paper relies on the NBER subcategories as they are well studied.

[7] One item, "gambling", only begins to be measured in 1999.

γ_t^j measures only the sensitivity of imports of good type-j to shocks in good type-j output.

This method makes several important assumptions, which can be relaxed or modified in other specifications. First, it assumes that imports of a good are randomly assigned across the input bundles of all industries. Second, it links intermediate imports to aggregate output rather than to consumption. This imposes structure on the relationship (i.e., identical proportional changes) between output and consumption that are not always reasonable. These issues are not addressed in this paper.

2.2.4 Cross-Sectional Methods

Estimating good-specific wealth and income elasticities is not easy. It requires consumption data divided into categories of goods sufficiently fine to inform the hypothesis.[8] More importantly, it also requires a difficult choice about identification strategy based on comparing (i) one individual to themselves at different times and wealth levels (panel data) versus (ii) different individuals with different wealth levels at the same time (cross sectional data). The former strategy is in many ways the most appealing, as it imposes no structure on the preferences of individuals relative to one another. However, over time an individual will face variation in both their income / wealth level *and* in the prices that they face. So any attempt to extract wealth elasticities from panel data will require detailed price data and assumptions or precise estimates about price elasticities.

Therefore, this paper opts for the second identification strategy. By imposing assumptions on individuals' preferences, elasticities can be estimated using cross sectional data. This method is desirable, not only because the data are available, but

[8] This issue makes the PSID and most other popular panel data-sets unusable as, even when they have wealth/expenditure data, they lack detail on the types of goods purchased. Also, monthly data are tremendously more desirable when examining business cycle behavior, though none of the specifications used in this paper exploit this feature.

also because it reduces or eliminates price variation between observations.[9] In order to compare observations, i.e., different individuals with different wealth/income levels in the same period, I will simply assume identical preferences. (This is extremely hard to justify across wealth levels as will be shown in several tables.)

In any case, the estimation of elasticities requires some assumptions about the structure of consumers' utilities or, equivalently, the functional form of their Engel Curves. While there have been interesting innovations in semi- and non-parametric estimation techniques for Engel Curves, it is not clear that this research project gains from this additional structural robustness. In order to gain insight into trade, knowledge about the Engel Curve must be combined with knowledge about the import share of each sector. If the later information is a discrete numerical value, then some structural assumptions will need to be made in order to turn the non-parametric Engel Curve estimate into a number. It seems that any assumptions made here would be just as offensive as parametric assumptions in the initial step. For this reason, and because of the computational constraints imposed by the need to compute 1600 elasticities for the data-set, I will choose the simplest possible formula for the Engel Curve – a constant elasticity hyperbola.

$$\ln(g_{it}^{j}) = \ \beta_t^j * \ln(I_{it}) \ + \ \alpha_t^j + \ \epsilon_{it}^j$$

In this estimation, g denotes expenditure on a good j, i is for the individual and t is the time period. Because both income and expenditure are in logs, β_t^j can be interpreted as an elasticity. Note the presence of a good-specific fixed effect, α. This both serves to allow for a "minimum level" of consumption to deal with situations where the Engel curve does not pass through the origin.[10] Also note that

[9] There has been substantial work examining variations in consumer prices based on both region and neighborhood. It is well documented that the poor pay higher prices for basic consumer goods, but in this paper the phenomenon is ignored.

[10] This value does have the problem of, for curves that do not pass through the origin, not being a conventional elasticity. However, elasticities have long been noted to have inconvenient mathematical

this functional form removes any wealth-related substitution effects between goods, i.e. imposes a constant elasticity across the population.

In order for the hypothesis of this paper to hold, a higher (lower) income / wealth elasticity for a good or service should correspond to higher (lower) import share for that good. Unfortunately, despite the size and complexity of the estimation process, I only have 28 data-points to examine in the final iteration. Due to the lack of degrees of freedom, the simplest regression method seems best.

$$\varphi_t^j = \rho * \beta_t^j + \alpha_t^j + \epsilon_t^j$$

Here ρ is a common coefficient across all of the $\beta_t^j s$, i.e., across elasticities. **Table I** reports the results. Regressions 1-4 use elasticities estimated over all of the years. Regressions 5-8 only look at data from 2002. In summary, the results are never significant and often have the opposite sign the hypothesis would predict. The regressions in **Table II** examine the role of constant elasticity in this outcome. It divides the population into low and high income individuals (bottom and top deciles) and computes the elasticities for each population separately. If the constant elasticity assumption made in the estimates in Table I were totally valid, then these estimates should be identical to what was found in the full population estimates. However, this does not seem to be the case. The point estimates seem to change notably, suggesting changes in curvature in the Engel curve that are not captured by a single elasticity parameter alone. Therefore, more robust methods seem to be required.

properties in the neighborhood of the origin. Roughly speaking, the origin-adjusted "elasticity" speaks to the curvature of the Engel curve and greatly improves the efficiency of OLS as an estimator. (The regression without the intercept, α, has extremely large errors for goods with substantially non-zero intercepts.) This curvature measure does speak to the type of expenditure response due to income increases or decreases (e.g., inferior or luxury goods). Finally, very few goods display substantial non-zero intercepts; including the intercept has little impact on most point estimates, and mostly effects the size of the errors.

2.2.5 Semi-Parametric Methods

The most obvious issue with the estimation method used in the previous section is that the model uses a single elasticity to describe all consumers, and there does not exist a non-homothetic, well-behaved utility function that rationalizes a constant income elasticity across all goods at all levels of income. The model utilized in the previous section is defensible as a local-linear approximation of the true aggregate demand function, but this simplification is necessarily blind to variations across the income distribution.

To get away from this issue, the next model utilizes a semi-parametric regression to sketch Engel curves that can then be used to describe the effects of wealth shocks. By producing an actual Engel curve the model also gains the advantage of allowing examination of distributionally asymmetric wealth shocks. That is, the previous model produced a single local linear approximation for the entire population, so if high earners suffered greatly and low earners not at all, or the reverse, either shock would necessarily predict the same aggregate change in consumption. By semi-parametrically estimating the Engel curve, the estimated curves allow for variations in shape across the distribution and give relevance to the structure of the change in income.

2.3 Estimation

There are several different strategies for creating semi-parametric curves. The simplest such method is local-linear regression. However, because the data grow sparse at the extremities of the income distribution, the local-linear regression (and most other "moving average" type smoothing algorithms) produce some very bad behavior. A countervailing pressure on any estimation strategy was that the more computationally demanding estimator would put a strain on the rather small (year-specific) N of

1400. It would be possible to use multiple years' data to solve this problem except that the estimation strategy is critically dependant on prices remaining fixed for the consumers in order to exclude substitution effects from our estimates.

In order to allow use many years of data by including a reasonable control for each time period, this paper instead uses a Yatchew partially linear regression (Yatchew 1997). This allows me to include fixed effects for each good in each year.[11] A Yatchew regression runs a regression of the independent variable (horizontal axis) on the dependent variable (vertical axis) while controlling for the effect of covariates based on a linear model. It requires that the independent variable be distributed continuously (which the data are, to a fairly strong degree, in this case). By including year fixed effects, and adjusting income and expenditure for inflation, it allows the inclusion of all the years of data plotting the adjusted Engel curve. The inflation adjustment is intended to make incomes comparable over time, and the annual fixed effect is intended to allow for shifts in taste, or price-driven substitution between goods, on a year by year basis. A weakness of these controls is that they do not allow for differential year-specific effects for different levels in the income distribution.[12]

The Engel curves produced are reasonably well behaved (*e.g.*, they are upward sloping). Certain goods show curves that are clearly inferior; Figures I A&B show that clothing is an inferior good for households, both for income and for wealth.[13] There are too many curves to attach to this paper, but the general results are that (i) most curves are primarily linear but (ii) some show clear signs of being income

[11] See Appendix II for the actual regression equation.

[12] It would, in theory, be possible to include more detailed controls for individuals. For instance, Engel curves could be adjusted for age, education, and family size. Unfortunately, as more and more controls are added, particularly if they are "noisy", poorly correlated controls, Yatchew plots become less and less comprehensible. Using a relatively short list of controls, the plots quickly become more or less flat lines with a value of zero.

[13] It should be noted that this paper elects to reverse the axes of the standard Engel curve. This was initially due to the fact that most software programs prefer to put the "running" (i.e., independent) variable on the horizontal axis when making graphs. But also, since the distribution of the individuals along the horizontal axis is important to the results, this seems to make the scatter plots more easily understood.

inelastic (in this case a diminishing slope of less than one) such as clothing and food. There is also an inferior good (tobacco). However, there seem to be no relatively clear income elastic goods with clearly upward bending Engel curves (perhaps air travel, housing additions, and laundry). Encouragingly, each good produces a wealth Engel curve and an income Engel curve with shapes similar to each other, which is consistent with basic consumer theory. [14]

2.3.1 Simulated Shocks

In order to test these curves against the research hypothesis, it is necessary to express the shape of these Engel curves in a numerical way that can be compared against the trade share data. While there do exist mathematical ways of describing the curvatures over the entire population, it is far simpler and more intuitive to simulate a recession by altering each agent's income. This has the further advantage of utilizing information about the distribution of agents across the domain of the Engel curve.

Put more precisely, each individual in the distribution of agents has an initial income or wealth. This income or wealth is then reduced by a given amount (5% in the example used in this paper). Then each agent has their consumption of each good recomputed at their new income based on the Yatchew-estimated Engel curve. These new consumption levels are then aggregated across consumers to produce an

[14] At this juncture it is appropriate to discuss an unpublished paper from 2012 by Hummels & Lee that examined a similar hypothesis to this paper. Five years later, this paper remains unpublished, but adopts a very similar methodology using a similar data set (developed contemporaneously and wholly independently). But in this paper the authors find a very large elasticity effect of trade using nearly identical methods and data (a traded goods income elasticity of nearly 1.5). This paper does not find consistent findings, despite producing nearly identical descriptive statistics. This is in part because most of the variation in their implied Engel curves arise from the uper few percentiles of the income distribution where the data are inconsistent. Also, the authors use a parametric technique (Stone-Geary preferences) for creating Engel curves that encourages non-unitary elasticities. But finally, the authors do not use the IO Matrix or any analogous technique for assigning the "tradability" of specific goods types. Therefore, car expenditures are treated as "traded" and housing and (strangely) food expenditures treated as "nontraded". I argue that these differences make my paper the superior assessment of micro-founded income elasticity and trade, and explain why their paper has remained unpublished.

implied change in consumption of each good for the entire economy. In the final step, these implied percentage changes in consumption are regressed on the trade-share percentages used above. The results of these regressions are given in Table III. The results show a clear positive correlation for both wealth and income elasticities, but without statistical significance.

