2018-09-04

Essays on Networks, Volatility, and Trade

Tweedle, Jesse

UNIVERSITY OF CALGARY

Essays on Networks, Volatility, and Trade

by

Jesse Tweedle

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE

DEGREE OF DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN ECONOMICS

CALGARY, ALBERTA

September, 2018

UNIVERSITY OF CALGARY

FACULTY OF GRADUATE STUDIES

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies for acceptance, a thesis entitled "Essays on Networks, Volatility, and Trade" submitted by Jesse Tweedle in partial fulfillment of the requirements for the degree of DOCTOR OF PHILOSOPHY.

Dr. Trevor Tombe
Supervisor
Department of Economics
University of Calgary

Dr. Arvind Magesan
Internal Examiner
Department of Economics
University of Calgary

Dr. M. Scott Taylor
Committee Member
Department of Economics
University of Calgary

Dr. Loretta Fung
External Examiner
Department of Economics
National Tsing Hua University

Dr. Eugene Beaulieu
Committee Member
Department of Economics
University of Calgary

Dr. Dan Gordon
Neutral Chair
Department of Economics
University of Calgary

Date

Abstract

This is a collection of essays on networks, volatility and trade, based on confidential micro-data on manufacturing, trade, and input-output networks in Canada. I have chosen methods and data to support or challenge arguments around these subjects, and aim to show how ownership, production, and trading networks affect firm characteristics, and how, in turn, those firm characteristics affect aggregate volatility.

These essays represent my contribution to the study of networks, volatility and trade in Canada. I show that firm input-output networks matter most for the firm size distribution and aggregate volatility; that firm ownership matters for volatility; that firm size distributions aren't always distributed like a power law; and that there are barriers to domestic trade.

Introduction

This is a collection of essays on networks, volatility and trade, based on confidential microdata on manufacturing, trade, and input-output networks in Canada. I have chosen methods and data to support or challenge arguments around these subjects, and aim to show how ownership, production, and trading networks affect firm characteristics, and how, in turn, those firm characteristics affect aggregate volatility.

First, I explore theories of microfoundations for aggregate volatility, or the variation in the economy's growth rate over time. Those theories rely on a skewed individual size distribution, which may be skewed because of a skewed productivity distribution, or because of skewed demand characteristics that result in an asymmetric production network. What matters more? To find out, I use detailed data on firm-firm trade (i.e., the transaction between a buyer firm and a supplier firm) in Canada to calibrate a model in which productivity and demand characteristics vary independently. This allows me to recover unobserved demand characteristics from the observed production network, which conflates productivity and demand. I find that the demand network accounts for 60% of the firm size distribution, productivity explains little, and that approximately half of the demand network effect is due to higher-order network interconnections. Microeconomic shocks can account for approximately 32% of aggregate volatility, and removing variation in the demand network would reduce aggregate volatility by 25%.

Continuing with the idea of volatility, the second chapter is a study of the correlations of growth shocks within firms. Due to its association with cross-country business cycles, propagation of idiosyncratic shocks, and even financial contagion, firm comovement is an important facet of macroeconomic research. However, we know little about whether pairs of establishments within firms comove more than pairs of establishments that are from different firms. Using a long panel of Canadian manufacturing establishments, which allows a more precise measurement of the covariances compared to existing research, I investigate the correlations and covariances of within-firm pairs of establishments and decompose them into labour inputs, intermediates and profit. I find that within-firm establishment pairs have correlations 0.0477 higher than between-firm establishment pairs (which have an average correlation very close to zero) after controlling for industry and region effects. Covarying intermediate input costs account for 49% of the within-firm comovement effect.

The third chapter examines the firm size distribution a bit more closely. Since research on gains from trade and volatility typically rely on the power law size distribution for certain results, I critically evaluate the data on size distributions using recent developments in power-law estimation. Given a firm size distribution, I test the null hypothesis that the distribution is best fit with a power law. Using data from Compustat and OSIRIS on several countries, as well as confidential microdata on Canadian establishments and firms. I find that a power law distribution fits the U.S. firm size distribution for most years, but France, Germany, and Canada confidently reject the null hypothesis that their firm size distributions are best fit with power laws. However, Canadian manufacturing *plants* are best fit with power laws, supporting the microfoundation of aggregate volatility argument above.

Fourth, I explore Canadian trade and geography. We use sub-provincial trade flows generated from a transaction-level transportation dataset to measure the effects of borders on trade. The results show that border effects fall as geographies are more fine-grained and uniform. In contrast to the U.S., where state border effects were eliminated using similar approaches, provincial border effects remain, with an implied 6.9% tariff equivalent.

These essays represent my contribution to the study of networks, volatility and trade in Canada. I show that firm input-output networks matter most for the firm size distribution and aggregate volatility; that firm ownership matters for volatility; that firm size distributions aren't always distributed like a power law; and that there are barriers to domestic trade.

Table of Contents

Abstract . . . iii
Table of Contents . . . v
List of Tables . . . viii
List of Figures . . . ix
1 The microfoundations of aggregate volatility: productivity or network asymmetry? . . . 1
1.1 Abstract . . . 1
1.2 Introduction . . . 1
1.3 Model . . . 7
1.3.1 Model Basics . . . 8
1.3.2 Important model features . . . 9
1.3.3 Outdegree and unbalanced production networks . . . 12
1.4 Data . . . 12
1.4.1 Skewed distributions: output, productivity and demand . . . 13
1.4.2 The importance of higher-order interconnections . . . 15
1.5 Calibration . . . 16
1.5.1 Parameters . . . 17
1.5.2 Productivity vs. demand . . . 17
1.5.3 Dynamic model . . . 18
1.5.4 Counterfactuals of the firm size distribution . . . 19
1.6 Results and discussion . . . 21
1.6.1 Counterfactual firm size densities . . . 22
1.6.2 Volatility . . . 24
1.7 Conclusion . . . 25
1.8 Appendix: Theory . . . 26
1.8.1 Full model . . . 26
1.8.2 Derivation of influence vector . . . 30
1.8.3 Asymptotic Theory . . . 32
1.8.4 Aggregate volatility depends on the product of the distributions of outdegree and productivity . . . 35
1.9 Appendix: Data and Empirics . . . 37
1.9.1 Data sources . . . 37
1.9.2 Intensive and Extensive Margins of Volatility . . . 38
2 Correlated shocks within firms . . . 42
2.1 Abstract . . . 42
2.2 Introduction . . . 42
2.3 Empirical framework and data . . . 45
2.3.1 Data . . . 47
2.4 Econometrics and Results . . . 49
2.4.1 Results . . . 51
2.5 Conclusion . . . 53
2.6 Appendix . . . 53

2.6.1 Data . . . 53
2.6.2 Robustness . . . 55
3 Looking for power-laws in all the wrong places: estimating firm size distribution tails across countries and datasets . . . 58
3.1 Abstract . . . 58
3.2 Introduction . . . 59
3.3 Methodology . . . 63
3.3.1 Upper-tail distribution estimation . . . 63
3.3.2 Tests . . . 65
3.4 Data . . . 66
3.5 Results . . . 67
3.5.1 Power-law estimations . . . 68
3.5.2 What do the estimated distributions look like? . . . 70
3.5.3 Direct and alternative hypothesis tests . . . 71
3.6 Application . . . 73
3.7 Conclusion . . . 76
3.8 Appendix–Robustness . . . 77
3.8.1 The truncated Pareto distribution . . . 77
3.9 Appendix—Theory . . . 78
3.9.1 Distributions . . . 78
3.9.2 Fitting the model, estimating $x_{\min}$ and α . . . 80
3.9.3 Likelihood-ratio tests . . . 81
4 Going the Distance: Estimating the Effect of Provincial Borders on Trade when Geography (and Everything Else) Matters . . . 82
4.1 Abstract . . . 82
4.2 Introduction . . . 82
4.3 Data development . . . 86
4.4 Model and estimation strategy . . . 90
4.4.1 Trade model . . . 90
4.4.2 Estimator . . . 92
4.4.3 Geography and estimation . . . 94
4.5 Model estimates . . . 96
4.5.1 Standard province-based estimates . . . 97
4.5.2 Estimates by sub-provincial geography . . . 99
4.5.3 Sub-provincial estimates robustness checks . . . 102
4.6 Tariff equivalent of provincial border effects . . . 108
4.7 Conclusions . . . 110
.1 Data Appendix . . . 112
.1.1 Valuing shipments . . . 112
.1.2 Geocoding shipment origins and destinations . . . 112
.1.3 Benchmark weights . . . 112
.1.4 Comparing network and great-circle intra- and inter-provincial distances 116
.2 Additional Robustness checks . . . 118
.2.1 Testing for the differential effect on distance on intra- and inter-provincial trade . . . 118

.2.2 Estimates by year 119
.2.3 Differential border effect estimates for Quebec 119

List of Tables

1.1 Skewness of main variables 40
1.2 Counterfactual firm density statistics 40
1.3 Microeconomics shocks and aggregate volatility in the data 40
1.4 Robustness of firm size counterfactuals to different measures of productivity 41
1.5 Table of Notation 41
1.6 Intensive vs. Extensive Margin Volatility 41

2.1 Summary statistics. 48
2.2 Firm correlation and covariance, u_{it}^s 51
2.3 Firm covariance decomposition, u_{it}^s, u_{it}^π, u_{it}^{pay}, u_{it}^{input} 52
2.4 Firm correlation and covariance, u_{it}^s, including extra covariates 56
2.5 DHS growth rates; firm correlation and covariance, u_{it}^s 56
2.6 DHS growth rates; firm correlation and covariance, u_{it}^s, including extra covariates 57

3.1 Summary statistics of firm sizes by country and dataset 67
3.2 Summary of p-values for different tests, by country and dataset. 72
3.3 Herfindahls for each predicted distribution and their differences from the data, by country 75
3.4 Summary of p-values for different tests, by country and dataset. 78

4.1 Benchmarked intra-/inter-provincial trade flows as a percentage of actual flows. 89
4.2 Manning and Mullahy (2001) test by province, economic region and census division. 94
4.3 Provincial border effect estimates based on provincial average flows (2004 to 2012). 98
4.4 Provincial border effect estimates based on flows between large and small hexagons (2004 to 2012). 101
4.5 Robustness of provincial border effect estimates to non-linear effects of distance and contiguity. 121
4.6 Provincial border effect estimates for Forward Sortation Areas (FSA). 123
.7 Network and great-circle distance. 125
.8 Provincial border effect estimates based on flows between province, economic regions and census division allowing coefficients to vary across time (2004 to 2012), selected variables. 126
.9 Test of the effect of Quebec on provincial border effects. 127

List of Figures and Illustrations

1.1 Counterfactual firm densities 22

2.1 Density of establishment-establishment correlations, within-firm (solid blue line) and across firm (dotted black line). 49

3.1 Difference in curvature on rank-size plot for the US and Canada. 59
3.2 Exponent distributions 68
3.3 Exponents vs. number of observations 69
3.4 Estimated and empirical distributions 70
3.5 Boxplot of p-values of the test of null hypothesis H_0^{pexp} for each Canadian micro-dataset. Each dot represents a p-value for one year of a dataset. 74
3.6 Estimated and empirical distributions, with truncated Pareto 79

4.1 Transformation of logistic to trade flows, full and broken sets. 88
4.2 Intra- and inter-provincial distance shipped, benchmark and survey weights. 91
4.3 Size and placement of hexagonal lattices. 96
4.4 Coefficient estimates for Own Province (4.4a), Own Region (4.4b), and log($distance$) (4.4c), by size (km per side) and placement of hexagons. 120
4.5 Coefficient estimates for Own Province (4.5a) and Own Region (4.5b) by model, hexagon size (km per side) and placement. 122
4.6 *Ad valorem* tariff equivalent by standard and hexagonal geographies. 124

Chapter 1

The microfoundations of aggregate volatility: productivity or network asymmetry?

1.1 Abstract

Theories of microfoundations for aggregate volatility rely on a skewed individual size distribution, termed *granularity*. If so, what causes granularity? I use detailed data on firm-firm trade in Canada to estimate a model in which productivity and demand characteristics vary independently to determine firm sizes. This allows me to recover unobserved demand characteristics from the observed production network, which conflates productivity and demand. I find that the demand network accounts for 60% of the firm size distribution, productivity explains little, and that approximately half of the demand network effect is due to higher order network interconnections. Microeconomic shocks can account for approximately 32% of aggregate volatility, and removing variation in the demand network would reduce aggregate volatility by 25%.

1.2 Introduction

Are idiosyncratic shocks sources of aggregate volatility? How do they propagate across the economy? The idea that aggregate demand and supply shocks are the only source of volatility in the economy leaves important mechanisms in the shadows. Previously, the possibility that idiosyncratic shocks to firms can cause aggregate fluctuations had been debunked by the law of large numbers; how can tens of thousands, or millions, of uncorrelated shocks average out to anything but zero? However, if the economy is structured in such a way that certain firms are disproportionately large, the law of large numbers argument may fail. Idiosyncratic shocks to these firms may propagate through the economy and make up a substantial portion of aggregate volatility.

Theories of microfoundations of aggregate volatility all rely on this skewness of the firm size distribution, called *granularity*. If granularity allows idiosyncratic shocks to cause aggregate fluctuations, what causes granularity? In this paper, I study the sources of skewness in the firm size distribution—productivity and the firm-firm production network—and how they affect aggregate volatility.

A skewed productivity distribution is the usual culprit in models with skewed size distributions: the standard source of firm heterogeneity is a Pareto productivity distribution in a simple Melitz (2003) model. However, the size distribution is also skewed if the firm-firm production network is skewed. A firm may produce goods that are required inputs in production for a substantial fraction other firms, which causes that firm to be very large. The first and second order effects of the interactions between these factors turn out to be very important. For instance, a firm may have low productivity and few customers in the production network, but if those customers are themselves large, it will be large in turn.

The key to differentiating these features is to use a model in which they vary independently, and, more importantly, data that allows me to calibrate and estimate the model. I extend a model of firm-firm trade with firm heterogeneity in not only productivity, but pair-specific demand characteristics that define the production network as well. The most important thing to note is that production networks are endogenous—using expenditures shares as measures of input-output requirements, such as those used in industry level input-output tables, conflates the three factors I study here. Recovering the true source of granularity requires data and a model that differentiate these things. After doing so, I perform counterfactuals on the parameters to see how changing the underlying productivity and demand would change the size distribution and aggregate volatility.

The model extends a simple firm-firm Cobb-Douglas production network model to incorporate productivity differences, trade costs and substitutability across firms. Each firm is in a region, and total regional income is the sum of all value-added in that region; regional income can be spent on goods from any firm, subject to trade costs. The market structure determines the skewness of the size distribution, which in turn affect the way idiosyncratic shocks propagate across the economy. Almost all of the shocks are transmitted through input-output links, though the reasons the production links exist in the first place are determined by productivity and demand characteristics.

One must note that it is not enough to use industry input-output characteristics to define the economy, in the model or in the data. First, in the model, using industry input-output shares as demand characteristics implies all within-industry firm heterogeneity cannot be due to demand characteristics, which is refuted by the data. Using industry-level IO data also implies that within a pair of industries with an input-output linkage (i.e., a positive direct requirement coefficient, which is the term for the expenditure share in the industry-by-industry input-output tables produced by national statistical agencies), all firms trade with each other. And in any industry with an input-output linkage with itself, all firms within the industry trade with each other, including itself. This is again refuted by the data, which I turn to next.

The microdata[1] are from several sources: the Annual Survey of Manufacturing (ASM), the Surface Transportation File (STF), the detailed-confidential Input-Output and Supply-Use tables (IOT), the Inter-Provincial Trade Flow file (IPTF), and the Import-Export Register (IER). For more details of each database and on data construction and benchmarking, see Appendix 1.9.

[1] Here I make the first distinction between firms and establishments. I use the term 'firm' to be consistent with previous work on firm-to-firm production networks, firm size distributions and aggregate volatility, and because it's shorter and easier to say and write than 'establishment.' This is convenient for the writer and reader when describing establishment-establishment trade. Nevertheless, the data are at the establishment level. Firm-level microdata are difficult to study geographically, because they typically do not have 'locations' in the physical sense used in models of economic geography. When using administrative data, the firm unit is defined by tax accounting standards, not economic or physical standards, and so firms are not guaranteed to have actual physical locations. Instead, they have corporate headquarters that may have complex legal and operational heirarchies and no geographic information on economic activity.

The data show clear skewness in the productivity and size distributions and a very asymmetric firm-firm production network. The empirical strategy works in two parts. First, I start with the observed data and use the model to uncover unobserved demand characteristics and the implied demand network. Next, after uncovering the parameters that govern the firm size distribution, I turn to calculating aggregate volatility. It is difficult to infer the parameters that determine idiosyncratic volatility due to the general equilibrium nature of shock propagation: if the granularity hypothesis is true, a large aggregate shock is the result of idiosyncratic shocks, not evidence against them. This is called the "reflection problem." I attempt to circumvent this problem by estimating uncorrelated productivity shocks and using those as a lower bound for the contribution of idiosyncratic shocks to aggregate volatility.

After calibrating the model, I can investigate the effect of each feature's skewness on the economy. For instance, what happens to aggregate volatility if we remove the variation in productivity across firms? What if we remove variation in the demand network instead? The relative changes in aggregate fluctuations after changing the distributions of each feature give important insights into the economy. Perhaps surprisingly, removing skewness in productivity actually increases skewness in the size distribution, which would increase aggregate volatility by 11% and highlighting the importance of the complexity of the network.

Research on idiosyncratic shocks and aggregate volatility restarted in earnest when Gabaix (2011) and Acemoglu et al. (2012) revived the debate between Horvath (1998, 2000) and Dupor (1999) on whether idiosyncratic shocks average out in aggregate. Gabaix (2011) proposes that the largest, granular firms are so big that their idiosyncratic shocks do not average out at the aggregate level. Acemoglu et al. (2012) suggest the reason for non-diversification of idiosyncratic shocks is an asymmetric input-output network, in which a shock to a sector that supplies a large number of other sectors propagates through the economy and generates aggregate fluctuations. I add an understanding of the connections between the two theories at an empirical level, specifically showing the complementarity between granularity and production networks and how idiosyncratic firm-level shocks rely on firm-level input-output variation within industries.

The most direct predecessors of this paper are empirical studies of aggregate fluctuations. Starting with Shea (2002), and continuing most recently with Foerster et al. (2011), Di Giovanni et al. (2014), Acemoglu et al. (2015). Foerster et al. (2011) combined factor analysis with structural model of industrial production in the US, finding common shocks are the source of the majority of volatility, with idiosyncratic shocks becoming more important after the great moderation. Di Giovanni et al. (2014) study fluctuations of French firm sales to individual countries and find idiosyncratic fluctuations account for the majority of aggregate volatility, and that much of it comes from covariances between firms. They suggest the firm covariances are due to firm-to-firm linkages, although they only observe industry-level IO data. In contrast to both papers, I use firm-level network data to establish the determinants of firm covariances, using deeper levels of disaggregation to examine both covariances (firm level to establishment level) and input-output networks (industry level to establishment level). As well, I study the determinants of the network itself, something taken as exogenous in previous empirical work.

Any study of granularity builds on a body of work on the determinants of firm size and the characteristics of its distribution, from specific applications in international trade (Di Giovanni et al., 2011; Di Giovanni and Levchenko, 2012, 2013), or studies on general characteristics and theories of the size distribution itself (Luttmer, 2007). I add an endogenous network perspective to this research and use it to further explore the determinants of the firm size distribution and the sources of granularity. My work also fits naturally with Hottman et al. (2016), who use detailed retail scanner data on consumer non-durables to suggest 'firm appeal' is the dominant source of firm heterogeneity, accounting for 50-70% of firm size. Holmes and Stevens (2014) also provide evidence that demand characteristics are the main source of firm heterogeneity, in contrast to standard Melitz applications. In my case, the input-output requirements of downstream firms translate into a dominant source of firm appeal, and therefore are a large determinant of firm size.

My argument is also related to recent work on customer-supplier relationships, especially Barrot and Sauvagnat (2016), who study the disruption of production networks after natural disasters. In addition, research on customer-supplier relationships in Japan (Bernard et al., 2015; Carvalho et al., 2014) and the US (Atalay et al., 2011) suggests larger firms have different input-output characteristics than smaller firms. Most recently, Lim (2016) studies creation and destruction of firm-firm relationships, although he notes the difficulty of matching geographic characteristics. Typically, customer-supplier relationship data only includes an indicator for whether a firm supplies another firm, not the strength of the relationship or the commodities made and used. In my case, I have measures of the strength of the interaction between firms. To this research, I add a characterization of the complexity of the production network in Canada.

These papers are also part of a recent wave of interest in the formation and effects of social and economic networks. Carvalho and Voigtländer (2014), Oberfield (2017) and Jones (2011) each apply these ideas specifically to production and growth, whereas other works focus on volatility and contagion in financial markets such as Acemoglu et al. (2015) or Elliott et al. (2014), or network formation and volatility in Anthonisen (2016). Other applications and background on network measures used in this paper can be found in Jackson (2010).

In Section 1.3, I present a simple, but necessary, extension to the Cobb-Douglas input-output model used in Acemoglu et al. (2012) to allow three features crucial to reconcile the empirical regularities in the economy: I incorporate productivity variation and substitutability across firms and unobserved demand network characteristics. The asymmetry of the production network and the productivity distribution combine to determine firm sizes, which is the key to evaluating the granularity of the economy and its effect on aggregate volatility. In Section 1.4, I present the firm-level volatility and production network data.

I document an unbalanced production network at a disaggregated level, with a few firms acting as central suppliers to the network.

In Section 1.5, I calibrate the model to uncover the underlying demand characteristics network from the endogenous, observed input-output network and evaluate the competing theories of the microfoundations of aggregate fluctuations. In Section 1.6, I present results. Previewing the main calibration results, the productivity distribution is not heterogenous enough to account for the asymmetry in the observed production network. The majority of the firm size distribution is due to the underlying demand network, consistent with results in Holmes and Stevens (2014) that challenge the reliance of the firm size distribution on productivity alone. In addition, higher order interconnections are economically significant determinants of the firm size distribution. Turning to the macroeconomy, I find idiosyncratic shocks can account for approximately 32% of aggregate volatility, and that removing variation in productivity would actually increase firm size skewness and aggregate volatility by 11%.

Section 1.7 concludes, and two Appendices follow, giving details on theory, measurement and development of the firm-to-firm production network, and other necessary but tedious details.

1.3 Model

To study the relationships between volatility, endogenous asymmetric production networks and the factors that determine them, I adapt the sectoral model of Acemoglu et al. (2012), which is itself based on Long and Plosser (1983). There are three key additions.

First, I study individual firms and not sectors. Although technically easy (e.g., relabeling sectors as firms), it puts the focus on the determinants of granularity—is it the production network or productivity? This becomes crucial as we turn to the study of a very disaggregated economy, which is the primary reason for studying microfoundations of aggregate volatility.

Third, and most importantly, I relax the assumption that the production network is exogenous. In my model, a firm may be a central supplier of the network because it is a required input in many other products (it has many high unobserved demand characteristics) or because it is so productive that many other firms substitute toward it.

To introduce these features, I need a model in which productivity and unobserved demand characteristics can vary independently to create an observed firm-firm production network that I can take to the data. I give a table of important notation in Table 1.5 in Appendix 1.8. In general, I use capital letters to refer to matrices, lowercase to refer to vectors and elements of vectors and matrices, latin characters for observed variables and greek characters for the equivalent unobserved variables. For example, $G = [g_{ij}]$ is the observed expenditure share matrix, $\Gamma = [\gamma_{ij}]$ is the unobserved demand matrix.

1.3.1 Model Basics

To start, there are R regions. A representative household in a specific region r inelastically supplies a labour L_r, and has Cobb-Douglas preferences over N different goods (I relax this assumption later, but it is useful to focus first on firm-firm demand characteristics),

$$u_r(c_r) = \prod_{i \in N} c_{ri}^{\lambda^{ri}} \tag{1.1}$$

where c_{ri} is region r's consumption of good i. There is free migration between regions, so that the wage w in equilibrium is constant across regions. Later, I normalize $w = 1$.

Each good is produced by a single firm using Cobb-Douglas combination of labour and a firm-specific intermediate input which is itself a CES aggregate of other products,

$$q_i = z_i l_i^{\beta} \left(\sum_{j \in N} \gamma_{ij}^{\frac{1}{\eta}} q_{ij}^{\frac{\eta-1}{\eta}} \right)^{\frac{(1-\beta)\eta}{\eta-1}} \tag{1.2}$$

where z_i is productivity, β is the labour share in production, q_{ij} is the quantity of firm j's product demanded by firm i, and η is the elasticity of substitution between intermediates. The crucial part of production is $\gamma_{ij} \geq 0$, which is the exogenous direct input coefficient. If

γ_{ij} is high, then independent of firm j's productivity, firm i requires a lot of firm j's input to produce. If γ_{ij} is low but positive, then firm i may still demand a lot of q_{ij} if firm j is very productive. In this way, the endogenous production network is determined jointly by productivity, substitutability and unobserved demand characteristics. Firm i can only draw labour from its region r. There can be multiple firms in any given region.

With perfect competition, prices equal marginal costs for firm i,

$$p_i = C z_i^{-1} \left(\sum_{j \in N} \gamma_{ij} p_j^{1-\eta} \right)^{\frac{1-\beta}{1-\eta}} \tag{1.3}$$

where $C \equiv \beta^{-\beta}(1-\beta)^{\beta-1} w^{\beta}$ is independent of i.

$$p_{ri} = \prod_{i \in N} \left(\frac{p_i}{\lambda_{ri}} \right)^{\lambda_{ri}} \tag{1.4}$$

The full derivation of the model, along with any extra notation needed, can be seen in Appendix 1.8.1.

1.3.2 Important model features

The model is simple, but it delivers several important results that are typically ignored when looking at models of production networks.

Remark 1 *Observed expenditure shares depend on productivity and unobserved demand characteristics.*

The input-output tables provided by statistical agencies give an expenditure share of industry i on goods from industry j. The firm production network I detail in the previous section is constructed in a similar way, an expenditure share of firm i on firm j. If we assume production is Cobb-Douglas, then the expenditure share parameter in production exactly determines the observed expenditure share. This is no longer true if the elasticity of substitution is not equal to 1. Define the observed expenditure share g_{ij},

$$g_{ij} = \frac{p_j q_{ij}}{p_i q_i} \tag{1.5}$$

In equilibrium, this simplifies to

$$g_{ij} = (1 - \beta) \left[\frac{\gamma_{ij} p_j^{1-\eta}}{\sum_{k \in N} \gamma_{ik} (p_k)^{1-\eta}} \right] \tag{1.6}$$

If $\eta = 1$, the observed expenditure share is exactly determined by the relative exogenous coefficient γ_{ij} (that is, if you rederive the solution starting with $\eta = 1$ in the production function). However, it is clear that the observed expenditure shares are jointly determined by the vector of direct input coefficients $\gamma_{i\cdot}$ and the vector of prices, which are themselves determined by the vector of firm productivities (and more complex interconnections). Again, the observed production network is endogenously determined by the vector of firm productivities and demand characteristics.

Remark 2 *Expenditure shares still "determine" size, but they say nothing about the underlying determinants of the size distribution.*

In an important result, Acemoglu et al. (2012) shows that the vector of industry sizes, normalized by total sales in the economy, which he calls the influence vector v, is the crucial link between the production network network and volatility. The influence vector determines the extent to which microeconomic shocks contribute to aggregate volatility, and the influence vector is determined by the characteristics of the exogenous production network. Hence their claim that the production network is the main determinant of aggregate volatility. Here I show that the same holds for the observed production network. That is, an empirical association between the influence vector and observed production network does not tell you the effect of the production network on volatility, because the observed network may be entirely determined by productivity. Write the system of market clearing equations,

$$\sum_{r \in R} p_{ri} c_{ri} + \sum_{j \in N} \tau_{ij} p_i q_{ji} = p_i q_i, \text{ for } i \in N \tag{1.7}$$

And rewrite in terms of g_{ij} using (1.5),

$$\sum_{r \in R} p_{ri} c_{ri} + \sum_{j \in N} g_{ji} p_j q_j = p_i q_i, \text{ for } i \in N \tag{1.8}$$

Then a similar derivation to Acemoglu et al. (2012) (see Appendix 1.8.2) gives you the influence vector as a function of the matrix of observed expenditure shares $G = [g_{ij}]$, observed demand shares $A = [a_{ri}]$, and regional labour $L = (L_1, \ldots, L_R)$,

$$v' = \beta \left(\frac{L'}{\mathbf{1}'L} \right) A(I - G)^{-1} \tag{1.9}$$

The influence vector, v, is always related to the observed production network, but the observed production network is endogenous. So observing the association between the influence vector and the production network does not give you any information on the importance of the underlying demand characteristics, $\Gamma = [\gamma_{ij}]$, or region demand characteristics, $\Lambda = [\lambda_{ri}]$.

Example 1 *Suppose $\gamma_{ij} = 1$ for all $i, j \in N$. Then there is no exogenous demand variation, and all of the observed production network characteristics are due to productivity.*

If $\gamma_{ij} = 1$, then all firms use the same intermediate bundle and face the same intermediate input price. This means the expenditure share equation (1.5) reduces to

$$g_{ij} = (1 - \beta) \left[\frac{z_j^{\eta-1}}{\sum_{k \in N} z_k^{\eta-1}} \right] \tag{1.10}$$

Which is determined solely by relative productivities. In this case, if productivities are distributed with a power law, we will still observe an influence vector consistent with the unbalanced production network, even though the underlying demand characteristics are homogenous.

Example 2 *Suppose $z_i = 1$ for all $i \in N$ for all $i, j \in N$. Then there is no productivity variation, and all of the observed production network characteristics are due to the exogenous demand characteristics.*

When productivities are identical across all firms, the expenditure share terms reduce to

$$g_{ij} = (1 - \beta) \left[\frac{\gamma_{ij} p_j^{1-\eta}}{\sum_{k \in N} \gamma_{ik} p_k^{1-\eta}} \right] \tag{1.11}$$

where the prices can be written as a recursive function of prices and demand parameters, which implies the expenditure shares are determined only by demand parameters.

1.3.3 Outdegree and unbalanced production networks

An unbalanced production network is one in which individual firms are central suppliers to the entire economy. The easiest way to ask how central a firm is by adding up the demand parameters of a firm's customers (unobserved outdegree, δ_i), or the observed expenditure shares of a firm's customers (observed outdegree, d_i),

$$\delta_i = \sum_{j \in N} \gamma_{ji}; \; d_i = \sum_{j \in N} g_{ji}, \tag{1.12}$$

Example 3 *Suppose $\gamma_{ij} = \delta_j / N$, for $j \in N$.*

Expenditure shares are

$$g_{ij} = (1-\beta)\left[\frac{\delta_j z_j^{\eta-1}}{\sum_{k \in N} \delta_k z_k^{\eta-1}}\right] \tag{1.13}$$

Observed outdegree is

$$d_i = (1-\beta)\left[\frac{\delta_i z_i^{\eta-1}}{(1/N)\sum_{k \in N} \delta_k z_k^{\eta-1}}\right] \tag{1.14}$$

And one element of the influence vector is

$$v_i = \frac{\beta}{N} + (1-\beta)\left[\frac{\delta_i z_i^{\eta-1}}{\sum_{k \in N} \delta_k z_k^{\eta-1}}\right] \tag{1.15}$$

This examples highlights the dependence of the influence vector on productivity and the unbalanced production network—the distribution of v_i is determined by the distribution of $\delta_i z_i^{\eta-1}$. Recall that the argument for microfoundations of aggregate shocks requires the distribution of v_i to have a thick tail even as the number of firms grows large. However, as the number of firms grows large, the thick tail of v_i will tend to be dominated by the thickest tail of the two distributions of outdegree and productivity.

1.4 Data

The microdata are from several sources: the Annual Survey of Manufacturing (ASM), the Surface Transportation File (STF), the detailed-confidential Input-Output and Supply-Use tables (IOT), the Inter-Provincial Trade Flow file (IPTF), and the Import-Export Register (IER). For more details of each database and on data construction and benchmarking, see Appendix 1.9.

The establishment data is from the ASM, a defacto census of industrial output in Canada. It is a long-running annual panel of manufacturing establishments, including data on shipments by destination province (and exports), and inputs and outputs by commodity.

I analyze volatility over the period from 1990 to 2010, covering several volatile periods in Canadian manufacturing, including in the early 1990s, as well as 2001 and the Great Recession. Aggregate volatility, measured by the standard deviation of the aggregate growth rate of total output, over this period was approximately 6% in manufacturing, slightly higher than the overall for Canada during the same period, around 4%.

The trade data is from the STF, a transaction-level database of goods shipments in Canada, including trade to and from the United States. Each shipment includes value, tonnage, commodity classification, mode, shipper and receiver names, addresses and postal codes. This allows the identification of origin and destination establishments from the ASM, as well as establishment origins and final demand destinations.

1.4.1 Skewed distributions: output, productivity and demand

In this section, I document the skewness in each feature of the economy. For one-dimensional firm measures (i.e., firm size and firm productivity), I measure skewness with the herfindahl, the 90/10 percentile ratio and the slope of the right tail of the distribution on a log-log rank-size plot. Herfindahls are directly related to the granular theory (see Gabaix, 2011), with more concentrated distributions supporting more aggregate volatility. To estimate the shape parameter of the tail of the distribution, following Gabaix and Ibragimov (2011), I trim the distribution to the top 20^{th} percentile of variable x and estimate

$$\log(\text{rank}(x_i) - 1/2) = \alpha - \beta \log x_i \tag{1.16}$$

The estimated shape parameter $\hat{\beta}$ is a measure of the strength of the asymmetry in the distribution—a shape parameter of 1 is Zipf's law.

In the Canadian manufacturing sector, a few industries play outsized roles in output, employment and value-added. Transportation equipment production alone accounted for 21.5% of total manufacturing output in Canada in 1997, and the top *ten* firms in that industry account for the vast majority of its output. The herfindahl of firm sales is 0.048, and the tail parameter of the log-log rank-size plot is 0.99. The firm size distribution is clearly skewed.

I measure firm productivity in several ways. First, labour productivity, defined as total value-added divided by employment. Next, labour productivity, defined as total value-added divided by total payroll. Next, naïve total-factor-productivity, measured as the residual of a log-linear regression of output on employment, capital and total input cost. Finally, the estimation procedure developed by Gandhi et al. (2013), which I refer to as TFP (GNR).

