University of California
Los Angeles

Firm Heterogeneity in Macroeconomics

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Economics

by

Allen Tran

2014

Abstract of the Dissertation

Firm Heterogeneity in Macroeconomics

by

Allen Tran

Doctor of Philosophy in Economics

University of California, Los Angeles, 2014

Professor Hugo Hopenhayn, Chair

Macroeconomic models are often estimated with aggregate data, aligning the aggregated behavior of firms and households in models to the data. However, using aggregate data alone can overlook important details of firm behavior that are crucial for understanding issues in macroeconomics. In this dissertation, I use data on firms at the micro-level to more accurately capture firms behavior and their interactions with one another. This approach is applied to answer questions that relate to the monetary policy transmission mechanism, economic growth from new entrants and welfare gains from new technology.

A substantial literature exists which suggests that imperfect information across firms is capable of generating large monetary non-neutralities. In Chapter One, the level of imperfect information is taken from micro-data and used to discipline a standard menu cost model augmented with information frictions. In the model, imperfect information has a negligible effect and real responses to a monetary shock are small and transient in contrast to the bulk of the imperfect information literature. The selection effect dominates the effects of imperfect information as the level of dispersion in inflation expectations in the data is tiny. This result still holds even when the level of dispersion is set to that of the maximal observed levels of dispersion.

Chapter Two presents data that suggests new entering establishments compete for customers, rather than inputs in order to grow. Consistent with the data, I present a model where customers satisfice in forming relationships with establishments in the presence of search frictions. The extent of these search frictions is a new margin that affects selection and allocative efficiency. As search becomes less random and more directed, customers are less willing to satisfice, improving allocative efficiency and inducing exit of slower growing firms. When search frictions in product markets are increased to match establishment dynamics in Chile, output falls by roughly 14 per cent relative to the model calibrated to the US, reflecting decreased allocative efficiency.

Chapter Three studies the impact of online retail on aggregate welfare. I develop a new measure of store level retail productivity and with a spatial model, calculate each store's equilibrium response to increased competitive pressure from online retailers. From counterfactual exercises mimicking improvements in shipping and increased internet access, I estimate that improvements in online retail increased aggregate welfare from retail activities by 13.4 per cent. Roughly two-thirds of the increase can be attributed to welfare improvements holding fixed market shares, with the remainder due to reallocation. Surprisingly, 8.2 percent of firms actually benefit as they absorb market share from closed stores. Finally, I estimate that the proposed Marketplace Fairness Act would claw back roughly one-third of sales that would otherwise have gone to online retailers between 2007-12.

The dissertation of Allen Tran is approved.

Andrew Atkeson

Ariel Burstein

Pablo Fajgelbaum

Hanno Lustig

Hugo Hopenhayn, Committee Chair

University of California, Los Angeles

2014

Table of Contents

1 Dispersion in Beliefs and Price Setting 1

1.1 Introduction 1

1.2 Model 4

1.2.1 Consumers 5

1.2.2 Monetary policy 6

1.2.3 Firms 7

1.2.4 Market clearing 9

1.2.5 Solution algorithm 10

1.2.6 Parameter selection 11

1.3 Results 17

1.3.1 Discussion 22

1.4 Conclusion 24

2 Product Market Frictions, Selection and Allocative Efficiency 26

2.1 Introduction 26

2.1.1 Related literature 32

2.2 Data 34

2.2.1 Rivalrous growth 37

2.2.2 Wages do not drive selection 42

2.2.3 Evolution of cumulative growth 44

2.3 Model 46

2.4 Estimation 56

2.5 Results 63
2.6 Conclusion 68
2.7 Appendix 70
2.7.1 Definition of equilibrium 71
2.7.2 Solution algorithm 71

3 The Aggregate Impact of Online Retail 72
3.1 Introduction 72
3.1.1 Related literature 75
3.2 Data 77
3.3 Model 78
3.3.1 Consumers 80
3.3.2 Retail stores 83
3.3.3 Aggregate welfare 86
3.3.4 Equilibrium 87
3.4 Estimation 88
3.4.1 Demand system 88
3.4.2 Fixed costs 98
3.5 Results 103
3.5.1 Improvements in online retail 104
3.5.2 Effects of the Marketplace Fairness Act (2013) 121
3.6 Conclusion 123
3.7 Appendix 127
3.7.1 Estimates of internet access by zip code 127
3.7.2 Proof of uniqueness and existence 129

3.7.3 BLP algorithm with endogenous prices 131
3.7.4 Entry and exit algorithm 133
3.7.5 Results at bounds of fixed costs 135

List of Figures

1.1 Dispersion in inflation expectations 14

1.2 Impulse response of the aggregate price to a monetary shock . . . 19

1.3 Impulse response of output to a monetary shock 20

2.1 Exit rates and spread in 80th-50th percentile cumulative growth (log units) 39

2.2 Evolution of cumulative growth in USA 45

2.3 Evolution of cumulative growth in Chile 46

2.4 Spread of growth estimated for US (top) and Chile (bottom) . . . 64

2.5 Effect of θ on moments relative to values at estimated parameters 65

2.6 Effect of θ on model relative to values at estimated parameters . . 66

3.1 Growth of online retail sales 73

3.2 Estimated improvements in internet access across zip codes (2007 to 2012) 106

3.3 Shipping and fulfillment expansion at Amazon.com 108

3.4 Productivity improvements at FedEx and UPS 109

3.5 Online market shares by industry 111

3.6 E-commerce sales growth across products (2007 to 2011) 112

List of Tables

1.1 Calibrated Parameters 11

1.2 Estimated Parameters 16

1.3 Moments from Data and Model 16

1.4 Law of Motion for Inflation 17

1.5 Variance decomposition 20

1.6 Signal to noise ratio 22

2.1 Fixed effects regression 40

2.2 Probit regression on exit 42

2.3 Regression for average wages (log units) 43

2.4 Entrant size distribution 45

2.5 Moments from the data and the model 62

2.6 Estimated and calibrated parameter values 63

3.1 List of moments 92

3.2 Summary of parameter estimates and standard errors 94

3.3 Relative demand to a store 0 miles away 95

3.4 Implied dollar costs of traveling to stores 95

3.5 Variation in prices across markets 96

3.6 Markups and demographics 97

3.7 Mean store-level surplus and demographics 98

3.8 Estimates of fixed costs 103

3.9 Results from counterfactual exercises 115

3.10 Regression on store sales 116

3.11 Welfare decomposition 119
3.12 Decomposing aggregate profits 120
3.13 Distribution of profit changes at surviving firms 121
3.14 Average state tax rates inclusive of city and county rates 122
3.15 Estimated potential effects of Marketplace Fairness Act (2007-12) 123
3.16 Internet access and demographics 127
3.17 Results from counterfactual exercises 136

Acknowledgments

I wish to thank my dissertation committee and the faculty at UCLA for help and guidance throughout my graduate studies. In particular, Hugo Hopenhayn for expanding my view of macroeconomics and inspiring my fascination with data, Pablo Fajgelbaum for honest criticism that made my research more rigorous and ambitious, and Andy Atkeson for insight that often improved my understanding of my own research. For her help at the California Census RDC, I thank Abigail Cooke.

I am grateful to Emily, her family and Yuki for making Los Angeles a second home, and my teammates at Wikipedia Brown for providing a constant distraction from research.

My deepest gratitude is to my family for their love and support. Kelvin and Tina, for demonstrating that the world was larger than Lalor, and my mother and father, who sacrificed everything for a better life for their children.

Any opinions and conclusions expressed herein are those of the author(s) and do not necessarily represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed.

Vita

2002–2007	Bachelor of Commerce, Bachelor of Information Systems, University of Melbourne.
2008–2009	Research Economist, Reserve Bank of Australia.
2012	Masters of Arts (Economics), UCLA, Los Angeles, California.

Publications

Reconciling Microeconomic and Macroeconomic Estimates of Price Stickiness (with Adam Cagliarini and Tim Robinson), *Journal of Macroeconomics*, 2011, 33(1), pages 102-120.

CHAPTER 1

Dispersion in Beliefs and Price Setting

1.1 Introduction

There exists a wide literature on models where nominal rigidities are the key friction in generating monetary non-neutralities. Using a menu cost model capable of fitting the micro-data on price setting, Golosov and Lucas (2007) show that the mechanism by which nominal rigidities are implemented is important. Real responses to monetary shocks are small and transient in the menu cost model compared to the large and persistent responses in an equivalent Calvo-based pricing model. The intuition is that in the menu cost model, the "selection effect" is present (Caplin and Spulber, 1987). Although few firms change prices in any given period, those firms that do change price are the firms, the "selected", whose price is currently furthest from the optimal price and therefore price changes are large, nullifying real responses. This effect is absent in the literature that focuses on time dependent pricing models where the selection is random.

Although no model of price setting is completely consistent with the stylized facts on prices at the micro level, Klenow and Kryvtsov (2008) show that the menu cost model in Golosov and Lucas (2007) comes closest in comparison to a selection of well known models of price setting. It fails to produce many small price changes as in the data but unlike time dependent models, it is consistent with higher repricing rates in higher inflation environments and the fact that the size of each price change is uncorrelated with the duration it remained fixed. Therefore

it seems reasonable that in any model where the key friction is nominal, menu costs should be used to implement the nominal friction.

The literature on price setting under imperfect information has thus far focused on the mechanisms that lead to a lack of perfect information across agents. These mechanisms are important for providing microfoundations but they are typically embedded in environments where the nominal friction is either not subject to the selection effect, or not consistent with the micro data on prices. For instance, in both Woodford (2001) and Mackowiak and Wiederholt (2009), firms have imperfect information for different underlying reasons but prices are otherwise flexible. This implies prices adjust every period which is in stark contrast to prices at the micro-level, which are fixed for various lengths of time. In Mankiw and Reis (2002), firms are randomly selected to receive updated information. Although prices may remain fixed at the micro-level, since firms are *randomly* selected, the selection effect is not present.

Another consequence of the focus on the mechanisms underlying imperfect information is that the mechanisms are calibrated to demonstrate that they are capable of generating monetary non neutralities. Many of the mechanisms in the literature are difficult to map to the data. For instance, in the rational inattention model, the tightness of the information flow constraint is governed by a parameter. Similarly in Angeletos and La'O (2009), firms receive signals of the aggregate state which have some variance. Without any obvious way to map the variance of signals to data, the variance of the signal is set to the variance of the aggregate state.

Broadly speaking, three factors are important in models of price setting with imperfect information. The dispersion of beliefs across agents at each point in time, the nominal friction (or lack thereof) that prevents frequent price adjustments and the existence of complementarities. Regardless of the exact mechanism used to generate the information friction, what matters is the dispersion of beliefs across agents. Whether agents have imperfect information because they cannot

process all the available information or because it is costly to acquire information is redundant if they lead to the same cross section dispersion in beliefs across agents.[1]

This provides a simple way to indirectly calibrate these mechanisms. Each of the models with imperfect information imply a dispersion of beliefs across agents. In addition, there exists data, albeit survey data, on the dispersion of expectations across agents. Therefore it is simple to calibrate the mechanisms in these models so that they are consistent with the data.

This paper answers a question that is largely quantitative in nature. Can a model of imperfect information subject to the selection effect and consistent with the micro data on both prices and dispersion in beliefs across agents generate sizeable monetary non-neutralities? To this end, I construct a menu cost model that is capable of generating dispersion in beliefs across agents and calibrate it to fit the data. The calibrated model exhibits real responses that are very similar to those in Golosov and Lucas (2007), that is, small and transient. Moreover, even extreme levels of imperfect information lead to similar real responses as the selection effect dominates. Although the firms that change prices are those likely to receive extreme signals of the aggregate state, the response of prices at the micro level are large, and correct on average which nullifies real responses as in the full information benchmark. This result is robust to six times as much dispersion in beliefs across agents than evident in the data.

The next section presents the model. Section 3 presents the calibration and details on the identification of the parameter which governs the extent of imperfect information. Results are presented in Section 4 and I conclude in Section 5.

[1] Of course, not all dispersion is the same. Two models could have the same variance of beliefs but the distribution of beliefs may differ (say in skewness or kurtosis) such that outcomes are vastly different. Here I focus on beliefs normally distributed around the mean.

1.2 Model

The model is essentially an imperfect information version of the single sector model described in Nakamura and Steinsson (2010). In the literature discussed in the introduction, the models are typically either linearized or a quadratic loss function is taken as the firm's objective. The model used in this paper is not linearized and imperfect information is embedded into a model with aggregate uncertainty where the distribution of prices is a state variable.

Linearization aside, the approach here is actually quite similar to the method of underdetermined coefficients used in most of the literature where a parameterized guess for the equilibrium price is made and then used to solve the model in closed form. The coefficients are then recovered from the solution. Here I use a similar method based on Krusell and Smith (1998) except that I need to solve the model numerically and recover the coefficients via regression from simulations of the model.

The exact form of imperfect information used in the model is simpler than most models that incorporate imperfect information since the focus is on calibrating the model to fit the micro-data. Firms receive a noisy signal of the current aggregate state but essentially have perfect information about the past. Any other mechanism could have been used to generate imperfect information since the dispersion of beliefs is what matters but it is much simpler to calibrate a model where there is a single parameter, the variance of the signals, that governs the extent of imperfect information.

1.2.1 Consumers

There exists a representative consumer who has access to a complete set of Arrow-Debreu securities and has preferences over a final consumption good and leisure

$$\sum_{t=0}^{\infty} \beta^t \sum_{h_t} \left[\frac{1}{1-\gamma} C_t(h_t)^{1-\gamma} - \frac{\omega}{1+\psi} L_t(h_t)^{1+\psi} \right] \pi_t(h_t) \tag{1.1}$$

subject to the budget constraint at each history h_t,

$$P_t(h_t)C_t(h_t) + \sum_{h_{t+1}|h_t} Q_t(h_{t+1})B_{t+1}(h_{t+1}) = W_t(h_t)L_t(h_t) + B_t(h_t) + \int_0^1 \Pi_t(z, h_t)dz \tag{1.2}$$

where C_t is a consumption aggregate, L_t is labor supply, P_t is a price index, $Q_t(h_{t+1})$ is the price of a bond that pays off one dollar in history h_{t+1}, $B_t(h_t)$ are holdings of securities that payoff in h_t, W_t is the nominal wage and $\Pi_t(x, h_t)$ are profits from firm z in state h_t.

The consumption aggregate is formed in the standard Dixit-Stiglitz manner,

$$C_t = \left[\int_0^1 c_t(z)^{\frac{\epsilon-1}{\epsilon}} dz \right]^{\frac{\epsilon}{\epsilon-1}}$$

where $c_t(z)$ is a differentiated consumption good produced by a single firm. The associated price index is

$$P_t = \left[\int_0^1 p_t(z)^{1-\epsilon} dz \right]^{\frac{1}{1-\epsilon}} \tag{1.3}$$

From cost minimization, their demand of a particular differentiated good is

$$c_t(z) = \left(\frac{p_t(z)}{P_t} \right)^{-\epsilon} C_t \tag{1.4}$$

From the consumer's optimality conditions, the price of a k period risk-free bond

is therefore[2] where μ is the steady state rate of inflation and σ^2 is the variance of the shock to money supply.

$$Q_{t,t+k} = \beta^k \sum_{h_{t+1}|h_t} \left(\frac{c_{t+1}(h_{t+1})}{c_t(h_t)} \right)^{-\gamma} \frac{p_t(h_t)}{p_{t+1}(h_{t+1})} \pi_{t+1}(h_{t+1}|h_t) \tag{1.5}$$

and labor supply is governed by

$$\frac{W_t(h_t)}{P_t(h_t)} = \omega L_t(h_t)^{\psi} C_t(h_t)^{\gamma} \tag{1.6}$$

1.2.2 Monetary policy

For simplicity, I assume the monetary authority adjusts the money supply to target a level of nominal aggregate demand such that

$$P_t C_t = M_t \tag{1.7}$$

where $\log M_t$ follows a random walk with drift

$$\log M_t = \mu + \log M_{t-1} + e_t \tag{1.8}$$

Note that with flexible prices, the aggregate price would move one-for-one with the money supply and the consumption aggregate would remain constant in every period. With frictions in price setting, the aggregate price does not fully adjust and therefore equation (1.7) can only hold with equality if there are offsetting movements in the consumption aggregate. Therefore equation (1.7) describes the relationship between nominal frictions and real movements.

[2] Using equilibrium conditions and the calibration, the one period price is

$$Q_t = \beta e^{-\mu} \mathbb{E}\left[e^{-\epsilon_{t+1}}\right] = \beta e^{-\mu + \frac{1}{2}\sigma_m^2}$$

1.2.3 Firms

There exists a continuum of firms on the unit interval, indexed by z, who are monopolistically competitive and produce a differentiated good with labor with the following production function

$$y_t(z) = e^z l_t(z) \tag{1.9}$$

where z is a idiosyncratic technology shock that follows an $AR(1)$. The profit function in units of consumption is

$$\Pi^R(z) = \frac{p_t(z)}{P_t} y_t(z) - \frac{W_t}{P_t}\left(\frac{y_t(z)}{e^z}\right) \tag{1.10}$$

At the beginning of a period, each firm receives a normally distributed signal, $X(z)$ of the money supply, M_t. Firms know their own productivity but do not know the aggregate price. At the beginning of a period, firms decide whether to change prices or not, and then the aggregate price and money supply are revealed. After the revelation of this information, firms supply output according to the demand that arises. Their decision can be summarized as

$$V_0\left(z, \frac{p_{-1}(z)}{P_{-1}}, \frac{M_{-1}}{P_{-1}}, X(z), G\right) = \max\left[V_S(.), V_C(.)\right] \tag{1.11}$$

subject to the law of motion for the distribution of firms over the states

$$G' = \Gamma(G) \tag{1.12}$$

where $p_{-1}(z)$ is a given firm's price from last period, $V_0(.)$ is the value at the beginning of a period, $V_C(.)$ is the value from resetting prices and $V_S(.)$ is the value of remaining at the price from last period. The value of changing prices is expected revenues minus expected costs (including the menu cost, κ) at the

optimal reset price and the continuation value of starting next period at the optimal reset price.[3]

$$V_C\left(z, \frac{M_{-1}}{P_{-1}}, X(z), G\right) = \max_{p(z)} \mathbb{E}^z \left\{ \pi_C + QV_0\left(z', \frac{p(z)}{P}, \frac{M}{P}, X(z)', G'\right)\right\} \quad (1.13)$$

where

$$\pi_C = \left(\frac{p(z)}{P}\right)^{1-\epsilon} \frac{M}{P} - \omega \left(\frac{M}{P}\right)^{\gamma+1} \left(\frac{p(z)}{P}\right)^{-\epsilon} \left(\frac{1}{e^z}\right) - \kappa\omega \left(\frac{M}{P}\right)^{\gamma}$$

Note that the expectation operator, $\mathbb{E}^z$, is the mathematical expectation conditional of the information set of firm z, which is just the state variables and knowledge of the stochastic processes. The value of not changing prices is expected revenues minus expected costs at last period's price and the continuation value of starting next period with last period's price.

$$V_S\left(z, \frac{p_{-1}(z)}{P_{-1}}, \frac{M_{-1}}{P_{-1}}, X(z), G\right) = \mathbb{E}^z \left\{ \pi_S + QV_0\left(z', \frac{p_{-1}(z)}{P}, \frac{M}{P}, X(z)', G'\right)\right\} \quad (1.14)$$

where

$$\pi_S = \left(\frac{p_{-1}(z)}{P}\right)^{1-\epsilon} \frac{M}{P} - \omega \left(\frac{M}{P}\right)^{\gamma+1} \left(\frac{p_{-1}(z)}{P}\right)^{-\epsilon} \left(\frac{1}{e^z}\right)$$

Note that agents need to keep track of the entire distribution of firms over the states since they need to forecast the current aggregate price as well as tomorrow's aggregate price. Since the entire distribution of firms is an infinite dimension object, this makes the computation intractable. To overcome this, I use the method in Krusell and Smith (1998) and guess that the law of motion for the aggregate

[3]The value function is written in terms of real profits which in turn can be written as a function of $\frac{M}{P}$ using conditions from household optimality, monetary policy and choices in the calibration as follows

$$\frac{p(z)}{P} y(z) - \frac{w}{P}\frac{y(z)}{z} = \left(\frac{p(z)}{P}\right)^{1-\epsilon} \frac{M}{P} - \omega \left(\frac{p(z)}{P}\right)^{-\epsilon} \left(\frac{M}{P}\right)^{1+\gamma} \frac{1}{e^z}$$

price is log-linear in money. Specifically, I use the guess (taken from Nakamura and Steinsson (2010)) that the aggregate price satisifies the following relationship

$$\log \frac{P_t}{P_{t-1}} = \xi_1 + \xi_2 \log \frac{M_t}{P_{t-1}} \tag{1.15}$$

As long as the guess is verified, it is valid to use this as the agents method of forecasting the current and future aggregate price. The law of motion will also be useful later to link the behavior of money and agent's inflation expectations.