2.4 Conclusion

The results of this inquiry are ultimately inconclusive. The alteration in consumption behavior implied by this analysis shows general consistency with the hypothesis that imported goods are relatively wealth elastic, but not in a decisively compelling way. A next step would be to perform the same analysis on a cross national basis, allowing one to examine both import and export implications to see if that can lend power to the analysis. Also, by looking at cross-sectional data over time, the elasticity estimates identify only long-run responses to income shocks not short-run ones. Therefore, this method is incapable of identifying differences in the "deferability" of purchases that do not effect long run consumption responses. So if there is a short run shock to consumption based on a goods ease of deferment, and if traded goods are relatively more likely to be easily deferred (due to their shelf-life et cetera), this paper cannot address that scenario.

Table 2.1: Regression Using All Consumers

	Imports, sans Intermediates	Imports, sans Intermediates	Imports, WITH Intermediates	Imports, WITH Intermediates	Imports, sans Intermediates	Imports, sans Intermediates	Imports, WITH Intermediates	Imports, WITH Intermediates
Income	-0.880		-0.541					
	(-0.19)		(-0.34)					
Inc + Wealth		-13.13		2.225				
		(-1.92)		(1.20)				
Inc + Wealth					-11.40		2.207	
, 2002 Only					(-1.98)		(1.29)	
Income,						-1.588		-0.790
2002 Only						(-0.24)		(-0.34)
Constant	-0.123	-0.116*	0.102	0.0832**	-0.139*	-0.114	0.0868**	0.104
	(-1.21)	(-2.37)	(1.71)	(3.27)	(-2.40)	(-0.99)	(3.22)	(1.61)
Observations	28	28	28	28	28	28	28	28
Adjusted R^2	-0.036	0.087	-0.034	-0.023	0.064	-0.034	-0.022	-0.033

t statistics in parentheses, * $p < 0.05$, ** $p < 0.01$,

Table 2.2: Comparing Rich And Poor Consumers

	Imports, sans Intermediates	Imports, sans Intermediates	Imports, sans Intermediates	Imports, sans Intermediates	Imports, WITH Intermediates	Imports, WITH Intermediates	Imports, WITH Intermediates	Imports, WITH Intermediates
Inc + Wealth	62.81				-26.01			
of **Rich**	(1.41)				(-1.95)			
Inc + Wealth		-2.185*				0.763		
of **Poor**		(-2.12)				(0.91)		
Income,			26.31**				-10.26**	
of **Rich**			(4.41)				(-4.55)	
Income,				0.365				-0.125
of **Poor**				(1.96)				(-0.90)
Constant	-0.302*	-0.100*	-0.410**	-0.104	0.132**	0.0856*	0.172**	0.0864*
	(-2.51)	(-2.06)	(-5.37)	(-1.97)	(3.59)	(2.24)	(5.18)	(2.21)
Observations	28	28	28	28	28	28	28	28
Adjusted R^2	0.125	0.047	0.291	0.047	0.123	-0.010	0.251	-0.012

Table 2.3: t statistics in parentheses, * $p < 0.05$, ** $p < 0.01$

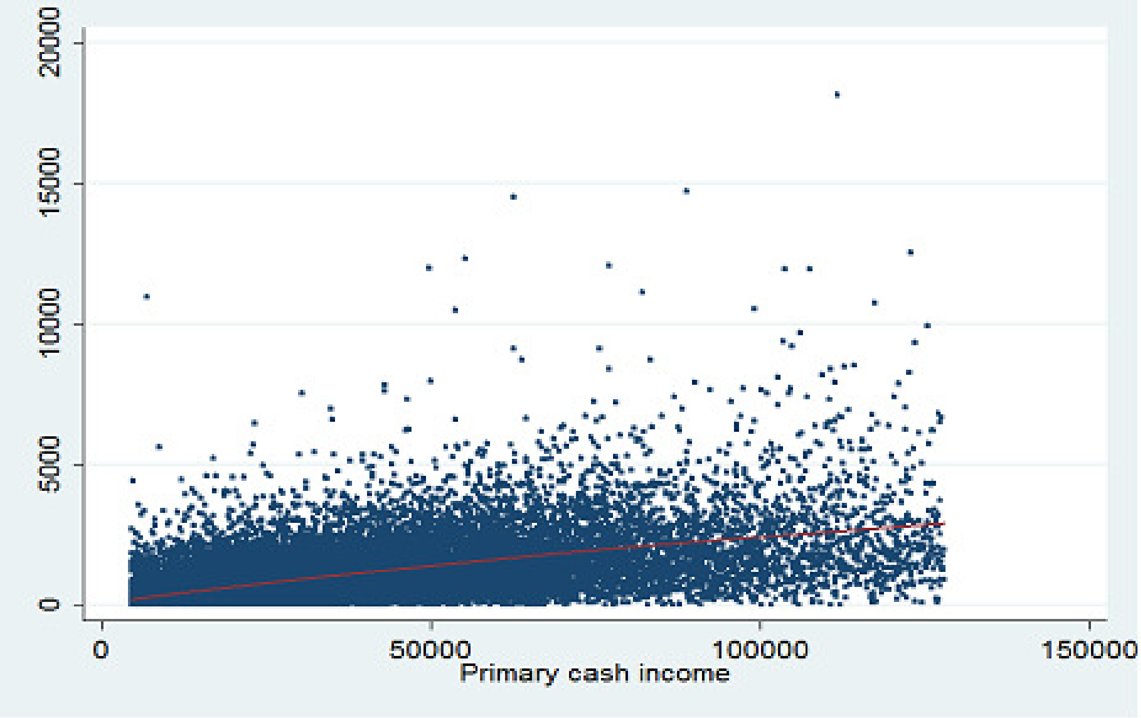


Figure 2.1: Adjusted Engel Curve: expenditure on clothing in terms of cash income

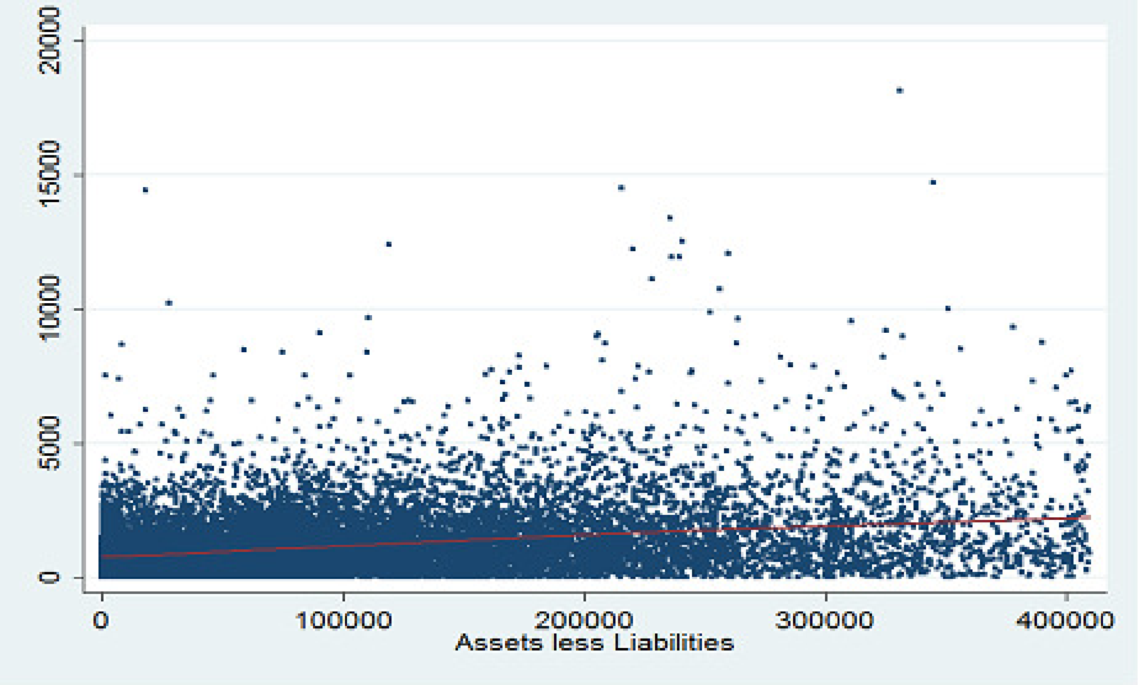


Figure 2.2: Adjusted Engel Curve: expenditure on clothing in terms of cash income

Table 2.4: Simulation Regressions

Change in good-specific consumption given 5% fall in household…	Regressed on *Import Shares*	
	Including Intermediates (and services)	*Excluding* Intermediates (and services)
Cash Income	6.821	4.930
	(9.363)	(4.654)
Wealth	5.650	0.971
	(5.645)	(2.866)

Table 2.5: CEX Categories Tables

Short Name	Full Name	Description
Air	Airfare	Air travel costs
Carservs	Car Services	Automotive repair and maintenance
Autos	New Cars	Newly produced automobiles
Clothes	Clothing	Manufactured apparel
Busiserv	Business Services	Legal, accounting, tax services, et cetera
Comp_Music	Computers & Music	Computers, audio-visual equipment, music and digital media
Drugs	Pharmaceuticals	Purchases of prescription and non-prescription drugs
Food	Food & Alcohol	Food & Alcohol purchased for home consumption
Elect	Electricity	Electricity bills to the home
Fuel	Fuel	Gasoline for any purpose & home heating coal
Furnish	Furnishing	Home appliances and furniture
Gambling	Gambling	Gambling expenses (only from 2001)
Housadd	Housing Addition	New construction on an existing home
Health	Health Expenses	Medical expenses other than drugs
Laundry	Laundry	Laundry and tailoring services
Home	Home Maintenance	Home maintenance expenditures
Natgas	Natural Gas	Natural Gas
Nurshome	Nursing Home	Expenditure on nursing home services
Parts	Car Parts	Parts without services
Telephon	Telecommunications	Telephone and internet services
Printing	Printed Media	Newspapers, books, and periodicals
Tobacco	Tobacco	Cigarettes and other tobacco products
Restrnt	Restaurant	Food & Alcohol purchased outside the home
Tv_Vacation	TV & Vacation	Television service, scenic travel, live entertainment*
Utilities	Utilities	Water, sewage, and waste services

CHAPTER III

Asymmetric Inflation

3.1 Introduction

Most modern theories of monetary economics are built on the long-run neutrality of money: the notion that monetary policy does not have an effect on real variables.[1] This view lends itself to the idea that money is basically homogenous throughout the economy: regardless of where a given dollar is spent, it should effect the economy in the same way. This helps to simplify the mysteries of monetary economics enormously by, for instance, allowing models to ignore locating stocks of money throughout the economy and instead focus on rates associated with the exchange of money (inflation, interest, et cetera). While it is easy to devise examples of localized price fluctuations that are significant in the economy, most ready examples are bubbles, crashes, or some other unstable activity.