Each method has benefits and drawbacks. The goal is to rely on the robustness of the results to a variety of different productivity estimates, rather than stick to a single productivity estimation procedure. First among the drawbacks, all estimates are of revenue productivity, not physical productivity (see Foster et al., 2008, for a discussion of the relevant differences). The drawback here is not as stark as it would be in reduced form studies that rely on the difference between revenue TFP and physical TFP, since I can recover the unobserved demand characteristics in the model, conditional on the assumptions. The productivity measures and results are consistent with previous work on firm heterogeneity, specifically that demand characteristics matter more for firm heterogeneity than physical productivity itself. Therefore, although I have no *a priori* justification for only using revenue productivity, the results suggest revenue productivity is a decent measure of productivity, as long as I account for the unobserved demand characteristics in the model.

In addition, both labour productivity measures have the obvious drawback of being partially determined by capital. Using payroll instead of employment tends to reduce this bias (since firms with higher capital-per-worker tend to pay higher wages, which reduces the variation in the payroll-based measure due to capital). Again, the real strategy is to show robustness across each measure.

The observed production network is defined by expenditure shares between firms, $G = [g_{ij}]$, and the firm-region expenditure shares $A = [a_{ij}]$. A directed link exists from firm j to firm i if i buys some positive amount of firm j's output. The intensity of the link is determined by the value of $g_{ij} \in [0,1]$. In this setting, observed (d_i) and unobserved (δ_i) outdegrees are

$$g_i = \sum_{r \in R} a_{ri} + \sum_{j \in N} g_{ji}; \quad \delta_i = \sum_{r \in R} \lambda_{ri} + \sum_{j \in N} \gamma_{ji} \tag{1.17}$$

The observed shares and show considerable asymmetry. As we saw in the model in Section 1.3, the asymmetry of the influence vector and the asymmetry of the observed production network do not necessarily let us infer anything about the underlying economic relationships between firms. We only know that a firm buys a lot of input from another firm, not why.

1.4.2 The importance of higher-order interconnections

Can we simplify the study of the complex firm-firm network to a one-dimensional measure? Acemoglu et al. (2012) cite outdegree as the main measure of network importance; can we focus on that one-dimensional firm measure and leave the complex network alone? Here, I show that one-dimensional measures do not explain much of the firm size distribution, and therefore higher-order interconnections are significant factors in explaining the economy—we cannot rely on one-dimensional firm measures alone.

Suppose the input-output connection is constant across firms, and equal to δ_j/N for firm j, as in Example 3. Then first-order outdegree and productivity alone explain the firm-size distribution,

$$\log v_i = \chi + (\eta - 1) \log z_i + \log \delta_i, \tag{1.18}$$

If this equation defines the firm-size distribution, estimating this equation with OLS should give an R^2 close to 1, subject to measurement and numerical error. However, the estimated R^2 is only 26.4% (about 5% when including productivity alone, and 21% when including outdegree alone). This leaves 73.6% of the firm-size distribution unexplained, which means the higher-order interconnections matter—it matters which firms you supply, and which firms they supply, and so on, and the complex effects of the network cannot be reduced to one-dimensional firm measures. Note that using a skewed distribution δ_j as demand parameters implies (skewed across suppliers j, constant across customers i within a given supplier) implies skewed distributions of second-order and higher-order outdegrees as well (see Acemoglu et al., 2012). This suggests it is not only the higher order demand connections, but how they interact with productivity as well.[2]

1.5 Calibration

In this section, I calibrate the model to match features of the data to further explore the relationships between productivity, the unbalanced production network and volatility. In addition, I add iceberg trade costs to the model to attempt to account for Canadian geographic characteristics. Instead of relying on asymptotic results to infer which factor dominates the size distribution (see Appendix 1.8.3), using the model described in Section 1.3, I use data on firm productivity z, trade costs $T = [\tau_{ij}]$, the observed region demand A, and the observed input share matrix G to solve for the unobserved region demand characteristics Λ and the unobserved technical requirement matrix Γ.

Although final demand did not add to the explanation of the model and asymptotic theory, it is important empirically. Therefore, to match the data better, I change the regional

[2] δ_i is calculated as the column sums of Λ plus the column sums of Γ. Using observed outdegree (via A and G) gives similar results. Variation in β also matters quantitatively for the firm size distribution, but this again suggests it matters which firms you supply, and who they supply, and so on, not just that you have a high outdegree.

consumer's utility function to a CES combination of each product,

$$u_r(c_{ri}) = \left(\sum_{i \in N} \lambda_{ri}^{\frac{1}{\epsilon}} c_{ri}^{\frac{\epsilon-1}{\epsilon}} \right)^{\frac{\epsilon}{\epsilon-1}} \tag{1.19}$$

Where c_{ri} is region r's consumption of firm i's output. Now the unobserved final demand characteristic λ_{ri} is similar to a γ_{ji} in firm j's production function, and the observed final demand share a_{ri} is similar to the observed expenditure share g_{ji}. In addition, variation in the value added share of output per firm matters for the distribution of output. After adding these features, the goal is to use the model to uncover the unobserved region-firm and firm-firm demand parameters from the data.

1.5.1 Parameters

There are several sets of parameters that determine the model. Most of the parameters I can select directly from data, a few I need to set, and the rest I use the model (and the given parameters) to solve. The observable set of parameters are: output s_i, the expenditure share matrices A and G, value added shares β_i, productivities z_i, regional income wL_r, and trade costs T. Next, I set the elasticities of substitution η and ϵ at 2. Finally, using the data and model, I solve for the unobserved demand parameters Λ and Γ. For a full description of the data sources, benchmarking, calibration and solutions to the model, see the Appendix.

1.5.2 Productivity vs. demand

Productivity and demand characteristics are tough to define. Productivity z_i is some technology specific to firm i that tells us how effective that firm is at turning inputs into outputs. However, with CES production technology, firm i is more productive (and is larger) if it uses more inputs (and $\gamma_{ij} = 1$ for all inputs j), even holding z_i constant. In this case, even though the demand characteristics are increasing its size, we'd like to associate that effect with productivity. In other words, if we normalize the demand characteristics γ_{ij} for each i, and associate that effect with productivity instead, we can more accurately describe the relative effects of demand and productivity:

$$q_i = C_i w^{\beta_i} \underbrace{\left[z_i \left(\sum_j \gamma_{ij}^{1/\eta} \right)^{\frac{(1-\beta_i)\eta}{\eta-1}} \right]}_{\tilde{z}_i} \left(\sum_j \underbrace{\left[\frac{\gamma_{ij}^{1/\eta}}{\sum_k \gamma_{ik}^{1/\eta}} \right]}_{\tilde{\gamma}_{ij}} q_{ij}^{\frac{\eta-1}{\eta}} \right)^{\frac{(1-\beta_i)\eta}{\eta-1}}, \tag{1.20}$$

and I refer to $\tilde{z}_i$ as augmented productivity, and $\tilde{\Gamma}$ as augmented demand. In the following empirical results, I use these augmented measures instead. The final results using the augmented measures suggest demand accounts for a significant portion of firm size, and using the raw productivity and demand measures only reinforce that result. Using the augmented measures serves to adjust for a producer's demand characteristics that results in higher productivity. It may also adjust for bias in raw productivity measures, since the model uncovers demand parameters that justify the size distribution—if a firm with low raw productivity z_i ends up with large measured demand characteristics, then the raw productivity measure was not enough to justify the firm's size, and the demand characteristics provide an augmented productivity measure $\tilde{z}_i$ that is correct and consistent with the model and data.

1.5.3 Dynamic model

To adapt the static model in Section 1.3 to include volatility, I use a strategy similar to Acemoglu et al. (2012). In each period, firms receive idiosyncratic demand shocks γ'_{ijt} and λ'_{rit}, as well as productivity shocks z'_{it}. In each period, the equilibrium is equal to the static model with the new parameters $\gamma_{ijt} = \gamma_{ij}\gamma'_{ijt}$, $\lambda_{ijt} = \lambda_{ij}\lambda'_{ijt}$, and $z_{it} = z_i z'_{it}$.

There are several important factors in the dynamic model that help us study the microfoundations of aggregate fluctuations, and the relative contributions of granularity, geography and exogenous production characteristics to aggregate volatility. Similar to the rest of the paper, the difference between the unobserved and observed parameters matters. The data are observed sales growth rates, but we would like to know the unobserved idiosyncratic shocks that gave rise to them. Furthermore, uncorrelated idiosyncratic shocks naturally result in correlated sales growth rates, depending on the linkages between firms and firms, and firms and regions.

Next, demand and productivity shocks may contribute differently to aggregate volatility. In previous work (see, e.g., Acemoglu et al., 2015; Shea, 2002), productivity shocks only propagate downstream, and demand shocks only propagate upstream. However, using a CES function in productivity and demand, both types of shocks can propagate in both directions. For example, a positive productivity shock can propagate upstream because it affects downstream expenditure for the product (positively, if the elasticity of substitution is greater than one).

The distinction between demand and productivity is an important factor in the literature on the firm-size distribution (see Foster et al., 2008, and Section 1.3 above), so it's reasonable to expect the same pattern in volatility. Demand variation by firm contributes significantly more to the firm-size distribution than does variation in productivity. Similarly, idiosyncratic demand shocks may contribute significantly more to volatility than does idiosyncratic productivity shocks.

The last important note: idiosyncratic shocks may or may not be correlated. First, I attempt to match aggregate volatility by using uncorrelated shocks, but if the simulations cannot match the data, I'll re-examine the assumptions, and see how far idiosyncratic shocks can go with reasonable parameter estimates.

1.5.4 Counterfactuals of the firm size distribution

To examine the effect of the demand network, productivity, and the interplay between these factors, I perform several counterfactuals on the data and model. The general idea is to remove variation in one or more of the parameters, solve the model, and (i) compare the true firm density with the counterfactual firm density, and (ii) regress the firm size from the data on the firm size implied by the counterfactual. The density comparison gives an effective visual comparison of the effect of each factor, but lacks sufficient detail to reject any hypotheses. Specifically, the firm density may be similar, but the rank of firm sizes may be scrambled, suggesting the distribution of parameters may give rise to similar aggregate effects, but the underlying parameters do not match the data well. In this case, it is better to compare the individual firm sizes with their corresponding counterfactuals. That is, compare firm i's actual size v_i with its implied size $\hat{v}_{xi}$ after performing some counterfactual x. That gives a better idea of what is truly determining the density by asking what determines the individual units that make up the density.

Demand network

In order to test the importance of the unobserved demand network to the firm size distribution, I eliminate variation in all other factors, recalculate the model and compare the resulting firm sizes with the firm sizes observed in the data. Specifically, I set $z_i = \bar{z}$ and $\tau_{ij} = \bar{\tau}$, and leave β_i, Λ and Γ at their original levels, and then recalculate the set of firm sizes v_i implied by the model.

The importance of higher order interconnections, counterfactual version

In Section 1.4.2, I found that higher order interconnections were significant determinants of the firm size distribution (or more specifically, one dimensional firm attributes like productivity and outdegree cannot explain much of the observed firm size distribution, which leaves the rest to be explained by the interactions betweent the two). Here, I offer similar evidence from a different method. Suppose the counterfactual firm demand networks Λ' and Γ' were such that the outdegrees were the same as the original networks, but the variation across customers for a given supplier was eliminated. Instead of variation across γ_{ij} for a given j, they are all set at a constant value of δ_j/N. The biggest source of change here is the extensive margin—setting the demand network to a constant adds all the firm connections that originally did not exist, turning the network from incredibly sparse to as dense as possible. Then, keeping productivity and value added shares as they are, recalculate the firm sizes.

This strategy keeps the outdegree centrality of a firm constant across the data and counterfactual, but eliminates the true variability in the higher-order interconnections betweens firm demand and productivity. Specifically, a firm that had been a central supplier to some subset of the economy, the same firm is of equal importance to the economy, but spreads the importance of its demand over the entire set of firms in the economy. This eliminates variation in the set of customers each firm has (and the set of customers those customers have), while keeping its 'importance' measures (and ranking thereof) intact. The resulting equilibrium sizes tell us how important the higher-order interconnections are for the economy.

Again, it is important to note that the skewness in the distribution of outdegree shown by Acemoglu et al. (2012) is not enough to explain the firm size distribution, since a skewed distribution of outdegree that results from a γ_{ij} that is constant across i will result in a skewed distribution of higher order outdegrees, but there's no guarantee that the resulting second order outdegree distribution explains firm sizes. In other words, there may be higher order variation in the data that does not match the pattern implied by a constant γ_{ij} across i. So, the evidence here will show not only that the higher order interconnections matter for the shape of the firm size distribution, but that the higher order interconnections matter for explaining the individual firm sizes themselves.

Productivity

Here, I ask whether productivity alone can match the firm size distribution. To remove the demand from the model, I eliminate all variation in demand and calculate the implied firm sizes. To be specific, I set $\gamma_{ij} = 1/N$ for all $i, j \in N$ and $\lambda_{ri} = 1/N$ for all $r \in R$ and $i \in N$, and $\tau_{ij} = \tau_{ri} = 1$ for all $r \in R$ and $i, j \in N$.

1.6 Results and discussion

There are several main results. The counterfactual firm densities are shown in Figure 1.1. The herfindahl, ratio of $90^{\text{th}}/10^{\text{th}}$ percentiles, and regression results are in Table 1.2. First, productivity accounts for very little, between 5-10%, of the existing firm size distribution. Second, the demand network accounts for much more, around 60% of the firm size distri-

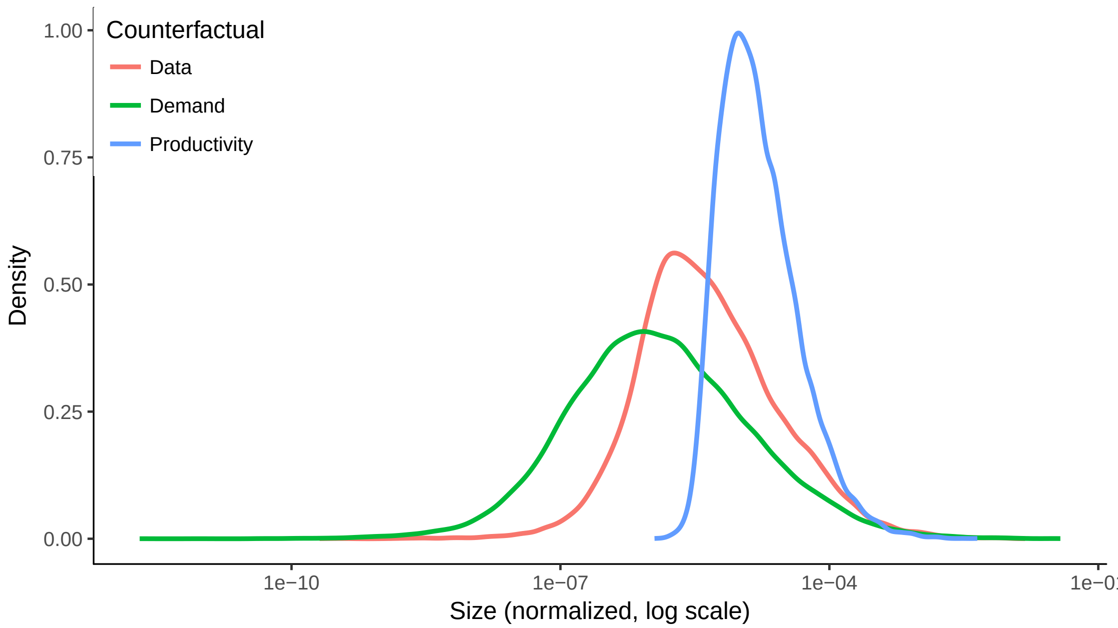


Figure 1.1: Counterfactual firm densities

Notes: 'x' is the resulting firm size density after removing all variation in the model except value added shares and the x parameters, where x is 'Demand,' or 'Productivity.'

bution, and much of that comes from higher-order interconnections. Finally, a reasonable calibration of idiosyncratic shocks can explain approximately one-third of aggregate volatility.

1.6.1 Counterfactual firm size densities

The firm density defined by productivity alone bears some resemblance to the empirical density but lacks the long right and left tails, suggesting there are demand characteristics that make some firms very small and very big relative to their productivity levels. This is reflected in the herfindahl, which is about 38% of the data, which would make aggregate volatility that much lower if productivity were the only source of variation in the data (see Section 1.6.2 for additional volatility results). Furthermore, the R^2 of a regression of $\log v_i$ on $\log v_{xi}$ is 0.092. This shows that productivity, although bearing visual similarities to the empirical distribution, cannot match the individual firm sizes themselves.

This result is robust to different measures of productivity, including different methods of estimating TFP and labour productivity. The fact that productivity does not vary enough or in the right directions to explain the firm size distribution accords well with other firm-level studies, including Holmes and Stevens (2014) and Hottman et al. (2016). Both show that demand characteristics explain much more of the firm size distribution than productivity, but in much different settings; Holmes and Stevens (2014) focuses on product differentiation and Hottman et al. (2016) focuses on scanner data for retail goods. Here, I show this same idea applies if you consider the input-output production network as defining demand characteristics.

That brings me to my main result: demand parameters explain much more of the firm size distribution. Visually, the shape of the Demand counterfactual distribution matches the data somewhat well, especially compared to the other counterfactuals. The mean is shifted left, with a slightly higher variance, with a similar right tail but longer left tail. Next, the herfindahl is slightly higher, 0.073 in the counterfactual to 0.046 in the data, implying volatility would increase if demand were the only firm variation in the economy. In addition, the percentile ratio is higher, with 54 in the data and 321 in the counterfactual, which is largely due to the long left tail of the Demand counterfactual distribution (see Figure 1.1). This significantly reduces the denominator of the 90/10 ratio. In spite of this drawback, the shape of the distribution of the demand counterfactual is very similar to the data. The Demand counterfactual does well explaining the individual firm sizes; the R^2 of a regression of $\log v_i$ on $\log v_{xi}$ gives an R^2 of 0.596, suggesting the demand measures alone explain 60% of the variation in the firm size distribution. In addition, removing higher order interconnections reduces the R^2 of the counterfactual sizes by 35 percentage points. Furthermore, the percentiles ratio increases substantially to an unreasonable number, again because of a very long left tail.

Although other studies of retail goods would consider the Λ and Γ parameters 'firm appeal,' and studies of production networks would call them direct-requirement or input-output parameters, they are conceptually the same. Here, an increase in γ_{ij} could mean an increase in preference by firm i for firm j's product, or a technical requirement for i to use j in production, and both are consistent with demand interpretations in other studies. The relevant distinction here is that these demand parameters are not constant within a firm j—different customers, both firms and final consumers, have different preferences for one firm's output. The interconnections between a firm's customer's preferences, and the preferences of their customers, and so on, have aggregate implications that single-firm measures cannot explain.

Note that each counterfactual has drawbacks, and cannot explain the firm size distribution alone. Specifically, demand explains a lot of the firm size distribution, but the herfindahl is actually higher after removing variation in productivity. This suggests the factors combine in complex ways, sometimes complementary (e.g., a firm with high demand characteristics is located close to its customers), sometimes not (e.g., a firm with higher than average productivity is in a remote area), to arrive at the final equilibrium.

1.6.2 Volatility

The contribution of microeconomic shocks to aggregate volatility depend on the skewness of the firm size distribution, and the skewness of the firm size distribution depends on the factors outined previously. Specifically, the contribution of idiosyncratic shocks to aggregate volatility can be calculated with the formula

$$\hat{\sigma}_{GDP} = \sum_i \left(\frac{s_i}{\sum_k \beta_k s_k} \right) \sigma_{zi}, \tag{1.21}$$

where the term in brackets is a firm-level Domar weight (sales over total value added), see Gabaix (2011) for a discussion of the justification Domar weights and Hulten's theorem. Using the weighted standard deviation of productivity as a measure of σ_{zi}, and writing β as the share of total value added in total output, this equation can be rewritten

$$\hat{\sigma}_{GDP} = \left(\frac{h}{\beta} \right) \sigma_z, \tag{1.22}$$

which provides an easy estimate of the contributions of microeconomic shocks to aggregate volatility. Using data on h, β, and σ_z, Table 1.3 shows the relative contribution of microeconomic shocks to aggregate volatility. These results are consistent with other studies of aggregate volatility.

In addition, the formula gives an easy calculation of aggregate volatility using counterfactual estimates of h and β. The sales herfindahl implied by the productivity distribution alone is very low, 0.018, and the aggregate value added share is higher at 0.70, giving an implied idiosyncratic volatility of 0.004, which lowers aggregate volatility by 25% (assuming the macroeconomic factors remain the same). However, using only variation in demand actually raises the herfindahl to 0.073, raising aggregate volatility by 11% (after accounting for a slight increase in the value added share).

1.7 Conclusion

In this paper, I ask whether productivity or network asymmetry provide better microfoundations for the propagation of idiosyncratic shocks. If granularity, a skewed firm size distribution, determines aggregate fluctuations, what determines granularity? Using detailed data on firm-firm trade in Canada, I study a firm-firm production network and its effect on aggregate volatility.

To differentiate between productivity and the unobserved demand network, I use a model in which these factors vary independently and use the production network data to uncover the model parameters. I find two main results: first, the demand network explains approximately 60% of the observed firm size distribution. One dimensional firm demand measures can only explain about 25%, which leaves higher order interconnections between firms to account for 35 p.p. of the firm size distribution. This suggests the complex demand *network*, i.e., your customers and the customers of your customers, is a significant determinant of the firm size. Second, I find that productivity only explains 10% of the firm size distribution. Productivity does not vary enough to explain the aggregate shape of the distribution and is not correlated enough with firm size to explain much of the individual sizes themselves.

Finally, reasonable levels of idiosyncratic shocks can account for approximately 32% of aggregate volatility. Counterfactual estimates suggest that removing cross-sectional demand variation in the economy would reduce aggregate volatility by 25%, while removing productivity variation would increase it by 11%.

The major conclusion to draw from this paper, and something that sets the stage for future work, is that the empirical results confirm the idea that the demand network significantly determines the firm size distribution and aggregate volatility. Futhermore, higher order interconnections between firms explain a large part of the firm size distribution. Firm-firm trade is complex, and studying the implications of the production network for aggregate volatility, trade, transaction costs, vertical integration, and many other subjects, will require much more theoretical and empirical work.

1.8 Appendix: Theory

See important model notation in Table 1.5.

1.8.1 Full model

Consumers

There are R regions, with a representative consumer in each with utility function $u_r(c_r)$,

$$u_r(c_r) = \left(\sum_{i \in N} \lambda_{ri}^{\frac{1}{\epsilon}} c_{ri}^{\frac{\epsilon-1}{\epsilon}} \right)^{\frac{\epsilon}{\epsilon-1}} \tag{1.23}$$

Labour is inelastically supplied given the stock of labour in region r, L_r. Consumer r's problem is

$$\max_{c_r} u_r(c_r) \text{ s.t. } \sum_{i \in N} p_{ri} c_{ri} \le w_r L_r \tag{1.24}$$

Consumer r must pay a trade cost τ_{ri} to buy from plant i, so that

$$p_{ri} = \tau_{ri} p_i \tag{1.25}$$

The solution gives r's price index

$$p_r = \left(\sum_{i \in N} \lambda_{ri} (\tau_{ri} p_i)^{1-\epsilon} \right)^{\frac{1}{1-\epsilon}} \tag{1.26}$$

Producers

There are N producers. Producer i's production function is

$$f_i(l_i, q_{i1}, \ldots, q_{iN}) = z_i l_i^{\beta_i} \left(\sum_{j \in N} \gamma_{ij}^{\frac{1}{\eta}} q_{ij}^{\frac{\eta-1}{\eta}} \right)^{\frac{(1-\beta_i)\eta}{\eta-1}} \tag{1.27}$$

Producer i's problem is to minimize cost

$$\min_{(l_i, q_{i1}, \ldots, q_{iN})} \sum_{i \in N} p_{ij} q_{ij} \text{ s.t. } f_i \geq \bar{q}_i \tag{1.28}$$

Producer i's input cost for one unit of the intermediate input is

$$p_{mi} = \left(\sum_{j \in N} \gamma_{ij} (\tau_{ij} p_j)^{1-\eta} \right)^{\frac{1}{1-\eta}} \tag{1.29}$$

Given perfect competition, plant i's price is (including wages),

$$p_i = \beta_i^{-\beta_i} (1-\beta_i)^{\beta_i - 1} z_i^{-1} p_{mi}^{1-\beta_i} \tag{1.30}$$

Market clearing

Labour is free to migrate between regions. Total labour in the economy is

$$\sum_{r \in R} L_r = L \tag{1.31}$$

Now, each plant i is in a region r, and the total value added produced by those plants in r add up to total income in that region,

$$\sum_{i \in r} \beta_i s_i = w L_r \tag{1.32}$$

For goods, producer i supplies the other producers $j \in N$, and each region $r \in R$, giving market clearing

$$\sum_{r \in R} c^s_{ri} + \sum_{j \in N} q^s_{ji} = q^s_i, \text{ for } i \in N \tag{1.33}$$

Iceberg trade costs mean producer i ships $c^s_{ri} = \tau_{ri} c_{ri}$ to region r and $q^s_{ji} = \tau_{ji} q_{ji}$. Replacing those terms and multiplying all terms by p_i,

$$\sum_{r \in R} p_i \tau_{ri} c_{ri} + \sum_{j \in N} p_i \tau_{ji} q_{ji} = p_i q^s_i, \text{ for } i \in N \tag{1.34}$$

Equilibrium

Equilibrium in the economy means two sets of prices $\{p_r : r \in R\}$, $\{p_i : i \in N\}$, wage w normalized to 1, and labour stocks by region $\{L_r : r \in R\}$, that solve the consumer's and producer's problems for each region and producer, and the labour and goods markets clear.

Solving the model given data

Given data on T, G, A, w, β, solve for Λ and Γ. We must also solve for prices of p_r and p_i that are incidental to the desired parameters, and normalize $w = 1$. In addition, I make assumptions about the elasticities η and ϵ. I have price equations:

$$p_r = \left(\sum_{i \in N} \lambda_{ri} (\tau_{ri} p_i)^{1-\epsilon} \right)^{\frac{1}{1-\epsilon}}, \text{ for } r \in R \tag{1.35}$$

$$p_{mi} = \left(\sum_{j \in N} \gamma_{ij} (\tau_{ij} p_j)^{1-\eta} \right)^{\frac{1}{1-\eta}}, \text{ for } i \in N \tag{1.36}$$

$$p_i = z_i^{-1} \beta_i^{-\beta_i} (1 - \beta_i)^{\beta_i - 1} w^{\beta_i} p_{mi}^{1-\beta_i}, \text{ for } i \in N \tag{1.37}$$

And share equations:

$$a_{ri} = \lambda_{ri} \tau_{ri}^{1-\epsilon} \left(\frac{p_i}{p_r} \right)^{1-\epsilon}, \text{ for } r \in R, i \in N \tag{1.38}$$

$$g_{ij} = (1 - \beta_i) \gamma_{ij} \tau_{ij}^{1-\eta} \left(\frac{p_j}{p_{mi}} \right)^{1-\eta}, \text{ for } i \in N, j \in N \tag{1.39}$$

$$\lambda_{ri} = a_{ri}\tau_{ri}^{\epsilon-1}\left(\frac{p_i}{p_r}\right)^{\epsilon-1}, \text{ for } r \in R, i \in N \tag{1.40}$$

$$\gamma_{ij} = (1-\beta_i)^{-1} g_{ij}\tau_{ij}^{\eta-1}\left(\frac{p_j}{p_{mi}}\right)^{\eta-1}, \text{ for } i \in N, j \in N \tag{1.41}$$

And region income equations,

$$\beta_i s_i = w l_i \tag{1.42}$$

$$\sum_{i\in R} \beta_i s_i = w L_r \tag{1.43}$$

$$\sum_{r\in R} L_r = L \tag{1.44}$$

And finally, sizes:

$$wA'\vec{L} + G's = s, \text{ or} \tag{1.45}$$

$$s = w(I-G')^{-1}A'\vec{L} \tag{1.46}$$

How many unknowns are there in this system? $p_r \to R$, $p_i, p_{mi} \to 2N$, $\Lambda \to RN$, $\Gamma \to N^2$, $s \to N$, $L_r \to R$, w. So $R + 2N + RN + N^2 + N + R + 1$. How many equations? $R + 2N + RN + N^2 + R + 1 + N$. The number of equations is the same as the number of unknowns.

Solving the model given parameters

Once we uncover the underlying parameters of the model, we'd like to simulate it. Given the same equations, and given with z, β, T (data), η (by assumption), Γ, Λ, solve the same equations for the outcome variables s, all p, A, G. That is, solve for firm sizes, prices, and observed input-output parameters.

Solve for Λ, Γ

Given data on T, G, A, w, β, solve for Λ and Γ. We must also solve for the prices p_r and p_i that are incidental to the model, and normalize $w = 1$. In addition, I make assumptions about the elasticities η, ϵ. We have price equations:

$$p_r = \left(\sum_{i \in N} \lambda_{ri} (\tau_{ri} p_i)^{1-\epsilon} \right)^{\frac{1}{1-\epsilon}}, \text{ for } r \in R \tag{1.47}$$

$$p_{mi} = \left(\sum_{j \in N} \gamma_{ij} (\tau_{ij} p_j)^{1-\eta} \right)^{\frac{1-\beta_i}{1-\eta}}, \text{ for } i \in N \tag{1.48}$$

$$p_i = \beta_i^{-\beta_i} (1-\beta_i)^{\beta_i - 1} w^{\beta_i} z_i^{-1} p_{mi}, \text{ for } i \in N \tag{1.49}$$

$$\lambda_{ri} = a_{ri} \tau_{ri}^{\epsilon-1} \left(\frac{p_i}{p_r} \right)^{\epsilon-1}, \text{ for } r \in R, i \in N \tag{1.50}$$

$$\gamma_{ij} = (1-\beta_i)^{-1} g_{ij} \tau_{ij}^{\eta-1} \left(\frac{p_j}{p_{mi}} \right)^{\eta-1}, \text{ for } i \in N, j \in N \tag{1.51}$$

To solve this system, propose initial values for Λ_0 and Γ_0, then solve for all unknowns. Given those unknowns and the data, solve back for new candidate solutions Λ_1 and Γ_1, then check how close the new solutions are to the previous solutions. If they're close enough, stop, if not, use the new solutions to generate another set of candidates. Repeat.

1.8.2 Derivation of influence vector

Using the definition of observed expenditure shares,

$$g_{ji} = \frac{\tau_{ji} p_i q_{ji}}{p_j q_j} \tag{1.52}$$

Rewrite the system of market clearing equations

$$\sum_{r \in R} \tau_{ri} p_i c_{ri} + \sum_{j \in N} \tau_{ji} p_i q_{ji} = p_i q_i, \text{ for } i \in N \tag{1.53}$$

as

$$\sum_{r \in R} \tau_{ri} p_i c_{ri} + \sum_{j \in N} g_{ji} p_j q_j = p_i q_i, \text{ for } i \in N \tag{1.54}$$

Then replace $\tau_{ri}p_i c_{ri} = a_{ri}p_r c_r = a_{ri}wL_r$ and define total sales as $s_i = p_i q_i$,

$$\sum_{r\in R} a_{ri}wL_r + \sum_{j\in N} g_{ji}s_j = s_i, \text{ for } i \in N \tag{1.55}$$

Rewrite in vector form, using $L = (L_1, \ldots, L_R)'$, write $a_{\cdot i}$ as the i-th column of A and $g_{\cdot i}$ as the i-th column of G,

$$wa'_{\cdot i}L + g'_{\cdot i}s = s_i, \text{ for } i \in N \tag{1.56}$$

Now stack those N equations on top of each other, which stacks the vectors $g'_{\cdot i}$ (now the *row* vectors of G'), which gives

$$wA'L + G's = s \tag{1.57}$$

Rearrange and factor out s,

$$s - G's = wA'L \tag{1.58}$$

$$(I - G')s = wA'L \tag{1.59}$$

Then pre-multiply by the Leontief matrix, the inverse of $(I - G')$,

$$s = w(I - G')^{-1}A'L \tag{1.60}$$

To get the influence vector, use $w\mathbf{1}'L = \beta\sum_{i\in N} s_i$ and $v_i = s_i/\left(\sum_{j\in N} s_j\right)$, and finally normalize wages to 1 ($w = 1$) and take the transpose of both sides:

$$v' = \left(\frac{\beta}{\mathbf{1}'L}\right) L'A(I - G)^{-1} \tag{1.61}$$

If value-added varies across plants, the relevant equation is

$$A'\overrightarrow{(\beta'v)_r} + G'v = v \tag{1.62}$$

Or,

$$A'\overrightarrow{(\beta'v)_r} = (I - G')^{-1}v \tag{1.63}$$

1.8.3 Asymptotic Theory

Asymptotic results are key to the arguments for and against the microfoundations of aggregate shocks.[3] The granular hypothesis relies on a thick tail of the size distribution. The unbalanced network hypothesis claims the reason *why* the size distribution has a thick tail is because of a thick tail of outdegree, a telling characteristic of an asymmetric production network. Only by combining the two approaches can we understand the forces that shape the observed centrality and size distributions.

In what follows, I rely especially on the following property of power law distributions:

Remark 3 *Suppose the random variables X and Y follow power law distributions with parameters ζ_X and ζ_Y. Then the distribution of $X + Y$ and the distribution of XY both follow power laws with parameter* $\min\{\zeta_X, \zeta_Y\}$*.*

The same result follows for many similar combinations of power law random variables (see Gabaix, 2009; Jessen and Mikosch, 2006). Using Remark 3, we are interested in explaining the tail parameter of the size distribution, β_v, given the tail parameters of the distributions of observed outdegree (ζ_d) and productivity (ζ_z).

Therefore, if the asymptotic results hold for this economy, network asymmetry cannot be the fundamental cause of the skewed firm size distribution because of the relative values of each tail parameter. But like so many other applications of power laws, the reality is not so black and white. In any case, we must understand the asymptotic argument first, and then ask if and when is it reasonable to apply it.

The network hypothesis relies on two sequential arguments. First, the tail of the distribution of the firm-level exogenous production network characteristics must determine the tail of the distribution of the observed firm-level production network characteristics. Second,

[3] In Appendix 1.8.4, I use Hulten's Theorem to show aggregate volatility depends on the herfindahl of the economy, and the herfindahl of the economy depends on the distribution of outdegree and productivity. These results are standard when applying the granular and network theories of aggregate fluctuations, so I omit them and focus on the new idea provided in this paper.

the tail of the distribution of the observed production network determines the tail of the firm size distribution. If either of these arguments fail, it is unlikely the underlying demand characteristics are the cause of the skewed firm size distribution.