1.2.4 Market clearing

The market clearing condition for the labor market equates labor supplied by households with labor demanded for production and price setting

$$L_t = \int_0^1 (l_t(z) + \mathbb{I}(z)\chi)\, dz \tag{1.16}$$

where $\mathbb{I}(z)$ is an indicator function that takes the value 1 when firm z has changed its price, bond markets clear

$$B_{t+1}(h_{t+1}) = 0 \qquad \forall h_{t+1} \tag{1.17}$$

and goods markets clear.

$$c_t(z) = y_t(z) \qquad \forall z \tag{1.18}$$

An equilibrium consists of:

1. functions V_0, V_S, V_C
2. a sequence of allocations $\{l_t(z), c_t(z), y_t(z), C_t, L_t, B_{t+1}\}_{t=0}^{\infty}$ and
3. prices $\{p_t(z), W_t, P_t, Q_t\}_{t=0}^{\infty}$

such that taking prices and the stochastic processes $\{M_t, X_t(z), z_t\}_{t=0}^{\infty}$ as given, the following hold:

1. $\{c_t(z), C_t, L_t, B_{t+1}\}_{t=0}^{\infty}$ solves the consumer's problem
2. $\{l_t(z), y_t(z), p_t(z)\}_{t=0}^{\infty}$ and V solves the firm's problem for each firm
3. markets clear

1.2.5 Solution algorithm

The assumption made in (1.15) is required to reduce the state space from one including an infinite dimensional object to one that can be solved using the method in Krusell and Smith (1998). To solve the model, I use the following algorithm:

1. Guess coefficients for the law of motion (1.15)
2. Solve the value function and save the policy function for prices
3. Simulate the model for T periods and F firms based on the policy function from Step 2
4. Run a regression on the simulated series
5. Check whether coefficients from regression match those from previous iteration, if yes, stop.
6. If not, use regression coefficients as guess in new iteration

Specifically, I simulate the model for 5000 periods and 3000 firms using value function iteration on an extremely fine grid for firms' relative prices to get the policy function. To reduce the state space, I rewrite the value function so that the signal is a deviation from the projection made last period. Therefore instead of requiring more grid points for the signals than the grid points for the aggregate

state, $\frac{M}{P_{-1}}$, I can use a small number, 21, which effectively represents 21 $\times N_{MP}$ potential signals of the aggregate state. This is extremely effective in reducing the number of grid points which is still just under 8.5 million.

1.2.6 Parameter selection

Since the focus of the paper is on quantifying the extent of imperfect information, I calibrate many of the parameters to values which are common in the literature and concentrate on selecting appropriate values for the variance of the signal, menu cost and variance of idiosyncratic shocks. In particular, I set most of the parameters to the values used in Nakamura and Steinsson (2010) as the complete information counterpart of our model is similar. These parameters are displayed in Table 1.4. The parameters relating to aggregate demand are inferred from the behavior of nominal GDP over the period 1947 to 2007. Note that the length of a period is a month. The rest of the parameters are estimated using Simulated Method of Moments.

Table 1.1: Calibrated Parameters

Parameter	Value
Discount factor	$\beta = 0.96^{\frac{1}{12}}$
Coefficient of relative risk aversion	$\gamma = 1$
Inverse of Frisch elasticity of labor supply	$\psi = 0$
Elasticity of demand	$\epsilon = 4$
Speed of mean reversion of idiosyncratic productivity	$\rho = 0.7$
Mean growth rate of nominal aggregate demand	$\mu = 0.0028$
St. deviation of the growth rate of nominal aggregate demand	$\sigma_m = 0.0065$

The key parameter in the model is the variance of the signal which determines how different the imperfect information model is from its complete information counterpart. It is not immediately obvious what data would be best at identifying the precision of signals. Here I use the information contained in firms' expectations of inflation. In the model, forecast errors of inflation are isomorphic to forecast errors of the current money supply so inflation expectations are informative re-

garding the precision of signals of the money supply. Since inflation forecasts are observable, I use the variation in current-period forecasts of the Consumer Price Index (CPI) across the cross section to help identify the precision of firms' signals. If signals are relatively precise, firms' inflation expectations would be relatively close to each other and vice versa if signals are relatively noisy. For the purposes of identifying the noise in signals amongst price setters, the best measure would be inflation expectations by firms. However, since firm-level data on inflation expectations is not available, I use data from the Survey of Professional Forecasters which includes forecasts made by individuals employed by both financial service firms and non-financial service firms.[4] In calculating the dispersion in inflation expectations, I exclude financial firms from the sample.

As the measure of dispersion in inflation expectations, I use the variance in the difference between inflation expectations across agents. This is proportional to the variance of demeaned inflation expectations. Differencing or demeaning is a cleaner measure of dispersion since part of the variance of inflation expectations comprises the variance of inflation itself, which the model cannot replicate with a single aggregate shock. To calculate the dispersion in inflation expectations, I calculate the standard deviation of pair by pair differences in inflation expectations for each quarter, and then take the average over the quarters over the period 1990:Q1 to 2010:Q4. The average difference between expectations over the whole sample is 0.0007 while the standard deviation (or dispersion) is 0.0011. For robustness, I also calculate an extreme level of dispersion which is the mean level of dispersion, 0.0011, plus 2 times the standard deviation of the dispersion in each quarter which gives a value of 0.0024 - there are only 2 quarters over 1990-2010 with larger levels of dispersion, 2008Q4-2009Q1 which are obviously abnormal periods post Lehman Brothers.

[4]Since the model is monthly and the data are quarterly, I assume that the current monthly forecast is one third of the quarterly forecast.

To calculate the counterpart of the dispersion in beliefs in the model, it is necessary to derive the conditional expectation of inflation for a given firm. Using the fact that the log of money supply follows a random walk with normal innovations and the law of motion for inflation in equation (1.15), I can derive the conditional distribution of inflation given a particular firms information set.

$$\begin{aligned} p_t - p_{t-1} | x_t^i, m_{t-1} - p_{t-1} \sim \\ N\left(\xi_1 + \xi_2\left(\mu + m_{t-1} - p_{t-1} + \frac{\sigma_m^2}{\sigma_v^2 + \sigma_m^2} x_t^i\right), \xi_2^2 \frac{\sigma_v^2 \sigma_m^2}{\sigma_v^2 + \sigma_m^2}\right) \end{aligned} \tag{1.19}$$

Define the signals as the sum of the innovation to the money supply and a normally distributed noise term with mean zero and standard deviation, σ_v.

$$x_t^i = \epsilon_t + v_t^i \tag{1.20}$$

Therefore the difference between any two forecasts made at the same date are

$$\mathbb{E}^i\left[p_t - p_{t-1}\right] - \mathbb{E}^j\left[p_t - p_{t-1}\right] = \xi_2 \frac{\sigma_m^2}{\sigma_v^2 + \sigma_m^2}\left(v_t^i - v_t^j\right) \tag{1.21}$$

and taking the variance of both sides provides the dispersion in inflation expectations in the model. Given that σ_m is calibrated, the dispersion in inflation expectations depends on the noise of the signals, σ_v and the second coefficient in the law of motion, ξ_2. This implies that the model needs to be solved in order to determine the actual value of dispersion.

$$var\left(\mathbb{E}^i\left[p_t - p_{t-1}\right] - \mathbb{E}^j\left[p_t - p_{t-1}\right]\right) = 2\left(\xi_2 \frac{\sigma_m^2 \sigma_v}{\sigma_v^2 + \sigma_m^2}\right)^2 \tag{1.22}$$

Using the law of motion from the full information counterpart of the model, figure 1.1 depicts the relationship between the variance of the difference in inflation expectations and the standard deviation in firms' signals fixing the law of motion for inflation to the full information case. The relationship is not monotone be-

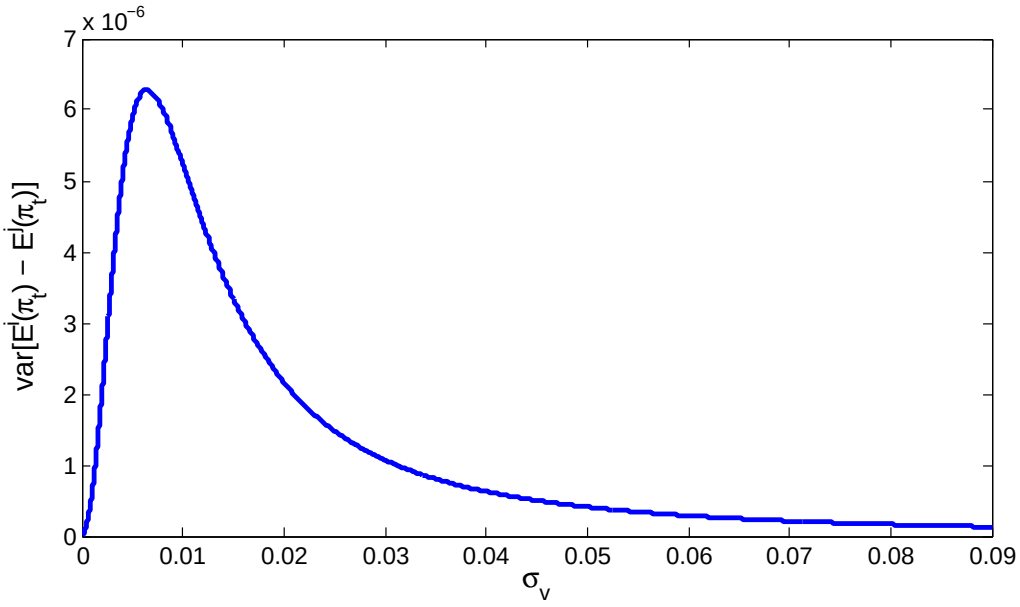


Figure 1.1: Dispersion in inflation expectations

cause firms can have homogeneous expectations for two vastly different reasons. If signals are infinitely precise, expectations are identical because all agents have correct expectations. On the other hand, expectations can be homogeneous because signals are so noisy that everyone forms expectations using the common prior alone. For these reasons, variance among inflation expectations is greatest when signals are moderately precise as agents place substantial weight on heterogeneous signals in forming expectations.

The other moments used in the estimation procedure are the dispersion in prices conditional on being changed, the mean size of price changes and the fraction of price changes in a period. The dispersion in prices provides information on both the variance of firms' signals and idiosyncratic shocks since both the idiosyncratic shock and the signal affect a firm's perceived target price. The last two moments are helpful in identifying the size of the menu cost and variance of idiosyncratic shocks. For instance, a small menu cost would lead to many firms changing prices in a period and the size of those changes to be small. Similarly,

large idiosyncratic shocks would lead to larger and more frequent price changes.

Note that for a given law of motion for inflation, the level of dispersion in beliefs would be consistent with two values for the variance of the signals if one focuses on the moment relating the variance in signals and inflation expectations alone. However, the two values for the variance of the signal that are consistent with the standard deviation in inflation expectations, 0.0011, imply vastly different dynamics which rules out observational equivalence. For the remaining moments relating to the behavior of prices, the relevant data set is the Bureau of Labor Statistics CPI Research Database described in Klenow and Kryvtsov (2008) and Nakamura and Steinsson (2008). The dispersion of prices reported in Golosov and Lucas (2007) is 0.087 and the size and frequency of price changes in Nakamura and Steinsson (2010) is 0.098 and 0.211.

To highlight the effect of introducing imperfect information into the model, results for three models are presented: the full information counterpart, the imperfect information model with the level of dispersion of beliefs from the data, and for robustness, the imperfect information model with an extreme level of dispersion in beliefs relative to the data. These three models correspond to three different parameterizations. Firstly, the menu cost, size of idiosyncratic shocks and standard deviation of the noise of signals are calibrated to match the four moments above. This corresponds to the calibration for the baseline model. The full information calibration sets the standard deviation of the noise of signals to zero, which implies agents have full information, while holding the remaining parameters at the baseline calibration. Finally, the high dispersion calibration also keeps the values for the menu cost and size of idiosyncratic shocks but chooses the standard deviation of the noise of signals to match the extreme level of dispersion described above.

To summarize, I present results for three models where the only difference is the variance of signals that firms receive. The complete information model $\sigma_v = 0$,

the baseline model for which the calibration strategy is based upon $\sigma_v = 0.0016$ and the model with an overly high degree of imperfect information $\sigma_v = 0.0053$. These values are summarized in Table 1.2.

Table 1.2: Estimated Parameters

Parameter	Full information	Baseline	High dispersion
Menu Cost - κ	0.0042	0.0042	0.0042
Std. dev of idiosyncratic shock - σ_z	0.0469	0.0469	0.0469
Std. dev of signal - σ_v	0	0.0016	0.0053

The standard deviation of the signal in the baseline calibration is quite small, roughly one-fourth the size of the monetary shock. The value for the menu cost is slightly larger than that reported in Golosov and Lucas (2007) because the model here includes aggregate shocks. If the menu cost were as small, the combination of idiosyncratic and aggregate shocks would lead to an overly high frequency of price changes relative to the data. Regardless, menu costs comprise roughly 1 per cent of revenues which is comparable to Levy et al. (1997) who find changing prices costs 0.7 per cent of revenues in supermarkets. The moments from the data and the three versions of the model under the calibration strategy described above are shown in Table 2.5. Even though the three models differ in the extent of noise in signals, their simulated moments except for the dispersion in inflation expectations are almost identical.

Table 1.3: Moments from Data and Model

	Data	Full information	Baseline	High dispersion
Frequency of price changes	0.211	0.210	0.210	0.212
Mean size of price changes	0.098	0.089	0.089	0.0089
St. dev of price changes	0.087	0.088	0.088	0.0872
Dispersion of inflation expectations	0.0011	0	0.0011	0.0020

Moments from the baseline and full information model are similar because increasing the standard deviation of signals from 0 to 0.0016 does not generate enough noise to have an effect. Note that although the moments from the high dispersion model are also almost identical, it is not the case that increasing the

standard deviation of signals has no effect. Rather, the high dispersion model is re-calibrated to match the moments and hence the model can match higher dispersion in inflation expectations to an extent. The target for dispersion in inflation expectations is 0.0024 but the empirical moment is 0.0020. Since there is a direct link between the variance in signals and the dispersion in inflation expectations, it must be the case that the other moments "move" too much in the minimization routine.

1.3 Results

The central equation governing monetary non-neutrality in the model is the law of motion for inflation which describes how aggregate prices respond to monetary shocks. If the aggregate price is relatively responsive to monetary shocks, reflected by a large coefficient in front on the money supply term in equation (1.15), prices adjust quickly and quantities do not need to adjust. Therefore, lower responsiveness in the law of motion directly indicates greater monetary non-neutrality. Table 1.4 shows the equilibrium law of motion for the three models.

Table 1.4: Law of Motion for Inflation

	Full information	Baseline	High dispersion
Constant	0.592	0.572	0.484
Response to monetary shock	0.540	0.522	0.441

Monetary neutrality would be reflected by a coefficient of 1 in the response to the monetary shock. That the full information response is much lower than 1 demonstrates that menu costs alone generate substantial price stickiness. Bear in mind the model is monthly so in log terms, in the full information model, the aggregate price absorbs 54 per cent of the monetary shock after one month

which translates into a quarterly response of 91 per cent.[5] This is consistent with the selection effect in Caplin and Spulber (1987) where the small number of firms changing prices in a period are those who change prices by a relatively large amount since they are furthest away from the optimum. While only 20 per cent of firms change prices in any given month, 91 per cent of the monetary shock is absorbed in the aggregate price after one quarter. For the baseline model and the model with an extreme level of dispersion in beliefs, 89 and 82 per cent of the aggregate shock is absorbed in the aggregate price after one quarter. Note that these responses are much smaller than New Keynesian models where responses monetary shocks typically take a year or more to dissipate.

To make the point graphically, Figure 1.2 shows the impulse response of prices across the three models and Figure 1.3 shows the responses of output to a one standard deviation monetary shock. The lower the level of dispersion, the larger the initial response of the aggregate price and the less persistent the deviations.

Conversely, the lower the level of dispersion, the larger the initial response and the more persistent the deviations are in output. In terms of cumulative deviation of real output, the difference between the full information and baseline model is 8 percent whereas the difference between the full information and the extreme dispersion model is 49 percent. This suggests that imperfect information can have a substantial effect on monetary non-neutrality but for a realistic degree of imperfect information, the effect is imperceptible. Therefore it is unlikely that imperfect information is an important channel for generating monetary non-neutrality even though it is capable of generating large effects in theory.

The difference in the size and persistence of impulse responses has implica-

[5]The equilibrium law of motion in log terms can be written as

$$p_t = c + \xi_2 \sum_{k=0}^{\infty} (1 - \xi_2)^k m_{t-k}$$

where c is a constant term.

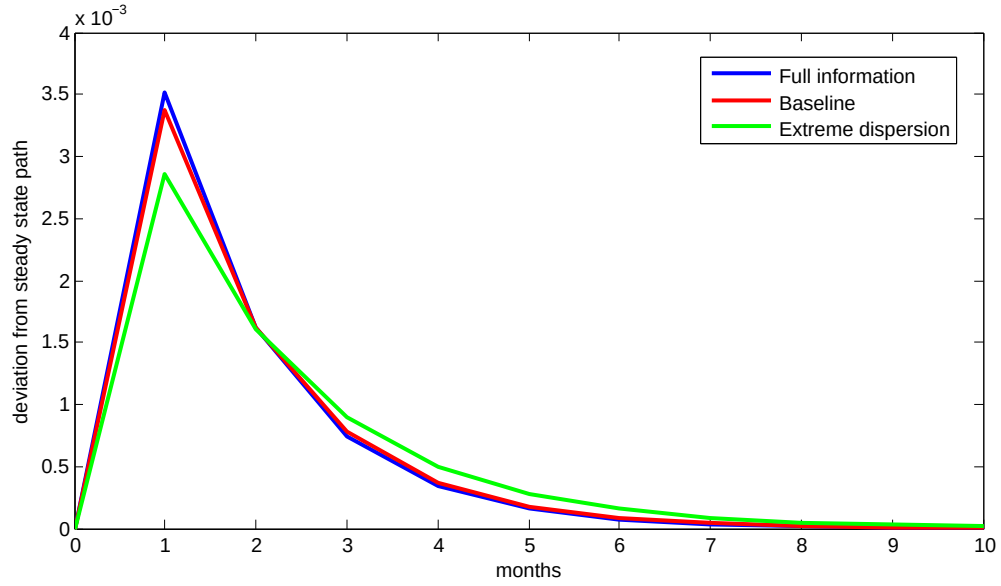


Figure 1.2: Impulse response of the aggregate price to a monetary shock

tions for the amount of variance that the model can account for. If monetary disturbances are quantitatively important in a world, they will account for a substantial fraction of the variance in output. The variance of log real GDP from a HP-filtered trend over the period 1947:Q1 to 2010:Q4 is 0.00029 and using the law of motion for inflation and the fact that money supply is a random walk, the variance of the log of real output in the model is[6]

$$var(c_t) = \frac{(1-\xi_2)^2}{1-(1-\xi_2)^2}\sigma_m^2$$

Table 1.5 shows the fraction of the variance that each of the three models can account for. The complete information and realistically calibrated imperfect information models can account for just 4 per cent of the variance in output. Even the imperfect information model with an overly high variance in the signal can

[6]Start from

$$c_t = m_t - p_t = \frac{(\mu-\xi_1)(1-\xi_2)}{\xi_2} + (1-\xi_2)\sum_{k=0}^{\infty}(1-\xi_2)^k\epsilon_{t-k}$$

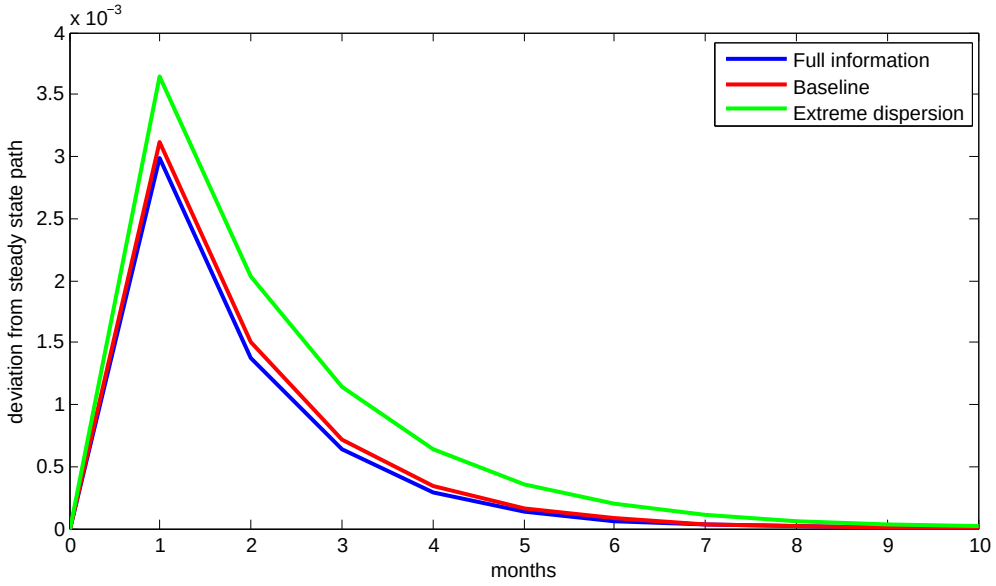


Figure 1.3: Impulse response of output to a monetary shock

only account for less than 10 per cent of the variance in output. Imperfect information is clearly incapable of generating monetary non-neutrality in this setting.