But in the quest for non-neutrality, there is a key disadvantage: non-neutral bubbles can only be seen when they burst. Price fluctuations can only be observed by comparing similar goods. Any price with a temporal component baked in can only be compared to its future or past self by using some discount rate that is itself inferred from the market. Therefore, there is no objective statement that time-dependent (usually meaning "asset") prices are incorrect until after the correction. Therefore, if

[1] Even in models that create non-neutrality, that effect is temporary and local.

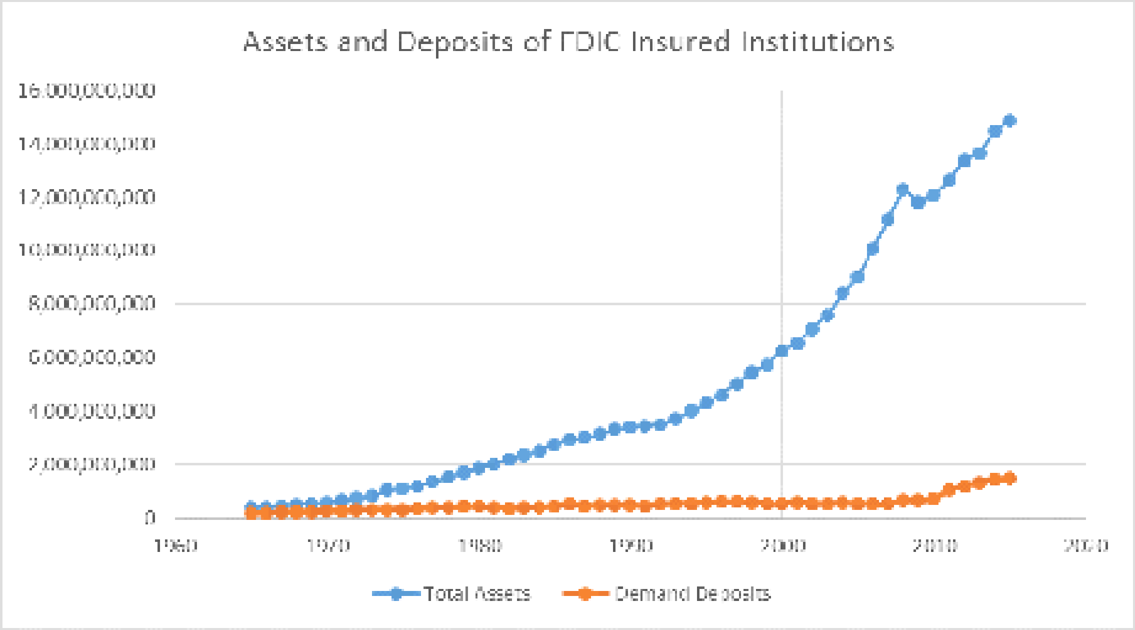


Figure 3.1: Total Bank Assets: FDIC Annual Report 2015

there were to occur a substantial long-run change in asset prices that was somehow stable but not determined by real factors, identifying such a phenomenon would be a largely philosophical question. First, some relevant stylized facts:

Over the past 40 years, the nominal GDP of the United States has grown by five-fold (per capita only four fold). In a similar time frame the asset value of FDIC-insured banks has increased 20-fold, whereas the value of insured deposits over the same time has only increased four-fold. This disparity in growth rates suggests that something has changed in terms of either (i) the desirability of assets or (ii) how cash is allocated for assets versus the consumption of real goods. Meanwhile, over the same time period, as asset value has been increasing the rate of gross capital formation has been flat, and the total stock of real capital in the US has increased approximately two and half fold (following a strikingly linear trend). The US and nearly all industrial economies seems to have experienced an increase in asset prices relative to total growth and real capital stock, and most explanations of this phenomenon relying on economic fundamentals (particularly arguments about changing demographics) seem

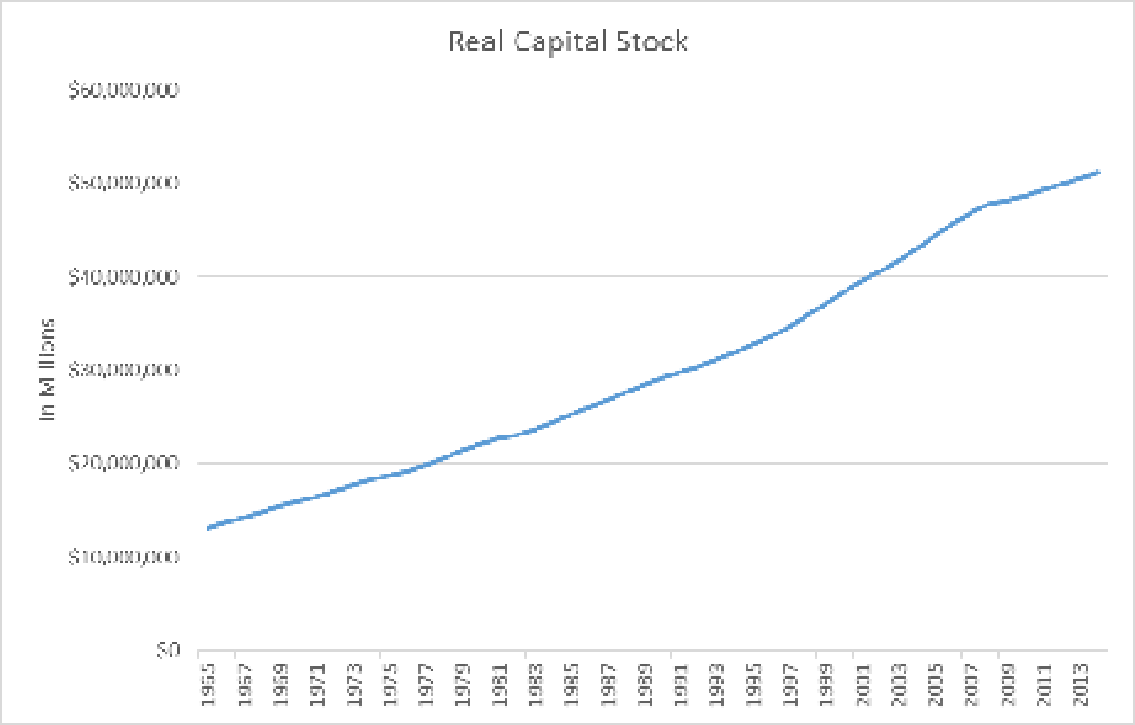


Figure 3.2: US Real Capital Stock: Kansas City Federal Reserve

to explain the data poorly (Chen et al (2007), Chamon & Prasad (2010), and Laibson & Mollerstrom (2010)). Simultaneously, the connection between consumer goods inflation and monetary policy seems to have been growing more tenuous (Mishkin (2009) and Bernanke (2005)).

This paper will examine an alternative possible explanation of this basic phenomenon based on structural features of monetary policy. It argues that monetary policy has different effects on different sectors of the economy, and because monetary policy is made without regard to price sectors outside of the CPI basket of goods, this differential effect can grow over time. The main reason for the differences in the behavior of monetary policy are structural features in the nature of underwriting loans (asset purchases are generally safer and therefore more generously funded than real goods purchases).

In order to explore this phenomenon, this paper will first present a simple formal model, then a more non-standard general model, discuss historical background and practical features, and finally discuss policy measures to solve these issues.

3.2 Illustrative Standard Model

Before proceeding, it is first necessary to illustrate that it is possible for prices in different sectors to grow at different rates for purely structural reasons. That is, the price of one good type might grow at a different rate than another type not because of changing preferences and production, but merely as an artifact of the monetary system. This is a large departure from standard models, in which such a structural deviation in price changes would normally have to be unstable (usually a bubble). So here, before later launching into the more general model at the heart of the paper, is a more conventional model that acheives a deviation in sector-price growth-rates for structural reasons, while remaining in a stable equilibrium.

3.2.1 Setup

Consider a simple economy without uncertainty with three types of infinitely-lived agents: consumers and firms and equity-holders. Let consumers be representative households that act as discounted lifetime utility maximizers subject to fixed identical labor supply. Consumers allocate one unit of labor across two types of employment, production of consumption goods (γ_t^C) and investment (γ_t^I). Consumers are also allowed to borrow (b_t^C), though total borrowing (B_{t-1}^C) must be repaid at the consumer borrowing rate (r_t^C) subject to an amortization function ($d^C\left(B_{t-1}^C, r_t^C\right)$).[2] Consumers distribute their purchases over a continuum of firms, but since firms are identical they will charge identically. This, combined with the fact that the consumer good(s) will be numeraire in this model, means consumer goods' prices are omitted.

[2] The reason for using a place-holder function here, rather than the explicit formula based on the interest rate, is due to the fact the model will break the standard logic on this point shortly.

Because wages will equalize across the two types of employment subject to demand, the representative household solves the following typical utility maximization for consumption and borrowing.

$$\max_{c_t, b_t^C} \left\{ \sum_{t=0}^{\infty} \left[\beta^t U \left(\{ c_t^i \} \right) \right] \right\} \quad s.t. \ \sum_i c_t^i + d^C \left(B_{t-1}^C, r_t^C \right) \leq w_t^c \gamma_t^C + w_t^I \gamma_t^I + b_t^C,$$

$$and \ \gamma_t^C + \gamma_t^I = 1 \ and \ D \left(B_{t-1}^C, b_t^C, r_t^C \right) = B_t^C$$

Firms, on the other hand, will face a slightly less standard problem. They are monopolistic competitors with identical market power.[3] The consumer good will be produced using a Cobb-Douglas production function $(k_t^{i\alpha} (\gamma_{t,i}^C)^{(1-\alpha)})$. Capital will depreciate at a constant rate of δ. The first structural oddity in the model will derive from the fact that new investment (I_t^i) must be financed through borrowing at a fixed amortization described by a function of the stock of borrowing $(d^I \left(B_t^I, r_t^I \right))$. New investment will then be used to create capital stock linearly using labor hired for the purpose $(l_{t,i}^I)$.[4]

These are the functions to describe capital and investment in a given period.

$$\pi_t^i = k_t^{i\alpha} (l_{t,i}^C)^{(1-\alpha)} - w_t l_{t,i}^c - d^I \left(B_{t-1}^I, r_t^I \right) - d^S \left(B_{t-1}^S, r_t^S \right)$$

$$k_t^i = (1 - \delta) k_{t-1}^i + I_t^i$$

$$D \left(B_{(t-1,i)}^I, I_t^i, r_t^I \right) = B_{(t,i)}^I$$

[3] Note how the fact that firms are identical and the fact consumer goods are numeraire again means no price in the profit function. It simply means that firm profits (π_t^i) will be non-zero unlike a perfectly competitive model.