I approach the second part of the argument first. For the observed network to matter asymptotically, the outdegree distribution must have a thick tail. If not, outdegree cannot be the ultimate source of the thick tail of the size distribution. If the outdegree distribution does have a thick tail, the parameter must match, or be "close" to matching (in a statistical sense) the tail of the size distribution. However, the measured tail parameter for the network is 1.21, about 20% higher than the firm size distribution's parameter of 1.04, which is consistent with a Zipf's law distribution of firm size. Therefore $\zeta_z < \zeta_d$ implies the degree distribution is dominated by some other firm characteristic, and thus does not determine firm size asymptotically or turn idiosyncratic shocks into aggregate fluctuations.

We can see this conclusion supported by prior research in different settings. A plethora of research on the firm size distribution conclude it is approximately described by Zipf's law in the upper tail (see Luttmer, 2007; Gabaix, 2009), while Acemoglu et al. (2012) measure the tail of the sector outdegree distribution at 1.38, much larger than the typical Zipf's law size distribution parameter of 1.

The first part of the argument, the required relationship between the observed and unobserved network characteristics is more problematic. The production network data are necessarily the observed shares, and so depend on both the underlying demand characteristics and other firm characteristics, especially productivity.

To establish this formally, I show that, under the assumptions of the model in the previous section, the tail of the size distribution is dominated by the thickest tail between productivity (adjusted for substitutability) and outdegree.

Proposition 1.8.1 *Suppose the distributions of outdegree and productivity both follow power* *laws with parameters* ζ_d *and* ζ_z,

$$P(d > x) = C_d x^{-\zeta_d} L_d(x), \tag{1.64}$$

$$P(z > x) = C_z x^{-\zeta_z} L_z(x) \tag{1.65}$$

Here, $L_d(x)$ *and* $L_z(x)$ *are slowly varying functions,* C_d *and* C_z *are constants, and* ζ_d *and* ζ_z *are positive. Then the size distribution also follows a power law with parameter* $\min\{\zeta_d, \zeta_z/(\eta-1)\}$,

$$P(v > x) = C_v x^{-\min\left\{\zeta_d, \frac{\zeta_z}{(\eta-1)}\right\}} L_v(x) \tag{1.66}$$

Proof 1 *One element of the influence vector,* v_i, *is*

$$v_i = \frac{\beta}{N} + (1-\beta)\left(\frac{d_i z_i^{\eta-1}}{\sum_{k\in N} d_k z_k^{\eta-1}}\right) \tag{1.67}$$

As $N \to \infty$, *the first term approaches zero, and the distribution of* w *is determined by the relative product term* $d_i z_i^{\eta-1}$, *which means*

$$v_i \to \chi d_i z_i^{\eta-1} \tag{1.68}$$

$$F_v(x) = F_v\left(\chi d_i z_i^{\eta-1}\right) \tag{1.69}$$

$$P(v > x) \to P(\chi d z^{\eta-1} > x) \tag{1.70}$$

$$= P(d z^{\eta-1} > \chi^{-1} x) \tag{1.71}$$

$$P(v > x) = P(d z^{\eta-1} > \chi^{-1} x) \tag{1.72}$$

$$= \int_{\underline{d}}^{\infty} P\left(z > \left[\frac{x}{\chi d}\right]^{1/(\eta-1)}\right) \mathrm{d}F_d(d) \tag{1.73}$$

$$= \int_{\underline{d}}^{\infty} C_z \left[\frac{x}{\chi d}\right]^{-\zeta_z/(\eta-1)} \mathrm{d}F_d(d) \tag{1.74}$$

$$= \chi^{\zeta_z/(\eta-1)} C_z x^{-\zeta_z/(\eta-1)} \int_{\underline{d}}^{\infty} d^{\zeta_z/(\eta-1)} \mathrm{d}F_d(d) \tag{1.75}$$

For the integral to exist, we need $\zeta_z/(\eta-1) < \zeta_d$. *If so, it is a constant (independent of* x*), so combine the other constants into* $C_v = \chi^{\zeta_z/(\eta-1)} C_z \int_{\underline{d}}^{\infty} d^{\zeta_z/(\eta-1)} \mathrm{d}F_d(d)$, *and write*

$$P(v > x) = C_v x^{-\zeta_z/(\eta-1)} \tag{1.76}$$

So v has a power law distribution with parameter $\zeta_z/(\eta-1)$. If $\zeta_z/(\eta-1) > \zeta_d$, we need to derive it the other way, and end up with a power law distribution with parameter ζ_d. Therefore the distribution can be expressed by

$$P(v > x) = C_v x^{-\min\{\zeta_d, \zeta_z/(\eta-1)\}} \tag{1.77}$$

Or,

$$\log P(v > x) = \log C_v - \min\{\zeta_d, \zeta_z/(\eta-1)\} \log x \tag{1.78}$$

The distribution of productivity has a tail parameter of approximately 1.98, so for a suitable choice of η, it is easy to match the empirical tail parameter of the firm size distribution. In particular, if $\eta \approx 2.89$, the size distribution will approximately satisfy Zipf's law. It also could satisfy both, if substitutability for final goods is higher than for intermediates. Note that similar studies on productivity and size, especially ones focusing on international trade models, (e.g., see Appendix 1.9.2 for an extension of the model with monopolistic competition and firm entry and exit) gives the same result—firm size is determined by a combination of productivity and substitutability, with the size tail parameter being very close to 1 (see, e.g., a series of papers by di Giovanni and Levchenko and their co-authors (Di Giovanni et al., 2011; Di Giovanni and Levchenko, 2012, 2013). The difference here is that they observe the size distribution and assume it *must* be because of productivity. For more on power laws and the determination of firm size, see Luttmer (2007) or Gabaix (2009).

Although the asymptotic theory gives clear cut answers as to which factor is responsible for the shape of the size distribution, the empirical results suggest the truth is somewhere between the two extremes.

1.8.4 Aggregate volatility depends on the product of the distributions of outdegree and productivity

Aggregate volatility scales according to $||v||_2$, according to Hulten's Theorem (Hulten, 1978) and Theorem 1 of Acemoglu et al. (2012). To add to those results, I characterize the behaviour of $||v||_2$ in terms of the distributions of outdegree and productivity.

Write an element of the influence vector v_i as

$$v_i = \frac{\beta}{N} + (1-\beta)\left(\frac{d_i z_i^{\eta-1}}{\sum_{k\in N} d_k z_k^{\eta-1}}\right) \tag{1.79}$$

Then the Euclidean norm of v can be written

$$||v||_2 = \sqrt{\sum_{i\in N}\left[\frac{\beta^2}{N^2} + (1-\beta)^2\left(\frac{d_i z_i^{\eta-1}}{\sum_{k\in N} d_k z_k^{\eta-1}}\right)^2 + 2(1-\beta)\left(\frac{\beta}{N}\right)\left(\frac{d_i z_i^{\eta-1}}{\sum_{k\in N} d_k z_k^{\eta-1}}\right)\right]} \tag{1.80}$$

$$||v||_2 = \sqrt{\frac{\beta^2}{N} + (1-\beta)^2\sum_{i\in N}\left(\frac{d_i z_i^{\eta-1}}{\sum_{k\in N} d_k z_k^{\eta-1}}\right)^2 + 2(1-\beta)\left(\frac{\beta}{N}\right)\sum_{i\in N}\left(\frac{d_i z_i^{\eta-1}}{\sum_{k\in N} d_k z_k^{\eta-1}}\right)} \tag{1.81}$$

Rewrite slightly,

$$||v||_2^2 = \frac{\beta^2}{N} + (1-\beta)^2\sum_{i\in N}\left(\frac{d_i z_i^{\eta-1}}{\sum_{k\in N} d_k z_k^{\eta-1}}\right)^2 + 2(1-\beta)\left(\frac{\beta}{N}\right) \tag{1.82}$$

$$||v||_2^2 = \frac{\beta^2}{N} + 2(1-\beta)\left(\frac{\beta}{N}\right) + (1-\beta)^2 h_g^2 \tag{1.83}$$

$$||v||_2^2 = \frac{\beta(2-\beta)}{N} + (1-\beta)^2 h_g^2 \tag{1.84}$$

$$||v||_2^2 \geq (1-\beta)^2 h_g^2 \tag{1.85}$$

Implying $||v||_2^2 = \Omega\left(h_g^2\right)$. In addition, $||v||_2^2 = \mathcal{O}\left(h_g^2\right)$. To see this, first note

$$h_g^2 \geq \frac{1}{N}\left(\sum_{i\in N}\frac{d_i z_i^{\eta-1}}{\sum_{k\in N} d_k z_k^{\eta-1}}\right)^2 = \frac{1}{N} \tag{1.86}$$

which we can rearrange to get $1/(Nh_g^2) \leq 1$.

$$||v||_2^2/h_g^2 = \frac{\beta(2-\beta)}{Nh_g^2} + (1-\beta)^2 \tag{1.87}$$

Meaning

$$\limsup_{N\to\infty}\frac{||v||_2^2}{h_g^2} = \limsup_{N\to\infty}\left[\frac{\beta(2-\beta)}{Nh_g^2} + (1-\beta)^2\right] \tag{1.88}$$

Using the result that $(Nh_g^2)^{-1}$ is bounded above by 1,

$$\limsup_{N\to\infty} \frac{||v||_2^2}{h_g^2} \leq \limsup_{N\to\infty} \left[\beta(2-\beta) + (1-\beta)^2\right] \tag{1.89}$$

$$\limsup_{N\to\infty} \frac{||v||_2^2}{h_g^2} \leq \beta(2-\beta) + (1-\beta)^2 < \infty \tag{1.90}$$

So $||v||_2^2 = \mathcal{O}\left(h_g^2\right)$, which combined with the Big-Ω result gives

$$||v||_2 = \Theta\left(h_g\right) \tag{1.91}$$

1.9 Appendix: Data and Empirics

1.9.1 Data sources

Additional descriptions of available data available at CDER: `http://www.statcan.gc.ca/eng/cder/data`.

Annual Survey of Manufacturing (ASM)

Also called the Annual Survey of Manufacturing and Logging (ASML). See `http://www.statcan.gc.ca/eng/survey/business/2103`, and an example survey at `http://www23.statcan.gc.ca/imdb-bmdi/instrument/2103_Q31_V3-eng.pdf`. Years available: 1961-2012.

Surface Transportation File (STF)

Based on the Trucking Commodity Origin and Destination File and Railway Universe File. Transaction-level trade database with shipper and receiver names, addresses and postal codes. Used to identify input shipments between establishments, and final demand shipments from establishments to regions. Includes information on carrier, mode, commodity classification (SCTG), value, tonnage, distance, and revenue to the carrier. Years available: 2004-2012.

Inter-provincial Trade Flows (IPTF)

CANSIM Tables 386-0001, 386-0002, 386-0003, 386-0004, `http://www5.statcan.gc.ca/cansim/a04`. I use the detailed-confidential versions of these tables in the paper. A province $\times$ province $\times$ commodity dataset of trade, including international imports, exports and re-exports. Years available: 2002-2012.

Input-Output Tables / Supply-Use Tables (IO)

CANSIM Tables 381-0033, 381-0034, 381-0035, `http://www5.statcan.gc.ca/cansim/a04`. I use the detailed-confidential versions of these tables in the paper. An province $\times$ industry $\times$ commodity dataset. Industry classification is IOIC, commodity classification is IOCC. Years available: 2002-2012.

Import-Export Registry (IER)

Records enterprise-product level imports and exports. I use this to impute the import share of each firm in order to generate an 'international' region. For more information, see `http://www.statcan.gc.ca/eng/cder/data#a2`.

1.9.2 Intensive and Extensive Margins of Volatility

In the main text, I assume there is no extensive margin of volatility. One may wonder how the results change if I allow for plant entry and exit. To test this empirically, I use a similar decomposition to Di Giovanni et al. (2014).

First, write sales of plant i at year t as s_{it}. Let I_t be the set of plants operating in year t, and $I_{t/t-1}$ be the set of plants operating in both years t and $t-1$. Then the log-difference aggregate growth rate of sales is

$$\tilde{g}_{At} \equiv \ln\left(\sum_{i\in I_t} x_{it}\right) - \ln\left(\sum_{i\in I_{t-1}} x_{it-1}\right) \tag{1.92}$$

$$= \ln\left(\frac{\sum_{i\in I_{t/t-1}} x_{it}}{\sum_{i\in I_{t/t-1}} x_{it-1}}\right) - \left[\ln\left(\frac{\sum_{i\in I_{t/t-1}} x_{it}}{\sum_{i\in I_t} x_{it}}\right) - \ln\left(\frac{\sum_{i\in I_{t/t-1}} x_{it-1}}{\sum_{i\in I_{t-1}} x_{it-1}}\right)\right] \tag{1.93}$$

$$= g_{At} - \ln\left(\frac{\nu_{t,t}}{\nu_{t,t-1}}\right) \tag{1.94}$$

where g_{At} is the intensive margin of growth and the other term is the extensive margin of growth. Now aggregate volatility is

$$\tilde{\sigma}_A^2 = \sigma_A^2 + \sigma_\nu^2 - 2\text{Cov}(g_{At}, g_\nu) \tag{1.95}$$

Calculating each of these in the data, we see that the extensive margin matters little (consistent with the results in Di Giovanni et al. (2014). Although large establishments do exit, it is more common for one to have large losses in one year, have a low value of output, and then exit the following year. This puts the volatility on the intensive margin, not extensive.

Table 1.1: Skewness of main variables

	Mean	Median	S.D.	90/10	Tail, $\hat{\beta}$
Output ($\$ \times 10^6$)	16.46	2.1	147.03	42.92	0.99
Value added ($\$ \times 10^6$)	6.42	1.1	36.69	49.80	1.05
Value added share	0.55	0.6	0.18		
TFP (Naïve)	1.10	1.0	1.46	2.12	1.98
TFP (GNR)	1.06	1.0	1.17	1.71	1.99
Labour prod. (Emp.)	1.23	1.1	0.85	4.89	3.99
Labour prod. (Pay.)	1.11	1.0	0.60	2.78	3.81
Outdegree	0.45	0.1	2.09	439.11	1.61

Notes: Output and value added are measured in millions of Canadian dollars. The 90/10 ratio is the ratio of the 90th percentile to the 10th percentile of the distribution of the variable. Output, value added, TFP and labour productivity are from the ASM, 2010. The tail parameter is estimated using the method of Gabaix and Ibragimov (2011). Outdegree d_i is calculated with the observed production networks A and G.

Table 1.2: Counterfactual firm density statistics

	Herfindahl	90/10	Coef.	R^2
Data	0.048	54.0		
Demand (§1.5.4)	0.073	321.92	0.536	0.596
Higher Order (§1.5.4)	0.052	14683.71	0.220	0.245
Productivity (§1.5.4)	0.018	27.72	0.379	0.092

Notes: The coefficient and R^2 are from a regression of $\log v_i$ on $\log v_{xi}$, where v_{xi} is the predicted value of the firm size in counterfactual scenario x, where x can be 'Demand', 'Productivity' or 'Higher Order'. 'Demand', and 'Productivity' counterfactuals are the resulting firm size density after removing all variation in the model except value added shares and x. 'Higher Order' is the resulting firm size density after *removing* the higher order interconnections between demand and productivity.

Table 1.3: Microeconomics shocks and aggregate volatility in the data

Productivity, z	σ_z	$\hat{\sigma}_{GDP}$	Rel. S.D.
TFP (Naïve)	0.17	0.019	0.32
TFP (GNR)	0.27	0.031	0.51

Notes: σ_z is the weighted standard deviation of productivity shocks. I remove industry and region shocks from z in an attempt to approximate idiosyncratic productivity shocks. The sales Herfindahl in the data is $h = 0.048$, the share of value added in aggregate sales is $\beta = 0.41$. The implied volatility is defined as $\hat{\sigma}_{GDP} = \sigma_z h/\beta$. Actual value added volatility is 0.06.

Table 1.4: Robustness of firm size counterfactuals to different measures of productivity

	Productivity					
	LP (Pay.)		TFP (Naïve)		TFP (GNR)	
	Coef.	R^2	Coef.	R^2	Coef.	R^2
Demand	0.536	0.596	0.581	0.603	0.549	0.619
Higher order	0.220	0.245	0.210	0.246	0.216	0.239
Productivity	0.379	0.092	0.475	0.146	0.388	0.089

Notes: productivity measures are described in Section 1.4.1. The coefficient and R^2 are from a regression of $\log v_i$ on $\log v_{xi}$, where v_{xi} is the counterfactual firm size in each case x, where x can be 'Demand,' 'Higher order,' or 'Productivity.' All coefficients are statistically significant with t-stats of less than 2×10^{-16}, so I omit standard errors from the table.

Table 1.5: Table of Notation

R	$\triangleq$	Set of regions. Abusing notation, R is also the number of regions.
N	$\triangleq$	Set of plants. Abusing notation, N is also the number of plants.
G	$\triangleq$	$N \times N$ matrix of observed plant input shares. An element g_{ij} is the share of plant j's input in plant i's sales.
Γ	$\triangleq$	$N \times N$ matrix of exogenous plant input demand characteristics. An element γ_{ij} enters plant i's demand for plant j's output.
A	$\triangleq$	$R \times N$ matrix of observed region-plant demand shares. An element a_{ri} is the share of region r's total expenditure on plant i's output.
Λ	$\triangleq$	$R \times N$ matrix of exogenous region input demand characteristics. An element λ_{ri} enters region r's demand for plant i's output.
T	$\triangleq$	$(R+N) \times (R+N)$ matrix of trade costs. An element τ_{ri} is the cost of trade between region r and i, and an element τ_{ij} is the cost of trade between plants i and j.
z_i	$\triangleq$	Productivity of plant i.
ϵ	$\triangleq$	Final demand elasticity of substitution.
η	$\triangleq$	Intermediate elasticity of substitution.
β_i	$\triangleq$	Share of value-added in plant i's production.

Table 1.6: Intensive vs. Extensive Margin Volatility

Volatility measure	S.D.	Rel. S.D.
Aggregate Volatility, $\tilde{\sigma}_A$	0.065	1.00
Intensive Volatility, σ_A	0.066	1.02
Extensive Volatility, σ_ν	0.009	0.14

Notes: Aggregate volatility is the standard deviation of total manufacturing output. Intensive volatility is the standard deviation of total manufacturing output from firms that are alive in periods t and $t-1$. Extensive volatility is the standard deviation of total manufacturing output from firms that entered or exited in period t.

Chapter 2

Correlated shocks within firms

2.1 Abstract

Due to its association with cross-country business cycles, propagation of idiosyncratic shocks, and even financial contagion, firm comovement is an important facet of macroeconomic research. However, we know little about whether pairs of establishments within firms comove more than pairs of establishments from different firms. Using a long panel of Canadian manufacturing establishments, I investigate the correlations and covariances of within-firm pairs of establishments and decompose them into labour inputs, intermediates and profit. I find that within-firm establishment pairs have correlations 0.0477 higher than between-firm establishment pairs (which have an average correlation very close to zero) after controlling for industry and region effects. Covarying intermediate input costs account for 49% of the within-firm comovement effect.

2.2 Introduction

This paper explores the determinants of comovement between and within firms. Firm comovement has recently been blamed for things like financial contagion, the propagation of idiosyncratic shocks, and explaining positive cross-country business cycle correlations. However, most of the microeconomic research has focused on individual-level measures and explanations (e.g., size, granularity) and have not focused on the (even more micro-) economic linkages between and within firms themselves. I use detailed establishment-level panel data to investigate the magnitude and significance of these comovements.

The research on firm comovement is broadly divided into macro and microeconomic areas. In the macroeconomic literature, aggregate GDP comovement is either associated with aggregate measures of integration (the country-country strategy) or individual firms and their integration with other countries (the firm-country strategy).

The country-country studies, such as Frankel and Rose (1998); Kose and Yi (2006), have little information on the actual mechanism driving the comovement—do establishments within multinationals really comove? Is it because of vertical linkages? Or financial dependence? Or common firm-level shocks? Is it because of capital, labour, or intermediate input comovement? Although the qualitative mechanisms seem obvious, we do not seem to understand them quantitatively (Johnson, 2014).

The firm-country studies in the microeconomic literature on firm comovement attempts to address and uncover these mechanisms, and have shown that shocks to parent firms are correlated with aggregate movements in regions where they have affiliate firms (see Kleinert et al., 2015; Cravino and Levchenko, 2016; di Giovanni et al., 2016, 2017). Overall, the firm-country research has aggregate results and calibrated models that strongly suggest firm-comovement can account for a significant amount of cross-country correlations. Previous empirical and theoretical work on firm comovement (or shock transmission and linkages in general) identified a myriad of possible causes: vertical linkages Burstein et al. (2008), internal capital markets Lamont (1997); Stein (1997), technology shocks Atalay et al. (2014), rent sharing Budd et al. (2005) or labour reallocation Giroud and Mueller (2016). I make identifying the within and between firm comovement itself the primary goal, and discover the components of those comovements.

To tackle this problem, I use data from the Annual Survey of Manufacturers, which is a defacto Census of manufacturing activity in Canada.[1] There are approximately 100,000 total plants in the sample, with around 30,000 alive per year. I focus on the period 1973-1999 to take advantage of consistent surveys and industry classifications. The ASM includes

[1]The data are available from CDER at Statistics Canada; see `http://www.statcan.gc.ca/eng/cder/data`.

establishment sales, value added, labour and intermediates, as well as the firms that own the establishments. One difference between this and other firm studies is that I study all the correlations of all establishments under a common parent, which is slightly different than a headquarter vs. affiliate analysis. Headquarters often house very different economic activities than their affiliates, such as mangerial and financial activities versus production in the establishments. Here, I test whether the production activities in establishments comove, rather than asking how managerial and production activities are correlated.

There are several advantages of using the ASM over other administrative data typically used in firm volatility studies. The main advantage is the long period of the sample required to efficiently estimate covariances between plants.[2] In addition, the ASM has detailed and consistent information on firm ownership, as well as product-level input and output by establishment to differentiate between possible vertical linkages and firm shocks, and also between industry shocks and product-level shocks. Note that industry level input-output measures and industry shocks may not correctly capture the relationships between establishments, due to the substantial diversity of plant input and output within industries.

This paper makes a significant contribution to the estimation of within-firm comovement by establishing it and decomposing it into its input components. The key to identifying within-firm comovement is applying a comprehensive econometric methodology to a long panel of detailed establishment level data. The strategy I employ is to first calculate correlations between the growth of total sales of each establishment pair in the data. To get a sense of the magnitude of the problem: tens of thousands of establishments mean hundreds of millions of establishment-pair correlations. Next, once I have a picture of the dependence within the economy, I decompose the firm component of shocks into capital and profit, labour and inputs.

Here, I delve deeper into the mechanisms and theory behind comovement. There is a

[2] A rule of thumb to judge significant correlations: using the Fisher transformation, the standard error is approximately $(T-3)^{-1/2}$. With only 9 or 17 periods (in some samples), the variability in the correlation estimates is substantial.

considerable amount of research around the relationships between parent firms, their subsidiaries and their operating establishments. Shocks may be transmitted through multiple mechanisms, including technology transfer, operating decisions, legal issues, intermediate contracts, vertical linkages, labour movement within the firm, other types of reallocation, or generic firm demand or productivity shocks. In addition, there are regional and industry shocks that may be attributed to firms, because firms are likely to own establishments with common characteristics.

As a first stab at the problem, one needs to remove industry and regional shocks at least and then ask whether establishments that belong to the same firm move together. To test for vertical linkages within firms, I include establishment level input-output measures. Furthermore, I decompose sales shocks into profit, labour input and intermediate input shocks to expose the sources of the within-firm correlation. The results show within-firm comovement is significantly higher than between-firm comovement, even after accounting for region and industry shocks, distance between establishments, vertical linkages, and common product-level inputs and outputs.

I find an establishment has a correlation 0.0477 higher with an establishment within the same firm relative to an establishment in a different firm. Approximately 49% of the within-firm effect is due to covarying intermediate input costs.

The rest of the paper is organized like so: Section 2.3 describes the framework I use to analyze the data and the data itself. Section 2.4 describes the econometric approaches to analyze the problems and the results, and Section 2.5 concludes. The Appendix follows in Section 2.6.

2.3 Empirical framework and data

Here, I outline the framework with which to examine establishment correlations. I'll start with an overview of the sales growth process for each establishment. There are two important ways to think of a sales growth rate: as a combination of industry, province, and idiosyncratic shocks, and also as a combination of the shocks of the components of sales, like payroll, profit and intermediate inputs. Once the different parts of the sales growth process are outlined, I can calculate comovement between establishments after accounting for different parts of the process.

To account for the possible industry and geographic sources of sales shocks, the establishment sales growth process, at a first approximation, follows

$$g_{it}^{s} = \theta_{it}^{s} + \mu_{pt}^{s} + \nu_{i}^{s} + u_{it}^{s} \tag{2.1}$$

Where $g_{it}^{s} = \Delta \log s_{it}$, the log-difference growth rate of sales of establishment i at time t, θ_{it}^{s} is an industry-time shock, μ_{pt}^{s} is a province-time shock, ν_{i}^{s} is a time-invariant establishment effect. $u_{it}^{s} \equiv \lambda_{ft}^{s} + e_{it}^{s}$ is a shock composed of two parts: a firm shock and idiosyncratic establishment shock. The questions: are g_{it}^{s}s are correlated across establishments within a firm? Are u_{it}^{s}?s

Of course, g_{it}^{s} may be correlated within firms because firms tend to own establishments within specific industries and provinces, and so are subject to common shocks. Removing those and examining u_{it}^{s}, are there common firm shocks that induce correlation within firms?

In addition, sales growth is composed of several components—how do these contribute to within firm correlation? I do so by recognizing that sales can be thought of as a combination of profit, payroll and intermediate inputs:

$$s_{it} = \pi_{it} + pay_{it} + input_{it}, \tag{2.2}$$

where s_{it} is sales, π_{it} is profit and capital services, pay_{it} is total payroll, $input_{it}$ is total value of intermediate inputs. The growth rate g_{it}^{s} can then be decomposed into the weighted sum of the growth of those components

$$g_{it}^{s} = w_{t-1}^{\pi} g_{it}^{\pi} + w_{t-1}^{pay} g_{it}^{pay} + w_{t-1}^{input} g_{it}^{input} \tag{2.3}$$

In turn, each one could be driven by industry, region, firm or other types of shocks. For instance, a strike in the auto industry could result in a negative labour shock across all establishments in that industry. Or a carbon tax in BC could affect the price of intermediates for all establishments in BC. In that case, we should also be interested in the residual growth of any variable $x \in V = \{\pi, pay, input\}$,

$$g_{it}^{x} = \theta_{it}^{x} + \mu_{pt}^{x} + \nu_{i}^{x} + \underbrace{\lambda_{ft}^{x} + e_{it}^{x}}_{u_{it}^{x}} \tag{2.4}$$

A natural question to ask is then, if u_{it}^{s} is correlated within firms, does it run through shocks to employment or something else? For instance, Giroud and Mueller (2016) shows firms respond to local shocks by reallocating labour across establishments within the firm (here, that would show up as a negative correlation between payrolls within the firm).

$$\mathrm{cov}(g_{it}^{s}, g_{jt}^{s}) = \sum_{x \in V} \sum_{y \in V} \mathrm{cov}(w_{it-1}^{x} g_{it}^{x}, w_{jt-1}^{y} g_{jt}^{y}) \tag{2.5}$$

What is the relationship between each component and common firm ownership?

2.3.1 Data

The data come from the Annual Survey of Manufacturers. The ASM is a long annual panel of manufacturing establishments in Canada, a defacto census of manufacturing activity from 1961-2011 (although I focus on 1973-1999 to get a more consistent sample with respect to establishment and firm identifiers, as well as industry classifications). I have observations on sales, value added, labour (wages and employment), intermediate inputs, and most importantly, firm ownership information. Industries are classified according to 4-digit SIC (1980). In addition, I construct establishment-level input-output linkages using the commodity survey that is included in the ASM for large establishments.

Table 3.1 shows relevant descriptive statistics for the manufacturing sector. Like many firm or establishment datasets, the distribution of sales and value-added are skewed to the right, with a relatively small number of establishments making the majority of sales in the

Table 2.1: Summary statistics.

	All estab.		Multi-estab.		Growth, g^x_{it}	
	(1) Mean	(2) SD	(3) Mean	(4) SD	(5) Mean	(6) SD
Sales	37.3	101.4	96.3	324.6	0.075	0.192
Payroll	5.5	15.1	13.7	46.3	0.067	0.240
Profit & capital	9.7	26.4	28.6	85.8	0.071	0.742
Input cost	22.1	60.2	67.5	212.5	0.074	0.282
N. obs.	4885		1797		125008	

Notes: The ASM data is at an annual frequency. 'Multi-establishment' is the subsample of establishments that are part of multi-establishment firms for at least one year. There are 76,857 plants in the full sample, and 11,092 in the multi-establishment sample. Both samples are restricted to establishments that are alive for at least 5 years in order to calculate reasonably accurate covariances. The static statistics shown [Columns (1–4)] are given for 1990, and are in millions of Canadian dollars. The growth (g^x_{it}) measure is for the multi-establishment subsample only, and calculated over 1974-1999. Profit & capital is everything left over after removing payroll and input cost from total sales. This includes capital costs.

economy. Before we get started on details, the most important question is whether within-firm comovements matter—is there enough within-firm "mass" for these shocks to matter in the aggregate? Yes; the total value added of multi-establishment firms averages around 90% of total manufacturing value added per year. The most important statistics for this paper are the annual sales growth rates g^s_{it}, with a mean of 7.5% and standard deviation of 19.2%. As a first look at within-firm versus between-firm correlations in sales growth rates, consider Figure 2.1, the distribution of correlations of within-firms (solid blue line) is pushed significantly to the right of the distribution of the between-firm correlations (dashed black line). The mean within-firm correlation is 0.019, which is significantly higher than the mean between-firm correlation, 0.0037. Establishments are more correlated with other establishments owned by the same firm. Establishments in the same 4-digit industry are only twice as correlated than ones in different industries (0.0064 vs. 0.0037) and ones in the same province are only slight more correlated than ones in different provinces (0.0039 vs. 0.0037). Of course, firmly establishing the importance of within-firm correlations requires addressing the econometric issues raised earlier.

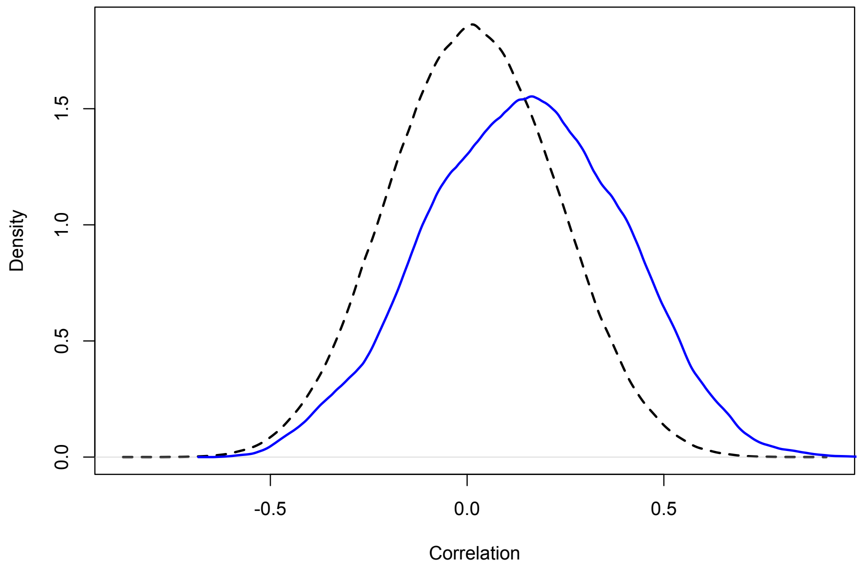


Figure 2.1: Density of establishment-establishment correlations, within-firm (solid blue line) and across firm (dotted black line).

2.4 Econometrics and Results

The economic goal is to understand within versus between firm comovement. The econometric strategy that achieves the economic goal requires the proper analysis of the growth processes in $\{\mathbf{g}_t^s\}_{t=0}^T = \{(g_{1t}^s, \ldots, g_{Nt}^s)\}_{t=0}^T$. I proceed in two steps. First, remove the effects of industries and regions and recover the residual growth estimates $u_{it}^s = \lambda_{ft}^s + \epsilon_{it}^s$. Next, calculate the matrix of correlations of the residual growth rates and compare within-firm establishment pairs to between-firm pairs. Then decompose the sales growth rates and see which elements drive the within-firm correlation.

I write the sample correlation of sales growth between two establishments as r_{ij} (dropping the superscript s unless otherwise needed), and use the Fisher transformation to change the sample correlation to a normally distributed variable,

$$z_{ij} = \frac{1}{2} \ln \left(\frac{1 + r_{ij}}{1 - r_{ij}} \right). \tag{2.6}$$

Note the standard error of each z_{ij} is $(T-3)^{-1/2}$ depends only on the number of periods T. In addition, since contemporaneous correlations are symmetric (and $r_{ii}=1$) means we only get $N(N-1)/2$ unique correlations. The typical approach to solving inference issues in these cases are to use multi-way clustering, which I apply at the establishment-pair level.

In order to perform the decomposition, I define

$$c_{ij}^{x,y} \equiv \mathrm{cov}(w_{it-1}^{x} u_{it}^{x}, w_{jt-1}^{y} u_{jt}^{y}), \tag{2.7}$$

where $w_{it-1}^{x} = x_{it-1}/s_{it-1}$. Note that $w_{it-1}^{s}=1$, and Equation (2.5) can be rewritten as $c_{ij}^{s,s} = \sum_x \sum_y c_{ij}^{x,y}$.

Next, consider the economic purpose of this paper. The main effect we want to understand is whether z_{ij} is higher or lower if i, j are both owned by a common firm f. To that end, I define firm$_{ij}$ as an indicator for common ownership. Similarly, industry$_{ij}$ and region$_{ij}$ are indicators of common industries and provinces. In addition, I use commodity input and output information at the establishment level to examine the effects of vertical integration and other IO measures on correlations. The relevant measures, outlined in Appendix 2.6, are output$_{ij}$, input$_{ij}$, and io$_{ij}$.

The first thing to do is estimate via OLS the equations

$$z_{ij}^{x} = \mathbf{X}_{ij}\beta + \beta_f^{x}\mathrm{firm}_{ij} + e_{ij}^{x} \tag{2.8}$$

$$c_{ij}^{x,y} = \mathbf{X}_{ij}\beta + \beta_f^{x,y}\mathrm{firm}_{ij} + e_{ij}^{x,y} \tag{2.9}$$

where $\mathbf{X}$ is a vector of controls that include common industry and region dummies. The common industry and region dummies in these regressions serve to check whether industry and region shocks were removed correctly after estimating Equation (2.1).