Table 1.5: Variance decomposition

	(1)	(2)	(3)
Fraction of variance in output accounted for	0.038	0.042	0.092

Moreover, the results above apply to welfare calculations since welfare losses from a flexible price baseline can be measured by the deviation of prices from the level required to maintain the steady state level of real money. By equation (1.7), this is equivalent to the deviation in output from steady state. Given the law of motion for inflation, the distribution of output is

$$c_t \sim N\left(\frac{1}{\xi_2}\left((1-\xi_2)\mu - \xi_1\right), \frac{(1-\xi_2)^2}{1-(1-\xi_2)^2}\sigma_m^2\right) \tag{1.23}$$

In terms of per period welfare, a Lucas style welfare calculation suggests that the cost of monetary disturbances is approximately 0.0005 per cent for the realistically calibrated model and 0.0012 per cent for the imperfect information model with

excessive variance in the signal. This result is purely driven by the inability of imperfect information to account for fluctuations in output and the standard result that fluctuations are not that costly in the first place.

To see exactly why the correctly calibrated value for the variance of the signal has such a small effect, it is useful to decompose the effect of imperfect information on a firm's decisions. For what follows, $\frac{M_t}{P_t}$ is referred to as the current aggregate state and the agent's information set at t is last period's aggregate state and a signal of the current aggregate state. Imperfect information of the current aggregate state affects decisions in three ways. It lessens knowledge of the current state, which implies firms do not know the relative price they will have tomorrow and the information set that will arise in the next period. Agents face uncertainty in the current period relative to no uncertainty in the complete information model, and even more uncertainty regarding next period than the uncertainty regarding next period in the complete information model. The variance in the signal pins down the level of precision agents have over these states.

The distribution over the current aggregate state, $\frac{M_t}{P_t}$, conditional on the firm's information set is related to

$$m_t - p_t | x_t, m_{t-1} - p_{t-1} \sim$$
$$N\left((1-\xi_2)\left(\mu + m_{t-1} - p_{t-1}\right) - \xi_1 + \frac{(1-\xi_2)\sigma_m^2}{\sigma_v^2 + \sigma_m^2} x_t, (1-\xi_2)^2 \frac{\sigma_v^2 \sigma_m^2}{\sigma_m^2 + \sigma_v^2} \right)$$

where I make use of the relationship between the lognormal and normal distributions. Note that the signal is defined in terms of the deviation from the projection of the aggregate state last period

$$x_t = v_t + \epsilon_t \tag{1.24}$$

The key term in this distribution and the distributions that follow is the signal

to noise ratio which dictates the weight agents place on the signal relative to last period's observation.

$$S(\sigma_v) = \frac{\sigma_m^2}{\sigma_v^2 + \sigma_m^2} \tag{1.25}$$

When the variance of the signal is zero, the distribution of the current aggregate state collapses to a single value which is equal to the true aggregate state as the signal to noise ratio, $S(\sigma_v)$, is one and the noise of the signal, v_t, is equal to zero. Table 1.6 shows the signal to noise ratio for the two level of variances used earlier, 0.0016 and 0.0053.

$$\begin{aligned}(1-\xi_2)\left(\mu + m_{t-1} - p_{t-1}\right) - \xi_1 + (1-\xi_2)S(\sigma_v)x_t = \\ = (1-\xi_2)\left(\mu + m_{t-1} - p_{t-1}\right) - \xi_1 + (1-\xi_2)\epsilon_t \\ = m_t - p_t\end{aligned}$$

Table 1.6: Signal to noise ratio

	$\sigma_v = 0.0016$	$\sigma_v = 0.0053$
Signal to noise ratio	0.942	0.601

When signals are relatively informative, agents almost rely on them exclusively whereas noisy signals are partly discarded. In the baseline and full information models, agents rely almost completely on signals which are either perfect or close to perfect, implying that each agent has near complete information about the aggregate state.

1.3.1 Discussion

For the model considered above, calibrated to reflect the behavior of prices at the micro level and the level of dispersion in inflation expectations in the cross section, imperfect information generates only small and transient real responses similar to Golosov and Lucas (2007). Even when the variance in firms' signals is increased to

a level close to the maximum seen in survey data, monetary disturbances can only account for 9 per cent of the variance in output. The result is entirely driven from the fact that the variance of inflation expectations across agents is small and that the selection effect leads the small group of price changers to neutralize monetary shocks. In the model used here, low dispersion in inflation expectations implies that the variance of firms' signals is small and that the imperfect information model resembles the complete information model which is incapable of generating monetary non-neutrality.

One mechanism that may seem promising is delays in learning. In the current model, firms learn the aggregate state perfectly after one period. However, the problem with introducing delays in learning is that it also increases the dispersion in firms' beliefs. It may be that firms do not learn the aggregate state so quickly but the length of delay is limited as the dispersion in firms' beliefs in small. Introducing delays in learning would need to be coupled with less variance in signals if the variance in inflation expectations is to remain consistent with the data.

The same argument applies to many of the mechanisms used to implement imperfect information in models of price setting. Introducing Calvo-type information updating as in Mankiw and Reis (2002), costs of acquiring information as in Alvarez et al. (2011) and Bonomo et al. (2010), and rational inattention as in Mackowiak and Wiederholt (2009) are all effective at generating monetary non-neutralities but they are likely to imply a dispersion in beliefs that is not supported in the data. It is precisely that these mechanisms generate dispersion in beliefs that enables them to generate monetary non-neutrality.

Clearly, the model considered in this paper is much simpler than the literature. Even though the models in the literature are quite varied, the common element in the literature is that dispersion in beliefs, caused by various frictions, is capable of driving monetary non-neutrality. Here the dispersion is tied directly

to the variance of firms' signals of the money supply. Given that the data suggest dispersion in beliefs is actually quite small, this implies a small variance for firms' signals. However, even in more complex models, the low dispersion in beliefs in the data will constrain the extent of those mechanisms that generate monetary non-neutrality. For instance, in a model with costs of acquiring information and noisy signals, low dispersion in beliefs implies a combination of relatively precise signals and low costs of acquiring information regardless of exactly how one identifies the parameters. In other words, a potential sufficient statistic for models of price setting with imperfect information is the dispersion in firms beliefs. Since the data suggests that dispersion in beliefs is small, it is unlikely that imperfect information operating through the dispersion of beliefs is quantitatively important in generating monetary non-neutrality.

1.4 Conclusion

Imperfect information seems a reasonable mechanism for generating monetary non-neutrality. The idea that firms don't have complete information is intuitive and much research on imperfect information in the area of price setting has focused on the mechanisms, studying models that incorporate learning, costs of acquiring information and rational inattention. The paper here focuses on an issue that has thus far been ignored, the quantitative importance of imperfect information. If imperfect information is simply not present in the data or quantitatively important, then the focus on imperfect information is likely unwarranted.

In order to make a quantitative assessment, I developed a parsimonious model that incorporates a simple form of imperfect information and is capable of being taken to the data. Firms face menu costs in price setting but only receive a private signal of the current aggregate state. The model implies a variance in inflation expectations across agents which I used to calibrate the degree of variance in the

private signals. The calibrated variance in the private signals is relatively small precisely because the variance in inflation expectations across agents in the data is small. Even if one doubts the source of inflation expectations, bounding the forecast errors by the maximum seen in survey data results in a small variance. Unless one is willing to allow an unrealistically high degree of variance in inflation expectations across agents, the implied noise in signals has to be small.

Since the variance of the signal is small, the imperfect information model behaves almost identically to the full information model where monetary non-neutralities are small, consistent with both Caplin and Spulber (1987) and Golosov and Lucas (2007). Impluse responses of output are small and transient, the variance accounted for by monetary disturbances are tiny and welfare losses are virtually non existent. In other words, the model suggests that imperfect information is not quantitatively important when menu costs are present and the variance of inflation expectations matches that in the data.

CHAPTER 2

Product Market Frictions, Selection and Allocative Efficiency

2.1 Introduction

Although selection effects play a prominent role in models of heterogenous firms, relatively little is known about how selection operates in practice. For example, selection effects generate part of the aggregate productivity gains from trade liberalization in Melitz (2003). Allowing firms access to export markets reallocates production to relatively productive firms, bidding up wages and inducing exit of relatively less productive firms. In this paper, I show that evidence of selection exists in the data but that it is not driven by variation in input prices. To reconcile these facts, I show that a simple model where firm dynamics are governed by product market frictions is capable of generating selection effects without any variation in factor prices.

The "up or out" pattern for young firms documented by Haltiwanger et al. (2012) is often interpreted as evidence of selection. It hints at a cutthroat-like selection process amongst entrants, with only the best entrants growing while the rest exit. However, the "up or out" pattern is a single data point stating that young establishments exit at a higher rate while young survivors grow faster relative to old establishments. A single data point cannot be taken as evidence of selection since selection effects involve changes in shares of production by the relatively productive and their effects on the exit of the less productive. Hence,

evidence of selection requires data on the relationship between exit and the shares of relatively productive entrants.

To arrive at the robust stylized fact that larger firms are also more productive, it is either the case that productive entrants grow faster than less productive entrants, or that less productive entrants grow faster initially and increase productivity later on. Regardless, some group of entrants which are "better", grow at the expense of other entrants. Therefore, data over time and across industries should show a positive relationship between exit rates and the growth of surviving entrants. Loosely speaking, selection effects lead to a rivalrous form of growth; establishments grow by accumulating resources from others in order to escape the hazard of exit. Cohorts born into industries or periods with more intense selection pressure, driven by the strength of its competitors, would be marked by higher thresholds for survival.

Alternatively, there are equally valid explanations for a negative relationship between the growth of surviving entrants and exit rates. For instance, industries that are in the process of growing as a whole could be facing rising demand. Establishments in a growing industry are not necessarily competing for a fixed set resources such as customers, but instead growing the market together. Hence growth in an expanding market need not be rivalrous, implying that growth rates of young establishments may be negatively related to exit rates. Hsieh and Klenow (2012) have a similar motivation in mind when using data from the US Manufacturing Census to show that productivity growth by young establishments may be linked with the growth of these young establishments. If growth in an industry is driven by productivity growth common to all, we should expect to see most establishments growing and potentially lowering the likelihood of exit. That is, a weakly negative relationship as higher growth is coupled with lower exit rates among entering establishments. Hence, there exists two forces behind the relationship between the growth of young establishments and exit rates.

When constructing the data for the "up or out" pattern, growth of young firms is defined as the employment weighted average year-on-year growth in employment of survivors relative to surviving older firms. However, this measure of growth conflates two very different forms of growth that map to the rivalrous and common distinctions of growth. Faster growth can be driven by a small number of rapidly growing firms with most other firms not growing at all or instead, a common growth rate for all entrants. The difference between these two different forms of growth is that the former is captured by the *difference* in growth rates between fast and slow growing firms, while the latter is captured by the *level* of growth rates. To see this more explicitly, suppose all entrants start with zero employment and denote $e_{i,t}$ employment of entrant i, $\bar{e}_t$ as the weighted average employment after t periods and N as the number of entrants. By definition, the size of the cohort after t periods is

$$E_t = \sum_i w_{i,t}(e_{i,t} - \bar{e}_t) + N\bar{e}_t$$

The size of the cohort can increase because of the growth of a few rapidly growing entrants, which is reflected in the first term, or by common growth of all entrants, reflected in the second term. These two measures of the growth of young firms mirror the distinction between rivalrous growth and growing the market together. A larger spread in the growth of fast and slow growing firms captures rivalry in growth while the level of growth captures overall market growth.

Since the "up or out" pattern is literally a single data point, it is uninformative about the relative roles of rivalrous and common growth in determining exit rates. In this paper, I use establishment-level data to uncover the relationship between exit rates and these two measures of the growth of young establishments. The data show that both forms of growth are closely linked to exit rates of young establishments, with an increase in the measure of rivalrous growth and a decrease

in the measure of common growth leading to greater exit of young establishments. I interpret these facts as evidence that rivalry in growth or selection, and common-to-all growth both matter in understanding the growth patterns of young establishments.

In particular, I use time series variation within 3 digit (SIC) manufacturing industries to establish these facts for young establishments. A 10 per cent increase in the spread of growth of fast and slow growing establishments is associated with at 1.5 per cent increase in exit rates. To put this in perspective, going from the 10th percentile observation in terms of this measure of the spread to the 90th percentile observation would entail a 27 per cent increase in exit rates. The effect from the level of growth goes in the opposite direction, but is much larger. A 10 per cent increase in the level of growth decreases exit by 4.1 per cent, with the 10th-90th percentile variation generating a 39 per cent decrease in exit. Both coefficients are statistically significant and control for year effects and fixed effects at the industry level.

However, in contrast to selection mechanisms that typically characterize models of heterogeneous establishments, stronger selection does not appear to be achieved through increases in factor prices. Industries which experienced an increase in the strength of selection, as defined by periods where there was an increase in the spread of growth and an increase in exit rates, did not experience a corresponding rise in wages. This is in stark contrast to dynamics in typical models of heterogeneous establishments when distortions or frictions are weakened.[1] The reallocation of resources towards more productive establishments increases the marginal product of factors, increasing factor prices and inducing exit of less productive establishments as they become unprofitable.

Instead, I focus on the response by customers in inducing exit. Since product

[1] Anything that reallocates production to relatively productive establishments will lead to an increase in factor prices.

markets are characterized by search frictions, customers often satisfice by accepting a less than desirable match to avoid the costly nature of search. The degree to which customers are willing to satisfice is a margin which affects selection. When customers believe the likelihood of a good match has increased or equivalently, the expected time until a good match will arise shortens, they satisfice less and reject matches with establishments they would have previously matched with. Hence selection forces can strengthen without corresponding increases in factor prices.

Understanding the growth of young establishments is important as entering establishments account for a significant proportion of productivity growth. Foster et al. (2008) use data on physical measures of inputs and outputs in a subset of manufacturing industries to show that just under a quarter of productivity growth is due to entry of new establishments. As is typical in the misallocation literature, the allocation of production across establishments directly affects allocative efficiency. The difference here is that establishments must match with customers instead of selling their output in spot markets. When customers are allocated inefficiently, production is reallocated to lower quality establishments, lowering welfare.[2] Hence spreads in growth rates between fast and slow growing establishments may be informative regarding allocative efficiency. An economy that can allocate the bulk of its customers to establishments producing high quality goods will be characterized by slower growth of establishments in the left tail of the quality distribution and much faster growth for those establishments in the thinner right tail.

As a cross country comparison, I compute the spread in growth rates between fast and slow growing establishments in the US and Chile. After 10 years, the 90th percentile (in terms of cumulative growth of employment) surviving Manufacturing establishment in the US has added 6.74 median establishments more worth of

[2]Although quality can alternatively be interpreted as productivity, the point here is that growth need not be correlated with productivity and supply side considerations. Instead, I focus on an environment with a homogeneous good produced with varying quality.

workers than the median 10 year old establishment while the spread in cumulative growth is only between 2.09-3.17 in Chile. In other words, the evolution of the cumulative growth distribution for entering manufacturing establishments fans out much more rapidly in the US than in Chile. Of course, inferences drawn from these differences should be interpreted as upper bounds since it is unlikely that the difference in spreads is caused entirely by variation in allocative efficiency.

To assess the quantitative effect of product market frictions and the role of customers in driving selection, I construct a model that is capable of matching establishment dynamics as well as pricing behavior by establishments. For example, establishments lower prices to attract customers leading to rising markups over the lifecycle, as in the data. Lessening product market frictions increases the ability of customers to find high quality/low price establishments, raising growth rates in the thin right tail of entrants endowed with a high quality good while slowing down growth of other establishments, leading to higher exit rates of young establishments. Moreover, as in the data, resources are continually being reallocated via endogenous entry and exit from incumbents to entrants, who have higher quality on average.

In the model, establishments match with customers via an establishment specific matching function which is increasing in the surplus offered to consumers while decreasing in other establishments surplus offered, as well as the mass of competitors. Depending on the calibration, the matching function can resemble purely random matching or directed search whereby the establishments offering the greater consumer surplus supply all the customers. Pure directed search is this manner resembles an economy without transport costs and zero search and information frictions. The degree to which search resembles directed search matters for allocative efficiency, since the allocation of customers reflects the allocation of production. A decrease in search frictions is beneficial for establishments producing higher quality goods since, in equilibrium, they offer the highest consumer

surplus. High quality producers find it easier to accumulate customers leading production to shift toward these establishments improving allocative efficiency.

Product market frictions lead to selection, without variation in factor prices, because customers' expectations of the surplus available in future matches affects their willingness to satisfice. When search frictions are high, customers realize that the likelihood of finding a match offering relatively large consumer surplus is relatively low implying that waiting for a new match is costly relative to accepting the match in hand. As search becomes more directed, it takes less time to find a high surplus match, increasing the value of search. Hence customers shun matches they would previously accept, in favor of returning to search. This effectively raises the exit threshold, as some low quality producers cannot make profits *and* lower prices enough to compensate customers for the quality of their goods.

To assess this new margin of selection and the effect of product market frictions on allocative efficiency, I calibrate the baseline model to a rich set of stylized facts on manufacturing establishments in the US and assess the effect on aggregates from an increase in search frictions in product markets that make search more random and less directed. To discipline the magnitude of this change, I match the slower fanning out of the cumulative growth distribution observed in Chile. Welfare falls by 14 per cent as customers are forced to satisfice more and production is reallocated towards lower quality producers who would otherwise have been forced to exit. Interestingly, part of this reallocation is offset by higher margins at higher quality producers whose market power improves with the reallocation.

2.1.1 Related literature

The paper is most closely related to Hsieh and Klenow (2012) in that it focuses on manufacturing establishment growth and aggregate productivity. Here, I focus on the evolution of the size distribution over the lifecycle and the role of customers

in driving selection and allocative efficiency, instead of productivity growth. The extent of product market frictions affects the intensity of selection and allocative efficiency. In contrast, Hsieh and Klenow (2012) is concerned with mean productivity growth over the lifecycle in determining aggregate productivity as opposed to allocative efficiency. As productivity is proportional to labor, the critical statistic in their model is the evolution of mean size over the lifecycle. Here, the model is agnostic regarding mean growth over the lifecycle while suggesting that a greater "fanning out" of the cumulative growth distribution over the lifecycle is evidence of higher allocative efficiency.

Within the strand of literature on firm or establishment growth, the process of establishment growth used in this paper is most closely related to Fajgelbaum (2013) who studies the effect of labor market frictions on export or technology adoption decisions via the constraint on firm growth. Firms hire workers subject to labor market search frictions until they have accumulated enough labor to make a once-off investment in exporting or technology adoption. The mechanism here is firms' rivalrous accumulation of customers, which is governed by product market frictions and customers' search behavior.

The mechanism of growth is also similar to Gourio and Rudanko (2011) and Dinlersoz and Yorukoglu (2012) where customer accumulation is the key channel of firm growth. Both papers introduce customers as a form of capital but for different purposes. Gourio and Rudanko (2011) focus on the sluggish nature of adjustment in customer capital which affects firms responses to shocks while Dinlersoz and Yorukoglu (2012) explore the role of information dissemination in attracting customers. In contrast, the model here features customer accumulation as the channel of rivalrous growth to explore allocative efficiency across heterogeneous establishments and the role of demand side considerations in selection.[3]

[3]In the model, customers can alternatively be interpreted as skilled labor where production is Leontief in skilled and final goods.

Models of establishment size or growth rates can broadly be categorized as relying on either exogenous characteristics that persist through the lifecycle (such as productivity in Hopenhayn (1992) and the likelihood of particular productivity draws Atkeson and Kehoe (2005)) or the accumulation of some resource (blueprints in Luttmer (2011) and knowledge capital in Atkeson and Burstein (2010)). Establishment growth in this paper relies on the accumulation of resources which are easier to accumulate for establishments born with a technology to produce relatively high quality goods.

The rest of the paper is organized as follows. In Section 3.2, I describe the data and document 3 stylized facts about establishment dynamics. The complete model is developed in Section 3.3. Section 2.4 describes the calibration and estimation strategy applied to US data. The results, which include the calibration to Chilean data and the subsequent effect on aggregates, are presented in Section 2.5. Finally, I conclude in Section 2.6.

2.2 Data

The data in this section are from the US Census Bureau's Synthetic Longitudinal Database (LBD) and the National Statistics Institute (INE) of Chile's Annual Manufacturing Census (ENIA). The synthetic LBD contains synthetic data between 1976 and 2000 created from the confidential version of the Census Bureau's LBD. Statistical models are fit to the data in the confidential LBD and then simulated to create the synthetic LBD (Kinney et al., 2011). Since the data is synthetic, there are obviously aspects of the data that the synthetic LBD will not capture, regardless of how comprehensive the statistical models used to create the synthetic data. However, in the dimensions explored in this paper, the synthetic LBD is representative of the confidential LBD as the results in this paper have been analytically validated. Hence aside from negligible quantitative differences,

the results may as well have come from the confidential version of the LBD.

The confidential LBD contains longitudinal data on all establishments on the Census Bureau's Business Register, a large list developed and maintained for the purpose of all federal statistical agencies (Jarmin and Miranda, 2002). The LBD is a particularly useful dataset for studying the dynamics of young establishments as the annual frequency and longitudinal nature of the data allows cohorts of establishments to be followed. For this paper, I focus on manufacturing establishments which are defined as those establishments with SIC code between 52 and 59.