[4] The most subtle trick made in this model with regard to the use of market power in determining prices comes from the fact wages will be less than marginal product. The market inefficiency means that there will be some mark up over marginal cost. However, the fact that the consumer good is numeraire means that this price "wedge" can only be achieved by reducing wages / marginal cost.

$$I_t^i = w_t l_{t,i}^I$$

Taken together with the consumers' problem, discounted present value profit maximization, and a fixed cost of entry to determine the number of firms,[5] this nearly produces a fairly typical macroeconomy merely with unusual notation. However, the large divergence of this example economy compared to more standard models comes because the firm has an unusual optimization problem. In this model the firm optimizes *share price* (P_t^i). Because profit maximizing pricing and labor demand are determined by the other variables, the firm achieves this by choosing two variables / courses of action: either borrowing to invest (I_t^i) in a period or borrowing to purchase the firm's own shares (S_t^i). Assuming that share prices describe net present value prices of future profits divided by shares, the firm solves the following optimization:

$$\max_{S_t^i, I_t^i} P_t^i \ = \frac{\sum_{t=0}^{\infty} \left(1/r_t\right)^t \pi_t^i}{\overline{S}_t^i} \ s.t. \ \overline{S}_t^i = \overline{S}_{t-1}^i - \frac{S_{t-1}^i}{P_{t-1}^i}$$

Just as in the case of consumer or investment borrowing, the stock of share purchase borrowing follows a similar simple law of motion:

$$D\left(B_{(t-1,i)}^S, S_t^i, r_t^S\right) = B_{(t,i)}^S$$

Finally, in order to simplify the equilibrium, equity holders are treated as a distinct type of consumer from standard wage earners who "eat" profits and share price appreciation. Therefore these goods contribute to these passive agents' utility without reentering the economy, similar to how consumption goods vanish each period.[6]

[5] The conspicuous absence of a determinate of firm entry is for simplicity. None of the subsequent analysis makes claim about the size of the economy, and is built on the assumption of ultimate money neutrality, so this aspect of the model is omitted.

[6] The logic of the model does not require anything quite so blunt. It merely requires two-types of agents with some reason for having very different preferences for goods between the two sectors. A theoretically simple way of achieving this is having agents with identical non-homothetic preferences for investment goods and consumption goods and have agents fall into two types of endowment (those with many shares who have proportionately less demand for consumption goods and those with no shares who have proportionately more). While this is simple and intuitive, it has all the

3.2.2 Modeling Borrowing and Monetary Policy

The largest deviation of this model from standard comes in the realm of how the lending market functions. In this model, there is no consumer good inflation, because the currency is fixed in value to the price of the consumer good (*i.e.*, all prices are real). It is, however, actually a nominal model because the need for access to fiat money is explicitly modeled because money must be borrowed to invest or buy shares. Instead of setting interest rates to manage consumer prices, in this model the central bank also acts as the central lender and borrowing-price-setter in the economy. The central bank is constrained by a strict amortization function,[7] and the bank must set rates so as to preserve full employment and keep wages stable.[8] It has access to a printing press and can print as much paper money as needed, but it cannot change consumer goods prices nor create consumer goods, instead only having power over borrowing. However, central bank lending is not completely free, as it must satisfy solvency-motivated boundary conditions so that the leverage of all types of borrowing is finite over time.[9] Stated more formally, the central bank's policy must be such that: consumer borrowing must be bounded by the growth of their wage,

$$\lim_{t\to\infty} B_t^C / w_t \; < \infty$$

investment borrowing must be bounded by the growth in consumer goods sales

$$\lim_{t\to\infty} B_{t,i}^I / c_t^i \; < \infty$$

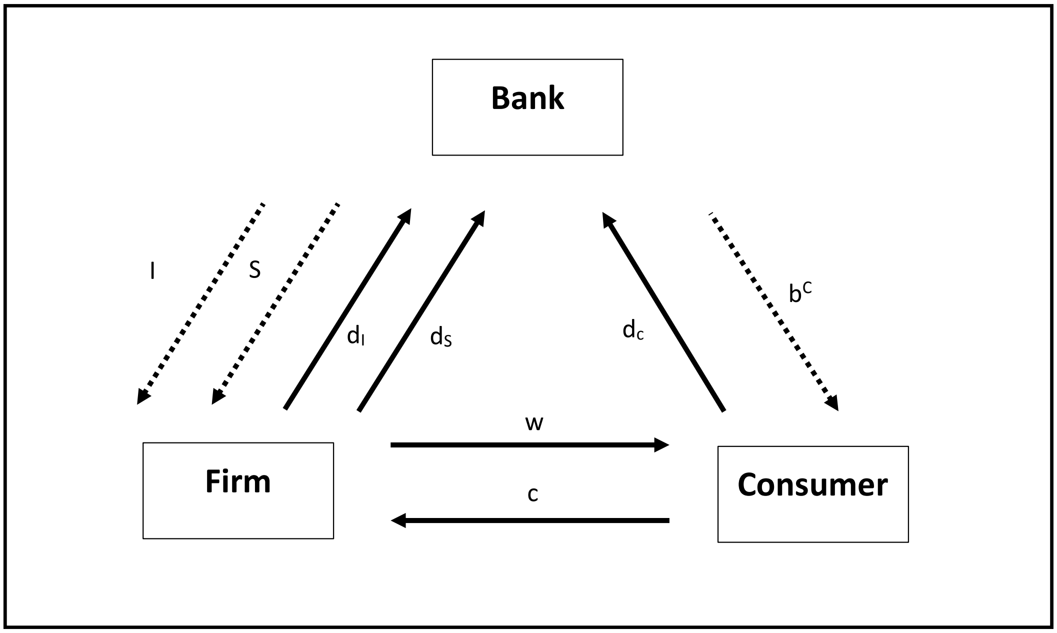
tractability pitfalls that go with having heterogenous agents with non-homothetic preferences.

[7] That is, they cannot tamper with the borrowing repayment functions ($d^X\left(B_{t-1}^X, r_t^X\right)$) except by choosing interest rates (r_t^X).

[8] Because there is no real goods inflation, the other prong of the Federal Reserve mandate is moot.

[9] Perhaps the most salient objection to this putative central bank is that, by setting rates and lending fiat currency in unlimited supply, there is really no market-clearing condition in this economy's bond market. However, because the central bank cannot create goods and must comply with the boundary conditions, as long as the real goods market-clearing conditions hold the bank is free to do as it likes.

Figure 3.3: Diagram of Standard Model Payments



and share price repurchase borrowing must be bound by the equity value of the firm

$$\lim_{t\to\infty} B^S_{t,i}/P^i_t\overline{S}^i_t < \infty.$$

This is further simplified by the fact the economy does not grow in this model, because there is no growth in household labor supply, population, or the number of firms.[10] Because the economy will not ultimately grow over time, this necessitates that household borrowing and investment borrowing be in their steady state solution from the first period onward.

A simple chart of the payment flows of the economy shows that only two types of transaction occur between private optimizing actors. It also shows why there is no easy way to redistribute the imbalance of prices that occurs in the equilibrium solution.

So now it only remains to solve the model. The consumer problem is simple and

[10] Several Kaldor facts do not hold in this model, and the absence of population growth is the reason why.

entirely typical, so all the interesting behavior of the model derives from the firms' share price maximization problem.[11]

$$\mathcal{L} = \frac{\sum_{t=0}^{\infty} \left(1/r_t\right)^t \pi_t^i}{\overline{S}_t^i} + \sum_{t=0}^{\infty} \lambda_t \left(\overline{S}_{t-1}^i - \frac{S_{t-1}^i}{P_{t-1}^i} - \overline{S}_t^i \right) + \sum_{t=0}^{\infty} \mu_t \left(\left(1-\delta\right) k_{t-1}^i + I_t^i - k_t^i \right) \ldots$$

$$\cdots + \sum_{t=0}^{\infty} \rho_t \left(D\left(B_{(t-1,i)}^I, I_t^i, r_t^I \right) - B_{(t,i)}^I \right) + \sum_{t=0}^{\infty} \sigma_t \left(D\left(B_{(t-1,i)}^S, S_t^i, r_t^S \right) - B_{(t,i)}^S \right)$$

Because we have asserted that the stock of capital and output must be in a steady state, the capital investment question ends up being trivial. Therefore we are only left with the share price purchase optimization. The first order conditions of the Lagrangian are as follows.

$$\frac{\delta \mathcal{L}}{\delta S_t^i} = -\lambda_{t+1} 1/P_t^i - \sigma_t D'$$

$$\frac{\delta \mathcal{L}}{\delta \overline{S}_t^i} = -\frac{P_t^i}{\overline{S}_t^i} - \lambda_t + \lambda_{t+1}$$

$$\frac{\delta \mathcal{L}}{\delta P_t^i} = 1 + \lambda_{t+1} S_t^i / {P_t^i}^2$$

$$\frac{\delta \mathcal{L}}{\delta B_t^S} = -\frac{\frac{1}{r_{t+1}}}{\overline{S}_{t+1}^i} d^{S'} - \sigma_t + \sigma_{t+1} D'$$

This gives us a system of four equations and four unknowns. Fixed amortization periods imply that the payment functions are linear in principal and logarithmic in interest rate. Together this implies that:

$$S_t^i = -P_t^2 / \lambda_{t+1} = P_t / \sigma_t D'$$

Because both the shadow price of borrowing (σ_t) and the payment size change from additional borrowing (D') are positive non-zero numbers, this shows that share pur-

[11] Note that dividing by the number of shares has the odd effect of precluding the discounted form of the Lagrangian.

chases will be non-zero and *increasing* in time across periods. So the share price will have to accelerate in order to keep the other variables in steady state, and the change of the real price of shares will be unbounded with time.

The final piece of the puzzle is to examine how each of the three lending rates (r_t^C, r_t^I, r_t^S) are modeled. It is interesting to note that, despite all the modifications at this point, this model still behaves more or less identically to a simple Solow model if all three interest rates are set equal to each other.[12] But the discussion of credit frictions above was directed at just the question of if these types of lending are treated equally. In the real world, these three types of lending are treated extremely differently.

3.2.3 General Implications

Many choices have been made to keep this model as simple as possible, which have implied the results of interest without a standard complete general equilibrium solution. The reason for this is that this model is an illustration of a much larger class of models for which it has proven difficult to provide an overarching standard model. There are three essential characteristics that this model sought to capture:

1. Detach the sale / asset price of a good / firm from its marginal / investment price, meaning the price of something can change without substantially effecting real economic activity. In this model the “detached” good was firm equity because of the different treatment of share buy-backs and further investment.

2. A monetary authority that will produce an arbitrarily large supply of money to stabilize general real parameters but only the nominal parameters of certain sectors / goods. In this case, the central authority set lending rates to stabilize capital stocks and real wages, without regard to share prices.

[12] I take this as evidence that the many simplifications and modifications have not broken the basic logic of a standard macroeconomic model.