To perform the decomposition, recover $\beta_f^{x,y}$ for each component $x, y \in \{s, \pi, pay, input\}$ after estimating Equation (2.9). Then the total firm effect on sales comovement can be decomposed into the firm effects on each component,

$$\hat{\beta}_f^{s,s} = \sum_x \sum_y \hat{\beta}_f^{x,y} \tag{2.10}$$

Table 2.2: Firm correlation and covariance, u^s_{it}

	Balanced panel		Full sample	
	Cov.	Corr.	Cov.	Corr.
Firm	0.00130***	0.0839***	0.00142***	0.0477***
	(0.00021)	(0.0130)	(0.00010)	(0.0030)
Industry	4.92×10^{-5}	0.00758	-9.97×10^{-5}**	8.15×10^{-5}
	(9.93×10^{-5})	(0.00517)	(3.87×10^{-5})	(0.00114)
Province	8.16×10^{-5}**	0.00416**	2.30×10^{-5}**	0.00100***
	(3.47×10^{-5})	(0.00176)	(1.10×10^{-5})	(0.00031)
Constant	0.000148***	0.00828***	-2.19×10^{-7}	0.000636***
	(2.37×10^{-5})	(0.00126)	(7.85×10^{-6})	(0.000207)
Observations	245,520	245,520	56,029,880	53,517,864
R-squared	0.000	0.000	0.000	0.000

Notes: Standard errors calculated using multi-way clustering at the establishment level (i and j). *** and ** denote significance at the 1% and 5% level, respectively.

In other words, the average increase in covariance resulting from a pair of plants being owned by a common firm can be decomposed into the firm effect on each component. Dividing both sides by $\hat{\beta}^{s,s}_f$ shows the contribution of each component to the total firm effect on comovement.

2.4.1 Results

The results, shown in Table 2.2, show within-firm correlations are positive and significant. An establishment has on average a correlation coefficient that is 0.0477 higher with establishments owned by the same firm than with other establishments. In other words, after accounting for industry and region shocks, residual growth rates are correlated within firms. This evidence lends support for the firm comovement theory. The results are stronger for the balanced panel, suggesting longer-lived plants are more likely than short-lived plants to covary within firms.

Next, I decompose the covariance coefficients into the within-firm effects of sales (s), profit and capital (π), total payroll (pay), and total intermediate input costs ($input$). The majority of sales covariances come from input cost comovement, which accounts for 49% of the total effect. The covariances of profit and capital within firms matters slightly more than

Table 2.3: Firm covariance decomposition, u^s_{it}, u^π_{it}, u^{pay}_{it}, u^{input}_{it}

	$\hat{\beta}_f^{x,x}$			$\hat{\beta}_f^{x,x}/\hat{\beta}_f^{s,s}$		
	Corr.	Cov.	S.D.	Corr.	Cov.	S.D.
Sales	0.081	0.0013	0.036	1.000	1.000	1.000
Profit & capital	0.013	0.0001	0.009	0.160	0.070	0.264
Payroll	0.066	0.0000	0.007	0.810	0.034	0.185
Input cost	0.085	0.0006	0.025	1.047	0.490	0.700

Notes: The remainder of the decomposition comes from covariance of the components with input cost (especially profit & capital, less so payroll). Profit and capital is everything left over after removing payroll and input cost from total sales. This includes capital costs. The relative correlation term is only to get a sense of relative magnitudes, it does not actually decompose (its components do not need to add to 1).

payroll covariances, but both are much lower than input costs. In fact, the majority of the leftover within-firm effect comes from the covariance of input costs with payroll and profit (i.e., the off-diagonal terms in the decomposition like $\hat{\beta}_f^{input,\pi}$).

Think of the effect of each component on overall volatility in two parts: size and individual volatility. The effect on overall establishment volatility will be greater if the component is a large share of sales, and will be greater if the component is more volatile. In this view, the decomposition results are consistent with the typical view of establishment inputs: labour and capital are tough to adjust in response to shocks, so they have low volatility and contribute less to overall establishment comovement. However, intermediate inputs vary a lot, and have a large share of sales for most plants, so they make up a lot of the overall firm comovement effect. Although recent studies have suggesting firms may reallocate labour across plants within the firm after suffering local shocks (which would induce negative comovement of payroll within firms), this suggests that overall, labour positively comoves. This may be because common firm shocks dominate local labour market shocks, either in magnitude or frequency, or both. This is bolstered by the fact that leaving in firm and regional shocks gives much higher coefficients on the comovement effect (see Figure 2.1).

2.5 Conclusion

In this note, I've shown that growth rates of establishments within firms are significantly more positive correlated than establishments that are not owned by the same firm. This comovement survives, but is reduced, after accounting for the fact that firms are likely to own establishments in the same industries and same geographic regions. On another note, about 49% of the covariance in growth rates within firms is due to covarying input costs, and not labour or profit and capital movements. This is consistent with the fact that labour and capital are tough to adjust in response to shocks, while intermediate inputs are more volatile.

This result is significant for the measurement of establishment and firm growth, and the application of those measures to understanding how volatility is transferred across borders within firms. Although this study uses domestic plants, it is suggestive that similar mechanisms operate across international borders, and contributes to the understanding of how MNCs contribute to global economic fluctuations.

2.6 Appendix

2.6.1 Data

Data available from CDER in Statistics Canada. The data need some processing in order to be analyzed in this paper. First, there are two choices for the meaning of "firm", either the "parent" or the "ultimate parent". An establishment may be owned and directly controlled by firm A (the parent), which is in turn owned by a firm, which is owned by a firm, which is owned by a much larger firm B (the ultimate parent). I use the ultimate parent to be faithful to work on MNCs, although an argument could be made that direct parents have much more control and are more likely to transmit shocks among their establishments, both more often and more strongly. In that sense, the results may be viewed as an lower bound. On the other hand, there would be fewer establishment pairs that have common ownership, which would reduce the overall effect. In any case, the definition of firm in this case does not seem to affect the results. In addition, firm identifiers may change over time for reasons unrelated to business activities. The firm may undergo an organizational change that results in a new statistical identifier; suppose firm A owns three establishments in the sample, and then all three establishments have a new firm identifier B at some year t; it's likely the firm identifier has changed, or a new firm acquired all of the former direct parent's establishments—in any case, the effect is likely to attenuate the effect of firm ownership if anything, but the proportion of these establishments is very low.

Growth rates have always their own issues with outliers. To deal with this, I experiment with winsorizing or trimming the log growth rates at different percentiles by year, (0.05,0.95), (0.025,0.975), (0.01,0.99), and (0.001,0.999). The results do not change. However, the payroll, profit and intermediate numbers are less reliable than sales, and so calculating the growth rates for the decomposition can result in a smaller sample size. For this reason, I use the balanced panel sample for the decomposition. I find little difference if I trim the growth rates to a greater degree (0.1,0.9) and run the decomposition on the full sample.

The input-output measures are defined in a similar way to industry IO measures, but using establishment commodity data instead of industry commodity data. The relevant measures are output_{ij}, input_{ij}, and io_{ij}. Output (output_{ij}) is a measure of how similar two establishments commodity outputs are,

$$\text{output}_{ij} = \frac{\sum_c s_{ci} s_{cj}}{\sqrt{\sum_c s_{ci}^2}\sqrt{\sum_c s_{cj}^2}}, \tag{2.11}$$

where s_{ci} is sales of commodity c by establishment i. This measure is called "cosine similarity". The Input and IO measures are defined as

$$\text{input}_{ij} = \frac{\sum_c inp_{ci} inp_{cj}}{\sqrt{\sum_c inp_{ci}^2}\sqrt{\sum_c s_{inpj}^2}}, \tag{2.12}$$

$$\text{io}_{ij} = \frac{1}{2}\left(\frac{\sum_c inp_{ci} s_{cj}}{\sqrt{\sum_c inp_{ci}^2}\sqrt{\sum_c s_{cj}^2}} + \frac{\sum_c s_{ci} inp_{cj}}{\sqrt{\sum_c s_{ci}^2}\sqrt{\sum_c inp_{cj}^2}}\right). \tag{2.13}$$

$$\tag{2.14}$$

2.6.2 Robustness

To check the robustness of the conclusions to different specifications, I perform the following checks. In all checks, I continue the practice of using both a balanced panel and the much larger unbalanced panel. First, I include measures of vertical integration and input and output competition to see if the intra-firm correlation is due to similar input-output structures that are not accounted for by industry specific effects. The results are shown in Table 2.4. Although some of the input-output measures are significant, the intafirm correlation coefficient does not change much, suggesting the product-level measures of establishment relationships are not causing the observed intra-firm comovement coefficient. In addition, there is no consistent finding of vertical linkages within firms associated with comovement. However, in the one case it is significant (covariances in the full sample), it is of the same magnitude of the within-firm effect itself. The balanced and full sample panels cannot reject the null of no within-firm vertical linkage effect.

Next, I calculate use a different growth rate measure for each sample to see if the log growth rate approximation is affecting the true relationship. The Davis-Haltiwanger-Schuh (DHS, Davis and Haltiwanger, 1992) growth rates are defined,

$$g_{it}^s = \frac{1}{2}\left(\frac{s_{it} - s_{it-1}}{s_{it} + s_{it-1}}\right) \tag{2.15}$$

the results are similar for both measures, in both the overall relationship and the decomposition.

Table 2.4: Firm correlation and covariance, u_{it}^s, including extra covariates

	Balanced panel		Full sample	
	Cov.	Corr.	Cov.	Corr.
Firm	0.00135***	0.0797***	0.00134***	0.0456***
	(0.00023)	(0.0137)	(0.00010)	(0.0031)
Industry	0.000184	0.00765	−0.000199***	−0.00428***
	(0.000131)	(0.00638)	(3.96×10^{-5})	(0.00107)
Province	8.16×10^{-5}**	0.00417**	2.32×10^{-5}**	0.00101***
	(3.48×10^{-5})	(0.00177)	(1.10×10^{-5})	(0.00031)
Output	-3.24×10^{-5}	0.0132	0.000629***	0.0268***
	(0.000213)	(0.0127)	(0.000141)	(0.0046)
Input	−0.000432**	−0.0185*	0.000311***	0.0136***
	(0.000183)	(0.0097)	(9.65×10^{-5})	(0.0032)
IO	1.27×10^{-5}	0.00372	-2.75×10^{-5}	0.000247
	(6.95×10^{-5})	(0.00400)	(3.93×10^{-5})	(0.001310)
IO × Firm	3.78×10^{-5}	0.0396	0.00108**	0.0167*
	(0.000473)	(0.0382)	(0.00045)	(0.0101)
Constant	0.000149***	0.00829***	-6.51×10^{-7}	0.000613***
	(2.38×10^{-5})	(0.00127)	(7.87×10^{-6})	(0.000207)
Observations	245,520	245,520	56,029,880	53,517,864
R-squared	0.000	0.000	0.000	0.000

Notes: Standard errors calculated using multi-way clustering at the establishment level (i and j). *** and ** denotes significance at the 1% and 5% level, respectively.

Table 2.5: DHS growth rates; firm correlation and covariance, u_{it}^s

	Balanced panel		Full sample	
	Cov.	Corr.	Cov.	Corr.
Firm	7.91×10^{-5}***	0.0659***	9.91×10^{-5}***	0.0452***
	(1.13×10^{-5})	(0.0087)	(7.11×10^{-6})	(0.0027)
Industry	2.11×10^{-5}***	0.0161***	-1.39×10^{-6}	0.00149
	(7.87×10^{-6})	(0.0049)	(2.58×10^{-6})	(0.00107)
Province	8.98×10^{-6}***	0.00734***	1.54×10^{-6}**	0.00122***
	(2.11×10^{-6})	(0.00161)	(7.17×10^{-7})	(0.00031)
Constant	1.41×10^{-5}***	0.0111***	9.38×10^{-7}	0.000898***
	(1.58×10^{-6})	(0.0012)	(5.57×10^{-7})	(0.000215)
Observations	624,890	624,890	63,024,530	60,777,540
R-squared	0.000	0.001	0.000	0.000

Notes: Standard errors calculated using multi-way clustering at the establishment level (i and j). *** and ** denotes significance at the 1% and 5% level, respectively.

Table 2.6: DHS growth rates; firm correlation and covariance, u^s_{it}, including extra covariates

	Balanced panel		Full sample	
	Cov.	Corr.	Cov.	Corr.
Firm	7.78×10^{-5}***	0.0646***	9.42×10^{-5}***	0.0433***
	(1.22×10^{-5})	(0.0093)	(7.11×10^{-6})	(0.0028)
Industry	2.34×10^{-6}	0.00329	-7.25×10^{-6}***	−0.00192
	(6.54×10^{-6})	(0.00450)	(2.60×10^{-6})	(0.00103)
Province	8.98×10^{-6}***	0.00734***	1.55×10^{-6}**	0.00123***
	(2.11×10^{-6})	(0.00161)	(7.17×10^{-7})	(0.00031)
Output	2.97×10^{-5}	0.0201	3.22×10^{-5}***	0.0216***
	(1.65×10^{-5})	(0.0107)	(8.38×10^{-6})	(0.0041)
Input	1.94×10^{-5}	0.0126	2.52×10^{-5}***	0.0117***
	(1.13×10^{-5})	(0.0074)	(6.20×10^{-6})	(0.0028)
IO	5.00×10^{-6}	0.00598**	-1.71×10^{-6}	0.000243
	(3.75×10^{-6})	(0.00289)	(2.60×10^{-6})	(0.001241)
IO × Firm	-2.96×10^{-5}	−0.0161	7.90×10^{-5}**	0.0183
	(2.78×10^{-5})	(0.0223)	(3.19×10^{-5})	(0.0099)
Constant	1.39×10^{-5}***	0.0110***	9.07×10^{-7}	0.000879***
	(1.58×10^{-6})	(0.0012)	(5.58×10^{-7})	(0.000216)
Observations	624,890	624,890	63,024,530	60,777,540
R-squared	0.001	0.001	0.000	0.000

Notes: Standard errors calculated using multi-way clustering at the establishment level (i and j). *** and ** denotes significance at the 1% and 5% level, respectively.

Chapter 3

Looking for power-laws in all the wrong places: estimating firm size distribution tails across countries and datasets

3.1 Abstract

Power-law distributions feature heavily in research that relies on heterogeneous firms. This paper applies recent developments in power law estimation ("Power-Law distributions in empirical data", Clauset, A., C. Shalizi, and M. Newman (2009), SIAM Review 51(4), 661–703) to reject or fail to reject the null hypothesis that the firm size distribution is best fit with a power law. I use data from Compustat and OSIRIS on several countries, and confidential microdata on Canadian establishments and firms. I fit power-law, log-normal and power-law with exponential cutoff distributions to each dataset, and test the following two hypotheses: does the data reject the power-law fit? If not, does the data reject a power law fit in favour of an alternative distribution (specifically, one with thinner tails). I find that a power law distribution fits the U.S. firm size distribution for most years, but France and Germany confidently reject the null hypothesis that their firm size distributions are best fit with power laws. Canadian firms, both public and private, reject the power-law null hypothesis, but Canadian establishments do not. As an application, I use the estimated power laws to calculate firm size herfindahls to estimate the implied contribution of idiosyncratic shocks to aggregate volatility in different countries. That the power law distributions produce herfindahl results that run strongly counter to the data is a consequence of a seemingly well-fit distribution failing exactly where it matters in economics—in the top 10 or 20 firms.

3.2 Introduction

We are always looking for simple laws to explain economic behaviour. The power law is a perfect example; the upper tail of distributions of several economic phenomena seem to obey a scale-free law relating the size and rank of individuals (Gabaix, 2009). The firm size distribution is an especially important application of this law; a scale-free firm size distribution has dramatic consequences for several fields of economics, including studies on gains from trade (Di Giovanni et al., 2011; Di Giovanni and Levchenko, 2012, 2013; Nigai, 2017), as well as idiosyncratic volatility (Gabaix, 2011). However, the empirical evidence

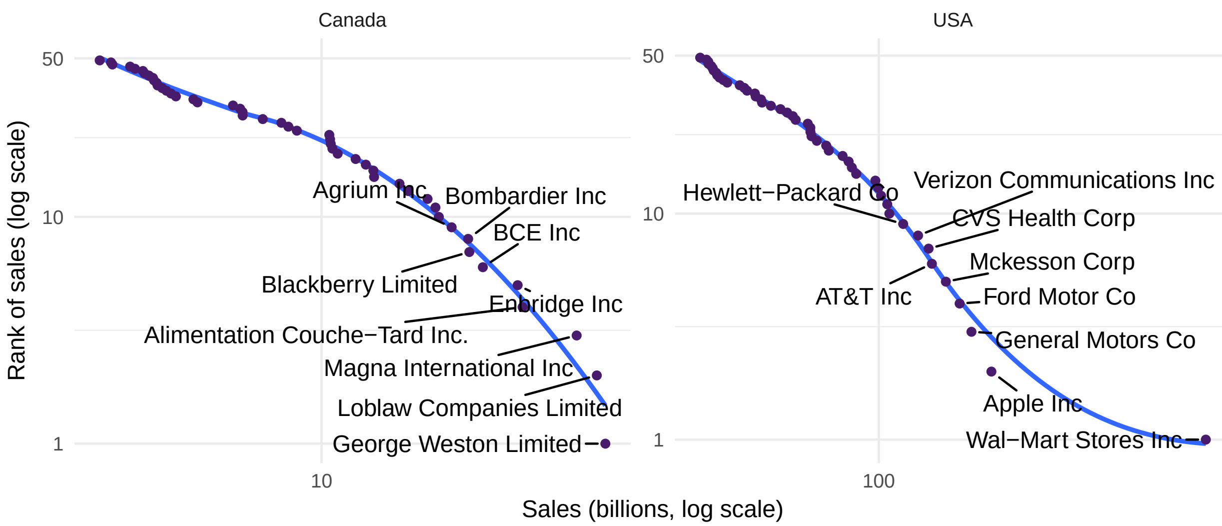


Figure 3.1: Difference in curvature on rank-size plot for the US and Canada.

Notes: Canadian firm data from OSIRIS, US firm data from Compustat. Petroleum and Financial firms have been removed. Note also the trouble with public company sales data: George Weston Ltd. has a controlling interest in Loblaw Companies Ltd. and reports Loblaw Co. Ltd.'s sales as its own (see, e.g., page 9 of the George Weston Ltd. Annual report (`http://www.weston.ca/en/pdf_en/gwl_2016ar_en.pdf`), which means Loblaw's $40B sales are reported twice. This double counting skews the relationship a bit; to correct this, I also look at confidential survey and administrative data that does not suffer from this issue.

supporting power law distributions is weak. In many applications, the methodology is: (1) eyeball the upper tail of the distribution on a log-log plot of rank vs. size and guess where it starts to look linear; (2) estimate via OLS the slope of the line (on the data above the eyeballed cutoff); (3) claim a high R^2 means a power law fits the data well. The estimated slope is used as the scale exponent for the power-law. For instance, in Figure 3.1, the shapes of the distributions are clearly very different, but an OLS fit to each plot will give a reasonable scale exponent and a high R^2. One should not conclude that both of these distributions are best fit with a power-law.

Despite the obvious drawbacks noted (even in the papers that use it), this methodology is used to justify many empirical power laws. The OLS methodology is so easy to use, and power laws are so simple and enticing to use in theory, that we do not apply the same statistical rigour that instrumental variables, for instance, would attract. In this paper, I apply recent developments in power law estimation (Clauset et al., 2009; Broido and Clauset, 2018) to reject or fail to reject the null hypothesis that the firm size distribution is best fit with a power law.

I use several sources of firm microdata: Compustat, OSIRIS and confidential microdata sources on Canadian establishments and firms. In each dataset, and each year, I estimate the upper tail cutoff x_{min}, the power law scale exponent α, as well as alternative distributions with thinner tails (the log-normal, and the power law with exponential cutoff). I calculate p-values for each important hypothesis: does the data reject the power law fit? If not, does the data reject a power law fit in favour of an alternative distribution (specifically, one with thinner tails). The conclusions matter for our understanding of firm heterogeneity—how skewed are our firm distributions?

This work contributes to two main strands of literature. First, the estimation of fat tailed distributions. Several methods have been proposed and refined, although all revolve around a rule-of-thumb for finding the cutoff of the upper tail. For distributions that are not truly power laws, the estimated scale exponent changes non-trivially with the cutoff, which means a researcher can easily draw the conclusions they'd like by varying the cutoff, and justify their choice with the resulting high R^2 given by the OLS estimate. Here, I adopt the much stricter methodology proposed by Clauset et al. (2009); Broido and Clauset (2018) that are used to estimate power laws in scale-free networks. A few estimation studies have changed track from estimating the upper tail to estimating the whole distribution. However, the conclusions of these papers still rely on a power law fitting the best on the upper tail, and then arguing for or against other distributions in the middle and lower tail of the distributions. If the upper tail is not truly a power law, these methods also have room to improve.

Secondly, this work is important for the application of power laws to economics. For an overall review of power laws in economics, see Gabaix (2009). Some of the proposed power laws are more robust than others. The city size distribution, for instance, seems robust to different specifications, as long as one adopts a sprawling definition of city: e.g., the US Census Bureau created 'combined statistical areas' (CSAs) to better represent the size of cities that sprawl over several municipal areas, and power law estimations of this type of city size distributions are more robust to estimation methods (see, e.g., the evolving research on Zipf's law for city sizes, including Gabaix (1999); Eeckhout (2004); Rozenfeld et al. (2011)).

However, research on power laws for firm sizes has not received the same attention. Theory suggests that trade increases the skew of the power law tail as size increases, because the most productive firms get access to more and more markets, which further increases their sizes. The data do not seem to support this. Since gains from trade in some models depends on the shape of the firm size distribution (Nigai, 2017; Head et al., 2014; Feenstra, 2018), correctly estimating the upper tail of the distribution is very important. An even stronger motivation comes from research on the microfoundations of aggregate volatility. The argument is that if the firm size distribution is skewed enough, the biggest firms are so big that idiosyncratic shocks to them are not washed out by random shocks to other firms in the economy (Gabaix, 2011; Acemoglu et al., 2012; Di Giovanni and Levchenko, 2012). Without a power law tail, the argument for idiosyncratic shocks causing aggregate volatility falls apart. Here, the estimation of the power law passes from important to strictly necessary.

However, even if the data reject a power law for a specific application, it does not mean that the power law is not a useful tool; for instance, it considerably reduces complexity in firm heterogeneity models like Melitz (2003), or models of preferential attachment in social networks (Jackson, 2010). However, the existence and importance and results of models of firm heterogeneity, for example, do not depend on the specific shape of the upper tail of the distribution; the power law just makes the algebra easier.

I find results that are in line with expectations (that a power law distribution fits the U.S. firm size distribution for many years), and some new results. Depending on the specification (e.g., including or excluding petroleum companies), France and Germany confidently reject the null hypothesis that their firm size distributions are best fit with power laws. Canadian datasets for firms reject power law distributions across the board, but when the unit of observation is an establishment, Canadian size distributions do not reject the power-law hypothesis.

As an application, I use the estimated power laws to estimate the implied contribution of idiosyncratic shocks to aggregate volatility in different countries. The crucial element is the herfindahl of the size distribution; I use the power law to calculate the implied herfindahl and the resulting aggregate volatility, and compare it to the empirical herfindahl and the herfindahl of the alternative tail distributions. As the tail of a dataset deviates farther from a power law, the power law herfindahl deviates much farther from the empirical herfindahl, vastly overstating the contribution to idiosyncratic shocks to aggregate volatility. That the power law distributions produce results that run strongly counter to the data is a consequence of a seemingly well-fit distribution failing exactly where it matters in economics—in the top 10 or 20 firms.

The paper proceeds as follows: Section 3.3 describes the methodology, Section 3.4 outlines the datasets used, and Section 3.5 gives results. Section 3.6 gives an application of how the estimates can affect economic phenomena, and Section 3.7 concludes. More estimation and dataset details can be found in Appendices 3.8 and 3.9.

3.3 Methodology

The methodology consists of two parts: (1) for each dataset, estimate each alternative distribution; (2) perform statistical tests to differentiate between possible hypotheses regarding the existence of power-laws. There are three distributions I consider: the power-law, log-normal, and power-law with exponential cutoff. The log-normal is a commonly proposed alternative to the power-law, and the power-law with exponential cutoff is an alternative proposed for degree sequences in social networks. Both have thinner tails compared to power-laws.

3.3.1 Upper-tail distribution estimation

Given a dataset of firm sizes $X = \{x_i : i \in 1, \ldots, N\}$, we need to estimate where the upper-tail $x_{\min}$ begins, and the shape of the distribution above it. We start with estimating the cutoff and scale exponent of the power-law, and then apply that same cutoff to the other distributions to ensure a fair and accurate comparison.

Power-law

A power-law distribution in the upper-tail follows the following density function

$$f(x) = Cx^{-\alpha}, \alpha > 1, x \geq x_{\min} > 0, \tag{3.1}$$

where α is the scale exponent and C is a constant, and $x_{\min}$ is the value that defines the upper tail. For details on the sources of power-laws in the world of economics, see Gabaix (2009). For our purposes, a power-law means a linear relationship between the size of an individual and the empirical counter-cdf (one minus the empirical cdf) on a log-log plot, *everywhere* in the upper tail. If the top 5 or 10 firms in the data diverge from the linear relationship, then the firm size distribution is not truly a power-law, in the sense that the important implications of power-law distributions do not hold (e.g., the granular hypothesis of aggregate volatility no longer applies).

Given $x_{\min}$, I use the MLE $\hat{\alpha}(x_{\min})$ as the estimate of the scale exponent. Then the estimate $\hat{x}_{\min}$ is the value of $x_{\min}$ that minimizes the Kolmogorov-Smirnov statistic D, the maximum distance between the cdf of the power-law fit and the ecdf, $E(x)$.

$$D = \max_{x \geq x_{\min}} |E(x) - F(x|\hat{\alpha})| \tag{3.2}$$

We use $\hat{x}_{\min} = \min_{x_{\min}} D$ as the cutoff for all distributions, and $\hat{\alpha} = \hat{\alpha}(\hat{x}_{\min})$ as the estimated power-law scale exponent. For more details, see Appendix 3.9.

Log-normal

The log-normal distribution is a common alternative to the power-law due to its skewness and association with Gibrat's law. The log-normal density is defined as:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{(\log x - \mu)^2}{2\sigma^2}}, \quad x > 0 \tag{3.3}$$

To compare it directly to the power-law distribution, I truncate the distribution at $x_{\min}$. Write the truncated distribution as:

$$h(x) = \frac{f(x)}{1 - F(x_{\min})} \tag{3.4}$$

so that the log-normal distribution is only defined in the upper tail and sums to 1 on the interval $[x_{\min}, \infty)$. MLE estimation of the parameters, after using $\hat{x}_{\min}$ from the power-law estimation to compare it fairly to the estimated power-law distribution.

Power-law with exponential cutoff

The power-law with exponential cutoff has a power-law-like tail up to a point, then has a exponential-like tail after that. The power-law is a special case of the power-law with exponential cutoff, which means it, by definition, cannot fit worse than a power-law. Nevertheless, except in that special case, it is not scale-free, and thus does not display the same economic properties as a true power-law. Its density is

$$f(x) = Cx^{-\alpha}e^{-\lambda x}, \tag{3.5}$$

where the constant $C = [e^{-x_{\min}\lambda}\Phi(e^{-\lambda}, \alpha, x_{\min})]$, and $\Phi(z, s, a) = \sum_{i=0}^{\infty} \frac{z^i}{a+i}^s$ is the Lerch Phi function. Again, it takes $\hat{x}_{\min}$ as given from the power-law KS statistic minimization.

When a distribution looks like a power-law everywhere but curves down at the end, one usually sees the justification "the distribution is a power-law but for finite-size effects" or "the power-law may hold only over a bounded range", meaning the very tip of the distribution *does not* extend to the very largest firms. These distributions are more likely to be power-law with exponential cutoffs than true power-laws, and since we care *most* about the very biggest firms, we would like to be able to statistically differentiate between the two types of distributions.

3.3.2 Tests

I start with the null hypothesis that the upper tail of the firm size distribution is a power-law, consistent with the literature. If the tests reject that hypothesis for a certain dataset, then that informs our understanding of the firm size distributions across countries or dataset types. There are two relevant tests: first, does the data directly reject the power-law? Second, does the data reject the power-law in favour of an alternative distribution?

Null hypothesis H_0^d (direct): the upper tail is a power-law

Clauset et al. (2009) proposes a semi-parametric bootstrap test to generate p-values for the power-law null hypothesis H_0^d. The idea is to simulate the data as if it were really a power-law, run the estimation procedure again. Simulating this 1000 times gives a null distribution of KS-statistics $Pr(D)$. If D^* is the KS-statistic for the best fitting power-law distribution, then the p-value for this model is defined as the probability of observing, under the null distribution, a KS-statistic at least as extreme as D^*. So $p = Pr(D \geq D^*)$ is the fraction of simulated datasets with KS statistics larger than that of the empirical dataset.

Null hypothesis H_0^a (alternative): the upper tail is fit equally well by a power-law and the alternative distribution

A set of likelihood-ratio tests can distinguish between the power-law and alternative distributions. Given the log-likelihoods of the power-law ($\mathcal{L}_{pl}$) and an alternative distribution ($\mathcal{L}_{alt}$), the test likelihood-ratio test statistic is

$$\mathcal{R} = \mathcal{L}_{pl} - \mathcal{L}_{alt}, \tag{3.6}$$

where the sign of R, if deemed significantly different from 0, gives evidence for or against the null hypothesis H_0^a, that the data are fit equally well by a power-law and the alternative distribution.

3.4 Data

I use data on firm sizes from several sources. First, Bureau van Dijk's OSIRIS database contains sales and identifying information on globally listed public companies, including 34,000 listed, 3,500 unlisted and 7800 delisted companies between 1900–2100. Second, the Fundamentals Annual section of the Compustat North America database. Both datasets come from Wharton Research Data Services (WRDS, 2018). I also investigate datasets with public and private firms in Canada. These are the Annual Survey of Manufactures, and the T2-LEAP administrative tax and employment dataset. For each, I use gross sales as a measure of size to be consistent with the other datasets.

Table 3.1 shows the summary statistics for each country. The US Compustat data is labelled "USA (Compustat)", while every other country is OSIRIS. OSIRIS data typically covers the period 1984–2016, with some exceptions (China's data starts in 1992 and Taiwan in 1995), while the Compustat data starts in 1961. Each country dataset is normalized by the median firm within each year, so the median for each dataset is 1. The data are clearly skewed right, with means up to 20 times higher than medians for some countries, and even 4 times higher than the 75th percentiles of the distributions.

Table 3.1: Summary statistics of firm sizes by country and dataset

Country	Period	N obs.	Mean	3rd Qu.	Max
Australia	2000–2016	510	35.5	19.9	1541.0
Bermuda	2001–2016	313	8.0	7.3	130.9
Canada	1996–2016	610	42.3	26.4	1043.9
Canada (ASM)	1973–1999	32622	9.0	4.0	11728.8
Canada (ASM)	2000–2011	52937	10.0	3.5	16518.5
Canada (ASM, firms)	1973–1999	28057	14.8	3.5	24122.6
Canada (ASM, firms)	2000–2011	47566	12.8	3.1	26131.9
Canada (T2, firms)	2001–2009	1412787	10.8	3.3	14035.5
Cayman Islands	2006–2016	452	7.6	7.5	152.4
China	2000–2016	1351	13.8	7.1	2861.8
France	1998–2016	363	58.4	25.4	2699.7
Germany	1998–2016	340	49.7	21.1	1674.5
Great Britain	1987–2016	685	44.7	18.2	6270.1
India	2000–2016	1288	29.9	14.7	3876.3
Japan	1996–2016	1159	10.6	7.0	567.2
Korea	2008–2016	288	19.5	5.6	2899.7
Malaysia	1996–2016	445	9.9	6.1	1047.3
Taiwan	2007–2016	297	16.0	7.7	762.3
USA	1984–2017	2544	26.3	18.2	3096.1
USA (Compustat)	1961–2014	2974	20.2	13.8	1675.8

Notes: the statistics are averages of all available years. E.g., 'N obs.' is the average number of observations per year. Each dataset is normalized within years by the median firm. Sources of datasets other than OSIRIS are indicated in parentheses. Non-OSIRIS Canadian datasets are confidential microdata. ASM is the Annual Survey of Manufactures (all manufacturing establishments with more than $30,000 in sales). The samples are divided into 1973–1999 and 2000–2011 because of a survey changes. If labelled 'firms', the establishment data are aggregated up to the ultimate parent (firm) level, to better compare with the public firm data sources. T2 is data from all firms in Canada derived from administrative tax records. The T2 sample is much larger and not restricted to manufacturing.

Later, I test robustness of the results to removing financial and petroleum firms from the datasets.

3.5 Results

In this section, I present the results of the power-law estimations, the alternative distribution estimations, and the tests that distinguish between them.

3.5.1 Power-law estimations

For a power-law to exhibit scale-free behaviour, the estimated scale exponent must satisfy $2 \leq \hat{\alpha} < 3$. A scale parameter less than 2 is not consistent with a stable distribution, but variations in the data can result in estimating a scale exponent less than 2. Therefore, for each country, one must consider all possible estimates across years to analyze the behaviour of a single country; Broido and Clauset (2018) use a similar approach to analyze separate components of a single network dataset.