Data on Chilean establishments are taken from the ENIA between 1995 and 2007. While there are less establishments in the sample and less years in the sample, the data on each establishment is much more detailed when compared with the LBD since the data come from a Manufacturing census. For example, detailed data on inputs used in the production process are available. The LBD is limited to industry code, payroll, employment and age for each establishment.

For each establishment in both samples, I calculate age as the time from when the establishment first appeared in the dataset. Therefore, for any calculation that requires the age of an establishment, I exclude establishments that first appeared in 1995 in the ENIA data and prior to 1976 from the LBD data since age is indeterminate for these establishments.

While both the LBD and ENIA samples both cover manufacturing establishments, they are not directly comparable. The ENIA excludes establishments that are (i) single-unit (not part of a firm with multiple factories) with less than 10 employees and (ii) part of a firm with less than 10 employees. In contrast, the LBD contains data on all establishments. Ideally, I would restrict the sample in the LBD on the same criteria as the LBD when directly comparing the two datasets. Although the synthetic LBD identifies establishments as being single unit or multi-unit (owned by a firm with more than one establishment), it does not contain parent or firm identifiers so it is not possible to identify firm size

(as opposed to establishment size). Given these constraints, I exclude single-unit establishments with less than 10 employees but leave establishments which are multi-unit and part of firms with more than one establishment because firm size is not identified. Fortunately, the number of multi-unit establishments at firms with less than 10 employees is tiny so leaving these establishments in the sample will not affect the results.[4]

To account for scale issues between the countries, I divide cumulative growth by the median employment for each industry. This should eliminate effects from the much larger size of the US economy and scale effects from different industry composition across the two countries. Given these adjustments, I calculate the change in cumulative size (by employment) over the lifecycle relative to the median sized manufacturing establishment for each country.[5]

Even after constructing the samples to make them as comparable as possible, it is likely that the industry composition of the manufacturing sectors in each country are fundamentally different. One expects a higher proportion of establishments to be classified as high tech relative to establishments in Chilean manufacturing. Because industries are likely to differ in terms of optimal size and returns to scale, comparisons between Manufacturing in the US and Chile need to account for these differences. To do so, I attach sampling weights to Chilean establishments based on their 4 digit ISIC Revision 3.1 codes that re-weight the sample such that industry composition mirrors that in the US.[6] Hence the results that follow

[4]According to data in the Census Bureau's Business Dynamics Statistics in 2010, multi-unit establishments at firms with less than 10 employees represent less than 0.05 per cent of all manufacturing establishments.

[5]The median establishment size in both countries is quite similar. Over their respective range of years, the median establishment size ranges from 27 to 30 workers in the US and 22 to 29 in Chile. Moreover, the median establishment size over the whole sample is 28 in the US compared to 24 in Chile.

[6]In particular, I used the Census Bureau's ISIC 3.1 to 2002 NAICS concordance. The mapping between ISIC to NAICS codes are often many to many. Since there is no way to make this mapping one-to-one, I attach weights such that an establishment with an ISIC code with many NAICS correspondences is treated as a separate establishment for each NAICS correspondence with sampling weights reflecting each NAICS code's relative weight in the US compared to the other NAICS correspondences. This ensures that the composition of manufacturing in

are not affected by differences in the composition of Manufacturing between the countries.

To summarize what follows, there are three new stylized facts that I observe in the data: (*i*) higher exit rates of young establishments are positively associated with larger spreads in growth between fast and slow growing establishments in the US, (*ii*) establishments that grow *relatively* slower exit at a faster rate and (*iii*) the right tail of the distribution of surviving establishments' cumulative growth fans out much faster in the US relative to Chile.

2.2.1 Rivalrous growth

As documented in Cabral and Mata (2003), the size distribution fans out over the lifecycle. Most establishments start small with the right tail dragging the median establishment rightwards over time. Here, I show that an increase in terms of the difference in cumulative growth of fast and slow growing establishments is associated with higher exit rates of young establishments. The measure of cumulative growth for an a-old establishment i, in industry j is

$$g_{a,i,j} = \frac{emp_{i,a} - emp_{i,0}}{median_{i'}\left(emp_{i',j}\right)}$$

I choose to focus on cumulative growth instead of growth *rates* because measuring growth by growth rates obscures the rivalrous nature of growth. An extra customer for one establishment implies that a customer is not available for another. This does not necessarily lead to an equivalent sized change in growth rates for the two establishments if they are different in size whereas the change in cumulative growth will be plus/minus one regardless of establishment size.

For each 3 digit SIC industry in manufacturing, I need a measure of the spread or fanning out of the cumulative growth distribution and exit rates for young

the adjusted ENIA sample mirrors that in the US.

establishments. To do so, I construct cohorts of establishments born in the same year for each industry, calculate the statistic of interest for each cohort, and then take the mean over the cohorts. Alternative weighting schemes over the cohorts, such as entrant or employment weighted generate qualitatively similar patterns.

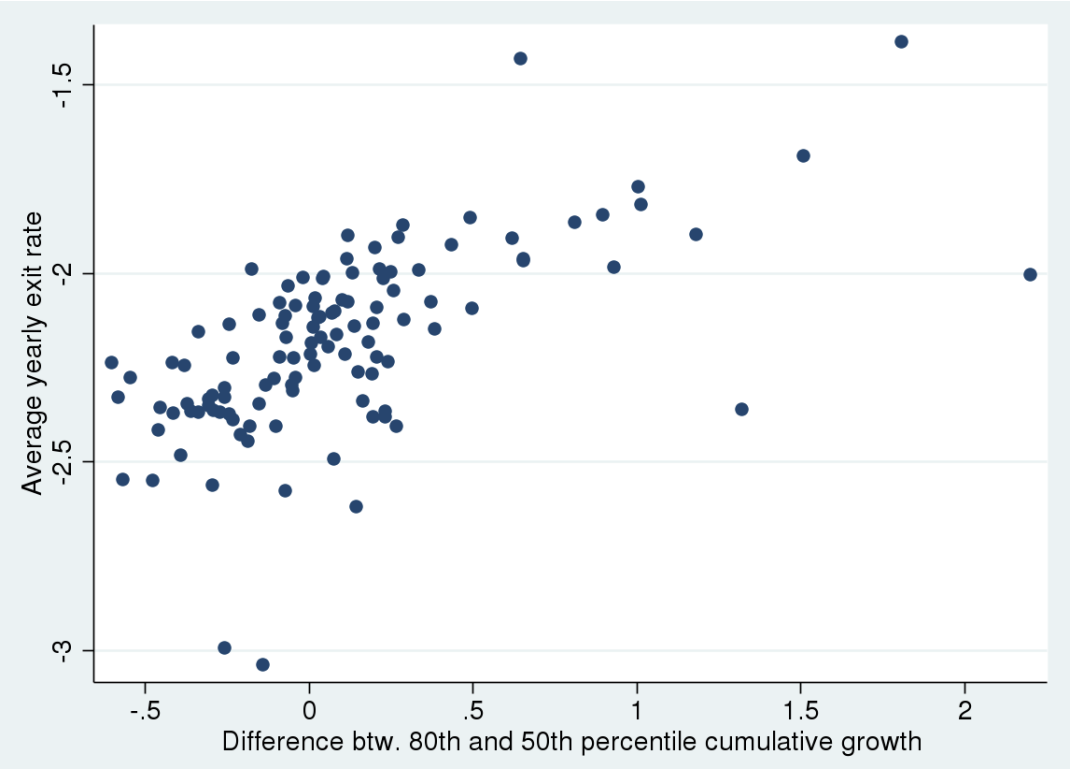
The exit rate for a particular industry is the average exit rate across the cohorts. That is, for every cohort in a given industry, I calculate the mean yearly exit rate for establishments aged 5 or less and then average across the cohorts. Similarly, for a given industry, I calculate percentiles of cumulative growth within each cohort up until age 5. Calculating percentiles within cohorts accounts for year effects at the industry level, since year effects will not distort relative growth within a cohort. For each cohort and at each age under 5, I then calculate the difference between the 80th and 50th percentile cumulative growth, accounting for exiters.[7] For example, I calculate the 50th-80th percentile spread for age 1, 2, ..., and 5 in a given cohort and then take the average of these spreads as a cohorts' value of the 50-80 spread. Finally, I take the average across the cohorts for each industry and am left with a 50-80 spread for each industry. So far, this calculation only uses the cross section for identification. To exploit the time dimension of the data, I perform the procedure above for without averaging over cohorts to generate a panel dataset where the cross sectional unit is an industry and new cohorts are born for each unit of time.

Figure 2.2.1 shows a scatter plot of this data. Each point represents an industry. There is a clear positive relationship between exit and the spread in cumulative growth of young establishments. However, the simple scatter plot analysis does not take into account fixed effects. Industries may have high exit rates for

[7]The identical calculation can be made only including survivors. However, this may mechanically generate a positive relationship between the spread in cumulative growth and exit. To see this, suppose exit disproportionally affects establishments that grow relatively slowly so that exit is essentially trimming the left tail of the distribution of cumulative growth. Higher exit is essentially larger truncation from the left. If the skewness of cumulative growth increases as you move rightwards along the distribution, then larger truncation from the left will mechanically increase skewness resulting in a positive relationship.

an entirely different reason that is correlated with larger spreads in cumulative growth. For example, it may be that an industries differ in their general level of competitiveness which is fixed over time yet correlated to larger spreads. In this case, it would be wrong to conclude that changes in the spread of cumulative growth would lead to changes in exit rates since the relationship is just identifying a relationship in the level of general competitiveness and exit rates.

Figure 2.1: Exit rates and spread in 80th-50th percentile cumulative growth (log units)



To account for these fixed effects at the industry level, I estimate the relationship on differences instead of levels. This eliminates the role of any variable that is constant at the industry level and uses within industry variation to identify the effect of the spread in growth rates on exit rates of young establishments. In addition to the spread of cumulative growth, I include cumulative growth at the 65th percentile which should capture the effect of the level of cohorts' cumulative

growth on exit rates and dummies for each year to account for year effects. I estimate the following equation using data on each industry j over the period 1976 to 2000 where u_j is the industry fixed effect eliminated by differencing.

$$\log\left(exit_{j,t}\right) = \left[\mathbb{I}_t \log\left(spread_{j,t}\right) \log\left(cumulativegrowth_{j,t}\right)\right]' \boldsymbol{\beta} + u_j + e_{j,t}$$

Table 2.1 displays the results of the estimation where reported standard errors are HAC robust. In all, there are 131-3 digit manufacturing industries representing 3259 industry-time pairs. The direction of each coefficient is consistent with the theory that the level of cumulative growth reflects growth of the cohort as a whole while the spread captures the growth of fast growing establishments relative to their slow growing counterparts. As the whole cohort grows, establishments are more likely to be performing well and hence less likely to exit. A larger spread in the relative performance of establishments, holding the level of growth across the cohort constant, implies that poor performing establishments grow even slower which is likely to lead to more exit. A 10 per cent increase in the spread of cumulative growth is linked to a 1.5 per cent increase in exit rates. To interpret this, going from the 10th percentile to 90th percentile industry-year observation in terms of the spread in cumulative growth is associated with a 24.8 percent increase in exit rates. The effect from the level of growth is even larger. A 10 per cent increase in the level of cumulative growth decreases exit by 40 per cent with the difference between the 10th and 90th percentile observation in terms of the level of cumulative growth representing a 39.7 per cent change in exit rates.

Table 2.1: Fixed effects regression

Dependent variable[1]	**Avg yearly exit rate < 5 yrs**
Spread btw 80th-50th pct of cumulative growth	0.150 (0.018)
65th pct of cumulative growth	-0.408 (0.062)

[1] Log-log units.

In the bulk of the literature, exit is either exogenous or the result of receiving

some shock that pushes an establishment under some exit threshold. The previous result seems consistent with the hypothesis that higher volatility in certain industries leads to a larger spread in growth rates and higher exit. However, this explanation is unlikely for two reasons. Firstly, more volatility in say, productivity shocks, increases the option value of remaining in operation, lowering the exit threshold. Therefore higher exit would be observed with lower spreads in cumulative growth. Secondly, industries with higher exit rates do not have higher volatility as measured by the standard deviation of employment. The correlation between exit rates and the standard deviation of employment is -0.02, indicating a very weak negative relationship. Including the standard deviation of employment in the previous equation results in a coefficient of 0.0001 which is not statistically different from 0. Therefore, it does not appear that differences in volatility underpin the relationship between exit rates and the spread in cumulative growth.

Instead, I interpret the relationship as evidence of some form of rivalry that affects establishment growth. If rivalrous resources are important, an increase in the speed of growth at fast growing establishments must come at the cost of slower growing establishments.[8]

Moreover, slower growing establishments are more likely to exit. To demonstrate this, I use a Probit regression which models the probability of exit as a function of an establishment's cumulative growth, their percentile in terms of cumulative growth relative to others in the cohort and dummies for industry and age. The coefficient on the level of cumulative growth is positive and significant, suggesting that the level of cumulative growth affects exit over and above an establishment's relative position in a cohort. Table 2.2 reports the average marginal effects and an interpretation based on the variation in cumulative growth. A move from the 10th to 90th percentile surviving establishment in terms of cumulative

[8]In the literal sense, all tangible factors of production are rivalrous when the supply of factors is fixed. However, for a group of establishments operating in a distinct market, resources do not need to be rivalrous within the market as resources can be reallocated from elsewhere.

growth is associated with a 6.7 percentage point decrease in the probability of exit. Taken together, these two facts are capable of explaining why larger spreads in cumulative growth are associated with higher exit rates of entrants.

Table 2.2: Probit regression on exit

	Effect on exit rates
Average marginal effect from change in cum. growth	-0.022 (0.001)
10 to 90th pctile-sized change in cum. growth	-0.067 (0.001)

[1] Dummy variables for age, industry and year included

2.2.2 Wages do not drive selection

Although these results are consistent with the effects of selection in typical models of heterogeneous establishments, it is unclear what the mechanism is that leads faster growing establishments to push out their slower growing counterparts. In standard heterogeneous firms models similar to Hopenhayn (1992), selection is driven by changes in wages or more generally, input prices. As relatively productive establishments are able to accumulate more resources, they increase the marginal products of factors of production bidding up factor prices, squeezing profits and inducing exit of less productive establishments.

To test this mechanism, I ask whether wages across industries are positively associated with exit rates or spreads in growth rates of fast and slow growing establishments. I deflate payroll at each establishment using the Consumer Price Index and calculate the employment weighted average wage across establishments. Weighting by employment is equivalent to finding the average wage across the labor force employed in a given industry. At each date, I take the mean employment weighted average wage over the next 5 years since exit rates for a given cohort are calculated up until age 5. To account for level differences in wages between industries, I regress differences in the log of wages against dummies for each year, and in logs, *(i)* differences in the spread of growth and *(ii)* the level of growth.

Table 2.3: Regression for average wages (log units)

	Effect on wages
Spread btw 80th-50th pct of cumulative growth	-0.009 (0.015)
Average yearly exit rate	-0.028 (0.022)

[1] Dummy variables for year included.

Theory would suggest a positive relationship between wages and both variables. A larger spread in growth rates between fast and slow growing establishments and higher exit, both reflecting greater intensity of selection should lead to increases in input costs and hence increases in wages. The results in Table 2.3 instead show a weak negative relationship suggesting that wages do not increase when the spread of growth or the exit rate increases. Of course, because members of a cohort in an industry are not necessary geographically concentrated, one may argue that wages in manufacturing are more or less set at the national level and the lack of correlation between wages and moments that reflect selection is unsurprising. Regardless, the conclusion is the same. Wages are not the channel by which fast growing establishments push out their slower growing counterparts.

One potential explanation that warrants discussion is that the link between the spread of growth and exit simply reflects a process where establishments draw different "trends" in productivity. In this environment, higher exit of a cohort occurs when these trends in productivity span a wider range, with some establishments experiencing larger growth and others declining into the exit threshold faster. While I cannot rule out this theory with the data used in this paper, there is supporting evidence that physical productivity is relatively unimportant in describing exit relative to demand side considerations. Foster et al. (2008) show that the implied 1 year persistence of demand is much higher than of physical productivity, 0.97 to 0.8 and that the level of productivity matters much less than demand in exit. Exiters are only 3 per cent less productive than incumbents but have 64 per cent less "demand" than incumbents. These facts imply that demand

side considerations are likely to explain the relationship between the spread of growth rates and exit, rather than a story based on productivity trends.

To provide a theory of selection that does not depend on factor prices, I develop a model in Section 3.3 where changes in demand-side behavior are an important channel for variation in selection. Instead of increases in factor prices inducing exit, customers can endogenously increase their expectations of the value of from matching and reject matching with some establishments, inducing exit. The idea is that in the data, average entrant quality varies from year to year altering the intensity of selection across time. In years with an above average quality cohort of entrants, the intensity of competition for customers increases inducing exit. Although the model does not feature this type of aggregate uncertainty, the selection mechanism is the same.

2.2.3 Evolution of cumulative growth

Figures 2.2 and 2.3 depicts the evolution of cumulative size over the lifecycle for both the US and Chile, where as discussed above, I use sampling weights to adjust the industry composition to mimic the composition in the US. Bear in mind that the units of measurement are each country's median size by employment. For example, suppose that the median establishment has 25 employees. The top half of Figure 2.2 suggests that the median establishment in the US adds roughly 0.75×25 workers after 10 years of operation. Viewed on their own, the graphs in Figure 2.2 show that the cumulative growth distribution fans out over the lifecycle, much like the size distribution does in Cabral and Mata (2003). However, what is interesting is the differences between this fanning out between the US and Chile. After 10 years, the 90th percentile establishment in the US has added 5.25 median establishments worth of workers more than their counterparts in Chile. This positive difference between the fanning out of the distribution in the US and Chile occurs at all percentiles and is monotonic in the percentile. For instance,

the 99th and 25th percentile difference between the US and Chile is X and Y respectively.

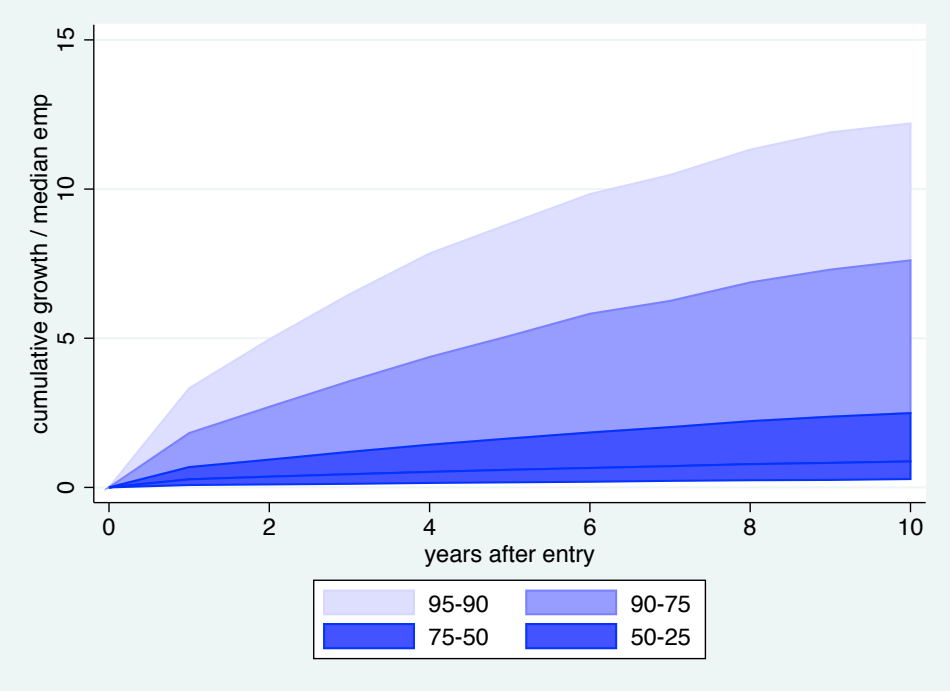


Figure 2.2: Evolution of cumulative growth in USA

There are one obvious potential explanation which goes against this interpretation. It may be that growth over the lifecycle may be less important in Chile because high quality establishments have a higher initial size relative to the US and selection forces upon entry are stronger in Chile. Hence observing that the right tail fans out less implies little about allocative efficiency since there is little to no growth to be done in Chile. To show that this is not the case, it is simple to check the size of entrants in both countries. In fact, the median US entrant is *smaller* than the median Chilean entrant.