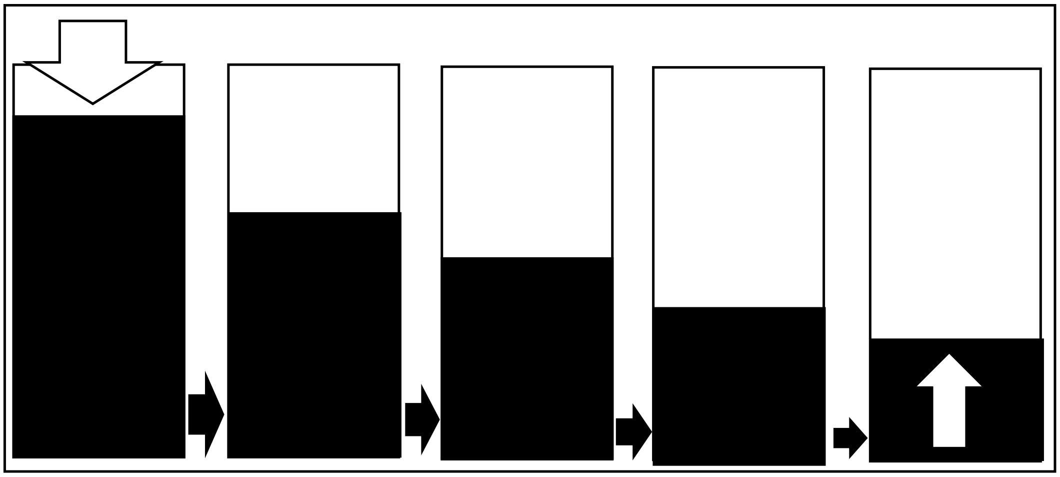
3. Some friction must prevent or slow equalization of untargeted parameters across sectors. This is done in this model by simply having consumers and shareholders be different types of agent who relied on each other for nothing, but this issue will be expanded more deeply later in the paper.

3.3 Abstract General Model

Consider a crude economy composed of many (in the illustration, five) different sectors of economic activity (shown below as vertical white rectangles). The amount of real economic activity in each sector is represented by the area of the entire rectangle, and it is treated as exogenous and uniform across sectors. In this economy, new fiat money is added to the system at a certain rate to produce a targeted level of inflation. The amount of money available in each sector is represented by the volume of the black rectangle (so the implicit "price level" would be the ratio of the areas of the smaller black rectangle divided by that of larger rectangel including the white portion). But there is a complication: new money (represented by white arrows) can be added in only one sector (the far left, in the example below), and the target inflation rate is measured in only one sector (vertical arrow at the far right). Critically, there are inefficiencies in how newly created currency travels between each of the relevant sectors (flows shown with black arrows), implying that less-than-equal shares of the new money arrive at each. Like water flowing between beakers through varying sized holes, the resulting system produces different price levels in different beakers.

This section will examine how this crude intuition might apply when looking at the systems governing monetary policy. Simply put, extant systems of monetary policy function by regulating the rate of endogenous money creation in asset markets, which is not measured directly in the aggregate or in terms of price level. And this rate is allowed to increase at whatever rate is necessary in order to produce a desired level of inflation in the sectors that produce the basket of goods used for the CPI. If there

Figure 3.4: Illustration of Concept of General Model



exist inefficiencies in how money flows between these sectors, is it possible for different persistent and *stable* levels of inflation to exist in these different sectors?

3.3.1 Background - the Unmeasurable Concept of Asset Price Inflation

Discussion of this growth in the asset value of the United States, and the other major industrial economies, has increased in prominence over the last 15 years. In the run up to the 2008 crisis, issues surrounding this trend and Bernanke's "global savings glut" led several researchers to examine the issue of "asset price inflation". Swirling around in this debate were two separate but related questions: (i) is it possible to identify a bubble ex ante and (ii) is it possible to measure the deviation of an asset's price away from the "true" value of the asset. The answer to both questions was ultimately the same:[13] there is no way to deduce the "correct" value of an asset as it is based on the future performance of that asset which is not yet revealed. As Greenspan was noted to say, "a bubble can only be detected when it bursts".

One conspicuous feature of the discussions swirling around at this time, though, was that the causal features examined in the debate were mostly rooted in the real,

[13] One facet of this discussion that is rarely examined is the possibility that these two questions need not be linked. It is important to note that this paper will examine a type of asset price inflation in the absence of a bubble.

rather than monetary, economy. Whether it was changes in savings behavior across countries or within countries based on demographic trends, or the borrowing behavior of national governments, there was virtually no discussion of the possibility it was rooted in the management of the money supply.

3.3.2 A Short History of Germane Monetary Systems

In its early days, the Federal Reserve existed primarily as a lender of last resort, in order to shore up the banking system and end the constant banking crises that had ravaged the US economy throughout the 19^{th} Century. Beginning in the 1920's, though, the bank moved away from thinking about the currency as merely a banking issue, and began conducting open market operations, direct interventions in the markets using its unlimited supply of fiat money, in order to influence interest rates in the entire economy. The Banking Act of 1935 formalized this process, and created the Federal Open Market Committee (FOMC) which was tasked with conducting open market operations in order to achieve targeted interest rates in the entire economy. The specific structure of these operations changed as markets transformed and grew more sophisticated, but the basic method was always the same: financed through seigniorage, the Federal Reserve would (i) use cash to buy assets when interest rates were too high or (ii) sell assets for cash when interest rates were too low in order to achieve the desired market rate.

Another part of the Federal Reserve's responsibilities at this time, in conjunction with its target interest rates, was managing the gold-convertibility or "gold coverage" of US currency. Prior to 1933, all dollars had explicit gold values and could be converted by private citizens into physical gold. When the Great Depression made it clear that having more flexible monetary policy was both valuable and necessary, convertibility by private citizens was ended but was left available to foreign countries in order to manage the exchange rate.[14] But throughout this time, there existed more paper money than the Federal Reserve held gold to back it, and like a retail bank managing the ratio of demand deposits to cash-on-hand, the "coverage" of the money in circulation was an important consideration in policy. Therefore, whenever making decisions about target rates and open market operations, the Federal Reserve was forced to consider the impact of creating new fiat money on their ability to convert the currency.

The demise of dollar-gold convertibility and the Bretton-Woods system from 1968 (the London Gold Pool Collapse) to 1971 (the Nixon Shock) thrust a major policy challenge on nearly all of the governments of industrialized countries at the same time. Because the Bretton-Woods system was built on gold, there was an explicit link between the exchange rates of different countries and the value of a specific investment good (gold). Of course, in many ways this connection was only notional, as the only counterparties it was available to were central banks and, at least late in the life of the system, all the major participants knew the valuation or conversion rate was totally unrealistic.[15] But however imperfect, the link between gold and money supply had two implications: it necessitated capital controls on financial flows between countries[16] and it imposed a clear link from the nominal money supply to the nominal asset supply. For all system members aside from the US, there existed a different but analogous difficulty: national currencies were backed by that country's supply of US dollars. So similarly, each of those nations' central banks needed to

[14] Intriguingly enough, the $35 per troy ounce conversion rate is actually dated to 1934. Bretton Woods kept the same exchange rate simply out of practical and political convenience.

[15] The acrimony and poor behavior associated with the late life of the Bretton Woods system is legendary, but it should be noted that several participants were not entirely respectful of the weakness or value of the system. France routinely converted huge dollar holdings, Germany was willing to defer conversion but not revalue the currencies, and earlier in the week that Nixon ended dollar conversion Britain shocked the Federal Reserve of New York by asking for $2 billion in gold. In the opinion of the author, the parallels between the policy behavior of the various countries at that time and during the Euro crisis today are striking and not discussed often enough.

[16] It should be noted that international capital controls had been incrementally relaxed since the 1950's, but still existed in several forms.

consider the money supply against their stock of dollars.

In the wake of the system's demise, the pressures on the policy challenge facing the Federal Reserve, and the central banks of all the western economies, changed markedly. Almost every economy in the Bretton Woods system had previously followed the United States in, as informed by the classic trilemma, (i) targeting interest rates, (ii) fixing exchange rates, and (iii) restricting international capital flows. The end of convertibility left each of those countries scrambling to find a new method of managing their money supply to achieve the common objective of stable prices and interest rates. And, as before Bretton Woods, nearly every western economy converged on the same general policy solution: (i) continued targeted interest rates with a new focus on inflation and price stability, (ii) floating exchange rates, and (iii) free flow of international capital.[17] However, there was another transformation in all these banking systems that is ultimately not addressed by their policy trilemma: now that they are liberated from having to back their currency with dollars or gold, the only consequence of a newly printed note is its effect on (i) consumer prices, (ii) interest rates, or (iii) exchange rates. There is no hypothetically "correct" volume of currency in circulation; the only relevant feature is how the extant money supply effects the relevant economic indicators. The reason this will prove vital in this section is: if a specific monetary system actually caused a major transformation in the money-supply allocation to non-consumer goods, and therefore demanded very large increases in overall money supply to maintain target inflation, this would have been noted as a cause for concern under a system that needed to consider the "coverage" of the currency by a fixed resource. In a post-Bretton Woods setting this would not be an issue.[18]

[17] This elides over Europe's many and repeated efforts to coordinate exchange rates on the continent. These are ultimately not relevant to the central point because none of those systems called on the central banks to consider anything resembling "coverage" again.

[18] One piece of evidence of this change in thinking is the decreasing attention paid to money supply by modern central bankers. The Federal Reserve stopped collecting data on M3, which was the smallest measure that included the vital repurchase agreement market, in 2005.

3.3.3 Quantitative Easing and the Missing Inflation

The creation of real capital / investment has not increased in proportion with asset prices. It is actually quite difficult to establish a theoretical justification as to why this should be. Newly created fiat money should, in most models, be distributed, through standard credit provision and bank activity, across the range of economic activities in the economy. Banks should not bias their loans toward any given segment of economic activity because if any one sector became inflated relative to the others the returns to capital in that sector would fall and the system would equalize.

In recent years, this does not seem to be consistent with what has actually taken place. In 2008, the Federal Reserve quadrupled the size of its balance sheet and undertook an aggressive program of asset purchasing ("quantitative easing") in order to stave off the effects of the financial crisis. Simultaneously, the Federal Government undertook a $750 billion dollar asset purchase program financed by deficit spending. Both of these activities, rather than being inflationary, came during a sustained period of near-zero inflation or in some cases even deflation. So rather than equalizing the distribution of new money across sectors, it seems that during this period that almost all of it was sucked into the banks.

The widely accepted explanation for this phenomenon is that the banks, because they were suffering a panic and were saddled with large stocks of mis-valued assets, needed to absorb cash in order to be healthy. But there are two interesting caveats to this: first, the fact that there can be sector-specific money absorption trends is proof that the "pathways" by which money is distributed between sectors are not flat or uniform. And second, quantitative easing continued on a massive scale until six years after the crisis, even after the banks had become well-capitalized and asset markets were healthy and growing. The program ended in 2014 after accumulating $4.5 trillion in assets without any appreciable impact on inflation.

Taken together, these two facts alone suggest that the process of distributing newly created money through the economy is not trivially simple, and that the classic model of equalizing returns leading to uniformity across sectors may not be true.