Figure 3.4 plot the densities of scale exponents for each country. In this figure, I only plot estimated exponents that later are not rejected by the statistical tests. The Zipf law's exponent of 2 is denoted with a vertical red line in each plot. In the left panel, all industries are included in the dataset, and in the right panel, finance, insurance, real estate and petroleum-related industries are removed. The results are striking—across countries, for all

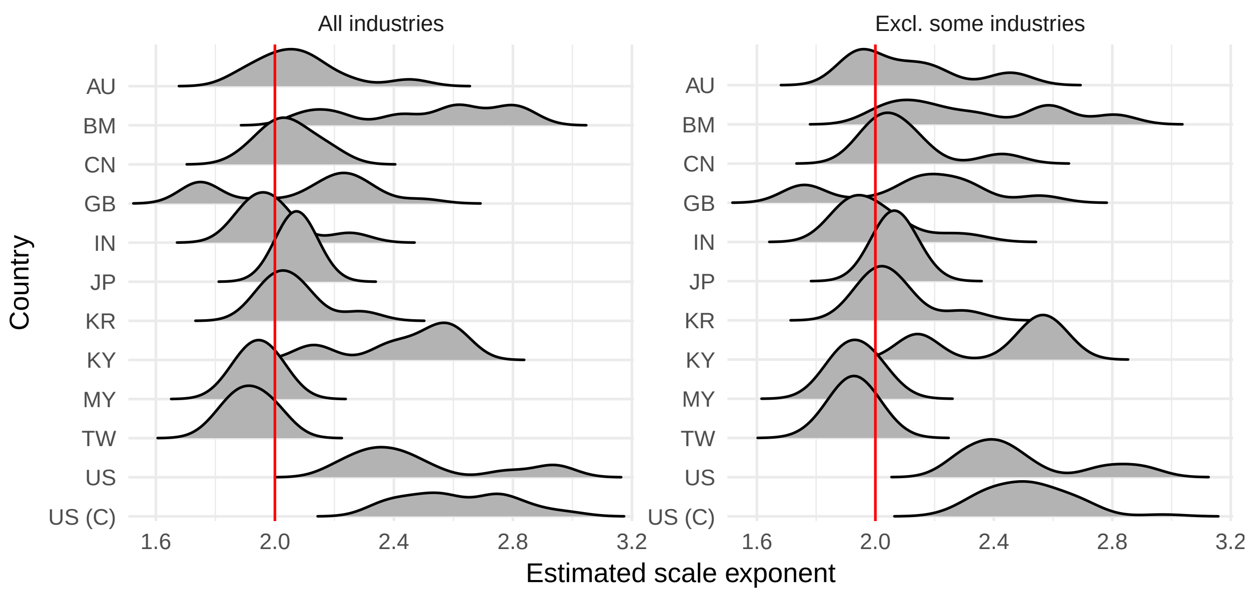


Figure 3.2: Exponent distributions

Notes: the left panel includes all non-bank firms in all industries. The right panel drops all firms in Finance, Insurance, Real estate and petroleum (NAICS 211, 52 and 53). The 'Zipf' power law exponent of 2 is denoted with a red vertical line in each plot. Estimated scale exponents are only shown for distributions that are not rejected in favour of an alternative distribution.

distributions that are not rejected in favour of alternative distributions, the mean estimated

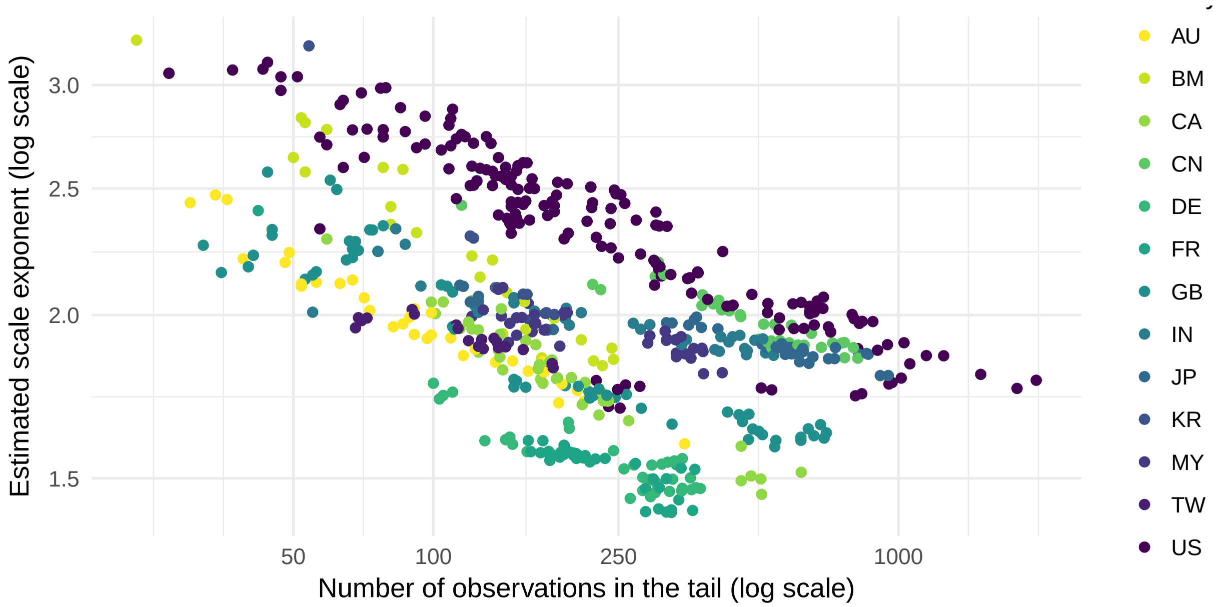


Figure 3.3: Exponents vs. number of observations

Notes: estimated scale exponents are shown for all datasets, whether or not they are rejected in favour of alternative distributions. There is a clear negative relationship between the number of observations in the tail (which depends on the estimated $\hat{x}_{\min}$ and the estimated scale exponent $\hat{\alpha}$.

scale exponent is around 2, with the U.S. coming in slightly higher at a little less than 2.5. For power-law estimations that are not rejected by the data, the results bolster theories that require scale-free firm size distributions.

However, there seems to be significant variation in the estimated exponents, and a closer look suggests there is a relationship between the estimated exponent and the number of observations in the tail of the distribution. In Figure 3.3, I plot the estimated scale exponents vs. the number of observations in the tail of the distribution (which depend on the estimated $\hat{x}_{\min}$). For smaller countries (or a strict $x_{\min}$, the scale exponent tends to be a bit larger, while larger countries (and countries with a smaller, less restrictive $x_{\min}$) tend to be closer to satisfying Zipf's law. This contradicts some ideas about the effect of trade on the firm size distribution—theory predicts that smaller countries would have as large or larger power-law exponents for their firm size distributions, because the most productive firms get access to markets much larger than other domestic firms, which skews the firm size distribution more

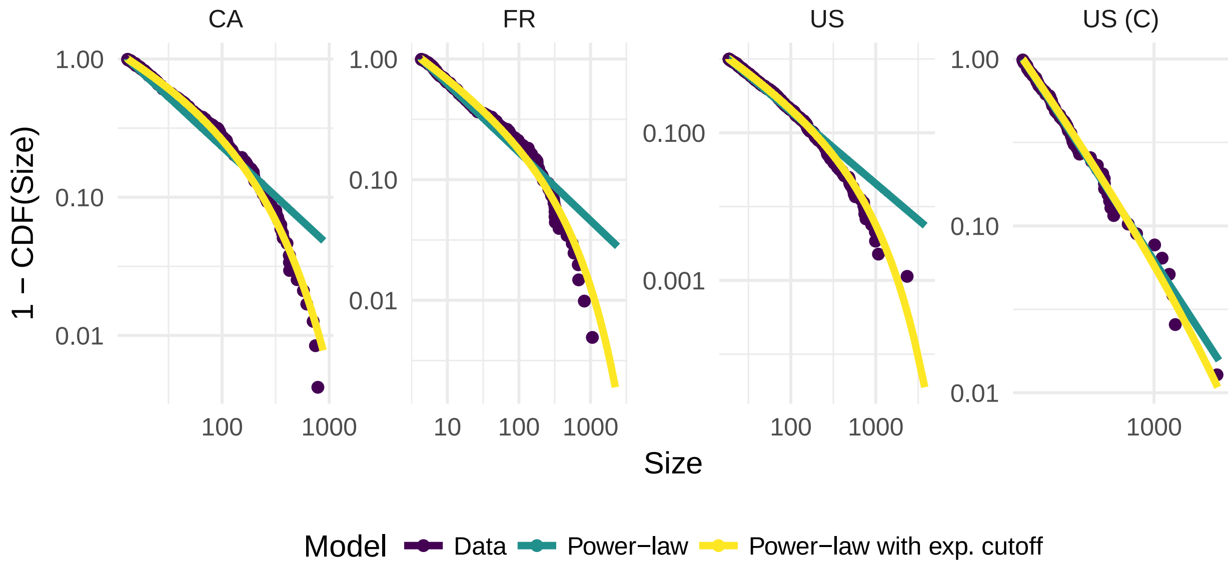


Figure 3.4: Estimated and empirical distributions

Notes: these figures for the year 2005. CA is Canada, FR is France, US is USA, US (C) is USA data from Compustat. The data are plotted as points, and each estimated distribution is plotted as a line.

than it otherwise would (Di Giovanni et al., 2011; Di Giovanni and Levchenko, 2012, 2013). Conditional on not rejecting the power-law fit, coutnries that satisfy Zipf's law tend to be larger than others.

3.5.2 What do the estimated distributions look like?

To get a feel for the shape of each possible distribution, I present visual representations of the differences between the alternatives (including the empirical cdf). For a few examples, I plot the estimated power-law, log-normal and power-law with exponential cutoff distributions, along with the empirical cdf. Firm size (log-scale) is on the x-axis, and the counter cdf is on the y-axis ($1 - F(\text{Size})$). The power-law with exponential cutoff typically fits the data very well. For the US, however, the dataset with more observations (OSIRIS, with 881 observations in the tail) is a power-law with exponential cutoff, while the other dataset (Compustat, with 78 observations in the tail) looks like a power-law. If one restricted the OSIRIS US dataset to a similarly high $x_{\min}$, one might succeed in fitting a power-law equally well as the power-law with exponential cutoff (or log-normal). However, the x_{min} was chosen to fit a power-law as well as possible; the only problem is there's a lot more data in the lower part of the upper-tail (i.e., closer to x_{min} than to the maximum size x), and *that* part of the distribution acts more like a power-law (a linear relationship in Figure 3.4) than the upper part of the upper tail. Keep in mind that $\hat{x}_{\text{min}}$ is chosen to match the empirical cdf $E(x)$, *not* the relationship between $E(x)$ and x.

To assess the fits statistically (instead of visually), I now move to the hypothesis tests.

3.5.3 Direct and alternative hypothesis tests

The results of the tests, shown in Table 3.2, seem to strongly separate countries into those with power-law firm size distributions and those that reject power-law firm size distributions. Looking at the direct tests first, the null hypothesis H_0^d is rejected by the median p-value in Canada, Germany, France, Japan, and the USA (OSIRIS dataset). Canadian microdata (as opposed to data on public firms only) give a slightly different answer: when the unit of observation is a firm, the results are consistent with using data on public firms only. However, when the unit of observation is an establishment, the size distributions do not reject the power-law directly or in favour of any alternative distribution I tested. For a more detailed look at Canadian microdata results, see Figure 3.5.

All other countries do not reject the null hypothesis that the upper tail of the data is generated by a power-law. However, this does not answer whether or not there's an alternative distribution that fits the data *better*. For that, we turn to the p-values for $H_0^{\text{log-n}}$ and H_0^{pexp}. The alternative distribution tests show similar results: the same countries reject power-laws for alternative distributions. A few more countries come closer to rejecting power-law distributions for alternatives, but p-values less than 0.1 (e.g., Australia, Bermuda, Great Britain, Cayman Islands, and the USA (Compustat)). The other countries still do not reject the null hypothesis that the power-law fits the data as well as the alternative distributions. To make an overall conclusion, I take into account the proportion of years a country rejects

Table 3.2: Summary of p-values for different tests, by country and dataset.

	Median p-value for:			% rejecting:		
Country	H_a^D	$H_a^{\text{log-n}}$	H_a^{pexp}	H_a^D	$H_a^{\text{log-n}}$	H_a^{pexp}
Australia	0.28	0.16	0.08	33	13	40
Bermuda	0.28	0.17	0.06	14	7	43
Canada	0.00	0.01	0.00	100	84	100
Canada (ASM 73–99)		0.46	0.19		0	26
Canada (ASM 00–11)		0.86	0.40		0	0
Canada (ASM 73–99, firms)		0.18	0.00		0	100
Canada (ASM 00–11, firms)		0.19	0.00		0	100
Canada (T2, firms)		0.00	0.00		100	100
China	0.14	0.20	0.19	33	0	40
Germany	0.00	0.02	0.00	94	94	100
France	0.01	0.01	0.00	100	94	100
Great Britain	0.20	0.16	0.07	46	39	43
India	0.54	0.38	0.29	20	0	13
Japan	0.03	0.02	0.00	74	63	84
Korea	0.34	0.70	1.00	14	0	0
Cayman Islands	0.13	0.18	0.07	44	33	44
Malaysia	0.18	0.18	0.14	21	5	21
Taiwan	0.51	0.41	0.46	0	0	0
USA	0.01	0.01	0.00	58	55	61
USA (Compustat)	0.38	0.37	0.42	29	29	39

Notes: H_a^D is the alternative hypothesis that the data are not well fit by a power-law; $H_a^{\text{log-n}}$ is the alternative hypothesis that the data are better fit with a log-normal; H_a^{pexp} is the alternative hypothesis that the data are better fit with a power-law with exponential cutoff. Sources of datasets other than OSIRIS are indicated in parentheses. Non-OSIRIS Canadian datasets are confidential microdata. ASM is the Annual Survey of Manufactures (all manufacturing establishments with more than \$30,000 in sales). The samples are divided into 1973–1999 and 2000–2011 because of a survey changes. If labelled 'firms', the establishment data are aggregated up to the ultimate parent (firm) level, to better compare with the public firm data sources. T2 is data from all firms in Canada derived from administrative tax records. The T2 sample is much larger and not restricted to manufacturing.

each null hypothesis. If 70% or more of the tests reject the null hypothesis, I classify the country as rejecting the power-law distribution. These countries are: Canada, Germany, France, and Japan. Most other countries reject the alternative distribution hypothesis less than 40% of the time, with less rejecting the direct null hypothesis H_0^D. The USA is the only one more difficult to classify. The Compustat dataset rejects the direct hypothesis only 26% of the time, while the OSIRIS dataset rejects it 58% of the time. The proportion rejecting in favour of the power-law with exponential cutoff distribution are closer at 55 and 61%; however, this is not conclusive evidence, especially with the existing literature supporting the power-law firm size distribution of the USA. One wonders whether a test incorporating the panel nature of the data could help differentiate the alternatives: for instance, can a power-law distribution suffer a few negative shocks at the very tip of the distribution to temporarily thin out the tail?

3.6 Application

The contribution of idiosyncratic shocks to aggregate volatility depend on the herfindahl of firm sizes in the economy (Gabaix, 2011; Di Giovanni and Levchenko, 2013; Di Giovanni et al., 2014),

$$h = \sqrt{\sum_{i}^{N} w_i^2}, \tag{3.7}$$

where w_i is the weight of firm i in the economy. The bigger the herfindahl, the bigger the contribution that idiosyncratic shocks make to aggregate volatility. The bigger the biggest firms are, the bigger the herfindahl, and finally, a power-law exponent between 2 and 3 is required for the biggest firms to be big in an economy with a lot of firms. I focus on the last point. If we claim the firm size distribution is best fit by a power-law, the estimated power-law distribution should imply a herfindahl close to the data. If not, the power-law distribution does not agree with the main statistic that governs the existence of the microfoundations of aggregate fluctuations.

Therefore, I use each estimated power-law and power-law with exponential cutoff distributions to simulate 1000 datasets for each country and year, and calculate the implied herfindahls, and take the mean across all the simulated datasets. The results are in Table 3.3. The empirical herfindahls are listed in the first column; the simulated herfindahls of the fitted power-law distributions are in the second column (labelled 'Power'), and the simulated herfindahls of the fitted power-law with exponential cutoff distributions are in the third column 'P-exp'. The p-exp distributions are much closer to the empirical herfindahls, with most being less than 10% different than the data, whereas the implied power-law herfind-

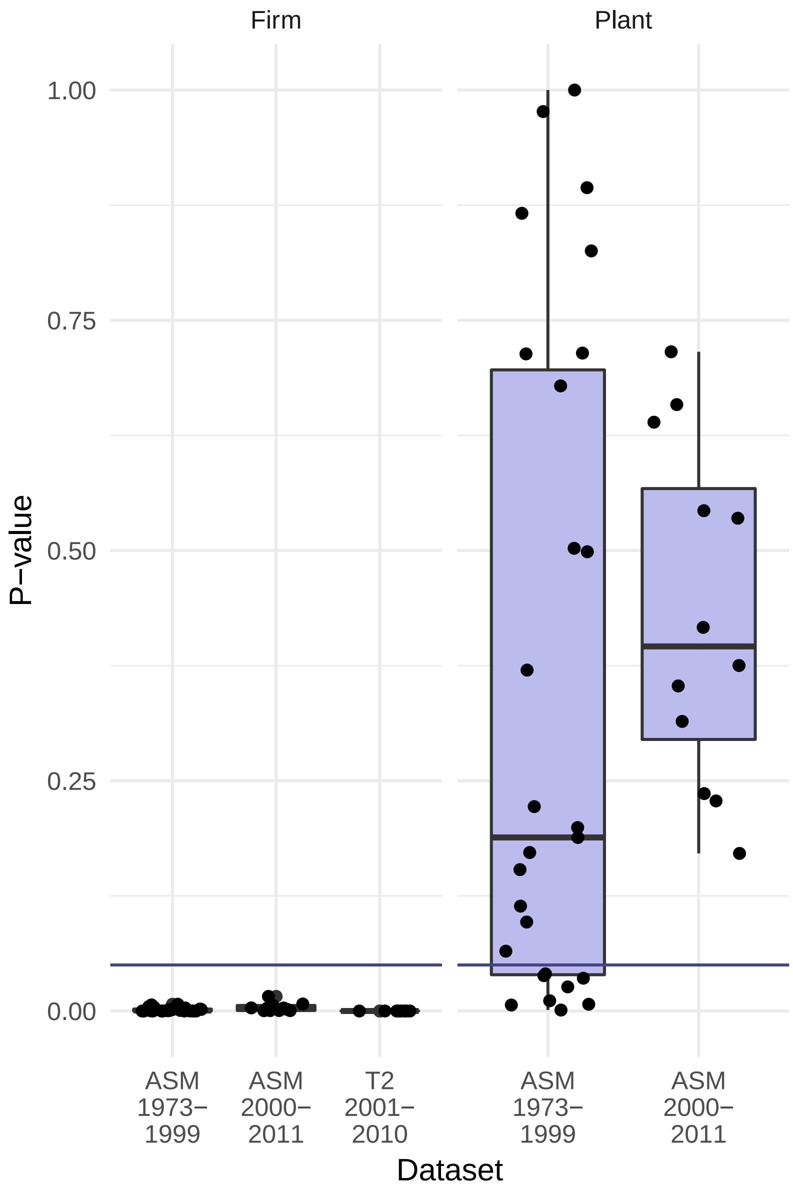


Figure 3.5: Boxplot of p-values of the test of null hypothesis H_0^{pexp} for each Canadian micro-dataset. Each dot represents a p-value for one year of a dataset.

Notes: These are boxplots of the p-values from the test of power-law versus the alternative hypothesis that the distribution is better fit by a power-law with exponential cutoff (H_a^{pexp}). The three datasets on the left use firms as the units of observation, and definitively reject the null hypothesis for all years. The two datasets on the right use establishments as the units of observation, and do not reject the null hypothesis that the data is fit well by a power-law.

Table 3.3: Herfindahls for each predicted distribution and their differences from the data, by country

Country	Herfindahls			% difference from data	
	Data	Power	P-exp	Power	P-exp
Australia	0.196	0.399	0.200	103.1	2.0
Bermuda	0.162	0.276	0.162	70.4	0.1
Canada	0.124	0.464	0.132	275.5	6.9
Canada (ASM 73–99)	0.100	0.156	0.092	55.9	-8.2
Canada (ASM 00–11)	0.087	0.167	0.100	92.2	15.8
Canada (ASM 73–99, firms)	0.102	0.381	0.094	274.8	-7.9
Canada (ASM 00–11, firms)	0.089	0.343	0.093	284.1	4.3
Canada (T2, firms)	0.019	0.176	0.023	820.6	20.1
Cayman Islands	0.138	0.288	0.137	107.9	−1.1
China	0.210	0.309	0.164	47.5	−21.7
France	0.176	0.643	0.182	265.4	3.6
Germany	0.181	0.635	0.199	250.4	9.7
Great Britain	0.271	0.434	0.210	60.3	−22.4
India	0.198	0.335	0.205	69.0	3.4
Japan	0.110	0.363	0.117	231.3	6.4
Korea	0.549	0.321	0.321	−41.5	−41.4
Malaysia	0.242	0.344	0.183	42.0	−24.3
Taiwan	0.274	0.391	0.285	42.7	3.7
USA	0.120	0.282	0.112	134.4	−7.3
USA (Compustat)	0.112	0.216	0.116	92.5	3.5

Notes: The first three columns are the herfindahls given by the distributions: data, the estimated power-law distribution, and the estimated power-law with exponential cutoff. The last two columns are the percentage differences between the estimated distribution's herfindahl and the data's herfindahl.

ahls can be almost 3 times higher than the herfindahls of the data. The Power and P-exp herfindahls are equally bad in some countries that did not reject the previous hypothesis tests, like Great Britain and China.

In Canadian microdatasets, the power-law with exponential cutoff distributions produce herfindahls much closer to the data than a power-law alone, even though the tests do not reject the fact that the data is represented by a power-law. For example, in Canadian establishment data from 2000–2011, of which all tests do not reject the hypothesis that a power-law generated the data, power-law herfindahls are 92.2% higher than the data on average, and the power-law with exponential cutoff generates herfindahls only 15.8% higher than the data. The results from Canadian microdatasets when the unit of observation is a firm are consistent with other public firm data, with power-law distributions generated herfindahls two to three times higher than the data on average, versus herfindahls of less than 10% in absolute value from power-law with exponential cutoff.

This highlights the importance of estimating distributions rigourously if one is to use the estimated distribution in crucial model calculations later. One should not rely solely on an OLS estimate of a power-law exponent as a justification for and basis of statistics that generate counterfactuals for economic effects and policies.

3.7 Conclusion

This paper applies recent developments in power law estimation (Clauset et al., 2009; Broido and Clauset, 2018) to reject or fail to reject the null hypothesis that the firm size distribution is best fit with a power law. I use several sources of firm microdata: Compustat, OSIRIS and confidential microdata sources on Canadian establishments and firms. In each dataset, and each year, I estimate the upper tail cutoff x_{min}, the power law scale exponent α, as well as alternative distributions with thinner tails (the log-normal, and the power law with exponential cutoff). I calculate p-values for each important hypothesis: does the data reject the power law fit? If not, does the data reject a power law fit in favour of an alternative distribution (specifically, one with thinner tails).

I find that a power law distribution fits the U.S. firm size distribution for most years, but France, Germany, and Canada confidently reject the null hypothesis that their firm size distributions are best fit with power laws. Results from confidential Canadian microdata support these results: when the unit of observation is a firm, Canadian size distributions consistently reject the hypothesis that the data are generated by a power-law distribution; on the other hand, when the unit of observation is an establishment, the Canadian size distributions cannot reject the power-law null hypothesis.

As an application, I use the estimated power laws to estimate the implied contribution of idiosyncratic shocks to aggregate volatility in different countries. The crucial element is the herfindahl of the size distribution; I use the power law to calculate the implied herfindahl and the resulting aggregate volatility, and compare it to the empirical herfindahl and the herfindahl of the alternative tail distributions. As the tail of a dataset deviates farther from a power law, the power law herfindahl deviates much farther from the empirical herfindahl, vastly overstating the contribution to idiosyncratic shocks to aggregate volatility. That the power law distributions produce results that run strongly counter to the data is a consequence of a seemingly well-fit distribution failing exactly where it matters in economics—in the top 10 or 20 firms.

3.8 Appendix–Robustness

To test robustness of the results to different specifications, I estimate and test all distributions again after removing petroleum and FIRE industries (fire, insurance and real estate). The results are very consistent with the results in the text, suggesting the shape of the firm size distribution does not depend on excluding certain types of firms.

3.8.1 The truncated Pareto distribution

The truncated Pareto distribution is a common alternative proposed to limit the upper tail of the size distribution.

$$f(x) = Cx^{-\alpha}, \alpha > 1, \infty \geq x_{\max} \geq x \geq x_{\min} > 0, \tag{3.8}$$

However, I did not find this distribution to fit all that well. Figure 3.6 reproduces Figure 3.4 with an estimated truncated Pareto distribution. The truncated Pareto distribution does poorly compared to the power-law with exponential cutoff, especially when the data are closer to a power-law (e.g., in the US).

Table 3.4: Summary of p-values for different tests, by country and dataset.

	Median p-value for:			% rejecting:		
Country	H_a^D	$H_a^{\text{log-n}}$	H_a^{pexp}	H_a^D	$H_a^{\text{log-n}}$	H_a^{pexp}
Australia	0.33	0.25	0.19	20	7	33
Bermuda	0.37	0.21	0.08	0	0	27
Canada	0.00	0.02	0.00	94	78	100
China	0.10	0.13	0.09	47	7	47
Germany	0.00	0.02	0.00	94	94	100
France	0.00	0.01	0.00	94	94	94
Great Britain	0.29	0.30	0.48	39	25	29
India	0.36	0.34	0.27	13	0	7
Japan	0.04	0.03	0.00	68	63	84
Korea	0.30	0.64	1.00	0	0	0
Cayman Islands	0.07	0.08	0.02	50	38	75
Malaysia	0.23	0.19	0.17	19	0	6
Taiwan	0.58	0.45	0.53	0	0	0
USA	0.01	0.02	0.00	65	55	68
USA (Compustat)	0.24	0.39	0.47	32	29	39

Notes: H_a^D is the alternative hypothesis that the data are not well fit by a power-law; $H_a^{\text{log-n}}$ is the alternative hypothesis that the data are better fit with a log-normal; H_a^{pexp} is the alternative hypothesis that the data are better fit with a power-law with exponential cutoff. Sources of datasets other than OSIRIS are indicated in parentheses.

3.9 Appendix—Theory

This appendix describes the empirical methodology in more detail. Much of this is also described in detail for a different application in Clauset et al. (2009) and Broido and Clauset (2018).

3.9.1 Distributions

Power-law distribution

A power-law of firm sizes above $x_{\min}$ follows this distribution:

$$f(x) = Cx^{-\alpha}, \alpha > 1, x \geq x_{\min} > 0, \tag{3.9}$$

where α is the scale exponent and C is a constant, and $x_{\min}$ is the value that defines the upper tail. On a log-log scale, this has the form

$$\log f(x) = \log C - \alpha \log x \tag{3.10}$$

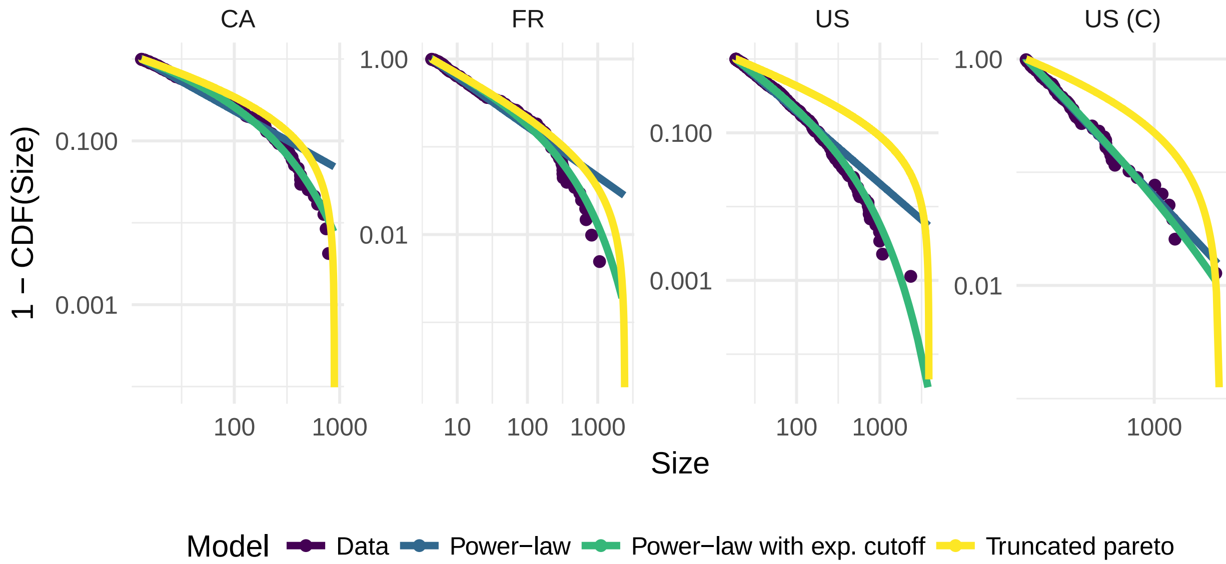


Figure 3.6: Estimated and empirical distributions, with truncated Pareto

Notes: these figures for the year 2005. CA is Canada, FR is France, US is USA, US (C) is USA data from Compustat. The data are plotted as points, and each estimated distribution is plotted as a line. The truncated Pareto distribution is estimated MLE as described in Aban et al. (2006).

which leads one to suggest OLS as an appropriate method to estimate α, after guessing an appropriate $x_{\min}$.

Log-normal distribution

The log-normal distribution is another alternative distribution that can have heavy tails that also happens to be consistent with Gibrat's law:

$$f(k) = \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{(\log x - \mu)^2}{2\sigma^2}}, x > 0 \tag{3.11}$$

Then write the distribution truncated below at $x_{\min}$ as

$$h(x) = \frac{f(x)}{1 - F(x_{\min})} \tag{3.12}$$

so that the log-normal distribution is only defined in the upper tail and sums to 1 on the interval $[x_{\min}, \infty)$. MLE estimation of the parameters, after using $x_{\min}$ estimated as if it were a power-law distribution, to compare it correctly.

Power-law with exponential cutoff

$$f(x) = [e^{-x_{\min}\lambda}\Phi(e^{-\lambda}, \alpha, x_{\min})]x^{-\alpha}e^{-\lambda x}, \tag{3.13}$$

where $\Phi(z, s, a) = \sum_{i=0}^{\infty} \frac{z^i}{a+i}^s$ is the Lerch Phi function.

3.9.2 Fitting the model, estimating $x_{\min}$ and α

Given $x_{\min}$, one can estimate the scale exponent α via maximum likelihood. A typical method to pick $x_{\min}$ is to plot rank-size on a log-log plot, and eyeball where the upper tail of the distribution "starts to look linear". In contrast, we use the Kolmogorov-Smirnov (KS) minimization method described in Clauset et al. (2009) and Broido and Clauset (2018).

The KS method selects the $x_{\min}$ that minimizes the maximum difference in absolute value between the empirical cumulative distribution (ecdf) $E(x)$ on the observed tail $x \geq x_{min}$ and the cdf of the best fitting power-law $F(x|\hat{\alpha})$ on those same observations. The $\hat{\alpha}$ is estimated via MLE as described above, given the $x_{\min}$ of the current step of the KS method. The KS statistic is defined as

$$D = \max_{x \geq x_{\min}} |E(x) - F(x|\hat{\alpha})| \tag{3.14}$$

$\hat{x}_{\min}$ is the value that minimizes D, and $\hat{\alpha}$ is the corresponding MLE estimate given $\hat{x}_{\min}$.

Testing goodness-of-fit

The power-law-fitting method will estimate $(\hat{x}_{\min}, \hat{\alpha})$ for any distribution, whether or not the data is from a power-law. To assess the fit, I estimate the p-value with a standard semi-parametric bootstrap approach (Clauset et al., 2009; Broido and Clauset, 2018).

Given firm size data, of which n_{tail} are $x \geq \hat{x}_{\min}$, with MLE $\hat{\alpha}$, I generate a synthetic dataset as follows. For each of n synthetic values, with probability n_{tail}/n I draw a random number from the fitted power-law model, with parameters $\hat{x}_{\min}$ and $\hat{\alpha}$. Otherwise, I choose a value uniformly at random from the empirical distribution below the upper tail, $x < \hat{x}_{\min}$. After n draws, this produces a synthetic dataset that closely follows the empirical distribution below $\hat{x}_{\min}$ and follows the fitted power-law model at and above $\hat{x}_{\min}$.

Then applying the previously defined power-law fitting procedure yields the null distribution of KS-statistics $Pr(D)$. Let D^* denote the value of the KS-statistic for the best fitting power-law model for the empirical distribution. The p-value for this model is defined as the probability of observing, under the null (power-law) distribution, a KS-statistic at least as extreme as D^*. Hence, $p = Pr(D \geq D^*)$ is the fraction of synthetic datasets with KS statistic larger than that of the empirical dataset. Following standard practice, if $p < 0.1$, I reject the power-law as a plausible model of the distribution, and if $p \geq 0.1$, then I fail to reject the model. Failing to reject does not imply the model is correct, only that it is a plausible data generating process.

3.9.3 Likelihood-ratio tests

Given two candidate distributions that fit the data, I use likelihood-ratio tests to differentiate between them. Let $\mathcal{L}_F$ be the log-likelihood of the fit of distribution F, where F could be pl (power-law), log-n (log-normal), or p-exp (power-law with exponential cutoff). The likelihood-ratio statistic (LRT) is given by the difference between the log-likelihood of the power-law and the log-likelihood of the alternative distribution, $\mathcal{R} = \mathcal{L}_{\text{pl}} - \mathcal{L}_{\text{alt}}$.

When $\mathcal{R} > 0$, the power-law is a better fit to the data, and when $\mathcal{R} < 0$, the alternative is a better fit to the data. When $\mathcal{R} = 0$, the data cannot distinguish between the two models. From here, I calculate a p-value against the null model of $\mathcal{R} = 0$, and reject the null hypothesis if $p < 0.05$ and interpret the sign of $\mathcal{R}$ as evidence for one distribution over another. A p-value of 0.05 is slightly more strict than the p-value of 0.1 used in Broido and Clauset (2018), but is more consistent with existing economics research on power-laws, and more conservative at the same time.

Chapter 4

Going the Distance: Estimating the Effect of Provincial Borders on Trade when Geography (and Everything Else) Matters

4.1 Abstract

In the presence of often-cited provincial non-tariff trade barriers, one should observe provincial border effects in Canada. However, using provincial trade data leads to upward biased estimates of the border effect, because intra-provincial trade is skewed towards short distance flows that are poorly estimated by gravity models. We overcome this bias by using sub-provincial trade flows generated from a transaction-level transportation dataset. The results show that border effects fall as geographies are more fine-grained and uniform. In contrast to the U.S., where state border effects were eliminated using similar approaches, provincial border effects remain, with an implied 6.9% tariff equivalent.

4.2 Introduction

It is well known that for some goods (e.g., dairy products and alcoholic beverages) there are significant (non-tariff) barriers to inter-provincial goods trade. Less well understood is the degree to which these barriers are reflected in the level of inter-provincial trade—that is, whether there are provincial border effects. The first and obvious objective of this paper is to assess the presence and magnitude of provincial border effects. To date, the estimation of these effects has been hampered by a lack of data with sufficient geographic detail. This paper overcomes the problem by using a dataset of transaction-level truck and rail shipments to develop estimates of intra- and inter-provincial trade. These fine-grained data permit the estimation of a rich set of models that account for many of the biases endemic to gravity model-based border effect estimates. Therefore, the second and subtler objective of this paper is to illustrate how these detailed data can be used to develop more accurate border effect estimates. The underlying motivation for this requires some context.