Table 2.4: Entrant size distribution

Initial size	**25th**	**50th**	**75th**	**90th**	**95th**
US	7	16	40	92	150
Chile	10	17	35	73	137

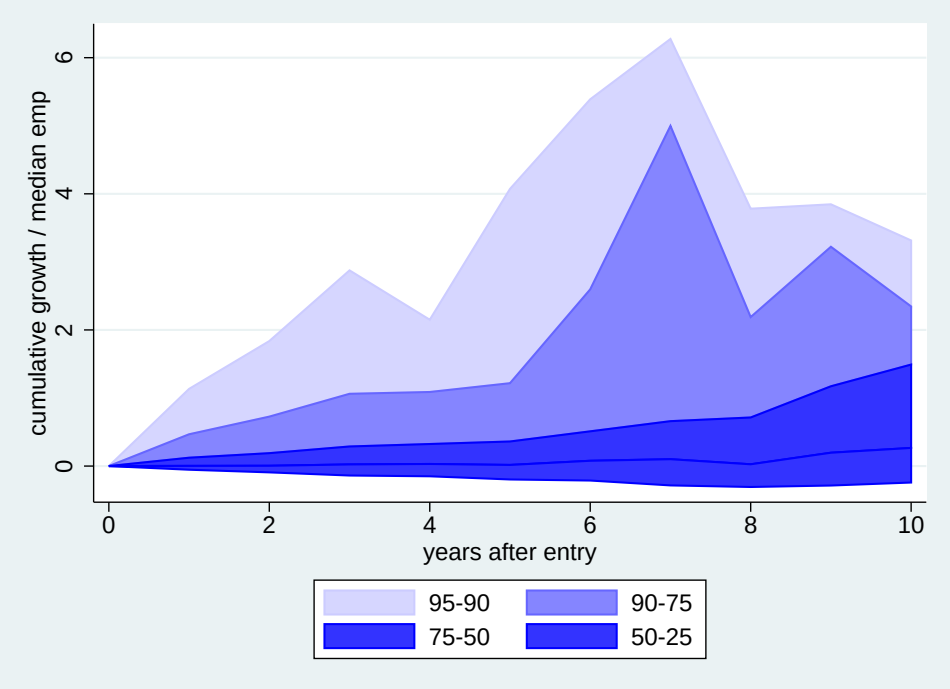


Figure 2.3: Evolution of cumulative growth in Chile

2.3 Model

Two stylized facts were presented in the previous section regarding the growth of young manufacturing establishments in the US : (i) increases in the difference between cumulative growth at fast and slow growing entrants is associated with higher exit rates and (ii) exit induced by a larger gap between fast and slow growing entrants is not caused by higher wages. In order to understand how these patterns arise and to quantify their effects on allocative efficiency, I develop a model which is capable of replicating these patterns while matching a rich set of stylized facts on establishment dynamics and price setting.

The model is essentially a continuous time version of a standard heterogeneous firms model where firms produce homogeneous goods of varying quality with two additional features. To mimic the productivity advantage that entrants have, the mean of the quality distribution from which entrants draw from is continually

growing. This induces endogenous exit via vintage capital effects whereby all establishments are eventually superseded by entrants. Secondly, establishments need to match with customers in long lasting relationships to sell their output. These frictions in product markets capture the idea that demand side frictions may be a source in reducing allocative efficiency. To increase the rate of matching, establishments can lower prices which increases consumer surplus.

Establishments produce a good using labor which differs in the amount of utility it delivers to consumers.[9] The simplest interpretation of this is as differential quality across establishments but technically boils down to output sold in bundles. There exist a mass, L, of individuals who derive utility from the quality of goods net of prices (linear utility) and earn a wage via a frictionless labor market. In order to consume, individuals need to match with establishments. Similarly, establishments need to be matched with a customer to sell a unit of output and hence set prices taking in account the effect on their own matching intensity.

There exists two sources of uncertainty that an entrant faces. Upon entry, entrepreneurs draw a permanent technology for producing goods of a given quality. Search frictions in the product market also induce uncertainty as matches are governed by a Poisson arrival rate. Establishments exit after enough time has passed for the growing technological frontier and the associated entrants to render their technology obsolete. After enough time, establishments can not jointly make positive profits while offering a positive surplus to continuing or new customers. New establishments continually drive out the old by increasing consumers' outside option in a continual process of creative destruction.

Growth is exogenous and comes from continual growth in entrants quality distribution. Entrants born at time t draw from a Pareto distribution with tail parameter ζ and a lower bound e^{gt}. Given the properties of a Pareto distribution,

[9]This is equivalent to an environment with homogeneous goods, where each establishment differs in the amount of goods produced by a single unit of labor with establishments forced to sell goods in bundles.

mean quality of entrants also grows at rate g.

$$G(z;t) = 1 - \left(\frac{e^{gt}}{z}\right)^{\zeta} \tag{2.1}$$

Later on, the model is solved around the balanced growth path to avoid keeping track of time as a state variable and establishments are identified by their "relative quality at entry", q. An establishment born with quality z and at time t_0 has relative quality at entry, or relative to its cohort, equal to

$$q = \frac{z}{e^{gt_0}} \tag{2.2}$$

Given the existence of search frictions in the product market, the key building block of the model is the consumer surplus from a match between a consumer and establishment. Upon matching, the establishment offers the potential customer a path of prices with termination of the match possible by either party at any time. The surplus of a match represents the discounted presented utility from consumption net of the agreed-upon path of prices, the discounted value of being an unmatched consumer upon termination of the match, accounting for the foregone value of being an unmatched consumer at the beginning of the match. Letting S denote the value to an unmatched consumer, the surplus of a match at an a-old establishment born at t_0 offering the path of prices $p(t)$ is

$$V_j(z,a,t_0|p) = \max_A \int_a^A e^{-r(t'-a)}(z - p(t_0+t'))dt' + e^{-r(A-a)}S(t_0+A) - S(t_0+a) \tag{2.3}$$

Continual entry by new establishments with on average better production technologies implies that as enough time elapses, the value of continuing the match becomes negative. With enough time, an establishment cannot simultaneously lower prices enough to prevent customers from returning to search and make positive profits. More formally, the value of search, S, increases continually while

utility net of prices, $z - p(t)$, cannot increase forever. Although fixed costs are absent and the establishment could choose to remain dormant in perpetuity, I define this as establishment exit.

Taking first order conditions, it is clear that the termination age of a match is the point in time at which the marginal utility net of prices. Define this common termination age as the date of exit for the establishment.

$$z - p(t_0 + A^*) = rS(A^* + t_0) - S'(A^* + t_0) \tag{2.4}$$

Since the mean of entrants' quality distribution is growing, I will solve the model around the balanced growth path (BGP). For simplicity, assume that the path of prices that establishments set are (i) customer specific and (ii) grow at the aggregate growth rate, g, from some initial price set upon initiation of the match, p. Note that since establishments match with many customers over their lives, the set of prices that an establishment charges will evolve over time. These assumptions on price setting make finding the BGP straightforward. To find the BGP, assume that the value of being unmatched, S, grows at some arbitrary rate $\hat{g}$ and normalize prices by this growth rate. That is,

$$S(t) = Se^{\hat{g}t} \tag{2.5}$$

Combining equations (2.2) and (2.5) gives the equation that pins down the termination date of a match as a function of an establishment's relative quality at entry and prices.

$$q = (p + (r - g)S)^{\hat{g}(t_0+A)-gt_0}$$

For a balanced growth path to exist, it must be the case that the date of birth is irrelevant for the termination date of a match, conditional on matching with an establishment of the same relative quality and birth, setting the same detrended

price. That is, two establishments born with the same relative quality at birth and setting the same inflation-adjusted prices but in different periods should have the same duration of a match which implies that $\hat{g} = g$. Note that all costs denominated in terms of the numeraire good have to grow at rate g, otherwise they will converge to zero, relative to the BGP. Using the fact that non-stationary variables on the BGP grow at rate g, it is simple to rewrite the consumer surplus from a match as

$$V_j(q, a|p) \equiv \frac{V_j(z, a, t_0|p)}{e^{gt}} = \int_a^{A(q)} e^{-(r-g)(t'-a)} \left(qe^{-gt'} - p \right) dt' + Se^{-(r-g)(A-a)} - S \tag{2.6}$$

where the age of the establishment at which the match is terminated as

$$A(q|p) = \frac{1}{g} \log \left(\frac{q}{p + (r-g)S} \right) \tag{2.7}$$

Holding prices fixed, the surplus from a match with an establishment is increasing in q as the flow utility from consumption depends linearly on q and decreasing in a as the level of quality relative to current new entrants declines holding q fixed. An establishment with a higher relative quality at entry exits later, while increases in the value of an unmatched consumer decrease the lifespan of all establishments. The lower bound for relative quality at entry, or the quality cutoff, is the value of q that leads establishments to exit instantly, when setting prices at marginal cost, $A(\underline{q}|mc) = 0$. That is, the point at which an establishment cannot simultaneously get customers to accept a match and make positive profits on the match.

$$\underline{q} = c + (r-g)S \tag{2.8}$$

Equation (2.8) states that selection is affected by two channels. The first is the standard input cost channel. As the cost of production increases, the lowest quality establishments are forced to exit.

Matching between consumers and establishments is governed by an establishment level matching function. Each establishment's matching intensity, $m(.)$, depends on the surplus offered to customers and a term that reflects the strength of competition in aggregate.

$$m(V_j(q,a|p),V) = \lambda u \frac{V_j(q,a|p)^\theta}{\mathbb{V}} \tag{2.9}$$

where $\mathbb{V}$ is defined as

$$\mathbb{V} = \int_{\bar{q}}^{\infty} \int_{0}^{A(q)} V_j(q,a|p)^\theta dF(q,a) \tag{2.10}$$

and $F(q,a)$ is the stationary measure of establishments over the quality-age space. $\mathbb{V}$ captures the rivalrous nature of growth. To increase their own rate of matching, an establishment can lower prices which slows down other establishments rate of matching. The degree to which an establishment can crowd out other establishments is governed by θ. In the limit, as $\theta \to 0$ approaches zero, establishments are unable to control their own rate of matching and matching is completely random. As $\theta \to \infty$, matching is completely directed. Since production is CRS, the establishment offering the maximum consumer surplus absorbs all matches with customers. Hence the parameter θ controls the degree to which search is directed. It is likely that θ is finite since in reality since factors such as geography limit establishments' ability to reach geographically dispersed customers, limiting the scope of establishments to expand instanteously.

Among a group of establishments of a common age, the largest establishments grow relatively faster; a seeming contradiction of the stylized fact that conditional on age, there is no connection between size and growth rates. However, growth here is defined as the change in customers, not as the percentage change in the number customers. While large establishments will increase their customer base the most in terms of raw customers, it is unclear whether growth rates are higher

as both the numerator and denominator are larger. Moreover, since young establishments begin with zero customers, their growth rate is relatively high compared to older establishments, matching the faster growth rates of surviving young entrants documented in Haltiwanger et al. (2012).

Upon matching, customers enter into a long term agreement with an establishment that specifies the path of prices. Each period prior to termination, they exchange payment for the good and when the match is terminated, they re-enter the search process. The establishment takes into account the date at which a customer will optimally terminate the match and sets prices accordingly to maximize prices. Producing a single unit of output requires c units of the final good which implies that expected discounted profits are simply the expected number of matches times the discounted net profit earned over the life of each match.

$$\pi(q,a) = \max_p m(V_j(q,a|p),V) \int_a^{A(q,a|p)} e^{-(r-g)(t-a)}(p-c)dt \tag{2.11}$$

Decomposing profits into net profit per match, matches and the effect of discounting and growth over the duration of the match

$$\pi(q,a) = max_p m(q,a,p)(p-c)\frac{1-e^{-(r-g)A(q,a|p)}}{r-g}$$

one can write the pricing rule as a markup rule which can be solved recursively.

$$p^{k+1} = c + \frac{1-e^{-(r-g)A(q,a|p^k)}}{-e^{-(r-g)A(q,a|p^k)}A'(p) + \frac{\theta}{V_j(q,a|p)}\left(1-e^{-(r-g)A(q,a|p^k)}\right)} \tag{2.12}$$

Theorem 1. *For a given S, there exists a solution to the pricing problem for every establishment.*

Proof. Note that the RHS of equation (3.9) is continuous in p. As $p \to 0$, the RHS of equation (3.9) is strictly positive. Similarly, as $p \to \infty$, the RHS of equation (3.9) converges to 0. By continutuity, there exists at least one solution. □

By inspecting equation (3.9), it is clear that high quality producers set higher prices while increases in the crowding out parameter, θ, lead to lower markups. The intuition is that as search becomes more directed, the marginal cost of raising prices increases as the elasticity of an establishment's demand effectively increases while the marginal benefit is more or less the same. Hence prices fall as establishments find discounting more effective in acquiring customers.

Given prices , the value of search to an unmatched customer represents the expected utility of consumption over the lifespan of the match net of prices accounting for the intensity of matching, the present discounted utility of re-entering the search process upon termination and finally, the increase in the value of search from growth in the value of search if one remains unmatched. Hence, the value to a customer of being unmatched is

$$rS = \lambda \int \int m(V_j(q,a|p(q,a)),V)V_j(q,a|p(q,a))dF(q,a) + gS \tag{2.13}$$

where r reflects the rate of time preference. The first term is the surplus from a match at a given establishment, adjusted by a term proportional to the probability of being matched to that establishment times the intensity of matching. The final term reflects growth of S. Note that if $r < g$, the value of being unmatched increases faster than the rate of time preference implying that it is optimal for customers to remain unmatched, a standard quirk of infinite horizon models.

It is worth noting that although all matches are random, customers are satisfied with their match. Ex-ante, they would prefer to be matched with establishments offering the largest consumer surplus but often accept a match delivering less utility since search takes time and is costly. Establishments exit when customers would rather search again than being matched with them. The degree of randomness in matching and the consequent satisficing is directly affected by the extent of search frictions. As search frictions lessen, establishments offering

relatively large consumer surplus can crowd out other establishments for matches more easily, leading to less satisficing in equilibrium. Search frictions in product markets may weaken as information constraints are relaxed or geographic barriers to trade are removed.

Since matching of customers is random and establishments are continually exiting, there exists a constant mass of individuals who are unmatched. Let u represent the mass of unmatched individuals which satisfies

$$L - u = \int_{\underline{q}}^{\infty} \int_{0}^{A(q)} \int_{0}^{a} m(q, a') da' dF(q, a) \tag{2.14}$$

where $m(q, a)$ denotes establishment matching intensities accounting for equilibrium prices. Equation (2.14) states that the mass of unmatched individuals is the mass of customers minus those that are currently in matches. The number of customers at an a-old establishment with relative quality at entry q is the integral of their matching function over the age domain from birth to age a. Therefore, the mass of matched customers is just the integral of the mass at each establishment over the age-relative quality space taking into account the stationary measure of establishments.

Theorem 2. *The mass of unmatched workers, u, that solves equation* (2.14) *lies between* $(0, L)$ *if there is some positive mass of entrants.*

Proof. Let $\hat{m}(q, a)$ denote matching intensities divided by u. Define

$$\kappa = \int_{\underline{q}}^{\infty} \int_{0}^{A(q)} \int_{0}^{a} m(q, a') da' dF(q, a)$$

and note that if there are a positive mass of entrants, $\kappa > 0$. It is straightforward to show that $u = \frac{1}{\kappa+1} L$. □

Labor market clearing equates labor demand with labor supply. Bear in mind

that the labor market is frictionless and always clears in equilibrium.

$$L = \int_{\underline{q}}^{\infty} \int_{0}^{A(q)} \int_{0}^{a} m(q, a')da'dF(q, a) \tag{2.15}$$

Establishment entry is governed by a free entry condition. To create an establishment, entrepreneurs require c_e units of the final good. Payment is made before receiving the permanent technology draw which leaves open the possibility of entry and instantaneous exit. Entrepreneurs who draw quality below the cutoff level decide to leave instantly where the cutoff satisfies equation (2.8). The free entry condition states that the expected value of creating a new establishment including the entry cost must be weakly negative. The expected discounted value of an establishment born with relative quality q is

$$V_f(q) = \int_{0}^{A(q)} e^{-(r-g)a}\pi(q, a)da \tag{2.16}$$

The expected value takes into account discounting, expected growth and matching at each age due to the randomness of matching and finally, the establishment's profits per match. The free entry condition can therefore be written as

$$w = \int_{\underline{q}}^{\infty} V_f(q)dG(q) \tag{2.17}$$

Note that the free entry condition equates expected profits to the cost of entry which implies that some entrants make losses. Because the entry cost is denominated in the numeraire good which must be purchased via search, there is an underlying assumption that individuals finance entrants in exchange for an equal portfolio of shares in all establishments. On average, individuals make zero profits from this investment which does not affect their decision making since they are risk neutral.

The final piece of the model is the stationary distribution over the quality

and age dimensions. Firstly, because exit is deterministic, there are equally as many young firms with a particular relative quality at entry q than there are old firms. That is, there are equal masses of 5 year old establishments born with some arbitrary quality q as there are entrants with q. Note that these two groups have different levels of non-detrended quality. There are relatively few old firms as in the data because the probability of drawing a technology for high quality output is low and diminishes faster than lifespans increase with q. Hence the model is capable of matching the data along the age dimension without exogenous exit of establishments. As age is a continuous variable and entry occurs continuously, the pdf of an a-aged firm born with q with a less than the maximum age $A(q)$ is the entry rate times the likelihood of drawing q at birth. If a is greater than $A(q)$, then $f(q, a)$ is 0. Otherwise for $a < A(q)$,

$$f(q, a) = eg(q) \tag{2.18}$$

2.4 Estimation

Before proceeding to discuss how the model's parameters are estimated, it is necessary to discuss the dimensions that the model will be unable to fit the data. Since quality draws are permanent, there is obviously no scope for the model to match any data on incumbent innovation. The model generates growth in the aggregate purely through innovation by entrants. As entrants draw quality from a distribution whose mean is continually growing, it may appear as though quality is declining with age, which is completely at odds with the data. However, because exit is endogenous and tied to establishment production quality, only establishments with high enough quality technology survive past a certain age. An establishment needs a relative quality at entry greater than $\underline{q}(a)$ to survive

past the age a.

$$\underline{q}(a) = \underline{q}e^{ga}$$

Hence as age increases, exit truncates members of the cohort from the left leaving behind relatively productive establishments. Exit exactly offsets the growth in entrants quality distribution such that the mean level of quality is flat across each age group. To see this, note that the level of quality for an a-old establishment born with relative quality q is

$$z(q, a) = qe^{-ga}$$

and convert the minimum level of relative quality at age a into non de-trended quality

$$z(\underline{q}(a), a) = \underline{q}(a)e^{-ga} = \underline{q}$$

which is the same lower bound as the current entrant distribution. Moreover, depending on the relative growth rates of high and low quality establishments, it may be the case that the *size-weighted* mean level of quality increases with age, even though there is no incumbent innovation. This provides an alternative interpretation of positive sloping productivity-age profiles that weight establishments or firms by employment. The positive slope may simply reflect greater weighting of establishments with high productivity over time coupled with continual exit of low productivity establishments.

In reality, exit is not deterministic with some low quality establishments surviving longer than expected and vice versa for some establishments producing high quality output. The absence of both relatively low quality establishments that persist in spite of their low productivity and the lack of failures among high quality establishments imply that the quality-age correlation in the model is too high. Note that this does not necessarily invalidate the earlier point made before that the mean quality-age profile is flat because it is a statement regarding the

composition of exiters and survivors.

Given the limitations of the model, the strategy in this section is to calibrate parameters that are standard and well identified in the data and then estimate the rest of the model to moments in the data that have a clear purpose in identifying parameters. The annual discount rate in the model is set to 3 per cent corresponding to a real interest rate of 3 per cent. Growth in the mean of the quality distribution is set to match the 1.2 percentage points entrants contribute to yearly productivity growth in manufacturing (Foster et al., 2008). By definition, since entrants draw quality randomly upon entry, some establishments detract from aggregate quality growth while others contribute most of the growth. Finally, I normalize the mass of customers and workers, L, to 1 since the mass of workers just scales the economy.[10]

The remaining parameters are estimated to minimize the distance between moments implied by the model and from the data (method of moments). To weight moments equally, I scale each moment such that it reflects the percent deviation from the moment in the data. Ideally, moments would directly identify a single parameter but because the model is highly non linear, most of the following moments identify more than one parameter.

A key parameter of the model is the parameter which reflects the degree to which the matching function resembles directed search. Increases in the crowding out parameter, θ, reflect more directed-like search which disproportionately benefit high quality producers who offer customers the largest surplus. Low quality producers receive a smaller share of matches with customers and hence, the spread or skewness in growth increases. These effects of the crowding out parameter map directly to three moments: the larger markups of older establishments relative to younger establishments, the dispersion in prices and the evolution of the size

[10]It is easy to show that the solution of the model is not affected by L. Although moments relating to levels such as output changes, relative differences such as the percent decrease in output from changes in parameters remain the same.

distribution over the lifecycle.

To capture the spread in growth across fast and slow growing establishments, I target the cumulative growth in employment of survivors after 10 years adjusted for median employment. Although it is easy to calculate cumulative growth or simply size as all establishments start at the same age, the median is slightly more difficult. The median establishment by employment is not simply the employment at establishments born with median quality technology because there is aggregate growth and even if aggregate growth were absent, establishments born with a relatively high quality technology last longer than those born with relatively low quality technology. For example, although many entrants are born with relative quality at or below the exit threshold, none of these establishments count in the calculation of the median since they exit instantly. To find the median, I calculate the stock of customers at each establishment within each group of establishments born with a particular relative quality and of the same age. Because matching is random, I take the mean size within each group and sort the groups by mean size. Finally, I accumulate groups in ascending size order and note their size until I reach a mass of groups that accounts for 50 per cent of all establishments. The mean size at the group that is just over 50 per cent is taken as median employment. The spread of growth is calculated as the difference between the median and 90th percentile establishment after 10 years, relative to median employment across all establishments. The spread in US data from Figure 2.2 is 6.82, implying that after 10 years, the 90th percentile surviving establishment is 6.82 median establishment's worth of workers larger than the 50th percentile surviving establishment.