3.3.4 Endogenous Money Creation Among Financial Institutions

A well-established theory of inflation and money creation is “endogenous money” by which, in the words of Keynes, “loans create deposits”. The textbook theory was that a certain given quantity of hard cash would be lent and deposited back and forth between consumers and banks to produce a much larger “effective” quantity of money in the economy. This phenomenon still occurs, despite the fact it has become enormously more complicated, in lending *between* financial institutions. A key feature of this process to note for the purpose of this paper, though, is that there is no particular reason why the money creation needs to occur only between consumers and banks. When banks lend to other banks or financial institutions, the borrowing party then has a pool of money with which to purchase or lend to other financial institutions (as well as lend to consumers). So when a bank lends to a financial institution, this could cycle through the financial system and lead to the bank having effectively lower costs of borrowing or higher share prices, which in turn will provide the lending bank with more money to lend than before. This process is, in fact, the cornerstone of monetary policy management, because the target lending rate focuses only on lending between banks.

This will ultimately prove relevant, because the rate of endogenous money creation should not necessarily decrease if a bank’s portfolio takes on an increasingly interbank character. It may even be possible that this would *increase* the rate of money creation, if the cycle of lending between banks was more highly leveraged and had a higher velocity.

3.3.5 The “Cash” and “Asset” Equivalence

When discussing the quantity of “money” endogenously created in the banking system, the introductory textbook story of the leverage of cash dollars begins to break down in the modern system. Under the modern system, the ability of banks and individuals to obtain liquid cash is limited only by their supply of either (i) demand deposits, (ii) assets that another party will purchase at market rates, and, in the case of some banks, (iii) certain US Treasuries which can be converted to cash through the Federal Reserve System. Now that transactions are virtually always entirely instant and electronic, it is extremely rare for a transaction to require the presentation of physical money. In fact, there are no physical money-holding requirements for banks save those attached to only their demand deposits. Therefore, cash for financial institutions exists primarily in digital form in their accounts, that of their lenders, and ultimately in the accounts of the Federal Reserve Member bank(s) that supplied them.

Because modern “cash” for banks is primarily digital, and transactions are concluded so swiftly, a very small amount can provide liquidity to an enormous number of transactions. So if a bank purchases a $100 asset, then sells it to a counterparty minutes later, it can repeat this process hundreds of times in a day while never having more than $100 notional dollars on its balance sheet. While it is theoretically possible during periods of major upheaval (such as in 2008) for this conversion process to be so greatly strained that instant, electronic conversion breaks down, this has rarely been an issue. Since the model presented in this paper is deliberately intended to describe an incremental, ongoing processes, crisis scenarios are disregarded. Therefore, for the purpose of this paper, the distinction between “cash” and “assets” will simply be disregarded. Both for interbank lending and lending to consumers, assets will be able to be converted into cash indifferently. This assumption may at first brush be controversial, but in practical terms it is almost impossible to think of a transaction in which a modern financial institution is constrained by a lack of cash rather than free unleveraged assets.[19] (It is interesting to note, that this is probably the point of greatest divergence with the Bretton Woods environment.)

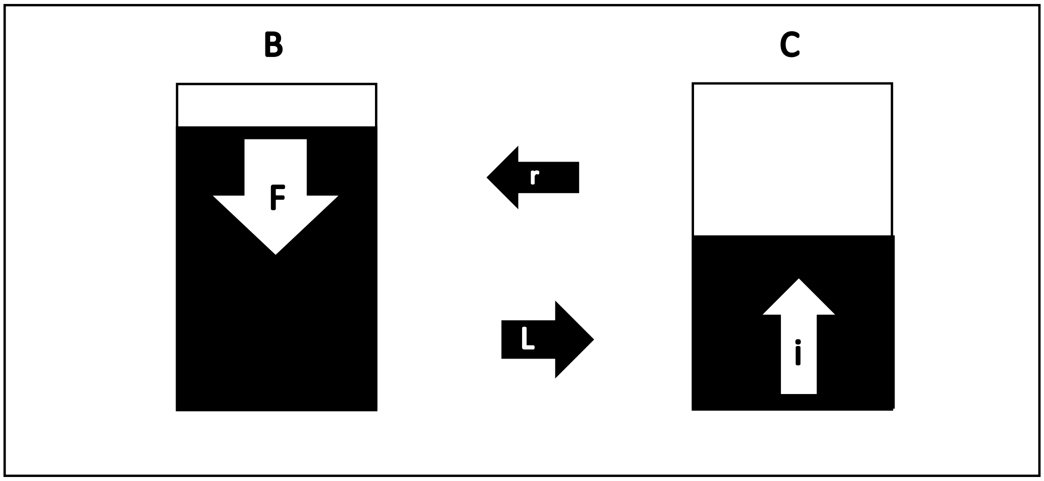
3.3.6 For Firms or Individuals Assets Are a Luxury Good

Another contributing factor in the inability of sectors to equalize their money-to-goods levels could be wealth effects. Therefore, it is worthwhile to note that assets are effectively luxury goods and furthermore, because they are primarily purchased in order to earn money later, if they were included in utility they would enter quasi-linearly. This is important to the modeling exercise, because it suggests that, as income levels of participants vary with sector level prices, there could be wealth effects in how they allocate their spending across sectors. One of the basic arguments against sector-specific inflation rates is that, as one sector grows inflated, substitution effects will reduce the funds flowing to it. Therefore, it is important to note that wealth effects also exist and could overcome the substitution effects.[20] Similarly, if the allocation is being made at the firm level, then the objective is profit maximization and there are no substitution effects (unless specified by the production function). When firms increase their profit, they either provide all of the windfall as dividends to their shareholders or they invest in expanding their business. Such investment would increase asset holding for the firm relative to other expenditures, again producing a behavior akin to luxury good wealth effect, in which purchase of assets disproportionately increases with income.

[19] One example of this is that financial institutions are often desperate to avoid holding cash on a short term basis, as evidenced by the speed and size of the "repurchase" market in which firms sell and repurchase low yield high security treasuries on an overnight basis for a tiny fraction of a percentage point of return. The effective return in this market is far below the risk free return, but it is simply a method of gaining some return on cash that would otherwise sit still overnight.

[20] This is analogous to, but not the same as, the old "stocks are a Giffen good" argument.

Figure 3.5: Illustration of Simple Model of Asymmetric Inflation



3.4 A Simple Model of Asymmetric Inflation

A simple motivating question: when a new "cash" dollar is created in, or given to, a bank, where does that dollar go? As a bank's balance sheet grows, it must choose between three broad categories of lending: (i) consumer loans, (ii) commercial loans, and (iii) lend money for transactions for the purchase of already existing assets. Consumer loans primarily consist of short term credit, student loans, and mortgages. Commercial loans, which are not for financial assets, are business loans used to undertake new economic activity, like hire workers, build facilities, or buy physical capital like machines. The final category, "transaction loans", are used to buy an existing asset, be that real estate, a trademark, or shares of stock. The central assumption of this paper is that these transaction loans do not contribute *directly* to consumer goods inflation but do cause asset price inflation.

To produce our simple model, let's define two sectors of activity: a banking sector representing a fixed set of assets such as real estate and shares of stock (B), and a consumer goods sector (C) which represents labor.

In this arrangement, the banking sector provides paper money to the consumer goods sector and the consumer sector pays debt service to the banking sector. The consumer sector (C) grows at the constant rate g_y. In this system, paper money is provided to the consumer goods sector as loans (L) and service on these loans is paid by the consumers (r). Central bank policy makers calibrate the input flow of new money (F) so that the money supply (M_j where $j = B\ or\ C$) produces a price level for consumer goods ($P_C = M_C/C$) that grows at a targeted rate of inflation (i). Money will be handed from sector to sector through an exogenous process based on the money supply in that sector, so $L = \alpha M_B$ and $r = \delta M_C$, such that $\delta,\ \alpha > 0$ and $1 > \delta,\ \alpha$.

If this system is in steady state equilibrium, both sectors must have identical growth rates. The first step in solving this problem is to recognize that the money supply in sector C in period t is implied by smooth constant growth at rate $g_y + i$ meaning

$$M_C = (1 + g_y + i)^t \mu_C = (1 + m)^t \mu_C$$

where $m = g_y + i$, which is the nominal growth rate of GDP, and μ_C is the initial money supply in sector C. Because the growth rate must be equal across sectors in steady state, it must be that $M_B = (1 + m)^t \mu_B$. This means that the growth, period on period of M$_C$ is $\Delta M_C = m\,(1 + m)^{t-1} \mu_C = L - r = \alpha M_B^{t-1} - \delta M_C^{t-1}$, where M_j^{t-1} is the money supply of sector j in the previous period. Substituting the functional forms for M_B^{t-1} and M_C^{t-1} this gives us that $(m + \delta)\,\mu_C = \alpha \mu_B$ implying that

$$\mu_C = \frac{\alpha}{m + \delta} \mu_B$$

Therefore, the ratio of the money supplies in the two sectors, if they are growing at identical rates, is fully determined by the transfer rates between sectors and the nominal rate of GDP growth. If it is true that transfer rate between banking sector and the consumer goods sector has a friction or inefficiency, this would imply α would be quite small, and certainly that $\alpha < \delta$ which would necessitate that the money supply in the consumer goods sector was, in all periods, lower than in the banking sector.[21]

The simplicity of the steady state solution makes it easy to compare new equilibria if the parameter values were to change over time. For instance, if the consumer lending rate (α) were to *decrease* over time (as it has empirically), this would cause the steady state M_B to increase relative to M_C. Similarly, an *increase* in real growth (g_y), target inflation (i), or the credit repayment rate (δ) would also cause a relative increase in M_B.

3.4.1 Dynamic Response Model

This very simple model can be used to develop a deeper and more nuanced dynamic response model. As this paper is still exploratory in nature, the dynamic model does not seek to be fully endogenized and the "policy" function mechanism is much more crude than standard monetary policy macroeconomic models. The total absence of the money supply in growth or investment decisions, or optimizing agents inside the model, are issues that would need to be addressed in further work. But the basic intuition of the policy problem it will examine is fundamental and informative to the issues addressed in this paper.

Going back to the simple two-sector constant-growth problem, consider in more detail the funding allocation problem. The central bank / policy maker seeks to provide cash to the banking sector at a rate necessary to achieve the desired rate of consumer goods inflation. To examine this, it will be necessary to define some policy function that the central bank uses to compute the correct F^t.[22] Also, in keeping with the discussion above, central bankers will only have the power to observe the money supply in the consumer goods sector (M_C^t) but be ignorant of the money supply in

[21] Here it should be noted that the constant growth assumption for M_C necessitates that the growth rates of the two sectors must be equal. This is sketched out briefly in Appendix 1.