Building on McCallum's (1995) initial work, a large literature has developed to measure border effects, be they national or sub-national. While much of the empirical literature has focussed on measuring national border effects, these same methods have also been applied to trade between sub-national regions. The arc of the sub-national border effects literature has been one where the application of more refined methods reduces estimated border effects. But, unlike the international literature,[1] this has led in some instances to the elimination of border effects altogether. In the United States, initially high estimates of inter-state border effects (Wolf, 2000) were reduced by developing more accurate measures of distance (Hillberry and Hummels, 2003; Head and Mayer, 2009; Crafts and Klein, 2015), restricting trade flows to shipments from manufacturers (Hillberry and Hummels, 2003), using a panel specification and controlling for internal migration (Millimet and Osang, 2007), and the use of more fine-grained geographies to define the sub-national trading units (Hillberry and Hummels, 2008, see also Coughlin and Novy, 2016).

Of particular importance are the effects of measured distance (Head and Mayer, 2009) and especially geography (Hillberry and Hummels, 2003, 2008) on border effect estimates. Head and Mayer (2009) show that the inaccurate estimate of distance substantially biases upwards estimates of the border effect, because intra-regional distances tend to be overestimated relative to inter-regional distances. Hillberry and Hummels (2008) demonstrate that

[1] At the national level, estimated border effects have been reduced as McCallum's initial specification was modified to take into account the effects of market access and competition on trade (see Anderson and van Wincoop, 2003; Anderson, 2011), estimates of distance have been refined (see Head and Mayer, 2009) and as new estimators have been applied (see Head and Mayer, 2014). Still, a consistent finding has been that trade is stronger within countries than between them.

estimated state border effects fall to zero as the size of the geographic unit of analysis is reduced, because short distance, large value flows are better estimated. State border effects are an artefact of the geographic scale at which the estimates are made. The sensitivity of border effect estimates to the scale of geographic unit chosen (Coughlin and Novy, 2016) is an instance of the modifiable areal unit problem (MAUP), which can only be addressed by developing a large set of estimates across a broad spectrum of geographies using units that are preferably of uniform shape and size (Arbia, 1989).

To be precise, the methodology requires data on trade between a fine-grained set of sub-provincial regions. The dataset developed here consists of shipments, where each shipment is characterised by its value, transportation cost, distance travelled, and origin and destination. Because origins and destinations have names, addresses and are geo-coded with a latitude and longitude, an almost limitless set of geographies can be applied, making it possible to test the sensitivity of border effect estimates to the geography chosen. The distance shipped is measured along the highway/railway network, eliminating the need to estimate the distance goods travel within and between geographic areas. Finally, because the cost to shippers (revenue to carriers) is measured, as well as the value of the shipment, transportation costs can be directly measured and used to estimate the *ad valorem* tariff equivalent of provincial border effects (see Head and Mayer, 2014). The main contribution of this paper, therefore, is the development of a transaction-level trade dataset that allows an arbitrary number of traditional trade datasets and a wide set of model specifications that address these econometric issues. It is after simultaneously addressing these problems the magnitude and significance of provincial border effects can be more confidently established.[2]

The analysis demonstrates that smaller geographic units typically result in lower border effects, but the adoption of uniform geographic units (hexagons) reduces border effects even

[2] Another strategy, complementary to this one, is to further refine the now standard estimators in order to mitigate issues of measurement error and missing variable bias. This is the approach taken by Agnosteva et al. (2014), who take advantage of the panel nature of current measures of intra- and inter-provincial trade to develop estimates of provincial border effects.

more. Their use helps to mitigate the effects of MAUP on the estimates, while also providing a means to test their sensitivity to the geography and model specification chosen. This is accomplished in the spirt of Briant et al. (2010) by randomly shifting and populating the hexagonal lattice and re-estimating the model each time. These simulations demonstrate the placement of the lattice matters more than size, with the variance of the estimates reduced by using smaller units. The obvious lesson is that border effect estimates are more reliable the smaller and more uniform the unit chosen.

Keeping this in mind, the analysis shows that intra-provincial trade is consistently stronger than inter-provincial trade after taking into account the distance between the trading regions, and the ability of the trading units to generate and absorb trade flows. When sub-provincial areas are used instead of provinces, the border effect tariff equivalent is almost halved, falling from 13.6% to 6.9%. The latter represents the estimate that held after applying an extensive set of checks to mitigating the (typically) upward biasing effects of model misspecification (e.g., non-linear effects of distance) and geography (i.e., the size and shape of the geographic units) on border effects. It stands in sharp contrast to the finding from the United States (Hillberry and Hummels, 2008), where state border effects are eliminated when similar approaches are applied.

The remainder of the paper is organized as follows. Section 4.3 (Data development) reviews the method used to estimate trade between sub-provincial geographic units. Particular attention is paid to explaining how these estimates are benchmarked to known intra- and inter-provincial trade totals and more broadly to the underlying validity of these estimates. Section 4.4 (Model and estimation strategy) outlines the structure of the trade model and the identification of an appropriate estimator. Section 4.5 (Model estimates) presents the estimates, starting from standard inter- and intra-provincial trade estimates, continuing through trade based on sub-provincial geographic units, finishing with a set of robustness checks that test for biases associated with misspecification and the Modifiable Areal Unit

Problem. Section 4.6 (Tariff equivalent of border effects) estimates the tariff-equivalent barriers to inter-provincial trade. Section 4.7 (Conclusions) finishes the paper with a summary of the results and their implications.

4.3 Data development

To date, analysis of Canada's internal trade has been limited to the provincial level, relying on trade tables from the provincial input-output accounts or from reported provincial trade patterns from the Annual Survey of Manufacturers (see Brown, 2003 for the latter). This paper develops a very flexible transaction-level point-to-point dataset. As such, it permits the measurement of trade flows between an almost limitless set of sub-provincial geographic units, providing a means to address many of the econometric issues raised in the borders and trade literature. Since this database is new, however, it is useful to begin by outlining how it was constructed and describing some of its basic characteristics before moving on to discuss the econometric strategy and results.

The data are derived from the Trucking Commodity Origin Destination Survey (TCOD) and railway waybills from 2002 to 2012, with the analysis limited to the 2004 to 2012 period.[3] As these data cover the two primary surface modes, the file is termed the Surface Transportation File (STF). The STF measures the movement of goods from the point where they are picked up to the point where they are dropped off. It is in essence a 'logistics file.' As such, these points do not necessarily represent locations where goods are made or where they are used. However, the analysis requires a database that captures the level of trade between sub-provincial regions, which is embedded as a concept in the gravity-based trade model applied here.

In order to transform the STF from a logistics file to a trade file, provincial trade flows from the input-output accounts are used to benchmark intra- and inter-provincial flows

[3]The discussion focuses on data from 2004 onward because 2002 and 2003 had more limited geographic detail, among other factors that affect comparability across years.

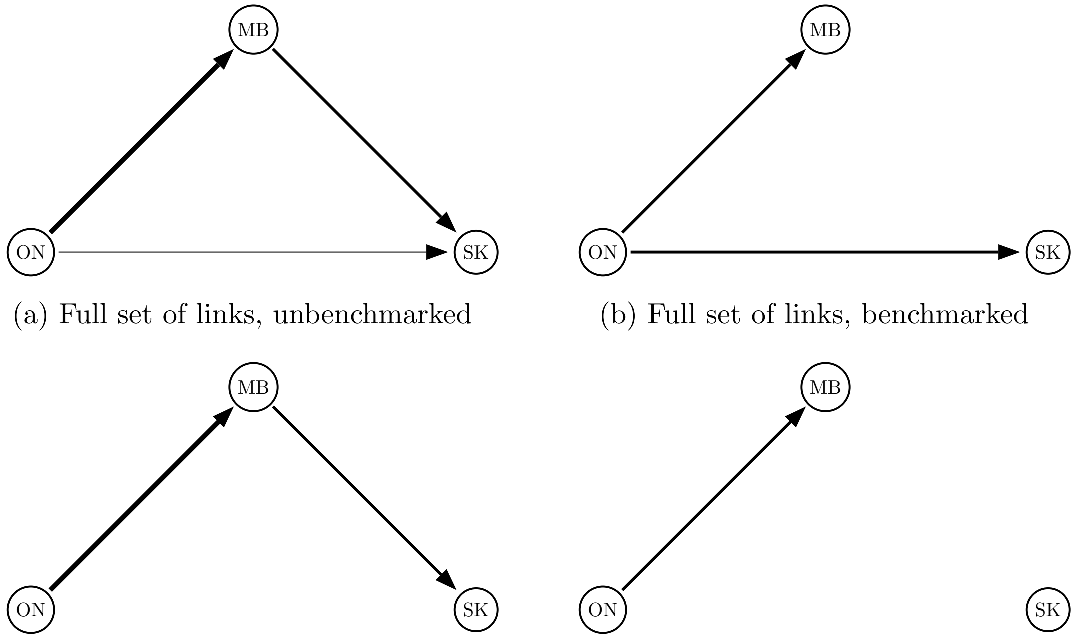
by commodity. That is, each transaction in the STF file is given a weight such that the aggregate adds to the total for the corresponding intra-/inter-provincial flow from the input-output tables. In formal terms, the nominal value of trade between sub-provincial regions (hereafter regions) i and j, X_{ij}, is the sum of the survey weighted value of shipment x indexed by l between origin region i and destination region j,[4] multiplied by the benchmark weight for shipment l, w_l^b:

$$X_{ij} = \sum_l w_l^b x_{lij}, \text{ where } w_l^b = w_l \times w^b. \tag{4.1}$$

The shipment benchmark weight is the shipment-based survey weight, w_l, multiplied by the province pair benchmark weight w_b for the commodity being shipped, with notation for the province pair and commodity suppressed in order to simplify the exposition. The benchmark weight is set such that trade between a given province pair (or within the same province) add to known totals from the provincial trade accounts by detailed commodity and year. The proof of this proposition and a more detailed discussion of the benchmarking procedure is developed in Appendix .1.3.

Conceptually, Figure 4.1 illustrates the benchmarking procedure. Consider the example of flows of vehicles made in various locations in Ontario and ultimately used at various locations in Manitoba and Saskatchewan. They may be first shipped to a distribution centre in Manitoba, with a portion of the shipment sent on to Saskatchewan, which is represented by the unbenchmarked flows in the upper left-hand quadrant of the figure. From a logistics perspective this is a correct representation of the flows, but from a trade perspective the flow from where the vehicles are made in Ontario to where they are used in Saskatchewan is underestimated and the flow from Manitoba to Saskatchewan is overestimated. As presented in Figure 4.1b, benchmarking to the input-output tables weights up at a micro-level the flow from Ontario to Saskatchewan and weights down (to zero) the flow from Manitoba to

[4]Shipments are geocoded by latitude and longitude. For shipments by truck, the latitude and longitude are derived from the postal code of the origin and destination, while for rail shipments is based on the latitude and longitude, and the Standard Point Location Code of the station (yard or siding) where shipments are picked up or dropped off.



(a) Full set of links, unbenchmarked (b) Full set of links, benchmarked

(c) Broken set of links, unbenchmarked (d) Broken set of links, benchmarked

Figure 4.1: Transformation of logistic to trade flows, full and broken sets.

Notes: Example flows for provinces Ontario (ON), Manitoba (MB) and Saskatchewan (SK).

Saskatchewan.

The weighting strategy relies crucially on there being a flow on the STF file between each province pair. If there is not, there is nothing to weight up (or down): $w^b = 0$. The result is no flow between the province pairs (see Figures 4.1c and 4.1d). The risk is that if these 'broken links' are too common and/or correlated with the distance between the province pairs, the benchmarking exercise will result in biased estimates. One source of bias is simply replaced by another.

Table 4.1 presents the ratio of the benchmarked STF inter-/intra-provincial flows to the actual flows from the input-output tables. Because the Atlantic Provinces were found to have a larger number of broken links, particularly with western Canada, they were aggregated together for benchmarking purposes. After doing so, there are relatively few pairs where there was a serious loss of trade. The overall proportion is 99%. There is a tendency for intra-provincial flows to have less of a loss, but this is small. Otherwise, there does not

Table 4.1: Benchmarked intra-/inter-provincial trade flows as a percentage of actual flows.

	Destination							
Origin	A.C.	Que.	Ont.	Man.	Sask.	Alta.	B.C.	Total
Atlantic Canada	99	99	89	94	77	94	89	95
Quebec	99	99	100	98	94	98	98	99
Ontario	100	100	100	99	98	100	100	100
Manitoba	93	97	95	96	95	97	95	96
Saskatchewan	87	96	96	95	98	97	97	97
Alberta	89	97	98	97	99	100	100	99
British Columbia	96	82	99	97	96	99	98	97
Total	98	99	99	97	98	99	98	99

Notes: A.C. stands for Atlantic Canada. Cells display the benchmarked trade value as a percentage of the total trade value given by the Interprovincial Trade Flow (IPTF) database.

appear to be large losses with distance. For instance, the loss for Atlantic Canada's exports to Alberta or B.C. is about the same as Ontario. The effect of these broken links are tested further below by estimating the gravity model with the input-output derived provincial flows and the benchmarked flows, with both sets of data providing qualitatively similar results (see Section 4.5).

While benchmarking adequately accounts for the level of intra- and inter-provincial trade, the pattern of trade especially within provinces may be affected by the functioning of the transport/distribution system—that is, shorter distance logistics driven flows may be more prevalent. This has important implications because, when pooled with inter-provincial flows, these shorter distance, intra-provincial flows may be underestimated, biasing upwards the estimated inter-provincial border effect.

The effect of benchmarking should be to stretch-out inter-provincial trade as short distance flows to/from distributions centres or wholesalers are weighted down and longer distance flows from points where goods are produced to where they are used are weighted up. This can be seen in Figure 4.2, which reports the shipment distance kernel densities with survey weights w_l and benchmark weights w_l^b, with shipment distances divided between intra- and intra-provincial flows. For inter-provincial shipments, as expected, benchmarking tends to reduce the importance of shorter distance flows (less than 1,000km) and increase the importance of longer distance flows, particularly those above 3,000km. For intra-provincial trade, after benchmarking short distance flows are reduced as imported commodities (e.g., shoes and apparel) that are distributed locally are weighted downwards. Still within provinces short distance logistics driven flows may be more prevalent. This effect can be tested more formally by observing whether distance has a stronger effect on intra- relative to inter-provincial trade. The results indicate that this is not the case (see Appendix .2.1 for a detailed discussion). As an additional check, in Section 4.5.3, we use name and address information to identify and remove shipments to and from wholesalers and transportation/logistics firms; the empirical results are similar, suggesting the benchmark procedure is effective at reducing the importance of wholesale and logistics flows)

4.4 Model and estimation strategy

The estimation of provincial border effects relies on the development of data of sufficient quality and richness to generate defendable estimates and a model and an estimator that is appropriate for the data at hand. This section addresses the latter concern.

4.4.1 Trade model

As is now standard in the literature (see Head and Mayer, 2014),[5] trade between regions i and j is treated as a multiplicative function of the capacity of i to serve export markets (S_i), the absorptive capacity of the export market in j (M_j), and a measure that captures the effect of trade costs between i and j (ϕ_{ij}):

$$X_{ij} = G S_i M_j \phi_{ij};\ 0 < \phi_{ij} < 1, \tag{4.2}$$

where G is a constant term. The export capacity can be defined by $S_i \equiv X_i/\Omega_i$, where $X_i \equiv \sum_j X_{ij}$ is the value of output in i and is the sum of exports across all trading partners (including itself). The absorbtive capacity can be defined by $M_i \equiv X_j/\Phi_j$, where $X_j \equiv$

[5]This basic exposition is borrowed from Head and Mayer (2014), albeit in a modified form.

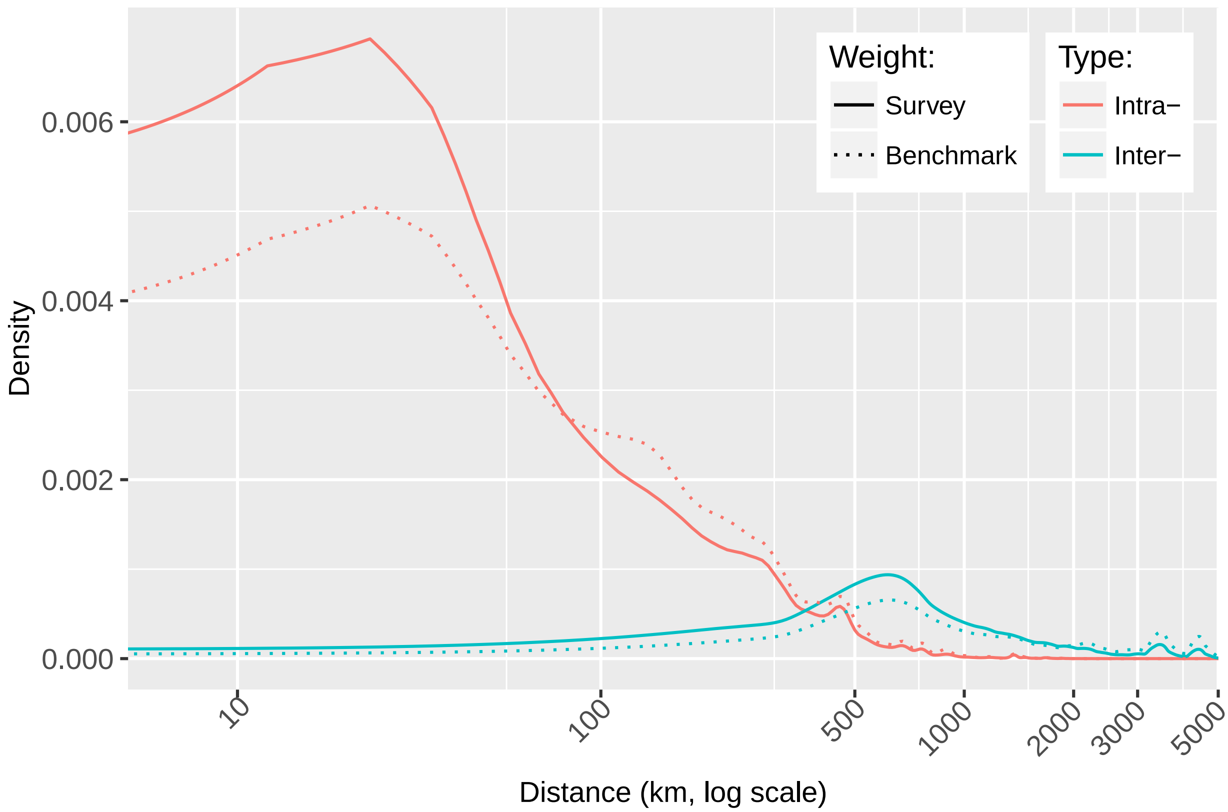


Figure 4.2: Intra- and inter-provincial distance shipped, benchmark and survey weights.

Notes: Distances are trimmed at 5000km for inter-provincial shipments to avoid the long right tail and focus on more-meaningful distance patterns. Epanechnikov kernel is computed in R using the levels of distance (in km), and then converted to a log-scale.

$\sum_i X_{ij}$ is the value of consumption in j and is the sum of imports across all trading partners (including itself). The terms Ω_i and Φ_j are multilateral resistance terms (Anderson and van Wincoop, 2003), where

$$\Omega_i = \sum_k \frac{\phi_{ik}}{\Phi_k} \text{ and } \phi_j = \sum_k \frac{\phi_{kj}}{\Omega_k}. \tag{4.3}$$

Ω_i is a measure of market access for exporting region i and Φ_j measures the level of competition in market j. Trade costs (ϕ_{ij}) are accounted for by the distance between i and j (d_{ij}), the effect of trading within-provinces (δ_p) and trading within sub-provincial regions (δ_r).

4.4.2 Estimator

Equation (4.2) can be estimated with OLS by adding a multiplicative error term, taking the natural logarithm of both sides:

$$\ln X_{ij} = \ln G + \ln S_i + \ln M_j + \underbrace{\beta \ln d_{ij} + \delta_p + \delta_r}_{\ln \phi_{ij}} + \ln \epsilon_{ij}, \tag{4.4}$$

with a set of origin and destination fixed effects to estimate $\ln S_i$ and $\ln M_i$, respectively. While this estimation strategy results in a loss of information regarding the underlying theoretically derived structure of the gravity model (Anderson and Yotov, 2010), it has become the standard means to estimate the gravity model[6] (see Anderson and Yotov, 2012), in part because of ease of estimation, but also because the fixed effects may pick up origin- and destination-specific unobservables that can bias full information-based estimates (Anderson and Yotov, 2010; Head and Mayer, 2014).

Missing variable bias is particularly important in the context of this work. While every effort is made to assign trade flows to where goods are made and used, there may be cases where a destination is acting as a distribution centre, inflating its level of exports and imports. In a similar vein, some provinces may have stronger ties with world markets than other provinces (e.g., British Columbia) reducing their role as a domestic trading partner. In both instances, the fixed effects should take into account these unobservables that affect the level of trade in and out of a region (Head and Mayer, 2014).

The standard approach to estimating Equation (4.4) is to use ordinary least squares (OLS), but it introduces two potential sources of bias. First, starting with Santos Silva and Tenreyro (2006), it has been recognized that OLS estimates of a log-linearized multiplicative model are biased in the presence of heteroscedastic errors. Second, OLS estimates are biased in instances with a larger number of zero flows, which are dropped when the gravity model

[6] This functional form is, in fact, a very well-known variant of a family of gravity models (for reviews see Sen and Smith, 1995 and Fotheringham and O'Kelly, 1989). These constrained gravity models recognize that origin-destination flows often depend not only on the sizes of each origin and destination, but also their relative locations. The economics literature (see Anderson and van Wincoop, 2003), however, provides a firm micro behavioral foundation for the model, particularly within the trade setting.

is estimated using OLS (see Head and Mayer, 2014). The latter is particularly important in this instance, because the models are estimated using flows between sub-provincial regions, resulting in many instances with zero flows between actively trading region pairs.

To address these problems, the first step is to assess whether the error term is heteroscedastic. To do so, the Manning and Mullahy (2001) test is applied using the following specification:

$$\ln \hat{\epsilon}_{ij}^2 = \alpha + \lambda \widehat{\ln X}_{ij}, \tag{4.5}$$

where $\widehat{\ln X_{ij}}$ is the predicted log-level of trade from the OLS estimation of (4.4) and $\hat{\epsilon}_{ij} = X_{ij} - \exp(\widehat{\ln X_{ij}})$ is the difference in levels between the data and the fitted values from the same estimator. Without zero flows, Head and Mayer (2014) find $\lambda \approx 2$ when the data generating process produces log normal errors, but $\lambda \approx 1.6$ when the data generating process produces (Poisson) heteroscedastic errors. In Table 4.2, the estimates of λ are presented for estimates by province, economic region (ER) and census division (CD), where each is a subunit of the other, respectively.

For provincial trade, the point estimate for λ is 2.11, suggesting log-normal errors. However, when the model is estimated by ER and CD the point estimates for λ are about 1.7. For ERs, where the number of zero flows is about 8%, the estimate is about what would be expected based on Monte Carlo simulations (see Figure 4 in Head and Mayer, 2014). For the CD estimates, where almost half of the pairs have zero flows, the expected value of λ is 1.6, with the actual estimate coming in again at about 1.7. However, this estimate is near what Head and Mayer (2014) obtain when they estimate λ from real data. The upshot is that in both instances the estimate for λ is significantly different from 2, suggesting the OLS estimator is inappropriate.

The second step is to assess the potential estimator in the presence of zero flows *and* heteroscedastic errors. Based on Monte Carlo simulation results, Head and Mayer (2014) find the Poisson Pseudo-Maximum-Likelihood estimator (Poisson-PML) of an appropriately

Table 4.2: Manning and Mullahy (2001) test by province, economic region and census division.

Geography	λ	95% c.i.	N
Province	2.11	(1.92, 2.30)	100
Economic Region	1.71	(1.68, 1.74)	5,069
Census Division	1.68	(1.67, 1.69)	47,156

Notes: λ is estimated using Equation (4.5) for Provinces, Economic Regions, and Census Divisions. When λ is significantly different from 2, the test can be interpreted as indicating ordinary least squares is not the appropriate estimator.

transformed version of Equation (4.4) tends to produce the least bias. Therefore, it is our preferred estimator, especially when estimates are based on flows between sub-provincial regions. It is also generally preferred because it perfectly replicates the Anderson and van Wincoop (2003) structural equation estimates (see Fally, 2015).

4.4.3 Geography and estimation

The analysis is based ultimately on the aggregation of point data into a set of geographic units of which the Standard Geographic Classification (hereafter standard geography) based units (e.g., ERs and CDs) are but one of an almost limitless number of geographies. As demonstrated by Hillberry and Hummels (2008), estimates of barriers to trade can be strongly influenced by the geography chosen. Hence, the sensitivity of the results to geography cannot be easily swept aside.

As noted above, Hillberry and Hummels' (2008) findings are an instance of the well-known and quite frankly terrifying Modifiable Areal Unit Problem (MAUP). MAUP is defined as "...the sensitivity of analytical results to the definition of units for which real data are collected" (Fotheringham and Wong, 1991, pg. 1025). MAUP is characterized by both a scale and zoning effect. That is, analytical results depend on the spatial resolution (scale effect) and the morphology (zoning effect) of the geography used to aggregate the data (Páez and Scott, 2004).

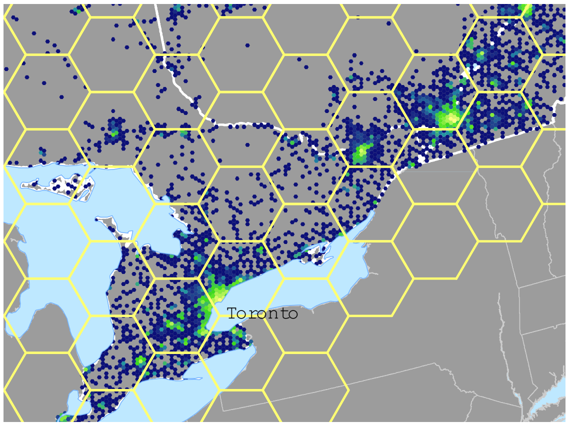
As has been shown elsewhere, these problems apply to multivariate statistics, including spatial interaction models like the gravity model (see Fotheringham and Wong, 1991; Amrhein and Flowerdew, 1992; Briant et al., 2010). In particular, Briant et al. (2010) show gravity model results are more sensitive to scale and to a lesser degree to zoning effects, but these are of secondary importance when compared to model specification problems (e.g., missing variable bias). Still, as Amrhein (1995) demonstrates, MAUP can emerge as a problem even when we abstract from model specification issues.

The effects of geographic aggregation need to be taken into account. This is accomplished by applying different geographies to the data. Here, two strategies are followed. The first is to see how sensitive the results are to the application of standard geographies, namely defining trading regions on the basis of Provinces, ERs, and CDs. The second strategy is to take advantage of the guidance provided by Arbia (1989) who shows analytically that biases resulting from the scale and zoning of the geography can be minimized by ensuring the geographic units are identical and spatially independent. Hence, a hexagonal lattice[7] is overlaid on the geocoded origin and destination points, creating an identical and spatially independent geography (see Figure 4.3). Hexagons that cross provincial borders are split and treated as discrete geographic units.

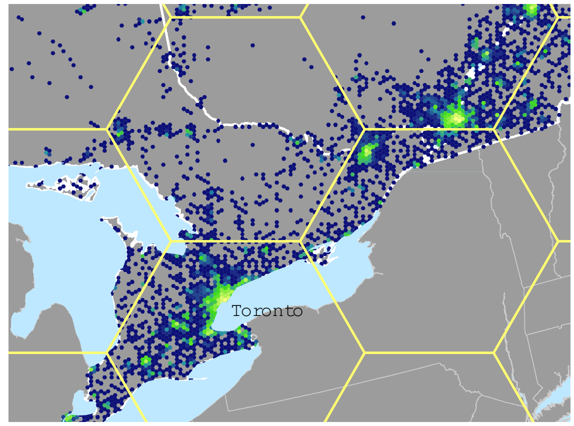
Of course, the use of a hexagonal geography, while perhaps minimizing the bias generated by aggregating data, does not eliminate it. Issues of scale and zoning remain. As there is no theoretically predetermined scale for the hexagons, the sensitivity of the results to size requires testing. For instance, compare the geographic coverage of the 75km and 225km per side hexagons in Figures 4.3a and 4.3b, respectively. The smaller hexagons cover portions of metropolitan areas, while the larger can envelop several. Similarly, while the hexagons do not change in shape, zoning still matters because they are arbitrarily positioned over the origin and destination points. For instance, in Figure 4.3a Toronto is split across two hexagons, while in Figure 4.3c it is split across three. Scaling and zoning effects will be

[7] Other geometries could have been used, such as squares or triangles, but hexagons are used because they would form trade market areas in an idealized world.

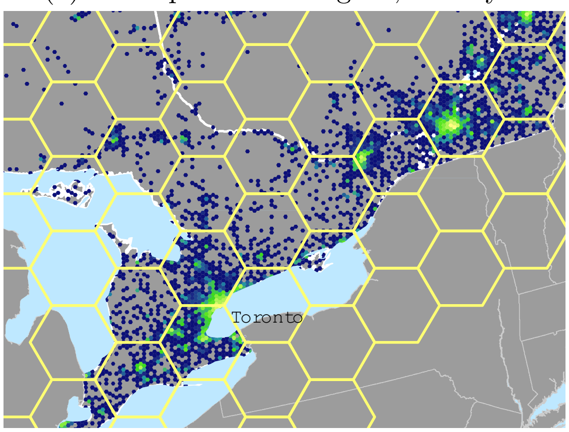
tested by running the model across different scales and zonings.



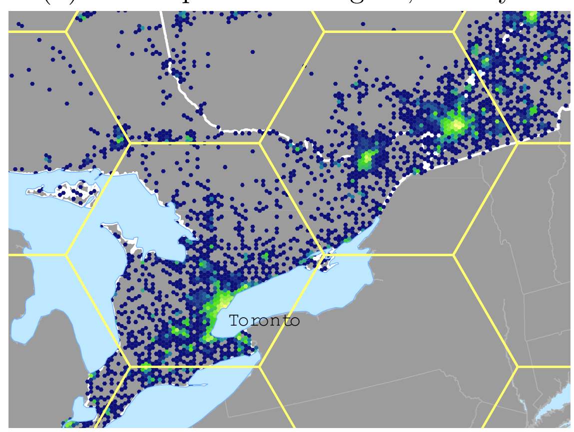
(a) 75km per side hexagons, overlay 1



(b) 225km per side hexagons, overlay 1



(c) 75km per side hexagons, overlay 2



(d) 225km per side hexagons, overlay 2

Figure 4.3: Size and placement of hexagonal lattices.

Notes: Figures 4.3a and 4.3c present two different overlays of hexagons with 75km sides on southern Ontario and Quebec, while Figures 4.3b and 4.3d do the same for hexagons with 225km sides. Hexagons must respect provincial boundaries and are split across provinces. Each ‘point’ in these maps is a 4km-sided ($42km^2$) hexagon with one or more origins/destinations (postal codes or railway terminals). The gradation in colour from blue to green to yellow denotes a greater number of origins/destinations. The 4km-sided hexagons are only for demonstration purposes, they are not used to determine which points fall into which hexagons in the econometric models.

4.5 Model estimates

The presentation of the estimates proceeds first by estimating border effects using province-level estimates of trade flows and therein providing a base case. The exposition then shifts to the estimation of border effects using sub-provincial geographies, which forms the core of the analysis. The remainder of the discussion focuses on a set of robustness checks, with particular attention paid to the sensitivity of the estimates to MAUP, alternative specifications of the model, or combinations thereof.

4.5.1 Standard province-based estimates

To begin, inter-provincial barriers to trade are measured by comparing intra- and inter-provincial aggregate trade levels. This serves several purposes. First, by comparing the actual level of inter-provincial trade to the benchmarked estimates the sensitivity of the results to the loss of trade from the benchmarking can be identified. Second, the OLS, Poisson-PML and Gamma-PML estimates can be compared absent zero flows. Based on their first-order conditions, the Poisson estimator puts more emphasis on the absolute deviation between the actual and predicted flows, while the OLS and Gamma-PML place more emphasis on the percentage deviation and as such are expected to give similar results (Head and Mayer, 2014). Third, the provincial results form a baseline to compare the estimated barriers to inter-provincial trade using trade between sub-provincial regions.

Table 4.3 (Panel A) presents the estimated effects of distance and own province on provincial trade using the input-output-based flows and those derived after benchmarking. The model is estimated using an appropriately transformed version of Equation (4.4) with the mean level of provincial trade from 2004 to 2012 as the dependent variable. There are several points to be drawn from the table. First, estimates based on the input-output and benchmarked flows are similar. There is a tendency for the own province estimates to be lower when using the benchmarked estimates, but this effect is relatively small, particularly when the Poisson estimator is used. There is relatively little loss of generality resulting from the benchmarking and so the remainder of the discussion will focus on these estimates.

Second, there is evidence of a border effect, regardless of estimator used. The one exception is the OLS estimator, which is not significant for the benchmarked flows. Using the input-output benchmarked estimates, the border effect ranges from 1.61 (OLS) to 2.26

Table 4.3: Provincial border effect estimates based on provincial average flows (2004 to 2012).

	Input-output			Benchmarked		
	OLS	Poisson	Gamma	OLS	Poisson	Gamma
Panel A: Network distance						
Distance	−1.025***	−0.661***	−0.999***	−1.077***	−0.686***	−1.078***
	(0.0458)	(0.0496)	(0.0453)	(0.0576)	(0.0522)	(0.0537)
Own province	0.607***	0.865***	0.775***	0.479	0.816***	0.634**
	(0.223)	(0.0807)	(0.190)	(0.289)	(0.0827)	(0.254)
Constant	12.31***	9.916***	12.42***	11.70***	9.515***	12.08***
	(0.410)	(0.559)	(0.373)	(0.630)	(0.877)	(0.535)
Border effect	1.83	2.38	2.17	1.61	2.26	1.89
N	100	100	100	100	100	100
Panel B: Great-circle distance						
Distance	−1.058***	−0.778***	−1.037***	−1.100***	−0.806***	−1.106***
	(0.0462)	(0.0571)	(0.0436)	(0.0613)	(0.0591)	(0.0564)
Own province	0.747***	0.780***	0.840***	0.653**	0.728***	0.743***
	(0.194)	(0.0907)	(0.171)	(0.274)	(0.0882)	(0.249)
Constant	12.01***	10.49***	12.17***	11.29***	10.12***	11.70***
	(0.405)	(0.547)	(0.360)	(0.644)	(0.848)	(0.535)
Border effect	2.11	2.18	2.32	1.92	2.07	2.10
N	100	100	100	100	100	100

Notes: Models include fixed effects for origins and destinations. ***, **, and * indicate significance at the 0.01, 0.05 and 0.10 levels, respectively. Robust standard errors are in parentheses. The border effect is given by exp(own province).