The evolution of the cumulative growth distribution after 10 years is also informative about the speed of growth. After 10 years, the median surviving establishment is 0.82 times the size of the median establishment across all establishments. I use this moment to help identify the scale of the matching function, λ. The

functional form of establishment-level matching functions imply that aggregate matches are equal to λu, with the matches formed in aggregate apportioned by the relative consumer surplus offered by an establishment. Therefore, increases in the matching function scale parameter scale up the growth of all establishments.

In the model, markups rise at older establishments due to the different composition of tenure among their customers. For a cohort of entrants, exit occurs via truncation from the left of the quality distribution. As truncation occurs at the rate of aggregate growth, all cohorts of entrants share the same distribution of the (non detrended) quality distribution, with the caveat that the mass of older establishments is smaller due to exit. Consequently, average prices set for new matches are identical across cohorts, no matter how old the cohort. However, average or customer weighted prices at older establishments are higher since the prices at an older establishment include the higher prices for customers who matched back when the older establishments were younger and offered larger consumer surplus.

Foster et al. (2008) use data that separates prices from physical quantities to show that entering establishments in manufacturing charge prices that are on average 3.3 per cent lower than their ten year old counterparts. The intuition being that entrants need to set lower prices to attract customers while incumbents have a customer base and have less need to offer discounts. To calculate the model analog, I follow Foster et al. (2008) and define the price at an establishment as the quantity/customer weighted average and then average across establishments within a cohort by revenue weights.

The crowding out parameter affects the dispersion of prices. As the crowding out parameter increases, equation (3.9) shows that markups fall, collapsing the support of prices as competition for customers becomes fierce. Ideally, the dispersion in prices would simply be the standard deviation in prices. However, prices change for many reasons that are completely unrelated to the pricing decision in the model. For example, idiosyncratic shocks to costs are likely to induce move-

ments in prices that increase the standard deviation. To account for this, I try to match half of the 18 per cent dispersion in prices documented in Foster et al. (2008) based on the observation in Kaplan and Menzio (2014) that roughly half of the variation in prices across retail stores due to store-area amenity effects are stable.

Because the quality distribution entrants draw their technology from is a Pareto distribution, a single tail parameter, ζ, characterizes the quality distribution. A higher tail parameter spreads out the quality distribution increasing both the mean and variance of quality among entrants. The first clear effect on moments in the model is in the skewness in growth and size. Larger variance in relative qualities increases the variance in the surplus of matches and hence growth rates. Part of the identification comes from the 90-50 spread discussed above, which is a measure of skewness in growth.

A less obvious effect of changes in the Pareto tail parameter is on exit rates. As the tail parameter increases, entrants are on average more productive and displace more incumbents. One can show that the mass of establishments is equal to

$$M = e\underline{q}^{-\zeta}\frac{1}{g\zeta} \tag{2.19}$$

Given that a mass e continually enters and $\bar{q}^{-\zeta}$ of a cohort does not instantly exit, $g\zeta$ reflects the exit rate of the cohort of establishments excluded those that do not instantly exit. Alternatively, $\frac{1}{g\zeta}$ is the average age of establishments that survive past birth. The growth rate of the entrant distribution affects exit as higher growth rates of entrants' productivity distribution increases the mass of establishments with better productivity than incumbents and hence displacement. In the model, there is a bunch of exit at birth and then a constant hazard of exit thereafter whereas the data show a smoothly declining hazard of exit. The smoothness in the decline in exit rates in the data probably reflects some element of learning

which would delay part of the instantaneous exit. Hence there are two measures of exit, depending on whether one includes instantaneous exit. I match the the constant hazard of exit in the model against the average exit rate of establishments in the data, 0.11 and the five year exit rate of a cohort which is a admittedly noisy estimate of instantaneous exit.

Table 2.5: Moments from the data and the model

Moment	**Data**	**Model**
Std dev. in prices	0.09	0.07
Δ% price at 10 yrs	0.03	0.04
Med. cum. growth at 10 yrs/med. emp.	0.89	0.88
90-50th pct cum. growth at 10 yrs/med. emp.	6.74	6.91
Avg. exit rate	0.11	0.09
5 yr exit rate	0.55	0.54

Sources: US Census Bureasu Synthetic LBD, US Census BDS, Foster et al. (2008) and Kaplan and Menzio (2014).

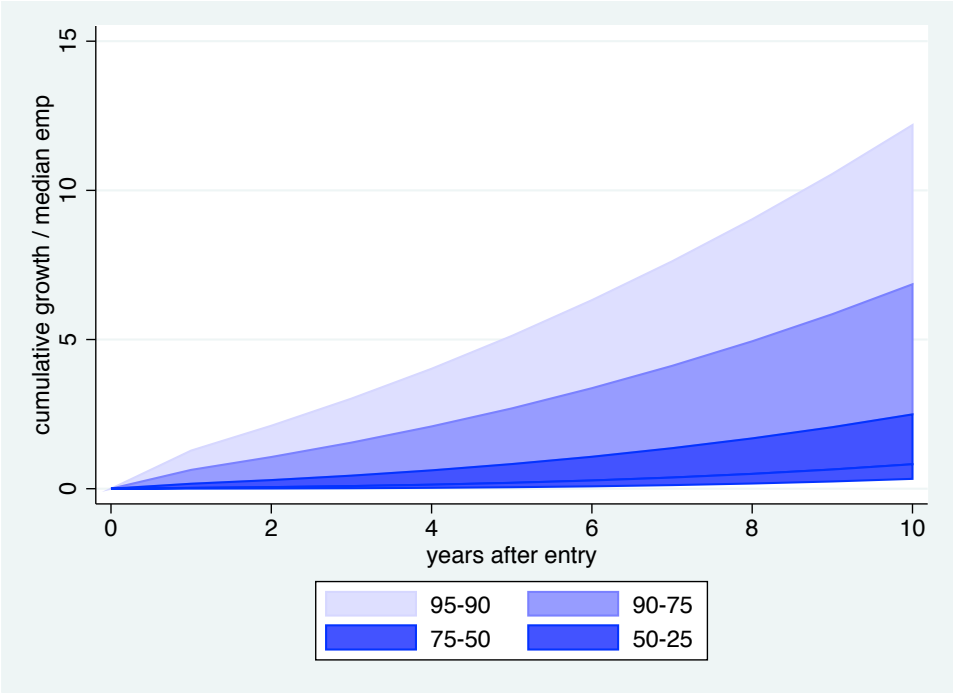
A summary of these moments from both the data and the model at the estimated parameters are shown in Table 2.5. Overall, the model does a good job of matching moments from the data. Both measures of prices from the model come close to matching their counterparts in the data. The standard deviation in prices in the model is 7 per cent, only slightly lower than the 9 per cent observed in the data. The difference between prices at medium-aged incumbents and entrants is slightly higher in the model, 3.8 per cent compared to 3 per cent in the data. The median surviving establishment is 0.88 times the size of the median establishment, almost identical to the value of 0.89 in the data suggesting that the model captures the median level of growth well. Similarly, the spread between the 90th and 50th percentile establishment in terms of cumulative growth is 6.74 in the data after 10 years, and 6.91 in the model. Finally, the model comes close to matching the data in terms of exit rates. The five year exit rate in the model, which includes instantaneous exit in the model, is equal to the value of 0.55 in the data while the average exit rate, excluding instantaneous exit is also close at 0.09 compared to 0.11 in the data. The calibrated and estimated parameters are

shown in Table 2.6.

Table 2.6: Estimated and calibrated parameter values

Parameter	Calibrated Value
r - real interest rate	0.03
L - mass of workers	1
g - growth of entrant Productivity distribution mean	0.012
	Estimated Value
ζ - tail parameter for Pareto distribution	7.72
β - entry cost	0.27
ϕ - marginal cost	0.61
λ - scale of matching process	7.48
θ - curvature in matching process	0.93

2.5 Results



In this section, I ask what implications do differences in the directed of customer search have for inferences about allocative efficiency? Differences in the directedness of customer search in product markets are likely to be driven by

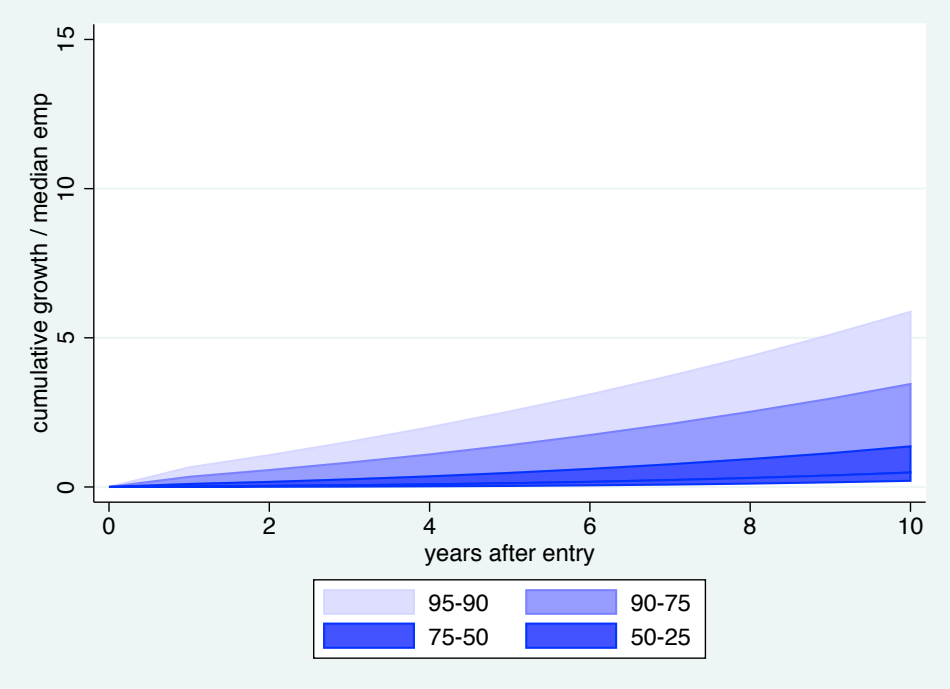


Figure 2.4: Spread of growth estimated for US (top) and Chile (bottom)

technology. One example of lessened search frictions is the ability of consumers to research products online and to purchase products from geographically distant locations without incurring any cost of travel. This technological improvement is driven by the availability of the Internet to both consumers and retailers and the existence of a well developed package transportation network, which is particularly streamlined in the US relative to developing countries.

To discipline the magnitude of the change in the directedness of customers' search for products, I reduce the degree of directedness in search to match the lower spread of growth between the 90-50th percentile of entrants in Chile. The 90th-50th percentile spread after 10 years in Chile is 3.05 median establishments which I match by reducing the parameter θ from 0.93 to 0.52. This counterfactual overstates the effect of search frictions in product markets since it is unlikely that differences in establishment lifecycles between Chile and the US are entirely due to frictions in product markets. Hence the results here represent upper bounds

for the effect that frictions in product markets can generate.

As stated earlier, reducing θ leads to a matching function for establishments that more closely resembles random search whereby matches with customers are generated regardless of the utility offered to customers. Although customers do not have to consumate these matches, reallocation towards lower quality establishments nonetheless because rejecting a match leads customers back into a search process that is random. Customers know that there are better establishments in existence, but choose to remain with suboptimal matches since it is difficult to find these better establishments. Hence customers choose to remain with lower quality establishments than if search were directed, leading to reallocation of production to low quality establishments. Moreover, this mechanism affects selection since some establishments who would otherwise be ignored by customers and forced to exit, now attract a small number of customers and remain in operation.

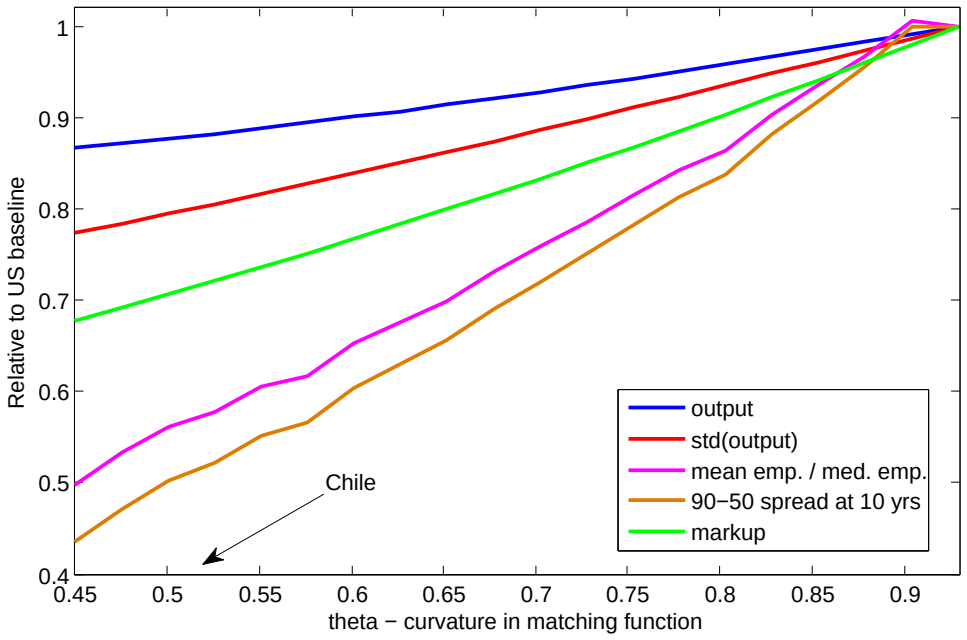


Figure 2.5: Effect of θ on moments relative to values at estimated parameters

Figure 2.5 depicts the moments of the model calculated at various values of

θ, starting at the estimated value of 0.93 and ending at the estimated value for Chile, 0.52. Each point on the line represents the value of a moment from the BGP, or stationary equilibrium, of the model at a given value for θ. It is clear from the figure that the reduction in θ generates a large response. As mentioned earlier, as the spread of growth relative to median employment falls, output net of entry costs falls as production and customers are reallocated from high to low quality establishments. This reallocation additionally manifests as a reduction the skewness in employment across establishments, which can be inferred from the shrinking in the gap between mean and median employment. The customer-weighted average markup falls as low quality producers, who set lower markups, receive more customers and the market and hence pricing power of high quality producers falls. Using the spread in growth rates in Chile to discipline the reduction in θ, the model suggests that output falls by roughly 14 per cent due to such a change.

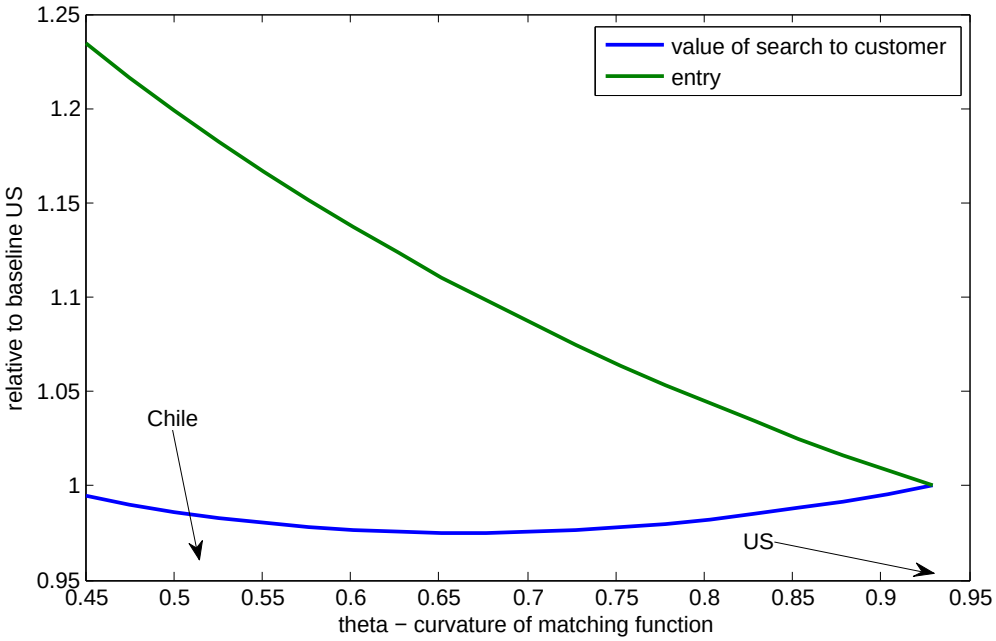


Figure 2.6: Effect of θ on model relative to values at estimated parameters

The reallocation of labor towards relative high quality producing establish-

ments is evident in the roughly 50 per cent reduction in the spread of growth rates in Figure 2.5. However, note that selection only plays a limited role as the exit threshold does not fall by much. To see this, remember that the exit threshold is an affine function of the value of search to an unmatched customer, which is U-shaped with respect to θ and only falls by 2.5 per cent at the most. The hump shape can be explained by two counteracting forces. As θ increases, customers find it easier/faster to match with relatively high quality establishments which is a force that increases the value of search. However, the flip side of this reallocation is greater market power for high quality producer and hence greater pricing power. When matching becomes easier for high quality producers, the tradeoff between customer acquisition and margins tilts towards increasing margins at the cost of accumulating customers leading to an initial decline in the value of search for customers. As θ increases further, the increase in the rate of matching with high quality producers overcomes the increase in markups and hence the value of matching rises. Another way of seeing this is to note that customers disutility from prices is linear in prices while the matching intensity essentially increases at an exponential rate with θ. Hence as θ becomes larger, the value from increased matching intensities at high quality establishments offsets the linear increase in markups.

Offsetting the negative effects on allocative efficiency are the responses of establishment entry. By construction, as the expected value of creating an establishment is finite, the Pareto distribution for qualities declines faster than the discounted value of operating an establishment with at birth relative quality, q, increases. Therefore, any reallocation from high to low quality establishments must increase the ex-ante value of starting a new establishment, increasing the mass of entry in equilibrium.

2.6 Conclusion

Selection plays a large role in models of heterogeneous establishments, but little is known about how selection operates in reality. In models of heterogeneous establishments, the removal of distortions or frictions that reallocate production to relatively productive establishments increases selection, inducing exit of less productive establishments. The increase in exit is typically generated by the bidding up of input costs by productive establishments. In this paper, using establishment level data in manufacturing, I identified a relationship between the spread of growth between fast and slow growing young establishments and exit rates of young establishments which I interpret as evidence of selection. However, wages do not appear to increase in periods of increased selection which is at odds with how selection operates in typical models.

I developed a model capable of matching a rich set of facts on establishment dynamics and show that changes in demand side behavior can drive selection in addition to changes in input costs. The estimated/calibrated model suggests that as much as 60 per cent of the movement in selection can be attributed to changes in demand side behavior. In particular, with search frictions in the goods market, customers willingness to refuse a match and return to search is a margin that affects selection. Establishments are forced to exit when customers reject matches with them. This result is consistent with the recent literature that emphasizes the importance of demand in understanding establishment/firm dynamics (Fishman and Rob, 2003; Foster et al., 2008, 2012).

A common feature of the early misallocation literature is that aggregate productivity essentially boils down into two components: the level of productivity, and the correlation between size and productivity which captures allocative efficiency (Bartelsman et al., 2009). Here, I take seriously the notion that establishments start small and analyse how growth patterns can be informative regarding

allocative efficiency. The main feature of the model is that large dispersion in growth rates between fast and slow growing establishments are indicative of resources are being allocated relatively efficiently. To obtain a large positive correlation between productivity and size, an economy needs relatively high dispersion in growth rates of young establishments. Note that this statement is agnostic regarding the mean or median growth rate of establishments, which may be informative about productivity growth, but less relevant in terms of allocative efficiency.

In the model, a increase in the curvature parameter (less curvature) of the establishment-level matching function, which governs the rate of establishment growth, reallocate of both production and customers to high quality establishments. However, because the reallocation towards high quality producers also has the effect of increasing their market power, markups increase at high quality producers which mitigates part of the reallocation and improvement in allocative efficiency. The estimated model suggests a reduction in the curvature parameter that matches the slower degree of fanning out of the cumulative growth distribution in Chile relative to the US would lead to a fall of 14 per cent in welfare.

2.7 Appendix

An equilibrium consists of establishments and individuals taking the following as given

- surplus from a match - $V_j(q, a)$: (2.6)
- matching technology for each establishment - $m(s, u; S)$: equation (2.9)
- aggregate marketing efforts - S: equation (2.10)
- value of owning an establishment of quality q - $V_f(q)$: equation (2.16)
- profits at establishment with quality q and age a - $\pi(q, a)$: equation (2.11)

while behaving optimally

- establishments and customers terminating matches when privately optimal - $A(q)$: equation (2.7)
- establishments exiting when they cannot match with customers - $\underline{q}$: equation (2.8)
- establishments' marketing efforts maximize profits - $s(q, a)$
- individuals reject matches and return to search optimally with full information - U: equation (2.13)

with "markets" clearing

- labor markets clear: (2.15)
- customer markets clear given search technology and marketing efforts: (2.14)
- free entry condition weakly holds: (2.17)

2.7.1 Definition of equilibrium

2.7.2 Solution algorithm

The model can be simplified into 4 equations, representing the equation that defines the mass of unmatched customers, the labor market clearing condition, the definition of the value of unmatched customers and the free entry condition.

$$u = L\frac{1}{1+\frac{\lambda}{V}\int\int\int V_j(x,a')^{\frac{\theta}{1-\theta}}da'dF(x,a)} \tag{2.20}$$

$$w = \frac{\lambda\theta\beta u}{(u-c_e e)V}\int\int V_j(x,a)^{\frac{1}{1-\theta}}dF(x,a) \tag{2.21}$$

$$(r-g)U = \lambda(1-\beta)\int\int\frac{s(x,a)^{\theta}}{S}V_j(x,a)dF(x,a) \tag{2.22}$$

$$w = \int_{\underline{x}}^{A(x)}\int_0^a e^{-(r-g)a}\lambda u\beta\frac{V_j(x,a)^{\frac{1}{1-\theta}}}{V}(1-\theta)\,dadG(x) \tag{2.23}$$

Note that there are 4 equations and 4 unknowns, U, u, e, w after substituting for the definitions of $V_j(.), A(x), S, s(x,a), \underline{x}$ and V. The algorithm is a simple fixed point algorithm

1. Guess U_0, w_0, e_0.
2. Update u_1, w_1, U_1 and e_1 via the above equations
3. Calculate difference in values and start at first step if difference large enough.