[22] Because timing will now be much more important, this model will use the notation M_j^t and F^t to denote specific time periods. The exogenous parameters do not need a time denotation.

the banking sector. From the reasoning in the static model, we know that

$$M_C^t = (1-\delta)\, M_C^{t-1} + \alpha M_B^{t-1} \quad \& \quad M_B^t = (1-\alpha)\, M_B^{t-1} + \delta M_C^{t-1} + F^t$$

under the transfer scheme of the simple model.[23] Using these equations, and focusing on the steady state, it is possible to derive a policy function for selecting the optimal choice of F^t given knowledge only of M_C^t, α, δ, and m, which is

$$F^t = \left[\frac{(m+\delta)(m+\alpha)}{\alpha} - \delta\right] M_C^t$$

and does not require knowledge of M_B^t which is treated as unobservable.

When considering the form that systemic shocks can take, it is next important to note which sorts of shocks are relevant. Since we are looking at a banking and monetary system, the most relevant shocks are: (i) sharp decreases in α, corresponding to an abrupt decrease in consumer credit and (ii) decreases in δ, corresponding to a decrease in repayment by consumers.[24] Increases in these two parameters are not considered, because increases in debt repayment by consumers or the provision of credit for the purchase of consumer goods rarely occur on a short timescale.[25] Because growth and inflation targets are not endogenized, changes to m would not be informative.

In order to understand the responses to these two types of shocks in the short run, simply consider the partial derivative of the policy function with respect to of each:

$$\frac{dF^t}{d\delta} = \left(\frac{m}{\alpha}\right), \quad \frac{dF^t}{d\alpha} = -\left(\frac{m}{\alpha}\right)\left(\frac{m+\delta}{\alpha}\right), \qquad \frac{dF^t}{d\delta d\alpha} = \frac{-m}{\alpha^2}$$

[23] This also crystalizes the payment timing which was left ambiguous in the earlier part.

[24] The fact that decreases in consumer repayment usually in reality necessitate large increases in the provision of new money to banks is also not modeled here, because in the current model "leverage" inside banks does not exist and therefore does not need to be addressed by policy makers.

[25] Note that increased credit for the purchase of real estate or assets does happen quickly sometimes, but these are asset purchases and therefore not part of the consumer goods category.

The first two derivatives are intuitive; the size of the funding stream to the banking sector must decrease if consumer repayment slows (δ decrease) because money is not exiting the consumer sector. If credit provision by banks slows (α decrease), then funding must increase to overcome the restriction in funding flow to consumers. What is noteworthy, though, is the relative size of the shocks. For simplicity at this stage, assume that initially $\alpha = \delta$. Given an equivalent shocks to α or δ, the magnitude of the change to the funding stream will be larger for the α shock. Similarly, the cross derivative tells us that if equivalent negative shocks happen to both parameters at the same time, the net effect will also be a proportional increase in the funding rate.

In either case, the shocks will slowly be absorbed and the system will return to the steady state defined by the parameters as described in the previous section. But this shows how, because the policy makers add funding in one sector and measure its effects in another with limited information, the policy making process will tend to maintain higher money supply in the banking sector simply due to lag inherent in their intervention.

3.4.2 Frictions

In the above model, the root cause of the inefficiency of the movement of money between assets and consumer goods is not generated endogenously. But there are several simple and intuitive explanations as to why such an inefficiency would occur.

If a sector experiences an increase in sector-specific money supply, there are several reasons funds might have trouble leaving that sector. The most obvious source of friction in the transfer of increased funds is the organizational cost of finding new counterparties in other sectors. In the specific example of new money provided to the asset market, consider if new reliable counterparties in the borrowing market can be found at the same rate. Quantitative Easing is an excellent example of this: in an unstable market in which consumers are already overleveraged and the economy is contracting, if banks are provided with trillions of new dollars, how will they allocate them to the consumer sector? The answer is, they won't. The ability of borrowers to take on debt is proportional to their ability to repay it, which fluctuates far more slowly than endogenous money creation.

Another source of friction may be transaction costs relative to transaction size. In the asset market, consumer credit involves very large administrative and management costs (*e.g.*, checking credit worthiness, providing disclosures, seeking legal enforcement) relative to the amount of money borrowed. On the other hand, the asset market can absorb millions, even billions, in minutes without even necessarily requiring any of the parties to sign a new contract and in which monitoring and enforcement costs are trivial compared to the sums involved.

A related friction, which differs from the preceding two, is the possibility that different sectors may naturally be more and less responsive to economic shocks. For instance, wages are historically considered "sticky" so a positive economic shock will take longer to translate into an increase in credit worthiness for borrowers than for firms that benefit from the shock.

3.5 Implications

3.5.1 Asymmetric Inflation Is Not a Bubble

It is important to note that the ongoing divergence in the nominal cost of different categories of goods is fundamentally different than a "bubble" existing in one or the other of those categories. In order for a bubble to exist, it must be possible for the said bubble to "burst", meaning that the relative prices would adjust back to a "correct" level. In the model laid out herein, no aspect of the system is unstable over time. The desire to hold assets does not change, regardless of how much their nominal value goes up, because the feedback process that causes them to increase in value will not change.

To explore this notion further, consider the case that asymmetric inflation drives one sectors value up an enormous number of times relative to the other sectors. The friction that has driven these prices is not variable, and will continue to be present in future periods. The underlying policy of adding new money until the target inflation rate is achieved will continue. And the interplay of substitution and wealth effects that allowed the asymmetry to arise will continue to be present. So in this model, there is no event in which agents decide to abruptly extract or move their funds.

To put this in practical context, consider a highly inflated asset market. In order for the system to be in equilibrium, those who earn in the asset market currently move money in and out of that market according to their preferences. Now suppose that market experiences a panic, or a bubble bursts, due to activity unrelated to asymmetric inflation. The value of assets plummets, the earnings from this market fall enormously, and a great deal of household income vanishes. But the question arises: does this precipitate a mass migration of money supply *out* of the asset market? From real world observation, this is never what occurs. During a crisis, the risk-averse financial firms tend to absorb money *out* of the consumer goods markets to stabilize their balance sheet. The actual money supply in the asset market tends to *increase* due to negative shocks. And, thinking at the consumer level, economic uncertainty virtually never leads to a *decrease* in savings rate.

3.5.2 Testability and Implications

One of the idiosyncratic features of this model is that, for all the reasons asset inflation is impossible to measure directly, asymmetric inflation is likewise nearly impossible to observe. Because, as with asset prices, the relative price levels in different sectors of the economy *should* transform over time and there is no theory to identify "correct" relative price level difference or rate of change. So this model is very difficult to test empirically, and the author has not yet devised a method.

Therefore, if managing the monetary system by creating new currency exclusively through the banking system is, in fact, creating distortions in price levels across sectors, the first question to ask is: would that be a problem? Nothing about this analysis suggests that this distortion is inherently unstable (*i.e.*, no bubbles), inflation, interest rates, and exchange rates are stable, and the economy is liquid and healthy. Why would the existence of this process be an issue? The comedian Henry Youngman was once asked, "how's your wife" and he replied "compared to what?" The question revolves around counterfactuals.

The two broad categories of alternatives are (i) a monetary system that somehow fixes or constrains the growth of money supply but continues to create new money primarily through the banks or (ii) a monetary system that continues not to regard aggregate money supply and instead distributes new fiat money across sectors. Narrowing back to the two sector, asset vs consumer goods example, when compared to either of the counterfactuals, those who earn through the asset sector (the money creation sector) would lose personally. In the first counterfactual, the asset earners would still be advantaged by the distortion, but new (or old) policy mechanisms would need to be introduced to improve the transfer of new money to the consumer good sector. The specifics of those policies could have widely different welfare implications. Alternatively, in the second counterfactual asset earners would unambiguously lose. But because the models set forth above do not allow for dynamic welfare, it is impossible to say which leads to overall better outcomes. If the ultimate goal of the asset market is to provide (i) a stable currency, (ii) allow for savings, and (iii) allocate capital for new economic activity, it is not at all possible based on the theoretical apparatus thus far to say which system is better.

However, one important caveat in this comparison is that the growing disparity is stable and persistent. Therefore, the income disparity produced by the system can only grow wider. So, moving out of the realm of economics into the field of political economy, if there are social or political costs to large-scale lack of equality, then either of the proposed counterfactuals would be preferred after some span of time.

3.6 Policy Implications & Conclusion

In order to prevent new money from being asymmetrically fed into assets, the process of money supply management would need to be modified to regulate the rate of money creation nearer to the sectors where inflation is observed. The simplest way to do this would be to, rather than manage money creation in banks, create new money in the real economy. An example might be to simply send checks for quantities of fiat money to ordinary households. Rather than relying on a lending process to be distributed, this new money would be spent on real goods almost immediately. Banks would extract new money out of the consumer sector based on factors like interest rates and disposable income.

If combined with the existing method of money creation through the banking system, this could serve to equalize the relative inflation rates across sectors and reach targeted inflation rates with less aggregate created money. Regardless, owing to the impossibility of measuring asset price inflation, it is still impossible to know what the relative gains might be without actual experimentation.

Regardless of the feasibility of any of these policy steps, the above models outline an alternative explanation of complex recent monetary trends based on relatively simple alterations of standard monetary theory. Much of our understanding and policy in monetary economics is based largely on theory and difficult to validate due to the paucity of experiments. But it is important to take time to consider our unexamined assumptions and the scope of their implications.

APPENDICES

APPENDIX A

Yatchew Regression

Yatchew partially linear smoothing is a way of adjusting a plot for a given set of linear controls. The model estimates a nonparametric function ($f^{\,j}$), relating individual income or wealth (I) to good specific expenditure (g^j spent on good j) subject to a set of linear controls (vector X_i) with an error term (ε).

$$g_i^j = \beta^j * X_i + \ f^j\left(I_i\right) + \ \epsilon_i^j$$

The algorithm is as follows: (i) the data are sorted in order in terms of the independent (horizontal axis) variable (I), (ii) then the dependent variable (g^j) is differenced between each observation in sorted order, (iii) the linear control variables (X_i) are similarly differenced between observations, (iv) then a ordinary least squares model is run on the differenced variables, and (v) finally the residuals of this regression (e^{OLS})[1] are then combined in a moving sum to approximate the nonparametric function.

$$\Delta g_i^j = g_i^j - g_{i-1}^j$$

$$\Delta X_i = X_i - X_{i-1}$$

[1] It is important to note that eOLS is distinct from the unobservable “true” error ε. In this case eOLS is the estimation residual for each observation produced by ordinary least squares.

$$\Delta g_i^j = \ \beta^j * \Delta X_i \ + \left(f^j\left(I_i\right) - \ f^j\left(I_{i-1}\right)\right) + \left(\epsilon_i^j - \epsilon_{i-1}^j\right)$$

The key insight is that, by differencing the two equations, and utilizing the fact that E$[\varepsilon_i - \varepsilon_{i-1}]$ converges to zero even more quickly than the undifferenced errors, the residuals from the regression describe the function $f^{\,j}$ more efficiently. Using the regression residuals (e^{OLS}) plotting $f^{\,j}$ is a straightforward of summing residuals.[2]

$$\Delta f_i^j = f^j\left(I_i\right) - \ f^j\left(I_{i-1}\right) = e_i^{OLS}$$

$$f^j\left(I_i\right) = \sum_{k=2}^{i} \Delta f_k^j = \sum_{k=2}^{i} e_k^{OLS}$$

The fact that the data are ordered then differenced places strict requirements of continuity on the independent variable. If the independent variable only assumes a few values, and many duplicates occur, then the estimate might be dependent on how these ordering “ties” are broken. Luckily, the data used in this paper are very continuous, save for a mass-point at zero. Different solutions to this problem exist in the literature, such as bootstrapping over alternative sorting of mass-points or using a parametric method to order based on the control covariates. An alternative specification was examined in this paper, where *any* value of the independent variable occurring more than once in data was dropped, and it had no discernible effect on the Engel curve estimates.