(Poisson)—that is, within province trade is between 61 and 126% higher than inter-provincial trade after taking into account distance and multi-lateral resistance.

One of the benefits of building the trade estimates up from shipment data is that it is possible to obtain a more accurate measure of the distance goods travel within and between provinces. The sensitivity of the results to the distance measure can be tested by comparing estimates based on the network distance to the great-circle distance typically used in the literature (see Appendix .1.4). In a nutshell, how distance is measured matters. On average, great-circle distances are 66% of the actual distance shipped. As a result of the compression of distance, the parameter on distance should be more negative for the great-circle distance-based estimates, which is true regardless of the estimator. It is also the case that the great-circle within-province distances are, in relative terms, over-estimated (see Appendix .1.4 Table .7). This over-estimation will have the effect of biasing upwards the own province effect. The OLS and Gamma estimators show this effect, but not the Poisson where the bias appears to be captured by the coefficient on distance.

4.5.2 Estimates by sub-provincial geography

Estimates of provincial border effects based on the comparison of intra-provincial to inter-provincial trade flows may still suffer from bias, if these units do not effectively capture the pattern of trade. As shown by Hillberry and Hummels (2008), if short distance flows predominate and these are not properly captured by the internal distance measure, the estimated border effect may be upward biased.

To further establish the presence and strength of provincial barriers to trade, intra- and inter-provincial trade flows are measured using sub-provincial geographies of different sizes and morphologies. Since trade can be both within and between sub-provincial geographic units, a binary variable is included for within unit trade (own region). It should capture non-linearities with respect to the effect of distance for these shorter distance flows and/or differences in the nature of own unit versus between region trade. Within region trade is more likely to include short distance flows between manufactures and distribution centres, between distribution centres and retail stores (Hillberry and Hummels, 2003) or between upstream suppliers and downstream users of intermediate inputs (Hillberry and Hummels, 2008).

Moving from the provinces down to the scale of sub-provincial units introduces the problem of zero flows between trading units. The set of trading units is defined as those units that either make or use the good. Excluded are units that do not engage in goods trade, either within themselves or with other units. This may result from no measurable goods production in the unit or because of sampling variability. Since the estimates are based on the average value of trade over 9 years, the effect of sampling variability is likely to be low. Of course, those units included in the trading set do not trade with all potential units, resulting in zero flows. Zero flows may be due to random chance (again sampling variability) or they may be structural (producers incur costs above the trading threshold). To permit the presence of zeros, the Poisson estimator is used. For zero flows, the distance between regions is measured using the out-of-sample predicted values of a regression of the network distance on the great-circle distance.

As noted above, there are five geographic units used for the analysis. Three are based on standard geographies, ERs, CDs and FSAs. The other geographic units are two hexagon lattices with sizes of 75km and 225km per side. Choosing hexagons with areas larger than 225km per side results in some smaller provinces having very few hexagons. On the other hand, using hexagons smaller than 75km per side results in such a large number of fixed effects that the estimation often fails to reliably converge, which is problematic for the simulations to follow.

Focussing first on ERs as the trading unit, the distance parameter tends to be less negative than the province-based estimates, with own region likely picking up the non-linear effect of short distance flows (see Table 4.4). More to the point, the own province estimate is smaller, resulting in an estimated border effect of 2.10. Using CDs—a fundamental building block of ERs—the number of potential trading pairs rise from 5,329 to 77,274. For this much larger set of smaller trading units, the border effect falls slightly to 1.97.

For both the small and large hexagons, the own region effects were not statistically significant, while the own province effect remained significant but notably smaller in magnitude than standard geographies. The result is an estimated border effect that falls in a narrow range from 1.60 (large hexagons) to 1.62 (small hexagons) (see Table 4.4).

On the surface, these results stand in contrast to Hillberry and Hummels (2008), who find

Table 4.4: Provincial border effect estimates based on flows between large and small hexagons (2004 to 2012).

	Geography				
	Standard		Hexagon		FSA
	Economic Region	Census Division	225km	75km	
Distance	-0.551^{***}	-0.573^{***}	-0.820^{***}	-0.742^{***}	-0.426^{***}
	(0.0461)	(0.0278)	(0.062)	(0.0357)	(0.0146)
Own region	0.408^{***}	0.467^{***}	-0.101	-0.0215	1.052^{***}
	(0.138)	(0.121)	(0.127)	(0.117)	(0.0966)
Own province	0.743^{***}	0.679^{***}	0.472^{***}	0.483^{***}	0.909^{***}
	(0.0951)	(0.0633)	(0.0872)	(0.0783)	(0.0421)
Constant	6.981^{***}	7.094^{***}	3.142^{***}	2.540^{***}	2.015^{***}
	(0.490)	(0.359)	(0.776)	(0.477)	(0.383)
Border effect	2.10	1.97	1.60	1.62	2.48
N	5,329	77,274	8,619	132,862	2,574,640

Notes: All models utilize a Poisson-PML estimator and include fixed effects for origins and destinations. ***, **, * indicate significance at the 0.01, 0.05 and 0.1 levels, respectively. Robust standard errors are presented in parentheses. Large hexagons are 225km per side while small hexagons are 75km per side. Own region refers to flows within the geographic unit of analysis: Economic Region, Census Division, hexagon or Forward Sortation Area (FSA). The border effect is given by exp(own province).

state border effects are an artefact of the geography used to measure internal trade. However, they found border effects only disappeared when using an even finer grained geography than applied here, namely 5-digit ZIP codes. To account for this, the model was re-run using Forward Sortation Areas (FSAs), which are the closest Canadian analogue to ZIP codes.[8] Importantly, the point estimates for own province remains positive and significant (see Table 4.4). Provincial border effects remain even with a very fine-grained geography, a finding, as will become apparent, that is robust to a wide set of specifications (see Section 4.5.3).

[8]FSAs are defined by the first three alphanumeric characters of a postal code. While the mean area of FSAs is much greater than ZIP code areas (5,894km^2 versus 229km^2), this is due to a few extremely large FSAs in Canada. In fact, the median FSA area is smaller than the median ZIP code area (41km^2 versus 96km^2) and FSAs remain smaller up to the 70th percentile. Given that these smaller FSAs are in dense metropolitan areas, they should be capturing the non-linear effect of distance on trade for these short distance flows.

The obvious conclusion to be drawn from the provincial- and sub-provincial-based estimates of the border effect is that the geography chosen matters, but at this point there is still insufficient information to draw strong conclusions. Two issues in particular need to be addressed. The first is the question of how sensitive the results are to the MAUP, namely scaling and zoning effects (i.e., the size and placement of the hexagons). It is unknown whether the variation in provincial border effects across hexagons of different sizes (or lack thereof) is outstripped by variability resulting from the placement of the hexagonal lattices. The second is whether there is a still unaccounted for non-linear effect of distance on trade that may, in turn, influence estimates of provincial border effects. The elasticity on distance varies considerably across geographies and estimators and, as Head and Mayer (2014) note, variation on the distance term between the Poisson and Gamma estimators may be an indication of model misspecification, which is observed in Table 4.3. The necessary next step, therefore, is to more rigorously assess how the geography and model specification, particularly non-linear effects of distance, influence estimated border effects.

4.5.3 Sub-provincial estimates robustness checks

To test the robustness of the estimates, the analysis proceeds in four steps. The first tests how sensitive the results are to the MAUP. The second step tests whether there is a non-linear effect of distance on trade that may, in turn, influence estimates of provincial border effects. The third combines the first two by asking how sensitive the results are to taking into account both MAUP and the non-linear effect of distance, and the fourth and final step returns to Hillberry and Hummels' (2008) results and asks whether provincial border effects remain using FSAs as trading units after applying their specification and estimator, as well as our fully-specified model.

Modifiable areal unit problem

The sensitivity of the results to the MAUP is tested by re-running the models on randomly shifted hexagonal lattices of varying sizes. Mechanically, the process is as follows. For a given size of hexagon, the lattice is superimposed on Canada's landmass, with each origin and destination point coded to their respective province and hexagon. The lattice is then perturbed by shifting the centroid of each hexagon to any random point within a circle circumscribed by its borders. The set of points is limited to the circumscribed circle, because shifting over more than one unit simply repeats the pattern. The origin and destination points are recoded to their province and hexagon. The lattice is randomly shifted 100 times,[9] resulting in a set of parameters that describes how sensitive the estimates are to the placement of the lattice (i.e., the MAUP zoning effect) for a given size of hexagon. This is repeated for seven sizes of hexagons increasing in 25km per side increments from 75km to 225km. This accounts for how sensitive the results are to the size of hexagons (i.e., the MAUP scaling effect).

To represent the distribution of coefficients resulting from the simulations for the main variables—own province, own region (hexagon) and distance—Figure 4.4 presents box plots by size of hexagon. The boxes represent the inter quartile range, with the line intersecting the box being the median coefficient value. The ends of the whiskers—the upper and lower adjacent values—represent the ranked coefficient value that is nearest to but not above (below) 1.5 times the inter-quartile range from above (below). The dots signify extreme values.

Regarding own province, the median coefficient values range from 0.50 for the smallest hexagons to 0.45 for the largest (scaling effect), with the coefficients converging towards the lower median value as the size of hexagons increase. This is consistent with Coughlin and Novy's (2016) analytical finding that if trade is particularly strong within small units as the

[9]It would have been preferable to randomly shift the lattices more than 100 times and increase the number of size categories used, but this is a computationally burdensome process, both in terms of geo-coding the flows to a given lattice and with respect to the Poisson PML estimations.

size of the unit expands, the border effect will tend to fall. The placement of the hexagonal lattice (zoning effect) has a larger effect on the estimates, with the difference between the box plot lower and upper adjacent values being greater than the difference in the medians across the size of hexagons, which contrasts with Briant et al. (2010) who find the scaling effect is more important. More broadly, the lesson to be drawn is that shifting to a uniform geography has a qualitative effect on estimated border effects, and this result holds after taking into account the effect of the size and placement of the hexagons on the estimates.

Non-linear effects of distance

Variation in the results across hexagons of varying sizes may stem from a non-linear effect of distance on trade, a telltale sign of which is the negative association between hexagon size and the distance coefficient (see Figure 4.4). As the hexagons become smaller the average distance shipped falls. If these more prevalent shorter distance flows are underestimated, the provincial border effect will be overestimated, because intra-provincial trade occurs over shorter distances more than inter-provincial trade (see Figure 4.2). This appears to be the case as there is a positive association between the own province and the distance coefficients (see Figure 4.4).

There are at least two reasons why the effect of distance on trade is expected to vary with itself. First, prices charged by trucking firms, for instance, include fixed and variable (line-haul) cost components. Since fixed costs per shipment are around \$200 and line-haul costs increase at about \$0.80 per km (see Brown, 2015), prices inclusive of transport costs will be (effectively) uniform over short distances. Second, the endogenous clustering of upstream suppliers and downstream firms[10] and hub and spoke distribution networks[11] (Hillberry and

[10] As shown in Behrens et al. (2015), plants tend to cluster geographically and this is negatively associated with distance from upstream suppliers and downstream intermediate goods users.

[11] If short distance trips from manufactures to distribution centres or from distribution centres to retail stores are captured by the data, which may still be the case despite the steps taken to adjust for these effects, the same pattern of trade will be observed as resulting from clustering. We address this in Section 4.5.3 by using the names and addresses of shippers and receivers to identify and remove shipments to and from wholesalers and transportation/logistics firms. The results are similar.

Hummels, 2008) may result in a large volume of trade over short distances with a steep drop as distance shipped moves beyond these 'just down the street' shipments. Uniform prices over short distances combined with clustering/distribution effects results in a complicated set of expectations. For very short distance flows, the effect of distance on trade may be very negative (or at least after a short plateau), but the negative effect of distance on trade beyond these very short distance flows is expected to be initially weak, but increasing as variable costs outstrip the effect of fixed costs on transportation rates. This pattern in the data requires moving beyond the standard quadratic form to account for non-linearities.

To account for these non-linear effects of distance, the model is re-estimated using a spline with knots at 25km, 100km and 500km (see Table 4.5) employing the same hexagonal lattices used for the estimates presented in Table 4.4.[12] Focussing on the smaller hexagon results, the distance elasticities are consistent with a steep drop in shipments over very short distances (reflecting the co-location of input-output linked plants, for instance), while the insignificant effect of distance for 25 to 100km distance flows is consistent with a relatively constant transportation rate charged by firms over short distances. Importantly, accounting for the non-linear effect of distance causes the coefficient on own province to become more similar across hexagon size classes. Still, given the sensitivity of the result to the placement of the hexagonal lattices, it remains unclear from this one set of point estimates how truly similar the border effect estimates are between the large and small hexagons.

Finally, as is standard in the literature, a binary variable is added for hexagons that share a border (contiguous regions). The expectation is that the contiguity measure will account for short distance flows across boundaries. For both the large and small hexagons, the contiguous region coefficients is insignificant, and the own province coefficient falls while remaining significant.

[12] These particular hexagonal lattices are used to maintain comparability across the models.

Non-linear effects of distance and the modifiable areal unit problem

This next check assesses whether accounting for the non-linear effects of distance reduces the degree of variation in results across different sizes and placement of hexagons. This is again accomplished by randomly perturbing the hexagonal lattices for the largest (225km per side) and smallest (75km per side) hexagons, but also across model specifications. The 'base' model estimates replicate those presented in Figure 4.5 (which use the specification presented in Table 4.4), while Model 1 and Model 2 match those in Table 4.5.

Taking into account the non-linear effect of distance reduces the median coefficient of the small hexagons, but increases that of the large hexagons (see Figure 4.5), effectively reversing the pattern in Figure 4.4. The addition of contiguity to the model (Model 2) produces large and small hexagon-based provincial border effects that are statistically indistinguishable. The coefficients on own hexagons also converge, but this only occurs when contiguity is taken into account. While the central tendencies of the small and large hexagon coefficient distributions are the same, their variances are not, with the large hexagons having more than double the inter-quartile range of the small hexagons. Hence, on this basis, the small hexagon border effects are the most reliable.

Provincial border effects based on Forward Sortation Areas

As a last robustness check, the analysis revisits Hillberry and Hummels' (2008) finding that state border effects are eliminated when trade is measured using 5-digit ZIP codes. This entails initially using the same estimator (OLS) and model specification (quadratic term on distance) used in their analysis and then applying the preferred estimator (Poisson) and model used above (distance effects estimated using a spline).

While the model and estimators can be equated, it should also be kept in mind results may vary because of differences in the underlying data. The benchmarking procedure weights the STF data towards longer distance flows, which, combined with Canada's geography, increases the average distance shipped relative to Hillberry and Hummels' (2008) Commodity

Flow Survey (CFS). The average distance shipped between FSAs being 1,679km (1049 miles), while the average distance between ZIP codes is 837km (523 miles) (see Hillberry and Hummels 2008). Since the distance elasticity increases (in absolute terms) with distance shipped, the expectation is that the effect of distance is likely to be stronger here. The CFS data are shipper-based, which allow Hillberry and Hummels (2008) to focus on manufacturing shipments only—wholesale and distribution shipments are not included in the data. While our data is carrier-based, we have the names and addresses of the shippers and receivers, allowing us to identify and remove shipments to and from wholesale and distribution centres, as well as transportation logistics firms.

Table 4.6 shows the estimates, with the first three columns of results based on the equivalent model used in Hillberry and Hummels (2008, Table 2). The first column presents the OLS-based estimates, while the second and third columns present the Poisson-based estimates with and without zeros included. Evaluating the effect of distance using the CFS mean distance of 837km (523 miles) the elasticity is -0.42, more than double the ZIP code-based estimate of -0.19. Also found is a much higher point estimate for own region (FSA). These results are in line with our expectations given the differences between the underlying data. Notably, the own province effect is positive and significant using the same estimator, model and equivalent geography as Hillberry and Hummels (2008), and estimates using the PPML estimator are in line with our other results.

The application of the Poisson estimator reduces the effect of distance, because larger, (typically) short distance flows are weighted more heavily. Evaluated at 837km, the elasticity on distance is -0.25, and only slightly lower when zero flows are added. The Poisson estimator also produces smaller but still significant own region and province effects. The inclusion of zero flows, results in a positive coefficient on distance up to 5km, and then a declining point estimate thereafter. The effect of adding zeros also raises the point estimates on own region and province. The highly non-linear effect of distance when the Poisson estimator is applied suggests the influence of distance on trade has to be treated in a very flexible manner. This is accomplished, as above, by estimating a spline on distance.

Model 2 uses the same structure as Model 1 in Table 4.5, with knots at 25km, 100km and 500km. The estimated provincial border effect is lower than when the quadratic is used on distance, but remains significant. Unlike when hexagons are used, there is no strong negative effect on distance between 0 and 25km. The effect of short distance flows is captured by the own region term instead, with a strong positive coefficient, because the vast majority of FSAs are small and located in metropolitan areas. Hexagons, whose size distribution by construction is not associated with the density of short distance flows, have a weaker relationship. The estimated provincial border effect is unchanged with these modifications to the specification. In short, unlike Hillberry and Hummels (2008), the adoption of very small trading units does not eliminate border effects. Therefore, provincial border effects, while sensitive to the specification of the model and geography, are never eliminated. The remaining question is whether they are economically meaningful.

4.6 Tariff equivalent of provincial border effects

To estimate the tariff equivalent of the provincial border effect the approach described in Head and Mayer (2014, pgs. 32–34) is applied. δ_p denotes the provincial border effect coefficient, which reflects the reduction in trade costs between sub-provincial regions by simply being part of the same province. Given that $\delta_p = \eta(\ln \rho^{inter} - \ln \rho^{intra})$, where ρ^{inter} and ρ^{intra} are inter-provincial and intra-provincial trade costs, respectively, and η is the trade elasticity with respect to transportation costs, if t is the tariff that must be removed to equate the cost of moving goods within and between provinces, then the inter-provincial trade tariff equivalent is

$$t = (1 + \nu) \left[\exp(\delta_p / \eta) - 1 \right], \tag{4.6}$$

where ν is the tariff equivalent of within-province barriers to trade, which are assumed to be zero. Hence the only missing information, is the trade cost elasticity of trade:

$$\ln X_{ij} = \phi_i + \xi_j + \eta \ln \tau_{ij} + \mu_{ij}, \tag{4.7}$$

where τ_{ij} is 1 plus the *ad valorem* transportation costs, ϕ_i and ξ_j are respectively origin and destination fixed effects and μ_{ij} is the error term. *Ad valorem* transportation costs are derived from the STF, which reports both the price charged to shippers and the estimated value of each shipment. The estimated[13] price elasticity based on (4.7) is -6.40, which is between the median (-5.03) and average (-6.74) price elasticities identified in Head and Mayer's (2014) meta-analysis.

For the median provincial border effect coefficient on the 75km per side hexagon (see Figure 4.5, Model 2), $t = \exp(0.426/6.40) - 1 = 0.069$, or 6.9%. To provide some perspective, using a very different methodology, Agnosteva et al. (2014) arrive at a lower, but statistically indistinguishable,[14] estimate of 5.6%.

The tariff equivalent of the border effect across the standard and hexagonal geographies are presented in Figure 4.6 and illustrate the impact of the trading unit chosen on the border effect. The hexagons use the median point estimates from the simulations presented in Figures 4.4 and 4.5. The provincial estimates are the highest at 13.6% followed closely by the ER- and CD-based tariff equivalents of 12.3% and 11.2%, respectively. It is the imposition of a uniform hexagonal geography that causes the most notable drop in the tariff rate. As the hexagons become larger, the point estimates converge to tariff equivalent of 7.3%. The tariff equivalent for the 75km and 225km per side hexagons that takes into account the non-linear effect of distance and contiguity (see Figure 4.5, Model 2) provides the lowest estimates that are essentially indistinguishable.

Therefore, in the fully specified model the size of hexagon chosen is of little consequence.

[13] Equation (4.7) is estimated using ordinary least squares. ERs are used as the trading unit because of the lack of zero flows that can bias the estimates.

[14] For instance, the 5.6% point estimate falls within the 90% confidence interval using the point estimates from Model 2 in Table 4.5 using the 75km per side hexagons.

At 6.9%, the 75km per side hexagons provide the preferred estimate, because of the smaller inter-quartile range relative to the 225km per side hexagons. Compared to this estimate, relying on provincial trade would increase border effect estimates by 6.7 percentage points. This difference is non-trivial. To put it into some perspective, this value is about the same as Canada's mean tariff rate (4.9%)[15] and larger than *ad valorem* transportation costs on internal trade (2.5%)[16].

4.7 Conclusions

Intra-national border effects have proven difficult to measure because of a lack of geographically detailed data on trade within and across provinces. Using a very flexible transaction-level transportation data file to generate regional trade flows within and across provincial borders, the analysis show that regardless of the model or geography chosen provincial border effects are always significant, with an implied *ad valorem* tariff equivalent of 6.9%. This stands in contrast to the U.S.-based estimates, where state border effects are eliminated when similar approaches—i.e., same model and geography—are applied (Hillberry and Hummels, 2008).

Beyond this substantive contribution, the paper's other contributions are methodological. The development of geocoded transaction-level data made it possible to test the effects of geography and model specification simultaneously through a set of simulations. From this several points can be drawn. First, while the results are sensitive to the size of geographic unit chosen (i.e., provinces, economic regions, census divisions and hexagons) there is no simple linear relationship between (average) size and border effects. In fact, choosing a uniform shape (à la hexagons) is more important than size, which speaks to Arbia's (1989) analytical finding that biases resulting from the scale and zoning of the geography are minimized when

[15] The estimate is based on the unweighted mean MFN tariff rate for the period 2004 to 2008. Source: `http://data.worldbank.org/indicator/TM.TAX.MRCH.SM.FN.ZS` (accessed: July 14, 2015).

[16] This is the average ad valorem transportation cost across 2-digit SCTG commodities for internal Canadian trade.

using identical units.

Second, after taking into account the non-linear effects of distance, the median smallest and largest hexagon's parameters converge. In other words, with a correctly specified model, the geographic scale of the unit does not matter. Finally, considerable variation in the estimates result from the simple shifting of the hexagonal lattice, even after applying the full model. These effects are, perhaps unsurprisingly, minimized by using smaller geographic units. In total, the most precise estimates come from a model that carefully accounts for the effect of distance on flows between small, uniform geographic units.

There is, of course, more work that needs to be done. Identifying the effect of provincial non-tariff barriers on estimated border effects will require direct information on the extent of these barriers, and other factors that influence inter-provincial trade (e.g., firm linkages and migratory flows across provincial borders). Furthermore, while this work is able to estimate provincial border effects and their tariff equivalents, there remains the question of the overall welfare implications if they were eliminated, which as Albrecht and Tombe (2016) demonstrate can be substantial.

Acknowledgements

We would like to thank Danny Leung, Bart Los, Dennis Novy, Trevor Tombe, Dan Trefler, Yoto Yotov, Afshan Dar-Brodeur, and participants at the Economic Analysis Division and UQAM seminar series and the North American Regional Science and Western Regional Science meetings for their helpful comments. We are grateful for research assistance from Javad Sadeghzadeh and Olena Melin. Authors' names appear in alphabetical order.

.1 Data Appendix

.1.1 Valuing shipments

The waybills on which the Surface Transportation File (STF) is based describe the commodity and tonnage for each shipment, but not its value. To generate an estimate of value, required is a measure of the value per tonne. This is derived from an experimental transaction level trade file that measures the value and tonnage of goods by detailed HS commodity in 2008. Since the trade file identifies the mode used for each shipment, the value per tonne for each commodity also varies by the mode used. Export prices indices are used to project the value per tonne estimates through time (see Brown, 2015 for a more detailed discussion).

.1.2 Geocoding shipment origins and destinations

Using postal code data from the Trucking Commodity Origin Destination (TCOD) survey and Standard Point Location Codes (SPLC) from the rail waybill file, each shipment is geocoded (given a latitude and longitude for the origin and destination) from 2004 to 2012. These are then used to give the file a 2006 Standard Geographic Classification. As a result, each origin and destination is coded to its ER and CD. Prior to 2004, the TCOD did not use postal codes to identify origins and destinations. For these years the flows are only coded to ERs and CDs. Note that because origins and destinations are given latitudes and longitudes other non-standard geographies can also be applied, such as the hexagonal lattices used here.

.1.3 Benchmark weights

When constructing the file, one of the primary goals is to ensure that the value of trade on a shipment basis in the STF adds to known trade totals by commodity from the interprovincial trade flow file. To do so there are two problems that need to be overcome as the files represent different trade concepts and use different commodity classifications.

In the interprovincial trade flow file, a source represents the point of production, while a destination represents a point of consumption. However, in the STF, a source represents the point at which the shipment is picked up, while the destination is the point at which the shipment is dropped off, including warehouses that act as transportation waypoints. A commodity that is produced in Quebec and consumed in BC would be recorded as a flow from Quebec to BC in the inter-provincial trade flow database, but that flow may have multiple sources and destinations in the STF if it stops at warehouses in different provinces along the way. For instance, a Quebec to BC trade flow might be counted as flows from Quebec to Ontario and then from Ontario to BC in the STF. This results in the STF overestimating the flows between close provinces and underestimating the flows between provinces that are farther away from each other, potentially biasing upwards border effect estimates. The benchmarking is an attempt to re-weight the surface transportation shipments to reflect the inter-provincial trade flow concept.

In addition to representing different concepts, the two files use related, but in practice, different commodity classification systems. Although both commodity classifications are built from the commodity-based Harmonized System (HS), the resulting aggregate classifications used are so different as to eliminate any possible one-to-one matching between them. The STF uses the Standard Classification of Transported Goods (SCTG 1996),[17] while the inter-provincial trade flow file uses the Input-Output Commodity Code system (IOCC). At every level of aggregation, there are SCTG codes that map to multiple IOCC codes, and vice versa. Since the number of multiple matches is large, no attempt is made to force a single IOCC code to any SCTG code. Instead, the goal is to benchmark the transportation file so that it represents the same values as the interprovincial flow file without taking a stand on which transported commodities represent which input-output commodities. That is, instead of forcing a one-to-one concordance between the files, we employ a strategy where the benchmark weights are set such that flows add total commodity flows generated by the

[17] http://www.statcan.gc.ca/subjects-sujets/standard-norme/sctg-ctbt/sctgmenu-ctbtmenu-eng.htm.

input-output system. The process for doing so is set out in a series of steps.

In the first step each file is aggregated to include values of flows by year, origin province, destination province and commodity (SCTG for the surface transportation file and IOCC for the inter-provincial flow file). This generates two vectors of the value of trade for IOCC commodity flows and SCTG commodity flows: X_I and X_S, respectively.

The second step builds a concordance between SCTG and IOCC by province pair and year. This is done through one-to-many mappings from SCTG to HS and from IOCC to HS, which combine to form a many-to-many map from SCTG to IOCC forming a concordance matrix C used in the third and final step.

In the final step the benchmark weights are calculated. To do so, for each year and origin and destination province pair, the two commodity vectors, X_I and X_S, are combined with the concordance matrix C, of which all values are either 0 or 1 (depending on whether a given SCTG commodity maps to a given IOCC commodity). Defining the number of IOCC commodities as M and the number of SCTG commodities as N, then X_I has length M, X_S has length N, and C is an $M \times N$ matrix. Then the benchmarking problem can be written:

$$(B \circ C)X_S = X_I, \tag{.8}$$

where B is the $M \times N$ matrix of benchmark values, and $\circ$ is the element-wise matrix product (Hadamard product). Any B that solves this system of equations will benchmark X_S to X_I. The problem is to find a solution to M equations given $M \times N$ unknowns. A typical solution is to force C to be one-to-one such that if $c_{mn} = 1$, then $c_{mo} = 0$ for all $o \neq n$ and $c_{on} = 0$ for all $o \neq m$, where i and j index elements of C. In that way, the matrix $B \circ C$ has only M non-zero values and the benchmark weight is $b_{mn} = V_{I_m}/V_{S_n}$. In this case, the concordance would be static. There would be no need to undertake a concordance by year let alone province pair. However, this approach throws away considerable amounts of information regarding the underlying trading relationships between provinces as the commodity profile of trade varies across province pairs. For instance, the commodity in a forced pairing may not be found in the trade between the two provinces. Hence, the benchmarking concordance should reflect and indeed take advantage of those differences.

In order to preserve information in the face of a particularly severe many-to-many concordance problem in C, each element of B is separated into two parts, $b_{mn} = b_m \hat{b}_{mn}$, where

$$\hat{b}_{mn} = \left(\frac{X_{S_n}}{\sum_o c_{mo} X_{S_o}} \right) \left(\frac{X_{I_m}}{\sum_o c_{on} X_{I_o}} \right). \tag{.9}$$

Equation (.9) is simply the product of the trade shares of the concorded SCTG- and IOCC-based flows. It is assumed that the SCTG- and IOCC-based flows are an accurate representation of the patterns of trade and therein provide appropriate splits against which to benchmark. b_m is the value that solves the equation

$$b_m \sum_n \hat{b}_{mn} c_{mn} X_{S_n} = X_{I_m}, \tag{.10}$$

for each equation in the system, with the convention that $b_m = 0$ if $X_{I_m} = 0$ or the sum on the left-hand-side of Equation (.10) is zero. The only remaining issue is to calculate a single benchmark value for one SCTG code given by

$$w_n^b = \sum_m b_{mn} c_{mn}, \tag{.11}$$

which is considered the benchmark weight for all shipments of SCTG commodity m in that year and province origin-destination pair. In other words, w_n^b is the sum of the values of column n of $B \circ C$.

Again, any B that solves this equation will be a benchmark, but the choice is made to maximize the information available. Specifically, $\hat{b}_{mn}$ is picked to use the value of an SCTG commodity flow relative to the total SCTG flows that point to the same IOCC code m, and also the value of the flow of that IOCC code relative to all of the IOCC codes that are pointed *at* by SCTG commodity n. In addition, although we cannot compare two commodities directly, we know the total value of benchmarked trade is that same as the total value of inter-provincial trade (for each year-province-province observation), because

$$\sum_n w_n^b X_{S_n} = \sum_m X_{I_m}. \tag{.12}$$

Hence, the procedure achieves the ultimate goal of ensuring trade flows add to know totals from the provincial accounts. Unfortunately, in some cases the sample of shipments will not cover all of the SCTG commodities between two provinces in a year (see Figure 4.1 in Section 4.3). In this case, for some IOCC commodity m, the i-th element of the vector $(B \circ C)X_S$ is zero because $X_{S_n} = 0$ for all the possible commodities that map to I_m (i.e., those for which $c_{mn} = 1$). In this case, the element X_{I_m} is included in the total interprovincial trade, but the corresponding X_{S_n} is zero on the right-hand-side, which means the total trade the STF is less than the total trade in the interprovincial flows,

$$\sum_n w_n^b X_{S_n} < \sum_m X_{I_m}. \tag{.13}$$

Finally, in the main body of the text the subscript n is supressed such that the benchmark weight is w^b.

.1.4 Comparing network and great-circle intra- and inter-provincial distances

The analysis relies on the network distance between geocoded origins and destinations, which is the average of transaction-level intra- and inter-provincial distances. Traditionally, intra- and inter-provincial distances are measured using the origin-destination population-weighted great-circle distance (hereafter great-circle distance) between sub-provincial units (see, for example, Brown and Anderson, 2002). This is calculated for the set of sub-provincial units (census divisions) within each province for intra-provincial trade and between the sets of sub-provincial units for each province pair:

$$d_{op} = \sum_{i \in o} \sum_{j \in p} \left(\frac{pop_i pop_j}{\sum_{i' \in o} \sum_{j' \in p} pop_{i'} pop_{j'}} \right) d_{ij}, \tag{.14}$$

where o and p index provinces, i and j index census divisions, pop is the population of the census division and d is the great-circle distance between the centroids of census divisions.

For intra-provincial trade ($o = p$), within census division distance is the radius of a circle of an area equal to that of the census division: $d_{ii} = \sqrt{area_i/\pi}$.

It might reasonably be assumed that network distance is always longer than great-circle distance. However, because the actual (network) distance travelled is skewed towards short distance trips, when short distance trips are more prevalent (e.g., for intra-provincial trade or trade between contiguous provinces), the measured network distance may be shorter. That is, for the great-circle distance, holding population constant, the distance between nearer census division pairs is weighted the same as between the more distant census division pairs. The network distance estimates, because they are derived from actual trips, will weigh more highly closer census division pairs.

This pattern in the data is evident in Table .7, which presents the network and great-circle distance within and between provinces. On average, network distance is 33% greater than the great-circle distance. However, there is a tendency for intra-provincial distances and distances between contiguous provinces to be closer to (or even less than) the network distance. For intra-provincial, contiguous province, and non-contiguous provinces network distance is 9%, 25% and 38% greater than great-circle distance, respectively. The exceptions are the Atlantic provinces, which form a *de facto* archipelago whose internal network distances quite naturally outstrip great-circle distances by a wide margin (see Table .7).

There are two implications that follow from these distance patterns for the econometric analysis. First, because great-circle distance is less than network distance, the elasticity on distance will be less when network distance is used. Second, the relatively shorter intra-provincial great-circle distances will tend to inflate the intra-provincial trade coefficient (border effect), because the underestimated intra-provincial trade given the actual distance travelled. Both effects are seen in the estimates.

.2 Additional Robustness checks

.2.1 Testing for the differential effect on distance on intra- and inter-provincial trade

If intra-provincial trade is populated with a large set of logistics-truncated flows, the distance parameter on intra-provincial flows should be more negative than inter-provincial flows, whose pattern results from benchmarking to the flows from the provincial input-output accounts. To test for this effect, a modified version Equation (4.4) is estimated,

$$X_{ij} = \exp\left[\ln S_i^{intra} + \ln S_i^{inter} + \ln M_j^{intra} + \ln M_j^{inter} + (\beta + \theta_p)\ln d_{ij}\right]\epsilon_{ij}, \qquad (.15)$$

with the distance parameter permitted to vary across intra- and inter-provincial flows using an indicator variable for intra-provincial flows (θ_p).[18] If the truncation effect predominates, the distance parameter on intra-provincial trade should be more negative than inter-provincial trade. To isolate this effect, the model is estimated with separate origin and destination fixed effects for intra- and inter-provincial trade, where p indicates the set of intra-provincial regions. Intra-region flows are excluded.[19] When estimated for ERs, the distance parameter was -0.769 for inter-provincial trade, but significantly *less negative* for intra-provincial trade -0.579 ($\hat{\theta}_p = 0.190; P > |z| = 0.037$). Using CDs, a subunit of ERs, the estimate was also positive but insignificant ($\hat{\theta}_p = 0.058; P > |z| = 0.235$). To the extent that it is present, the truncation of intra-provincial flows does not appear to be sufficient to bias the estimates.