CHAPTER 3

The Aggregate Impact of Online Retail

3.1 Introduction

Probably the largest technological innovation of the past two decades is the internet, yet we know surprisingly little about its effect on welfare. A clear example of this is the development of online retail which has transformed the retail sector. Nominal sales at online retailers grew by an average of 17.5 per cent per year over the period 2000-13, compared to 3.3 per cent for the entire retail sector. As a consequence, online retailers' market share expanded from 2.7 to 17.8 per cent of total retail sales. The rapid adoption of online retail is obvious in the data yet the impact on economic welfare is unclear. In this paper, I study the impact of online retail, the embodiment of new internet technologies combined with improvements in logistics and shipping technology, on both retailers and consumers.

Understanding the impact of online retail is important for a number of reasons. First, the retail and wholesale trade sectors are large. Value added in the retail and wholesale trade sectors accounted for half of personal consumption expenditures on goods post 2000. Over and above interest in the retail sector, it offers insight into the effect of internet based innovations on the services sector, an increasingly common form of creative destruction driven by innovation from technology firms. Proposed legislation such as the Marketplace Fairness Act, a nationwide online sales tax, make understanding these issues particularly pertinent for policy.

There are two primary margins that matter when assessing the impact of online

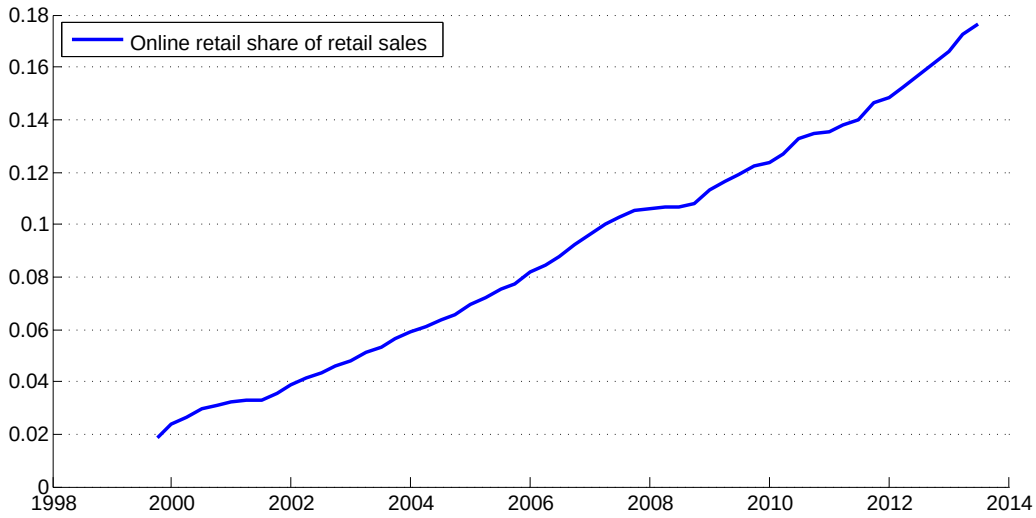


Figure 3.1: Growth of online retail sales

retail. First, consumers are likely to gain from both lower prices and the ability to shop at home, without incurring the cost of traveling to stores. Second, entry or improvements by online retailers alters the market structure of retail. Markups and sales per store fall, inducing store closures and a reconfiguring of the retailers participating in any local retail market. This extensive margin adjustment affects aggregate welfare through reallocation and also has ambiguous effects on consumer surplus as customers who are forced to shop at second-best options lose out.

Quantifying the impact of online retail requires the measurement these margins. However, both of these margins vary across markets since demand for online purchases varies across markets. Consumers without easy access to the internet and relatively high quality brick and mortar stores are less likely to seek online avenues to purchase goods. Therefore, an aggregate approach such as one based on a stylized two-sector model will not suffice. Here I explicitly account for geography and the heterogeneity of the US retail industry in order to accurately capture these margins.

In order to account for these margins properly, I estimate a spatial equilibrium model of retail with a geography based on the US retail industry. Stores exist in

their real world locations and compete for customers described by zip code level demographic data. The model is estimated on confidential store-level data from the US Census of Retail Trade, demographic data from the Current Population Survey and Economic Census, and shortest-route data collated from Google Maps.

A central component of the model involves the measurement of the consumer surplus and profits, or total surplus, generated per purchase for each store. Total surplus per store reflects the potential welfare from a particular store. The actual surplus or welfare generated by a store depends on the allocation of consumers across stores.[1] When aggregated across stores, this measure captures the value of services generated in the process of transferring goods from retailers to final consumers, holding fixed the quantity of goods. To estimate store-level consumer surplus, I rely on the spatial nature of the model to separate effects due to the competition from nearby stores and the composition of customers, from a store's sales. This methodology can be applied to infer establishment-level productivities from other service industries where geographic proximity defines sets of competitors.

To quantify aggregate gains in welfare from online retail, I use counterfactual exercises which measure the effects from shocks to the equilibrium state. Counterfactuals isolate the causal effect from online retail as they hold other factors fixed by design. The counterfactuals are based on likely drivers of the increase in the online sales share from 2007-12. They include increases in internet access, reductions in shipping time and diversification into new industries by online retailers. Increased internet access and improvements in shipping that map to the data can account for roughly half for the observed increase in online retailers' share of sales. Although improvements by online retailers lead to a large amount of store closures, the gains in consumer surplus and the savings in store operating

[1]Prices divide the surplus between store and consumers and do not affect the size of the surplus from a purchase. However, prices do affect aggregate welfare through the allocation of consumers across stores.

costs are enough to offset these losses. The 5.2 percentage point increase in online retailers' market share is associated with a 13.4% increase in aggregate welfare. While many traditional retail firms experience a reduction in profits, 8.2 per cent of traditional retail firms (inclusive of exiters) gain by absorbing part of exiting stores market share.

The final counterfactual quantifies the likely effects from the implementation of the Marketplace Fairness Act by imposing measures of state sales taxes on online purchases. I calculate the effect of the Marketplace Fairness Act by introducing the taxes after first projecting the model out from 2007 to 2012. The tax has a substantial effect in reallocating sales back to brick and mortar sales and mitigating store closures. Without the tax, sales at online stores more than double, leading to roughly 78,000 store closures over 2007-12. The tax has the effect of diverting one third of the lost sales back to brick and mortar stores preventing 18,000 store closures.

The paper is organized as follows. Section 3.2 describes the data that underlies most of the analysis in the paper. Details of the structural model are presented in Section 3.3. The description of the estimation routine for both the demand system and fixed costs and parameter estimates follow in Section 3.4. Results from the counterfactual exercises are reported in Section 3.5. Finally, I conclude in Section 3.6.

3.1.1 Related literature

3.1.1.1 Retail productivity

Both Foster et al. (2006) and Lagakos (2009) study retail productivity between "Mom and Pop" type stores and larger superstores using standard measures of productivity such as labor productivity and value added per worker. While these measures are informative at the aggregate level, they are less useful in evaluating

store level retail productivity as store level versions of these measures are clouded by effects from competition. Markups are likely to be lower in relatively competitive markets, lowering value added per worker and hence estimates of productivity in the area. Because much of the reallocation caused by online retail occurs between stores, as opposed to firms, I use a different measure of productivity which neutralizes competitive effects. This measure of store-level retail productivity captures the consumer surplus net of costs from purchasing at a particular store, a measure that is independent of prices.

3.1.1.2 Reallocation and productivity growth

In the model, technological change comes in the form of entry or improvements by online retailers. Therefore, the driving force of technological change is reallocation rather than across the board improvements in productivity. This mirrors the literature on reallocation from technological improvements for manufacturing such as Collard-Wexler and Loecker (2013) who study innovations in the U.S steel industry. However, measuring gains from reallocation is more complicated in the retail sector. Unlike manufacturing, retail productivity at the establishment level is not readily available in the data as physical units of output are not defined in retail. Moreover, geography matters. Establishments compete in many local markets rather a single aggregate market which requires one to model substitution patterns by consumers.

3.1.1.3 Spatial competition

The structural model that is presented in Section 3.3 borrows heavily from the literature in industrial organization on spatial competition which feature models of discrete choice with a geography that reflects real world locations (Davis, 2006; Chiou, 2009; Holmes, 2011). Spatial competition is generated by assuming that

customers face costs of traveling and hence prefer to shop at stores closer by, leading to a gravity based demand structure. A key difference here is that the model here is one of industry equilibrium. Firms and stores alter their behavior in response to changes in their operating environment. This is particularly useful when performing counterfactuals, which require equilibrium responses.

3.2 Data

The bulk of the data used in the paper come from the Census of Retail Trade (CRT) in the years 1997, 2002 and 2007. Although the CRT is available for every year ending in "2" or "7" dating back to 1977, I restrict the sample to 1997-2007 since online sales are virtually non existent before 1997 and the NAICS classification in the dataset is most stable across these years. Data from the CRT are derived from responses to forms mailed out to retail stores and administrative records. All establishments of multi-unit retail firms (chain stores), single unit establishments with approximately more than 3 employees and a sample of small employers receive forms requesting information. Information for stores not selected in this sample or not responding to mailed forms are derived from administrative records from other federal agencies including the IRS. Roughly 10 per cent of records are based on administrative records. The combined data represent the universe of retail stores (NAICS: 44-45 and 72) that exist in the US in a given Census year.

Data is at the store level and includes information on employment at March, annual sales, 6 digit NAICS code, geographic location (by zip code) and longitudinal firm and store identifiers. As the focus is on retail stores that can potentially be affected by online retailers, I restrict the data to those in retail industries that are traditional inventory-holding type stores. This excludes the following industries: accommodation and food services, non-store retailers, gasoline stations,

automotive dealers, art dealers and mobile home dealers.

The data used to construct measures of internet access come from the Current Population Survey (CPS) Internet and Computer Use supplements in Dec. 1998, Oct. 2003, Oct. 2007 and Oct. 2012 which I will map to internet access in 1997, 2002, 2007 and 2012 respectively. Each survey asks a variant of whether anyone within a household has access to internet at home. I describe the estimation of zip code level internet access from individual level estimates in the Appendix.

Data on commuting distances and times were collated from Google Maps in June 2013. Distances reflect the shortest path by road network, including highways, between an origin and destination pair. Travel times reflect a combination of factors including posted speed limits and average traffic levels. In all, the sample of data I use has distance and travel times for 3.2 million unique origin-destination combinations in the US.

3.3 Model

The model that follows is centered on a discrete choice framework with consumers choosing where to purchase some real unit of consumption. Consumers receive utility from making purchases with preferences dictating the utility attached to each purchase option. Taking consumers' preferences as given, stores set prices to maximize profits. The model is estimated at the 6 digit industry which implies that parameters are industry specific and agents' decisions are within industry. For instance, the utility of purchasing online is industry specific and stores only worry about competition from other stores within the same industry. Book stores are only concerned about competition from other book stores.

Stores and households exist in their real world locations with households able to shop anywhere. In practical terms, a store or household's location is the center

of the zip code tabulation area in which they reside.[2] Geography matters since consumers face a cost of traveling to stores. From this assumption, two patterns will arise. First, all else being equal, consumers will prefer to shop at nearby stores. Second, the effective set of competitors for any store are other nearby stores. Therefore, a store's market power is largely determined by its location. Being surrounded by many other stores will lower market power for a store lessening optimal markups.

The online retailer that competes with brick and mortar stores is exogenous with characteristics that are inferred from Amazon.com annual reports.[3] The utility that consumers get from shopping online is a parameter that is estimated for each year and each industry while cost and markup information is inferred from annual reports. Treating the online retailer as exogenous is a deliberate and conservative approach that makes the fewest assumptions regarding Amazon.com's behavior.

One limitation of the model is that it does not allow for cross industry substitution. For example, the choice set of consumers deciding to purchase from bookstores are limited to bookstores alone which does not explicitly allow for substitution towards, say, general merchandise stores. While it would be straightforward to extend the model to all industries and to allow for rich substitution patterns via consumer-industry interactions, it is simply not computationally tractable given the size of my dataset. To do so, it would be necessary to reduce the number of parameters that reflect firm and store level heterogeneity so that the model is closer to a model of industry rather than store choice. Instead, I lean towards estimating a simpler model incorporating within industry choices to make use of the rich store level data.

[2]Zip code tabulation areas represent contiguous geographic areas whereas zip codes are defined by the USPS and are collections of addresses.

[3]Amazon.com accounts for a sizable chunk of online retail sales, 11.2 percent of total retail e-commerce sales in 2011. Compare this to Walmart which accounts for only 5.4 percent of total retail sales in 2013.

3.3.1 Consumers

Consumers need to make multiple purchases of some real unit of consumption throughout a year. The real unit of consumption is fixed across stores in a given industry and represents a bundle of goods. For example, the real unit of consumption for grocery stores should be interpreted as a bundle of groceries. Consumers choice set includes each brick and mortar store, the online retailer (if they have access to the internet) or the option to undertake home production. Once they have chosen where to make their purchase, consumers can only purchase a single unit of consumption. This abstracts from intensive margin adjustment in the size of purchases at a store, although the decision to move from home production to a store/online retailer is akin to intensive margin adjustment.

Each consumer has preferences among these options which describe the utility conferred from purchasing at each option. Because the real unit of consumption is fixed across the options, differences in utility implicitly reflect differences in prices and the level of services, broadly defined, between options. A supermarket may be preferred to others because prices are lower and the customer service, opening hours and quality of produce is better.[4]

Denote the number of purchases per person in location l as $m_{l,t}$ which for estimation purposes is a function of time and demographic characteristics in the location.

$$m_{l,t} = \lambda_t e^{\lambda_i hhinc_{i,t} + \lambda_f family_{i,t} + \lambda_e educ_{i,t} + \lambda_w white_{i,t}} \tag{3.1}$$

A fraction of customers in each location, $\nu_{l,t}$, have access to the internet and hence access to online shopping while the remaining fraction do not.

Prior to shopping, consumers receive idiosyncratic taste shocks for each purchase option and choose the option that gives them the highest utility. Choosing

[4]This assumption regarding the same real unit of consumption across all stores is necessary as the data only reveals nominal sales at stores. It is impossible to identify differences in real purchase sizes from the data.

home production delivers utility u_h normalized to zero before the taste shock while shopping online delivers utility equal to $u_{o,t}$ if the customer has access to the internet. Denote the utility or consumer surplus that a consumer at l receives from shopping at store i as $u_{l,i}$.

$$u_{l,i} = \theta_{f(i)} + \mathbf{x_i}'\theta - c^d_{l,i} - \eta p_i + \xi_i + \epsilon_{l,i} \tag{3.2}$$

Because it is necessary to estimate $u_{l,i}$, I assume that it is comprised of some observable factors: a firm fixed effect, $\theta_{f(i)}$, fixed effects relating to unobservable store characteristics, $\mathbf{x_i}$, disutility from distance, $c^d_{l,i}$, disutility from prices, p_i, store fixed effect, ξ_i and the taste shock, $\epsilon_{l,i}$ which abusing notation, is specific to individuals.

The firm fixed effect is shared by all stores owned by the firm in the same NAICS category. It captures the common element of a firm's stores such as effects that are generated by the firm's brand and marketing. The store fixed effect captures effects specific to the store but not captured by other observables. As in Holmes (2011), store age is included as the store characteristic. I also include economic density, measured as establishments per square mile, as a store characteristic which captures utility from being nearby workplaces and other stores.

The disutility from traveling between locations consists of two components: a component that depends directly on the distance between locations and another component that depends on the opportunity cost of time, which is captured by the time taken to travel between locations interacted with household income (or household income divided by speed). Distances and travel times are measured by data on the distance and travel times by road between two zip code centoids from Google Maps. To control for heterogeneity in the size of zip codes, I set the

within zip code travel distance to half of the nearest zip code.

$$c_{l,i}^{d} = \left(\xi_0 + \xi_1 \frac{\log\left(hhincome\right)}{speed_{l,i}} \right) d_{l,i} \tag{3.3}$$

The final term is the taste shock which each customer draws before choosing where to shop. I assume that the support of the shock is unbounded, implying that all stores face a positive probability of being chosen by a customer, no matter how poor their offering to customers. This reflects the randomness in reality that drives people to sometimes visit stores that are much worse on average than others.

Although the model is simple at this stage, I further simply by assuming that the taste shock is distributed i.i.d extreme value type 1. This generates an analytic expression for the probability that a customer from a certain location shops at a given store. With a smaller dataset or one based on an aggregate market, it is computationally feasible to assume a more general form for the error term and use simulation based methods to derive market shares to avoid the well known problems of using logit errors. However, this is not possible here as I estimate the model on the universe of retail stores in the US, which makes simulating market shares for each store computationally intractable. Using the properties of the extreme value type 1 distribution and the independence assumption of the error terms, I can express the probability that customer j with internet access shops at store i as $s_{l,i}^{j}$ with internet access

$$s_{l,i}^{j} = \frac{e^{u_{l,i}}}{e^{u_h} + \mathbb{I}_j e^{u_{o,t}} + \sum_j e^{u_{l,j}}} \tag{3.4}$$

where $\mathbb{I}_j$ is an indicator for whether customer j has internet access or not.

Equation (3.4) shows the rival nature of competition. Any improvement in the utility that a particular store delivers to customers increases that stores share of customers while decreasing everyone else's. One property of this demand structure

is that entry by a competitor leads to reallocation from all retailers, not only stores offering low utility due to the random component of utility. For example, a McDonalds that opens offering worse utility than all incumbent restaurants will take market share from all incumbents as the entry of McDonalds endows households with another opportunity for a large taste shock.

3.3.2 Retail stores

Given household preferences, each store sets a price that applies to all consumers in order to maximize profits. I assume that stores face a constant marginal cost of serving a consumer as well as a fixed operating cost. Denote π_i as gross profits for store i.

$$\pi_i = \max_p (p - c) y_i(p) \tag{3.5}$$

Total purchases at a store, y_i, are equal to purchases aggregated across locations.

$$y_i = \sum_l y_{l,i}^{access} + y_{l,i}^{noaccess} \tag{3.6}$$

Purchases from a type of consumer from a particular location are equal to the store's share of purchases multiplied by the total number of purchases in that location from that particular type of customer. Prices affect the purchases from consumers through the effect on the share of consumers or allocation across stores.

$$y_{l,i}^{access} = m_l n_l \nu_l s_{l,i}^{access} \tag{3.7}$$

$$y_{l,i}^{noaccess} = m_l n_l (1 - \nu_{l,t}) s_{l,i}^{noaccess} \tag{3.8}$$

The optimal price balances the extra revenue gained per purchase against the foregone profits from consumers that choose to shop elsewhere and satisfies the

following markup rule over marginal costs.

$$p_i = c + \frac{1}{\eta\left(1 - \bar{s}_i\right)} \tag{3.9}$$

where $\bar{v}_i$ represents a store's purchases-weighted average market share across consumers with and without access to the internet.

$$\bar{v}_i = \sum_l \frac{y_{l,i}^{access} s_{l,i}^{access} + y_{l,i}^{noaccess} s_{l,i}^{noaccess}}{y_i}$$

Equation (3.9) shows that markups are driven by a common component and a component specific to stores. The common component is the extent to which customers dislike higher prices, captured by the parameter η. More sensitivity to higher prices will lead to lower markups. The variation in markups come from variation in stores' average market shares. Stores with large market shares have a lower own-price elasticity and hence set higher markups as the tradeoff between increasing margins and acquiring more customers tilts in favor of customer acquisition. Entry by a new establishment in an area lowers incumbents' market shares putting downward pressure on markups.

Average market shares can differ across markets for two primary reasons: market size and quality of retailers. An increase in market size, assuming that the number of stores is roughly proportional to market size, will lead to decreases in market shares. Each store will serve the same number of consumers but their market share is smaller. Hence the model predicts that markups are lower in larger markets, all else being equal. The quality of retailers also matters since customers can always choose the outside or online option. Competition amongst brick and mortar retailers in two areas may appear equal (in terms of each store's share of brick and mortar sales), yet if average quality differs between the two groups, actual market shares (measured against total sales including both home production and online sales) will differ. As an example, suppose a mediocre restaurant is the

sole restaurant within 100 miles. Although that restaurant may appear to have a monopoly and hence unlimited pricing power, the fact that customers can cook at home and view the restaurant relative to this outside option implies that the restaurant's market power is in fact quite limited.