[2] Note that the summand begins at k=2 because ΔI1 is not defined.

BIBLIOGRAPHY

Ahn, JaeBin, Mary Amiti and David E. Weinstein 2011 Trade Finance and the Great Trade Collapse. *American Economic Review: Papers & Proceedings* 101:3, 298–302.

Alessi, Lucia, and Carsten Detken. "Quasi real time early warning indicators for costly asset price boom/bust cycles: A role for global liquidity." *European Journal of Political Economy* 27.3 (2011): 520-533.

Anderson, James E., and Eric Van Wincoop. "Gravity with gravitas: a solution to the border puzzle." *The American Economic Review* 93.1 (2003): 170-192.

James E. Anderson; Eric van Wincoop "Trade Costs" *Journal of Economic Literature*, Vol. 42, No. 3. (Sep., 2004), pp. 691-751

James E. Anderson, 2011. "The Gravity Model," *Annual Review of Economics, Annual Reviews*, vol. 3, pages 133-160

Amiti, Mary and David Weinstein 2009. "Exports and Financial Shocks." *NBER Working Paper.*

Armington, Paul S. "A theory of demand for products distinguished by place of production." Staff Papers 16.1 (1969): 159-178.

Baier, Scott L., and Jeffrey H. Bergstrand. "Bonus vetus OLS: A simple method for approximating international trade-cost effects using the gravity equation." *Journal of International Economics* 77.1 (2009): 77-85.

Baldwin, Richard. "Heterogeneous firms and trade: testable and untestable properties of the Melitz model." No. w11471. National Bureau of Economic Research, 2005.

Balistreri, Edward J., and Hillberry Russell H. "Trade Frictions and Welfare in the Gravity Model: How Much of the Iceberg Melts?" *The Canadian Journal of Economics / Revue Canadienne D'Economique* 39, no. 1 (2006): 247-65.

Berentsen, Aleksander, Gabriele Camera, and Christopher Waller. "The distribution of money balances and the nonneutrality of money." *International Economic Review* 46.2 (2005): 465-487.

Bos, Len and Adonis Yatchew 1997. Nonparametric Least Squares Estimation and Testing of Economic Models, *Journal of Quantitative Economics,* 13, 81-131.

Bernanke, Ben S. *The global saving glut and the US current account deficit.* No. 77. 2005.

Berthou, Antoine and Charlotte Emlinger, 2010. Crises and the Collapse of World trade: The shift to Lower Quality. *CEPII Working Paper* 2010-07.

Borio, Claudio, and Haibin Zhu. "Capital regulation, risk-taking and monetary policy: a missing link in the transmission mechanism?." *Journal of Financial Stability* 8.4 (2012): 236-251.

Boivin, Jean, Marc P. Giannoni, and Ilian Mihov. "Sticky prices and monetary policy: Evidence from disaggregated US data." *The American Economic Review* 99.1 (2009): 350-384.

Broda, Christian and John Ramalis 2008. Inequality and Prices: Does China Benefit the Poor in America? *NBER Working Paper*. (**Retracted**)

Broda Christian and David Weinstein 2006. Globalization and the Gains from Variety. QJE 121.

Case, Karl, John Quigley, and Robert Shiller 2005. Comparing Wealth Effects: Stock Market Versus the Housing Market. *IBER Working Paper* 01-004.

Chamon, Marcos D., and Eswar S. Prasad. "Why are saving rates of urban households in China rising?." *American Economic Journal: Macroeconomics*2.1 (2010): 93-130.

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

Chen, Kaiji, Ayşe İmrohoroğlu, and Selahattin İmrohoroğlu. "The Japanese saving rate between 1960 and 2000: productivity, policy changes, and demographics." *Economic Theory* 32.1 (2007): 87-104.

Chen, Natalie, and Dennis Novy. "Gravity, trade integration, and heterogeneity across industries." *Journal of International Economics* 85.2 (2011): 206-221.

Coughlin, Cletus & Novy, Dennis. "Estimating Border Effects: The Impact of Spatial Aggregation." CEPR Discussion Paper No. DP11226.

Costinot, Arnaud, and Andres Rodriguez-Clare. Trade theory with numbers: Quantifying the consequences of globalization. No. w18896. National Bureau of Economic Research, 2013.

Deardorff, Alan. "Determinants of bilateral trade: does gravity work in a neoclassical world?." *The regionalization of the world economy.* University of Chicago Press, 1998. 7-32.

Disdier, Anne-Clia, and Keith Head. "The puzzling persistence of the distance effect on bilateral trade." *The Review of Economics and Statistics* 90.1 (2008): 37-48.

Dixon, Peter, Michael Jerie, and Maureen Rimmer. "Modern trade theory for CGE modelling: the Armington, Krugman and Melitz models." *Journal of Global Economic Analysis* 1.1 (2016): 1-110.

Eaton, Jonathan, Samuel Kortum, Brent Neiman, and John Romalis, 2010. *Trade and the Global Recession.* Penn State University and University of Chicago.

FDIC. *Annual Report, 2015.* Washington, D.C.

Feenstra, Robert C., James R. Markusen, and Andrew K. Rose. "Using the gravity equation to differentiate among alternative theories of trade." *Canadian Journal of Economics/Revue canadienne d'conomique* 34.2 (2001): 430-447.

Fehr, Ernst, and Jean-Robert Tyran. "Money illusion and coordination failure." *Games and Economic Behavior* 58.2 (2007): 246-268.

Galí, Jordi. "Monetary policy and rational asset price bubb." *The American Economic Review* 104.3 (2014): 721-752.

Helpman, Elhanan, Marc J. Melitz, and Stephen R. Yeaple. "Export Versus FDI with Heterogeneous Firms." *American Economic Review* 94, 1 (March 2004): 300-316

Elhanan Helpman & Marc Melitz & Yona Rubinstein, 2008. "Estimating Trade Flows: Trading Partners and Trading Volumes," *The Quarterly Journal of Economics*, MIT Press, vol. 123(2), pages 441-487, 05.

Hirano, Tomohiro, and Noriyuki Yanagawa. "Asset bubbles, endogenous growth, and financial frictions." *The Review of Economic Studies* 84.1 (2016): 406-443.

Hummels, David and Lee, Kwan Yong 2012, "Income Elastic Goods and the Great Trade Collapse: Evidence from MicroData." Unpublished working paper.

Hummels, David, and Peter J. Klenow. "The variety and quality of a nation's exports." *The American Economic Review* 95.3 (2005): 704-723.

Hummels, David L. "Toward a geography of trade costs." Available at SSRN 160533 (1999).

Imbs, Jean, and Isabelle Mejean. "Elasticity optimism." *American Economic Journal: Macroeconomics* 7.3 (2015): 43-83.

Kimura, F. & Lee, "The Gravity Equation in International Trade in Services", *H. Rev. World Econ.* April 2006, Volume 142, Issue 1, pp 92–121

Laibson, David, and Johanna Mollerstrom. "Capital flows, consumption booms and asset bubbles: A behavioural alternative to the savings glut hypothesis." *The Economic Journal* 120.544 (2010): 354-374.

Lawless, Martina, and Karl Whelan. "A note on trade costs and distance." (2007).

Levchenko, Andrei, Logan Lewis, and Linda Tesar 2011. *American Economic Review: Papers and Proceedings*, 101:3 293-297. 2010. The Collapse of International Trade During the 2008-2009 Crisis: In Search of the Smoking Gun. *IMF Economic Review,*58:2 214-253.

Lazonick, William. "Profits without prosperity." *Harvard Business Review* 92.9 (2014): 46-55.

Melitz, Marc J. and Stephen J. Redding. 2015. "New Trade Models, New Welfare Implications." *American Economic Review*, 105(3): 1105-46.

Melitz, Marc J., and Gianmarco IP Ottaviano. "Market size, trade, and productivity." *The review of economic studies* 75.1 (2008): 295-316.

Melitz, Marc J. "The Impact Of Trade On Intra-Industry Reallocations And Aggregate Industry Productivity," *Econometrica*, 2003, v71(6,Nov), 1695-1725.

Miroudot, S., R. Lanz and A. Ragoussis (2009), "Trade in Intermediate Goods and Services", *OECD Trade Policy Papers*, No. 93, OECD Publishing.

Miroudot, Sbastien, Jehan Sauvage, and Ben Shepherd. "Measuring the cost of international trade in services." *World Trade Review* 12.04 (2013): 719-735.

Mishkin, Frederic S. *Is monetary policy effective during financial crises?*. No. w14678. National Bureau of Economic Research, 2009.

Novy, Dennis. "International trade without CES: Estimating translog gravity." *Journal of International Economics* 89.2 (2013): 271-282.

Obtsfeld & Rogoff "The Six Major Puzzles of International Macroeconomics: Is there a Common Cause?". Chapter in *NBER Macroeconomics Annual 2000*, Volume 15. 2001 (p.339-412).

Rauch, F. (2016), The Geometry of the Distance Coefficient in Gravity Equations in International Trade. *Review of International Economics*. doi:10.1111/roie.12252

Simonovska, Ina & Waugh, Michael E., 2014. "The elasticity of trade: Estimates and evidence," *Journal of International Economics*, Elsevier, vol. 92(1), pages 34-50

Sulku, Seher Nur. “Testing the long run neutrality of money in a developing country: Evidence from Turkey.” *Journal of Applied Economics and Business Research* 1.2 (2011): 65-74.

Taylor, John B. *Housing and monetary policy.* No. w13682. National Bureau of Economic Research, 2007.

Tinbergen, Jan. 1962. “Shaping the World Economy: Suggestions for an International Economic Policy.” New York: Twentieth Century Fund.

Walsh, Keith. “Trade in services: does gravity hold?.” *Journal of World Trade* 42.2 (2008): 315-334.

Yatchew, Adonis 1997. An Elementary Estimator of the Partial Linear Model. *Economics Letters*, 57, 135-43. 1999. *Differencing Methods in Nonparametric Regression: Simple Techniques for the Applied Econometrician.* (Unpublished)