[18] If there are significant barriers to inter-provincial trade, the dampening effect of distance on trade would be expected to be less, as the lower level of competition would raise the cost cut-off (see Melitz and Ottaviano, 2008; Baldwin and Gu, 2009) at which firms would engage in trade across sub-provincial units. While this effect may be accounted for by multilateral resistance terms, the distance parameter may also be affected and, therefore, when both the effects of the transportation system and provincial barriers to trade are present, they will have confounding effects on the distance parameter.

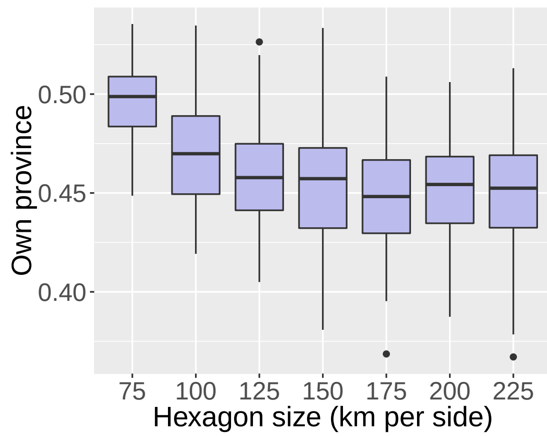
[19] These flows are excluded in order to have a comparable set of inter-regional flows. Trade between region i and j within the same province can be compared to trade between i and k across provinces.

.2.2 Estimates by year

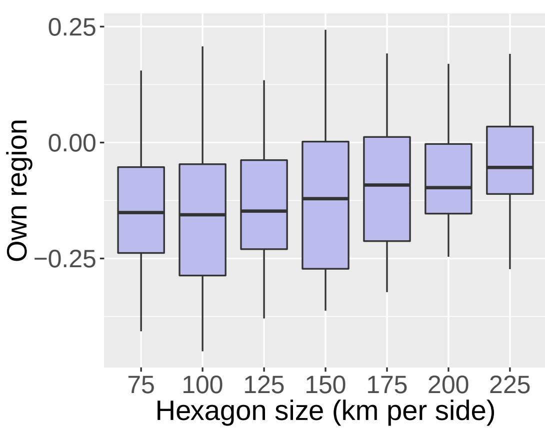
The estimates are presented for trade averaged across the 9-year study period stretching from 2004 to 2012. This is a long enough period to potentially observe changes in provincial border effects, such as from changes in policy or shifts in the macro-economy. To account for these effects, the baseline model was estimated with all of the variables interacted with time fixed effects, with 2004 being the excluded year. Whether the model is estimated using provinces, ERs, or CDs, as the trading units, there is no significant difference in the coefficients across years. Hence the average trade level-based estimates reported in the main body of the paper provide a reasonable picture of provincial border effects over the entire period.

.2.3 Differential border effect estimates for Quebec

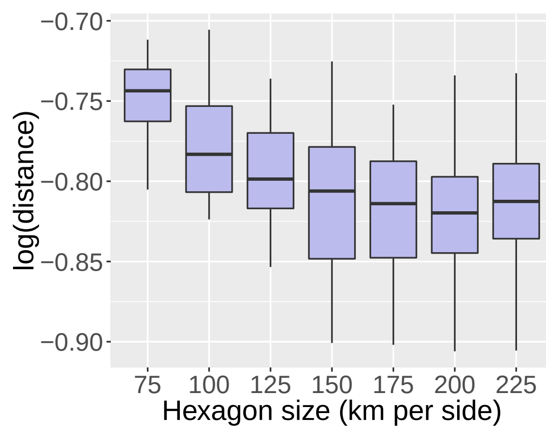
To test for the effect of Quebec on internal trade, own province is interacted with an indicator variable for internal Quebec trade flows. While the point estimate on the interaction term is positive, it is not significantly different than zero (see Table .9).



(a)



(b)



(c)

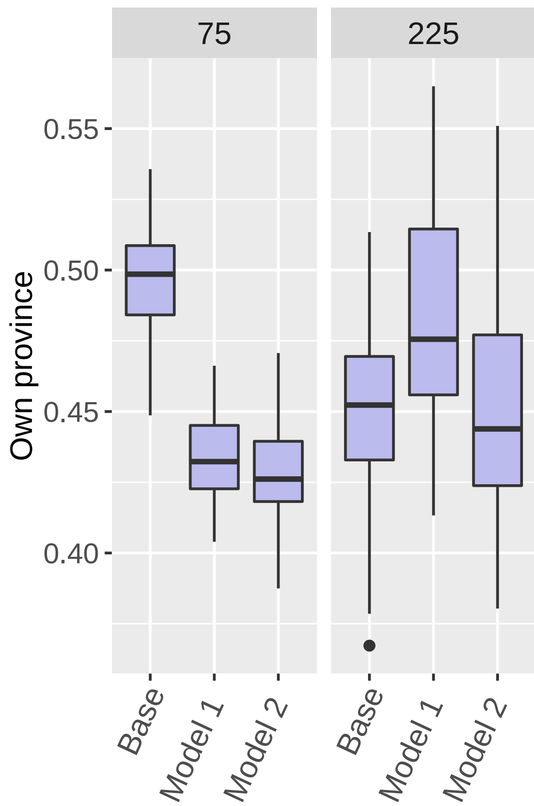
Figure 4.4: Coefficient estimates for Own Province (4.4a), Own Region (4.4b), and log($distance$) (4.4c), by size (km per side) and placement of hexagons.

Notes: The boxes represent the inter quartile range, with the line intersecting the box being the median coefficient value. The ends of the whiskers—the upper and lower adjacent values—represent the ranked coefficient value that is nearest to but not above (below) 1.5 times the inter-quartile range from above (below). The dots signify extreme values.

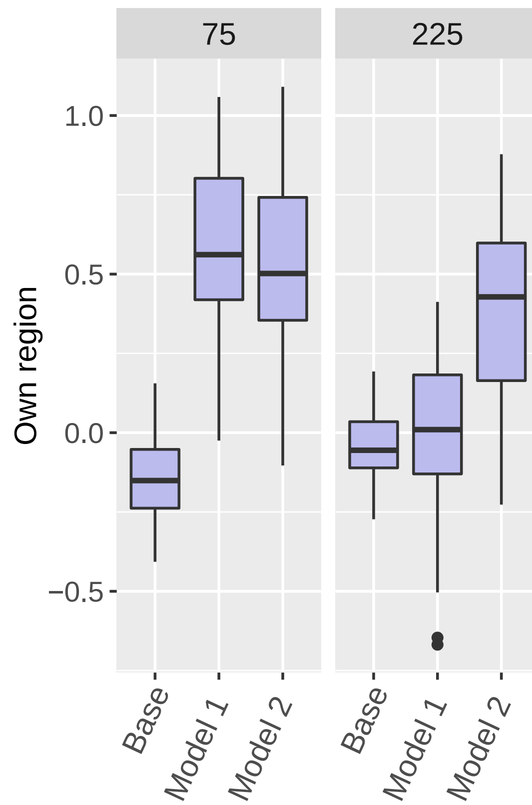
Table 4.5: Robustness of provincial border effect estimates to non-linear effects of distance and contiguity.

	Hexagons			
	225km per side		75km per side	
	Model 1	Model 2	Model 1	Model 2
Distance				
0 to 25km	-1.356^{***}	-1.338^{***}	-0.932^{***}	-0.923^{***}
	(0.284)	(0.281)	(0.122)	(0.122)
25 to 100km	-0.544	-0.561	-0.268	-0.276
	(0.471)	(0.469)	(0.227)	(0.227)
100 to 500km	-0.836^{***}	-0.720^{***}	-0.711^{***}	-0.801^{***}
	(0.119)	(0.120)	(0.0598)	(0.0915)
greater than 500km	-0.818^{***}	-0.772^{***}	-0.862^{***}	-0.858^{***}
	(0.0854)	(0.109)	(0.0684)	(0.0689)
Own region	-0.0608	0.233	0.312^{*}	0.179
	(0.173)	(0.237)	(0.161)	(0.199)
Own province	0.458^{***}	0.412^{***}	0.431^{***}	0.418^{***}
	(0.0839)	(0.0755)	(0.0713)	(0.0709)
Contiguous regions		0.195		-0.132
		(0.138)		(0.0972)
Constant	4.513^{***}	4.184^{***}	2.638^{***}	2.746^{***}
	(0.807)	(0.833)	(0.494)	(0.501)
Border effect	1.58	1.51	1.54	1.52
Observations	8,619	8,619	132,862	132,862

Notes: All models utilize a Poisson-PML estimator and include fixed effects for origins and destinations. ***, **, * indicate significance at the 0.01, 0.05 and 0.1 levels, respectively. Robust standard errors are presented in parentheses. For the 75km per side hexagons (Model 4), origins and destinations with very few flows were dropped in order to estimate the standard errors. The point estimates remain qualitatively unchanged compared to the full-sample results. The border effect is given by exp(own province).



(a)



(b)

Figure 4.5: Coefficient estimates for Own Province (4.5a) and Own Region (4.5b) by model, hexagon size (km per side) and placement.

Notes: The 'base' model estimates replicate those presented in Figure 4.4 (which use the specification presented in Table 4.4), while Model 1 and Model 2 match those in Table 4.5. The boxes represent the inter quartile range, with the line intersecting the box being the median coefficient value. The ends of the whiskers—the upper and lower adjacent values—represent the ranked coefficient value that is nearest to but not above (below) 1.5 times the inter-quartile range from above (below). The dots signify extreme values.

Table 4.6: Provincial border effect estimates for Forward Sortation Areas (FSA).

	OLS	Poisson				
	Model 1	Model 1		Model 2	Model 3	Model 4
Distance (log)	-0.500^{***}	0.00374	0.232^{***}			
	(0.0223)	(0.0536)	(0.0586)			
Distance (log)^2	0.0120^{***}	-0.0361^{***}	-0.0676^{***}			
	(0.00180)	(0.00459)	(0.00492)			
Distance						
0 to 25km				-0.0359		
				(0.0608)		
0 to 10km					0.323^{***}	
					(0.111)	
0 to 5km						0.613^{**}
						(0.253)
5 to 10km						-0.0219
						(0.362)
10 to 25km					-0.426^{***}	-0.341^{**}
					(0.122)	(0.138)
25 to 100km				-0.310^{***}	-0.213^{***}	-0.222^{***}
				(0.0697)	(0.0700)	(0.0697)
100 to 500km				-0.497^{***}	-0.507^{***}	-0.505^{***}
				(0.0381)	(0.0378)	(0.0379)
500km to ∞				-0.770^{***}	-0.767^{***}	-0.767^{***}
				(0.0245)	(0.0244)	(0.0244)
Own region	2.322^{***}	1.331^{***}	1.522^{***}	1.507^{***}	1.591^{***}	1.583^{***}
	(0.106)	(0.105)	(0.104)	(0.107)	(0.102)	(0.102)
Own province	1.202^{***}	0.468^{***}	0.624^{***}	0.601^{***}	0.601^{***}	0.602^{***}
	(0.0156)	(0.0372)	(0.0400)	(0.0407)	(0.0407)	(0.0407)
Constant	-3.140^{***}	1.634^{***}	1.166^{***}	1.405^{***}	0.807^{*}	0.514
	(0.218)	(0.406)	(0.418)	(0.425)	(0.449)	(0.481)
Observations	633,835	633,835	2,574,492	2,574,492	2,574,492	2,574,492
Border effect	3.33	1.60	1.87	1.82	1.82	1.83
Dist. elasticity, 832km	-0.42	-0.25	-0.23			
Includes zero flows	No	No	Yes	Yes	Yes	Yes

Notes: All models utilize a Poisson-PML estimator and include fixed effects for origins and destinations. ***, **, * indicate significance at the 0.01, 0.05 and 0.1 levels, respectively and are based on robust standard errors. The border effect is given by exp(own province).

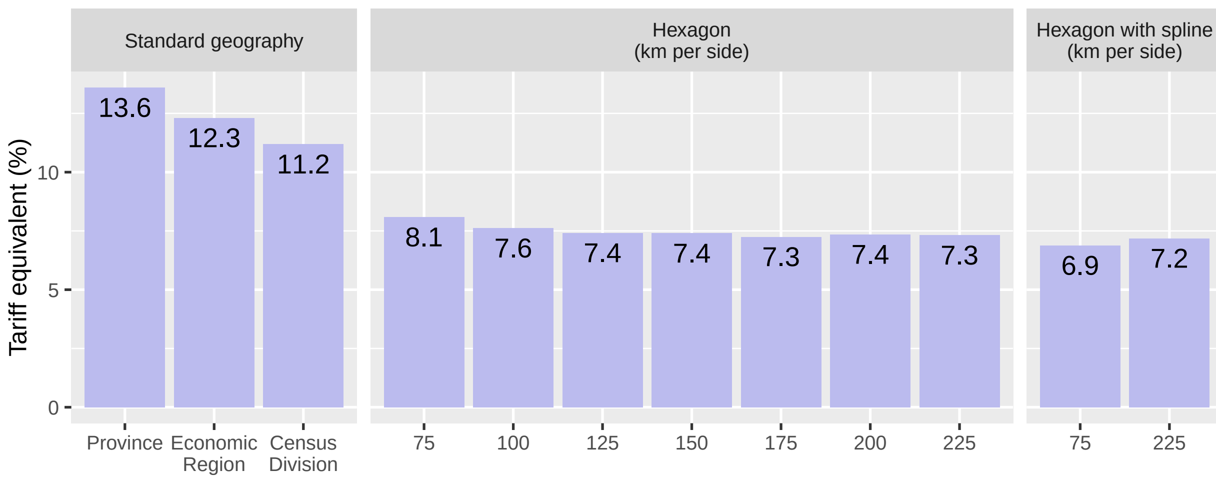


Figure 4.6: *Ad valorem* tariff equivalent by standard and hexagonal geographies.

Notes: All tariff equivalents are estimated using a price elasticity on transportation costs of −6.40. The standard geography *ad valorem* tariff equivalents are based on the provincial border effect estimates from Table 4.3 (Poisson estimate of the benchmarked flows using the network measure of distance) and Table 4.4. The hexagon-based tariff equivalents are based on the median point estimate from Figure 4.4, while the hexagon with spline-based tariff equivalents use the median point estimate from Figure 4.5 based on Model 2 from Table 4.5.

Table .7: Network and great-circle distance.

	N.L.	P.E.I.	N.S.	N.B.	Que.	Ont.	Man.	Sask.	Alta	B.C.
Panel A: Network distance (km)										
Newfoundland & Labrador	386	1,364	1,226	1,344	1,567	2,789	4,650	5,223	6,074	6,902
Prince Edward Island	1,412	61	333	272	1,115	1,706	3,584	4,209	4,810	5,696
Nova Scotia	1,326	324	136	389	1,173	1,815	3,616	4,308	4,977	5,802
New Brunswick	1,359	240	396	153	692	1,357	3,293	3,946	4,588	5,307
Quebec	1,478	1,095	1,222	728	280	584	2,459	3,114	3,734	4,607
Ontario	2,818	1,730	1,819	1,456	599	191	2,026	2,798	3,429	4,320
Manitoba	4,573	3,526	3,627	3,236	2,410	1,707	213	654	1,340	2,207
Saskatchewan	5,249	4,127	4,322	3,929	3,118	2,692	621	221	683	1,570
Alberta	5,806	4,907	4,908	4,578	3,720	3,248	1,316	660	219	905
British Columbia	6,873	5,750	5,872	5,476	4,640	4,283	2,244	1,631	1,010	204
Panel B: Great-circle distance (km)										
Newfoundland & Labrador	261	715	762	894	1,407	1,987	3,056	3,539	4,056	4,717
Prince Edward Island	715	39	193	211	756	1,322	2,547	3,071	3,621	4,274
Nova Scotia	762	193	143	290	805	1,344	2,635	3,167	3,723	4,374
New Brunswick	894	211	290	140	578	1,134	2,377	2,909	3,464	4,115
Quebec	1,407	756	805	578	208	615	1,893	2,442	3,010	3,648
Ontario	1,987	1,322	1,344	1,134	615	226	1,541	2,107	2,688	3,292
Manitoba	3,056	2,547	2,635	2,377	1,893	1,541	145	604	1,173	1,780
Saskatchewan	3,539	3,071	3,167	2,909	2,442	2,107	604	234	628	1,233
Alberta	4,056	3,621	3,723	3,464	3,010	2,688	1,173	628	221	709
British Columbia	4,717	4,274	4,374	4,115	3,648	3,292	1,780	1,233	709	213
Panel C: Difference between network to great-circle distance (percent)										
Newfoundland & Labrador	48	91	61	50	11	40	52	48	50	46
Prince Edward Island	98	59	72	29	47	29	41	37	33	33
Nova Scotia	74	67	-5	34	46	35	37	36	34	33
New Brunswick	52	14	37	9	20	20	39	36	32	29
Quebec	5	45	52	26	34	-5	30	28	24	26
Ontario	42	31	35	28	-3	-15	31	33	28	31
Manitoba	50	38	38	36	27	11	47	8	14	24
Saskatchewan	48	34	36	35	28	28	3	-6	9	27
Alberta	43	36	32	32	24	21	12	5	-1	28
British Columbia	46	35	34	33	27	30	26	32	43	-5

Table .8: Provincial border effect estimates based on flows between province, economic regions and census division allowing coefficients to vary across time (2004 to 2012), selected variables.

	Geography		
	Province	Economic Region	Census Division
Own province	0.756***	0.752***	0.747***
	(0.113)	(0.128)	(0.093)
2005	−0.021	0.0273	0.0716
	(0.149)	(0.194)	(0.150)
2006	−0.0449	0.0256	0.0595
	(0.143)	(0.180)	(0.134)
2007	0.0548	0.117	0.117
	(0.144)	(0.189)	(0.128)
2008	0.028	0.128	0.146
	(0.151)	(0.187)	(0.145)
2009	−0.0173	−0.031	−0.0495
	(0.158)	(0.173)	(0.131)
2010	0.0663	0.147	0.0619
	(0.180)	(0.196)	(0.134)
2011	0.0192	−0.0619	−0.142
	(0.174)	(0.190)	(0.132)
2012	0.263	−0.0598	−0.103
	(0.248)	(0.171)	(0.127)
N	900	47,961	713,480

Notes: All models utilize a Poisson-PML estimator and include distance, fixed effects for origins and destinations, own region (when applicable) own province, and year. All variables are interacted with the year fixed effects, with the excluded year being 2004. ***, **, * indicate significance at the 0.01, 0.05 and 0.1 levels, respectively and are based on robust standard errors. Own region refers to flows within the geograhic unit of anlaysis (Economic Region and Census Division).

Table .9: Test of the effect of Quebec on provincial border effects.

	Hexagons			
	225km per side		75km per side	
	Model 1	Model 2	Model 3	Model 4
Distance				
0 to 25km	-1.329^{***}	-1.315^{***}	-0.939^{***}	-0.931^{***}
	(0.282)	(0.280)	(0.122)	(0.122)
25 to 100km	-0.596	-0.605	-0.265	-0.273
	(0.468)	(0.466)	(0.225)	(0.225)
100 to 500km	-0.833^{***}	-0.723^{***}	-0.707^{***}	-0.803^{***}
	(0.120)	(0.121)	(0.0602)	(0.0916)
greater than 500km	-0.835^{***}	-0.788^{***}	-0.880^{***}	-0.877^{***}
	(0.0795)	(0.103)	(0.0657)	(0.0661)
Own region	-0.0718	0.209	0.316^{**}	0.176
	(0.176)	(0.237)	(0.160)	(0.198)
Own province	0.396^{***}	0.361^{***}	0.362^{***}	0.346^{***}
	(0.0947)	(0.0919)	(0.0832)	(0.0844)
Quebec $\times$ Own province	0.195	0.169	0.216	0.224
	(0.222)	(0.221)	(0.208)	(0.209)
Contiguous regions		0.185		-0.139
		(0.137)		(0.0973)
Constant	4.543^{***}	4.224^{***}	2.702^{***}	2.819^{***}
	(0.812)	(0.835)	(0.494)	(0.503)
Observations	8,619	8,619	132,862	132,862

Notes: All models use a Poisson-PML estimator and include fixed effects for origins and destinations. ***, **, * indicate significance at the 0.01, 0.05 and 0.1 levels, respectively and are based on robust standard errors. The border effect is given by exp(own province).

Bibliography

Aban, I. B., M. M. Meerschaert, and A. K. Panorska (2006). Parameter estimation for the truncated Pareto distribution. *Journal of the American Statistical Association 101*(473), 270–277.

Acemoglu, D., U. Akcigit, and W. Kerr (2015). Networks and the Macroeconomy: An Empirical Exploration. In *NBER Macroeconomics Annual 2015*, Volume 30. University of Chicago Press.

Acemoglu, D., V. Carvalho, A. Ozdaglar, and A. Tahbaz-Salehi (2012). The Network Origins of Aggregate Fluctuations. *Econometrica 80*(5), 1977–2016.

Acemoglu, D., A. Ozdaglar, and A. Tahbaz-Salehi (2015). Systemic Risk and Stability in Financial Networks. *The American Economic Review 105*(2), 564—-608.

Agnosteva, D. E., J. E. Anderson, and Y. V. Yotov (2014). Intra-national trade costs: Measurement and aggregation. *NBER Working Paper Series, no 19872*.

Albrecht, L. and T. Tombe (2016). Internal trade, productivity, and interconnected industries: A quantitative analysis. *Canadian Journal of Economics 49*(1).

Amrhein, C. G. (1995). Searching for the elusive aggregation effect: Evidence from statistical simulation. *Environment and Planning A 27*(1), 105–119.

Amrhein, C. G. and R. Flowerdew (1992). The effect of data aggregation on a Poisson regression model of Canadian migration. *Environment and Planning A 24*(10), 1381–1391.

Anderson, J. E. (2011). The gravity model. *Annual Review of Economics 3*(1), 133–160.

Anderson, J. E. and E. van Wincoop (2003). Gravity with gravitas: A solution to the border puzzle. *American Economic Review 93*(1), 170–192.

Anderson, J. E. and Y. V. Yotov (2010). The changing incidence of geography. *American Economic Review 100*(5), 2157–2186.

Anderson, J. E. and Y. V. Yotov (2012). Gold standard gravity. *NBER Working Paper Series, no 17835*.

Anthonisen, N. (2016). Microeconomic shocks and macroeconomic fluctuations in a dynamic network economy. *Journal of Macroeconomics 47*(January), 233–254.

Arbia, G. (1989). *Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems*. Dordrecht: Kluwer Academic Publishers.

Atalay, E., A. Hortaçsu, J. Roberts, and C. Syverson (2011). Network structure of production. *Proceedings of the National Academy of Sciences of the United States of America 108*(13), 5199–5202.

Atalay, E., A. Hortaçsu, and C. Syverson (2014). Vertical Integration and Input Flows. *American Economic Review 104*(4), 1120–1148.

Baldwin, J. R. and W. Gu (2009). The impact of trade on plant scale, production-run length and diversification. In T. Dunne, J. B. Jensen, and M. Roberts (Eds.), *Producer Dynamics: New Evidence from Micro Data, Studies in Income and Wealth 68*. Chicago: University of Chicago Press.

Barrot, J.-N. and J. Sauvagnat (2016). Input Specificity and the Propagation of Idiosyncratic Shocks in Production Networks. *The Quarterly Journal of Economics 131*(3), 1543–1592.

Behrens, K., T. Bougna, and W. M. Brown (2015). The world is not yet flat: Transportation costs matter! *CEPR Working Paper Series, no. 10356*.

Bernard, A. B., A. Moxnes, and Y. U. Saito (2015). Production Networks, Geography and Firm Performance.

Briant, A., P.-P. Combes, and M. Lafourcade (2010). Dots to boxes: Do the size and shape of spatial units jeopardize economic geography estimations? *Journal of Urban Economics 67*(3), 287–302.

Broido, A. D. and A. Clauset (2018). Scale-free networks are rare. pp. 26–28.

Brown, W. M. (2003). Overcoming distance, overcoming borders: Comparing North American regional trade. *Statistics Canada Working Paper, no. 008*.

Brown, W. M. (2015). How much thicker is the Canada-U.S. border? the cost of crossing the border by truck in the pre- and post-9/11 eras. *Research in Transportation Business and Management*.

Brown, W. M. and W. P. Anderson (2002). Spatial markets and the potential for economic integration between Canadian and U.S. regions. *Papers in Regional Science* (81), 99–120.

Budd, J. W., J. Konings, and M. J. Slaughter (2005). Wages and International Rent Sharing in Multinational Firms. *Review of Economics & Statistics 87*(1), 73–84.

Burstein, A., C. Kurz, and L. Tesar (2008). Trade, production sharing, and the international transmission of business cycles. *Journal of Monetary Economics 55*(4), 775–795.

Carvalho, V., N. Makoto, and Y. U. Saito (2014). Supply Chain Disruptions: Evidence from the Great East Japan Earthquake. *RIETI Discussion Paper Series* (14-E-035).

Carvalho, V. and N. Voigtländer (2014). Input Diffusion and the Evolution of Production Networks.

Clauset, A., C. Shalizi, and M. Newman (2009). Power-Law distributions in empirical data. *SIAM Review 51*(4), 661–703.

Coughlin, C. C. and D. Novy (2016). Estimating border effects: The impact of spatial aggregation. Technical report. CEPR Discussion Paper 11226, and FRB St. Louis Working Paper 006A.

Crafts, N. and A. Klein (2015). Geography and intra-national home bias: U.S. domestic trade in 1949 and 2007. *Journal of Economic Geography 15*(3), 477–497.

Cravino, J. and A. A. Levchenko (2016). Multinational Firms and Business Cycle Transmission. *The Quarterly Journal of Economics 132*(2), 921–962.

Davis, S. J. and J. Haltiwanger (1992). Gross Job Creation, Gross Job Destruction, and Employment Reallocation. *The Quarterly Journal of Economics 107*(3), 819–863.

Di Giovanni, J. and A. Levchenko (2012). Country size, international trade, and aggregate fluctuations in granular economies. *Journal of Political Economy 120*(6), 1083–1132.

Di Giovanni, J. and A. Levchenko (2013). Firm Entry, Trade, and Welfare in Zipf's World. *Journal of International Economics 89*(2), 283–296.

Di Giovanni, J., A. Levchenko, and I. Méjean (2014). Firms, Destinations, and Aggregate Fluctuations. *Econometrica 82*(4), 1303–1340.

Di Giovanni, J., A. Levchenko, and R. Rancière (2011). Power Laws in Firm Size and Openness to Trade: Measurement and Implications. *Journal of International Economics 85*(1), 42–52.

di Giovanni, J., A. A. Levchenko, and I. Mejean (2016). The Micro Origins of International Business Cycle Comovement. *NBER Working Paper* (21885).

di Giovanni, J., A. A. Levchenko, and I. Mejean (2017). Large Firms and International Business Cycle Comovement. *American Economic Review: Papers and Proceedings*, Forthcoming.

Dupor, B. (1999). Aggregation and irrelevance in multi-sector models. *Journal of Monetary Economics 43*(2), 391–409.

Eeckhout, J. (2004). Gibrat's Law for (All) Cities. *American Economic Review 94*(5), 1429–1451.

Elliott, M., B. Golub, and M. O. Jackson (2014). *Financial Networks and Contagion*, Volume 104.

Fally, T. (2015). Structural gravity and fixed effects. *Journal of International Economics 97*(1), 76–85.

Feenstra, R. C. (2018). Restoring the product variety and pro-competitive gains from trade with heterogeneous firms and bounded productivity. *Journal of International Economics 110*, 16–27.

Foerster, A. T., P.-D. G. Sarte, and M. W. Watson (2011). Sectoral versus Aggregate Shocks: A Structural Factor Analysis of Industrial Production. *Journal of Political Economy 119*(1).

Foster, L., J. Haltiwanger, and C. Syverson (2008). Reallocation, Firm Turnover, and Efficiency: Selection on Productivity or Profitability? *American Economic Review 98*(1), 394–425.

Fotheringham, A. S. and M. E. O'Kelly (1989). *Spatial Interaction Models: Formulations and Applications*. Dordrecht: Kluwer Academic Publishers.

Fotheringham, A. S. and D. W. Wong (1991). The Modifiable Areal Unit Problem in multivariate statistical analysis. *Environment and Planning A* (23), 1025–1044.

Frankel, J. A. and A. K. Rose (1998). The endogenity of the optimum currency area criteria. *The Economic Journal 108*(449), 1009–1025.

Gabaix, X. (1999). Zipf' s Law for Cities: An Explanation. *The Quarterly Journal of Economics 114*(3), 739–767.

Gabaix, X. (2009). Power Laws in Economics and Finance. *Annual Review of Economics 1*(1), 255–294.

Gabaix, X. (2011). The Granular Origins of Aggregate Fluctuations. *Econometrica 79*(3), 733–772.

Gabaix, X. and R. Ibragimov (2011). Rank 1/2: A Simple Way to Improve the OLS Estimation of Tail Exponents. *Journal of Business & Economic Statistics 29*(1), 24–39.

Gandhi, A., S. Navarro, and D. Rivers (2013). On the Identification of Production Functions: How Heterogeneous is Productivity?

Giroud, X. and H. M. Mueller (2016). Redistribution of Local Labor Market Shocks Through Firms' Internal Networks. *NBER Working Paper Series.* (July).

Head, K. and T. Mayer (2009). Illusory border effects: Distance mismeasurement inflates estimates of home bias in trade. In P. van Bergeijk and S. Brakma (Eds.), *The Gravity Model in International Trade: Advances and Applications*, pp. 165–192. Cambridge: Cambridge University Press.

Head, K. and T. Mayer (2014). Gravity equations: Workhorse, toolkit, and cookbook. In G. Gopinath, E. Helpman, and K. Rogoff (Eds.), *Handbook of International Economics vol. 4*, pp. 131–195. Cambridge: Cambridge University Press.

Head, K., T. Mayer, and M. Thoenig (2014). Welfare and Trade without Pareto. *American Economic Review: Papers and Proceedings 104*(5), 310–316.

Hillberry, R. and D. Hummels (2003). Intranational home bias: some explanations. *Review of Economics and Statistics 85*(4), 1089–1092.

Hillberry, R. and D. Hummels (2008). Trade responses to geographic frictions: A decomposition using micro-data. *European Economic Review 52*, 527–550.

Holmes, T. J. and J. J. Stevens (2014). An Alternative Theory of the Plant Size Distribution, with Geography and Intra- and International Trade. *Journal of Political Economy 122*(2).

Horvath, M. (1998). Cyclicality and Sectoral Linkages: Aggregate Fluctuations from Independent Sectoral Shocks. *Review of Economic Dynamics 1*(4), 781–808.

Horvath, M. (2000). Sectoral shocks and aggregate fluctuations. *Journal of Monetary Economics 45*(1), 69–106.

Hottman, C., S. J. Redding, and D. E. Weinstein (2016). Quantifying the Sources of Firm Heterogeneity. *The Quarterly Journal of Economics*.

Hulten, C. (1978). Growth Accounting with Intermediate Inputs. *The Review of Economic Studies 45*(3), 511–518.

Jackson, M. O. (2010). *Social and Economic Networks*. Princeton University Press.

Jessen, A. H. and T. Mikosch (2006). Regularly Varying Functions. *Publications de L'Institut Mathematique 80*(94), 171–192.

Johnson, R. C. (2014). Trade in intermediate inputs and business cycle comovement. *American Economic Journal: Macroeconomics 6*(4), 39–83.

Jones, C. (2011). Intermediate Goods and Weak Links in the Theory of Economic Development. *American Economic Journal: Macroeconomics 3*(2), 1–28.

Kleinert, J., J. Martin, and F. Toubal (2015). The few leading the many: Foreign affiliates and business cycle comovement. *American Economic Journal: Macroeconomics 7*(4), 134–159.

Kose, M. A. and K.-M. Yi (2006). Can the standard international business cycle model explain the relation between trade and comovement? *Journal of International Economics 68*(2), 267–295.

Lamont, O. (1997). Cash Flow and Investment: Evidence from Internal Capital Markets. *The Journal of Finance 52*(1), 83–109.

Lim, K. (2016). Firm-to-firm Trade in Sticky Production Networks.

Long, J. B. and C. I. Plosser (1983). Real Business Cycles. *Journal of Political Economy 91*(1), 39—-69.

Luttmer, E. G. J. (2007). Selection, Growth, and the Size Distribution of Firms. *Quarterly Journal of Economics* (August), 1103–1144.

Manning, W. and J. Mullahy (2001). Estimating log models: To transform of not to transform? *Journal of Health Economics 20*(4), 461–494.

McCallum, J. (1995). National borders matter: Canada-U.S. regional trade patterns. *American Economic Review 85*(3), 615–623.

Melitz, M. J. (2003). The impact of trade on intra-industry reallocations and aggregate industry productivity. *Econometrica 71*(6), 1695—-1725.

Melitz, M. J. and G. I. P. Ottaviano (2008). Market size, trade and productivity. *Review of Economic Studies 75*, 295–316.

Millimet, D. L. and T. Osang (2007). Do state borders matter for U.S. intranational trade? the role of history and migration. *Canadian Journal of Economics 40*(1), 93–126.

Nigai, S. K. (2017). A Tale of Two Tails: Productivity Distribution and the Gains from Trade. *Journal of International Economics 104*, 44–62.

Oberfield, E. (2017). A Theory of Input-Output Architecture. *Econometrica* (Forthcoming).

Páez, A. and D. M. Scott (2004). Spatial statistics for urban analysis: a review of techniques with examples. *GeoJournal 61*(1), 53–67.

Rozenfeld, H. D., D. Rybski, X. Gabaix, and H. A. Makse (2011). The Area and Population of Cities: New Insights from a Different Perspective on Cities. *The American Economic Review 101*(5), 2205–2225.

Santos Silva, J. and S. Tenreyro (2006). The log of gravity. *Review of Economics and Statistics 88*(4), 641–658.

Sen, A. and T. Smith (1995). *Gravity Models of Spatial Interaction Behaviour*. Berlin: Springer.

Shea, J. (2002). Complementarities and comovements. *Journal of Money, Credit, and Banking 34*(2), 412–433.

Stein, J. C. (1997). Internal Capital Markets and the Competition for Corporate Resources. *The Journal of Finance 52*(1), 111.

Wolf, H. C. (2000). Intranational home bias in trade. *Review of Economics and Statistics 82*(4), 555–563.

WRDS (2018). Wharton Research Data Services.