The final piece of the model concerns establishment exit. For simplicity, each firm is assumed to have a fixed cost of operating a store in a location which represents the fixed cost of labor, rent or opportunity costs of holding property and other implicit costs such as an entrepreneur's foregone wages or shareholder's required return on assets. I assume that firms own stores and are responsible for the entry and exit decisions of stores. Denote $V(i, \Omega)$ as the value of firm i which encompasses net profits at each of its stores given the aggregate state encapsulated in Ω.

$$V(i, \Omega) = \max_{S_i} \sum_{j \in S_i} \pi_j(S_i; \Omega) - \varphi_j + \beta \mathbb{E} V(i', \Omega') \tag{3.10}$$

where S_i is the set of stores operated by firm i and $\pi_j(S_i; \Omega)$ are the gross profits from store j making explicit their dependence of the aggregate state and the set of stores operated by the firm. Optimality regarding the extensive margin of store choice therefore requires that perturbations to the set of stores to be suboptimal. That is, any store expansions or closures in the current period must not generate additional net profits to the *firm*. This rules out new stores that may be profitable on their own but not to the firm due to cannibalization effects. More concretely, $\forall i, \Omega, S'$,

$$V(i, \Omega) \geq V(i, \Omega; S') \tag{3.11}$$

where $V(i, \Omega; S')$ is the firm's value with the set of stores S'.

3.3.3 Aggregate welfare

Aggregate welfare is the total surplus (consumer surplus and profits) generated from all purchases net of the operating costs of stores.

$$W = \sum_{l} \sum_{i} y_{l,i} \left(\frac{u_{l,i}}{\eta} + p_i - c \right) - \sum_{i} \varphi_i \tag{3.12}$$

Note that consumer surplus from a purchase is divided by η which scales utility in dollar terms implying that prices have no bearing on the surplus at a store, beyond the allocation of households across stores. Equation (3.12) shows that what matters for welfare in the model are the services that are provided in the process of consumption. Retail stores generate welfare by providing services to consumers that have value over and above the value from the goods purchased. The value of these services are large in reality, with value added from Retail and Wholesale Trade accounting for 51 per cent of personal consumption expenditures on goods on average from 2000-2012.

Normalizing aggregate welfare by the number of purchases generates a new measure of retail productivity. This new measure captures an economy's ability to allocate consumers efficiently such that in the process of consumption, a relatively valuable amount of services are generated at relatively little cost. Other measures of retail productivity such as value added or sales per worker get at a similar notion except that they suffer from issues due to the nominal measure of output. A social planner using these standard measures of retail productivity would prefer low cost monopolies charging high prices. Instead, using the measure in this paper would lead the social planner to prefer low cost firms that provide relatively valuable services.

3.3.4 Equilibrium

Equilibrium in the model is optimality by consumers, which is satisfied by equation (3.4), optimality by stores, which entail a set of prices for every store that satisfies equation (3.9) and extensive margin choices by firms which satisfy equation (3.11) and market clearing. Regardless of prices, goods market clearing in each location is satisfied since aggregate demand is fixed with any residual demand not satisfied by stores going to either the outside option or an online retailer. Another way to think of this is that rather than carrying inventory, stores can instantaneously produce the good and supply it to the consumer. Market clearing holds since any pattern of customer choices across stores, home production and online retail are instantly fulfilled by that option, guaranteeing market clearing.

Theorem 3. *There exists a unique equilibrium to the model.*

Proof. See Appendix. □

Because the primary goal of the model is to Improvements by online retailers increase the utility from purchasing online and have two effects. The first of these channels is straightforward. Holding fixed allocations, consumer surplus increases for those already shopping online leading to aggregate welfare gains. Second, the market structure of retail changes as consumers substitute away from brick and mortar stores towards online retail. Sales at brick and mortar stores fall in addition to markups as stores' market power falls. Both of these lead to store closures which has an ambiguous effect on aggregate welfare. Consumers who initially substitute to online retail clearly gain but those consumers who still preferred to shop at the subsequently close store are now forced to find the next best alternative. Offsetting some of these potential losses are the savings in the closed store's fixed operating costs.

3.4 Estimation

The estimation procedure follows two steps. First, I estimate the demand system which uncovers parameters that govern consumer surplus on offer at each store and the online retailer. This requires estimates of internet access by zip code, demographic data by location, distance and travel times between zip codes, data on store characteristics and local market sales shares. When estimating the demand system, I do not need estimates of fixed costs since I merely need to assume that fixed costs are such that they rationalize the network of stores observed in the data. In the final stage of estimation, I estimate bounds for each store's fixed operating costs using a revealed preference approach based off the estimated demand system.

3.4.1 Demand system

Each store's consumer surplus is comprised of a firm fixed effect, store fixed effect, store characteristics, disutility from prices and the cost of distance from consumers. The basic strategy is to match each store's observed share of sales within its zip code to that generated by the model. A store observed with a high market share will have its consumer surplus parameters tweaked until the model generated market share matches the data.

More precisely, each parameter is identified by assessing the impact of variation in the variable of choice on stores' market shares. For instance, the parameter associated with store age is identified by the change in market share caused by variation in store age. However, note that since market shares are calculated within zip codes, this strategy only identifies parameters that rely on within zip code variation. To identify parameters associated with variables that vary only across zip codes, I use variation in the level of sales. For example, distance from households only varies for stores in different zip codes. The cost of distance is

inferred from the effect on zip-code aggregated sales from being located in different areas, perhaps for instance, further away from a major population center.

The remaining component of stores' consumer surplus that needs to be identified are the store fixed effects. Here I follow Berry et al. (1995) in setting store fixed effects such that together with the observable component of store utilities, the model generates market shares equal to the observed market shares in the data. These residual store fixed effects account for the portion of sales not accounted for by observable characteristics. Note that finding a set of residual store fixed effects that match observed market shares is a fixed point problem as changing a single store fixed effect alters all market shares. Moreover, this task is made more difficult than the standard Berry et al. (1995) algorithm as for any set of store fixed effects, a fixed point for prices needs to be found.

Formally, I want estimates of fixed effects and prices for each store such that (i) stores estimated shares of sales among stores located in the same zip code are equal to that in the data and (ii) the pricing equation holds.[5] Note that this is a non trivial problem because solving for equilibrium prices is itself a fixed point problem. While it appears that some form of a nested fixed point algorithm is necessary to find a solution, it is possible to find a solution that iterates on a transformation of the store fixed effects alone. I provide a more formal description of the algorithm in the Appendix and a proof that the operator defined by the algorithm is a contraction and provides optimal prices and store fixed effects that match observed market shares.

Because the data covers the universe of retail stores in the US, I make some restrictions in estimating the model to make the estimation procedure computationally tractable. I reduce the parameters that the estimation procedure needs to evaluate by only estimating firm fixed effects for firms with greater than 10

[5]Because there are multiple years in the data, I choose to match market shares aggregated over the 3 separate years in the data. While it is possible to match market shares in each year, this raises the risk of overfitting.

unique stores in a 6 digit NAICS industry in a single year over the sample. For the remaining small firms, I estimate a fixed effect term for each separate year, which reflects the average utility a small firm delivers relative to the outside option in a given year. To compensate for the loss in precision for these smaller firms, I include firm size as an observable characteristic.

3.4.1.1 Instruments and identification

The demand system is estimated via two-step GMM which requires as many moments as parameters. There are three broad classes of moments that I use. The first are moments based on the assumption that the residual store fixed effects, ξ_j, is mean independent of the variables that vary across stores within a location, $\mathbf{z}_1$.

$$\mathbb{E}(\xi_j|\mathbf{z}_1) = \mathbb{E}(\xi_j) \tag{3.13}$$

The residual store fixed effects reallocate market shares within locations to match observed market shares. Hence, these moments can only identify parameters that relate to variables that are capable of changing market shares within locations. The moments that I use in estimation require that the covariance between the unobservable store term and the variable k is zero. The variables used with this moment include firm size by number of stores for small firms, an indicator variable equal to one if the store is more than a year old, an indicator that selects small firms' stores for each year and an indicator that selects all stores owned by a firm, for each firm.

$$\mathbb{E}_j(\xi_j z_{k,j}) - \mathbb{E}_j(\xi_j)\mathbb{E}_j(z_{k,j}) = 0 \tag{3.14}$$

Each of these moments identifies a single parameter. As an example, consider the moment that selects stores for a given firm. Suppose that the moment were positive for a given parameter vector. This implies that stores owned by the firm require a higher than average store fixed effect to match market shares. To move

the moment closer towards zero, it is necessary to increase the firm's fixed effect parameter to soak up the positive covariance.

The next set of moments are based on measurement error between the level of sales in an area and the model's estimate of an area's sales. Measurement error still exists even when market shares are equal to that in the data because market shares are calculated within, not across locations. The moments based on measurement error are based on the assumption that the conditional mean of the measurement error with respect to the variables that vary across locations is zero.

$$\mathbb{E}_j(e_j|\mathbf{z}_2) = 0 \tag{3.15}$$

Moments based on measurement error are of the form

$$\mathbb{E}_j(e_j z_{k,j}) = 0 \tag{3.16}$$

The moments used are based on: distance to customers, distance to customers times population density and population weighted demographic variables. As an example, consider the moment based on the distance to customers. If the moment is negative, it implies that stores relatively close by to customers have larger measurement errors. To remove the correlation, it is necessary to increase the constant parameter relating to distance in the utility function which has the effect of reallocating customers and consequently sales towards stores relatively close by to customers. Similar arguments can be made for the other moments based on measurement error.

With this set of moments, the model remains under-identified. To complete the set of moments, I use moments based on more aggregated data. The demand shifters in each year are identified by yearly differences in total brick and mortar retail sales between the model and the data. Customers' disutility from prices are

identified via the markup rule for prices and the log difference between average markups in the model and in the data. The data on average markups are calculated from the average markup between 1997 and 2007 for each industry from the Annual Retail Trade Survey.

The remaining estimated parameters in the model are the utilities from the online retailer. I use moments that state that the model's estimate of the online retailer's share of total sales is equal to that in the data. Bear in mind this parameter is 6 digit industry specific/ To determine the online market share for an industry, I calculate the online market share for each product and then use the weight of each product in an industry's sales to compute a weighted average of the online market shares across products. There is substantial heterogeneity in these market shares, with online market shares for book stores and meat markets 24 and less than 1 per cent respectively. The full list of moments and the parameters they each identify are listed in Table 3.1.

Table 3.1: List of moments

Moment	**Parameter identified**
Local market shares	
indicator for groups of stores by firm	β_f
indicator for stores of small firms by year	$\beta_{s,t}$
firm size (# stores)	θ_{fs}
dummy for store age ≥ 2	θ_{age}
Measurement error across locations	
household income	β_i
family share of households	β_f
education	β_e
white share of pop.	β_w
distance	ξ_0
distance $\times$ log household income / speed	ξ_1
Aggregate moments	
average markup	η
level of sales for each year	λ_t
online market shares for each year	$u_{o,t}$

Finally, marginal costs for brick and mortar stores are normalized to 1 since the Census of Retail Trade is uninformative regarding costs at each store. Since prices

follow a markup rule over marginal cost, normalizing costs shifts any cost-driven variation in prices between stores into the consumer surplus term as discounts. Marginal costs for the online retailer are set to operating expenses over the cost of goods sold from Amazon.com annual reports, relative to the same calculation for the entire brick and mortar retail sector from the Annual Retail Trade Survey.

Normalizing costs implicitly fixes the level of prices which is likely to distort the amount spent per trip in some industries. Instead, variation in the amount spent per trip is soaked up by the cost of distance parameter. Estimates of the cost of travel will be lower in industries where consumers spend a large amount per trip. As a result, the model will suggest that consumers in these industries make many trips in order to make up the dollar volume of sales. Regardless, in aggregate, the total expenditure and total costs of travel are the same.

3.4.1.2 Demand system parameter estimates

While I estimate 17 industry-specific parameters for each of the 29 industries and firm fixed effects for every firm with at least one year of operation with more than 10 stores in some retail industry, I am prevented from reporting estimates at the firm, industry or year level due to disclosure restrictions imposed by the US Census Bureau. Instead, I report the mean of these parameters estimates and their standard errors across industries and across years where applicable. To help get a sense of the dispersion in these parameters and their standard errors across industries, I also report the 25th and 75th percentiles of the parameter estimate and the respective standard error.

Most of the parameters are relatively well identified since the values over which the standard errors range are relatively small compared to the parameter estimates. The estimated parameters relating to firm size and age are consistent with the notion that firms with a greater number of stores offer greater utility

Table 3.2: Summary of parameter estimates and standard errors

	Estimate (std. err.)[1]		
Parameter	**Mean**	**25th**	**75th**
Utility parameters			
online shopping constant - u_o	-2.25 (0.38)	-4.21 (0.80)	-0.55 (0.03)
distance constant disutility - ξ_0	0.120 (0.004)	0.058 (0.0020)	0.187 (0.007)
distance opportunity cost - ξ_1	0.007 (0.001)	0.001 (0.0001)	0.013 (0.001)
disutility of prices - η	2.34 (0.28)	1.29 (0.15)	2.62 (0.32)
firm size - θ_{fs}	0.51 (0.11)	0.27 (0.06)	0.66 (0.12)
age dummy - θ_{age}	0.40 (0.07)	0.28 (0.05)	0.54 (0.09)
economic density - ϕ	0.046 (0.003)	-0.19 (0.02)	0.042 (0.005)
Demand parameters			
demand ('000 visits per year) - λ_s	0.30 (0.05)	0.001 (0.0002)	0.13 (0.02)
household income demand - β_i	0.38 (0.06)	-0.44 (0.11)	1.36 (0.22)
family share demand - β_f	0.06 (0.01)	-1.22 (0.17)	1.69 (0.25)
white demand - β_w	1.10 (0.15)	0.41 (0.03)	1.69 (0.25)
education - β_e	0.08 (0.05)	- 0.90 (0.14)	1.49 (0.31)

[1] Percentiles represent the parameter estimate and standard error for the industry representing the percentile.

to customers relative to smaller stores and that stores less than 2 years old have lower sales than older stores. To assess the validity of these parameter estimates, I calculate moments implied from the model that were not targeted in the estimation process and compare them with data where possible. These include moments regarding shopping related travel of consumers, the variance of prices and the relationship between markups and competition.

Table 3.3 shows the effect of distance on the demand of an arbitrary store using the mean of estimated parameters and in a location with the average driving speed of Los Angeles, 26.8 miles per hour.[6] The table calculates the demand of a store located a certain distance away from a customer relative to the same store located next door to a customer. Demand falls both with distance from a customer and household income. Holding household income at $25,000, demand falls by 25 and 76 per cent as distance increases by 1 and 5 miles. In an area with $100,000 in household income, demand falls slightly more over those distances, 26 and 78 per

[6]This is calculated by taking the mean of all unique zip code pairs with one side of the pair in Los Angeles.

cent respectively as opportunity costs are larger.

Table 3.3: Relative demand to a store 0 miles away

	Household income		
Miles from customer	**$25,000**	**$50,000**	**$100,000**
1	0.75	0.74	0.74
2	0.57	0.55	0.54
3	0.43	0.41	0.40
4	0.32	0.31	0.29
5	0.24	0.23	0.22
10	0.06	0.05	0.05
25	0.00	0.00	0.00

[a] Calculated at 26.8 miles per hour average travel speed, average speed for Los Angeles.

Table 3.4 shows various moments regarding travel behavior of consumers in the model, averaged across industries. The median distance travelled per year is estimated to be 359.8 miles with a median cost of travel of 14 cents per mile with roughly 5 cents per mile of that cost coming purely from the cost of driving. Five cents per mile is lower but comparable to 14.9 cents per mile, the estimate of variable per mile driving costs provided by AAA for a medium sedan in 2007. The model's estimate of average distance per trip is also slightly higher than in the data. The average shopping trip in the Department of Transportation's National Household Travel Survey 2009 was between 7.2 to 7.63 miles whereas the median trip was 8.05 miles in the model.[7]

Table 3.4: Implied dollar costs of traveling to stores

	Percentiles across population weighted zip codes[1]						
	5	**10**	**25**	**50**	**75**	**90**	**95**
Median across industries							
Avg distance per trip (miles)	3.41	4.29	5.91	8.05	12.38	16.8	19.95
Yearly distance (miles)	30.2	66.3	230.0	359.8	518.8	729.9	899.4
Cost per mile ($)	0.12	0.13	0.14	0.14	0.15	0.16	0.16

[1] Percentiles are calculated for each industry and then averaged across industries.

In the model, variation in prices comes purely though variation in competition

[7] I restrict the sample to only include trips undertaken by automobile and from stores to home or vice versa.

across geographic space. To show that the estimated model is quite sensible with regards to price setting, I benchmark the model against a comparable estimate of the variance in prices across a geographic area. The median standard deviation in log prices, shown in Table 3.5, is 2.6 per cent which is slightly lower than the estimates of 3.3 to 5 per cent estimated in Kaplan and Menzio (2014) for the variance in prices within MSAs due to slow moving store characteristics such as store quality and location. Kaplan and Menzio (2014) interpret the variance in prices as evidence of search frictions, whereas the results here suggest that a large part of the variance can be explained by variation in competition within MSAs.

Table 3.5: Variation in prices across markets

	Percentiles across unweighted zip codes[1]						
	5	**10**	**25**	**50**	**75**	**90**	**95**
Std. dev (log prices)	0.000	0.001	0.011	0.026	0.045	0.074	0.101

[1] Percentiles are calculated for each industry and then averaged across industries.

To assess the relationship between the *level* of markups and demographics, I regress the average log markup paid in a 50 mile area on demographic variables describing the market and report the results in Table 3.6. Bear in mind that these are model estimates of markups, not actual markups from the data. Regardless, the model suggests that markups fall by population density, household income and education level across a broad range of industries. The effect of the non-latino white share varies across industries. A doubling of the population density, household income and the education level of an area is associated with an average 5.5, 4.4 and 2.7 per cent fall in markups. This is consistent with empirical evidence that competition is stronger in markets with greater population density (Bresnahan and Reiss, 1991; Campbell and Hopenhayn, 2005).

Table 3.6: Markups and demographics

	Regression[1]		
	Avg	**25th**	**75th**
Log population density	-0.055	-0.091	-0.006
Log household income	-0.044	-0.058	-0.011
Log frac. pop non latino white	0.026	-0.013	0.081
Log frac. pop > bachelors degree	-0.027	-0.064	-0.006

[a] All coefficients are significant at the 1% level.

3.4.1.3 Estimates of store-level productivities

Compared to a standard heterogeneous firms model, a key difference is that the distribution of store surpluses (or)productivities varies across space. The mix of stores does not repeat itself across geographic space so that the distribution of stores in each market is identical to one another. Some areas are populated with higher quality flagship stores of retail firms while others have less well maintained stores. Since estimates of these store level surpluses are available, it is possible to assess whether the distribution of store surpluses varies systematically. Define a store's surplus per purchase as the consumer surplus in dollar terms plus profits, per purchase by a household.

$$z_{l,i} = \frac{u_{l,i}}{\eta} + p_i - c \tag{3.17}$$

Prices have no effect on a store's surplus since prices enter negatively in consumer surplus and only serve to split the surplus between consumers and the store. To detect systematic variation in the distribution of store-level surpluses across space, I relate the mean store surplus in an area to demographic variables in an area. Table 3.7 describes the correlation between the mean store surplus and demographics at the zip code level and results from regressing the mean store surplus in an area on demographic variables. To facilitate comparisons between industries, the measure of mean store-level surplus in the regressions is demeaned at the industry level.

Table 3.7: Mean store-level surplus and demographics

	Correlation			Regression[a]		
	Avg	**20th**	**80th**	**Avg**	**20th**	**80th**
Constant	-	-	-	0.13	-0.86	0.84
Log population density	0.41	0.30	0.56	0.25	0.06	0.44
Log household income	0.06	-0.00	0.12	-0.12	-0.17	0.05
Log frac. pop non latino white	-0.16	-0.22	-0.10	-0.05	-0.09	-0.01
Log frac. pop > bachelors degree	0.24	0.14	0.33	0.21	0.05	0.35

[a] All coefficients are significant at the 1% level.

The results show a strong correlation between the mean store surplus and both the education level and population density in an area. Correlations for both population density and education are positive, even in the 20th percentile industry, suggesting a robust effect across industries. A doubling of population density and education is associated with an increase in the mean store surplus of roughly 25 cents per purchase (6.7 per cent of the nationwide average store surplus) and 21 cents per purchase (5.6 per cent). In contrast, household income and the fraction of white households have an ambiguous relationship with the mean store surplus in an area. This perhaps reflects the industry-specific relationship between demand and these other demographic variables.

3.4.2 Fixed costs

For a set of estimated parameters, the model generates corresponding estimates of profits for each store in equilibrium. Hence, by varying parameters which reflect changes in the operating environment, one can measure the equilibrium effect on the retail sector. In calculating these responses, it is necessary to obtain estimates of fixed costs as establishment exit is likely to be important when considering the effect on online retail.[8]

One way to measure fix costs is directly from the data, as Holmes (2011) does

[8]Entry is also likely to play a role but it is difficult to introduce entry into a model that covers the geography of the entire US and has a large number of potential entrants. Exit is less troublesome since it involves stores that exist and hence are observed.

for Walmart. Unfortunately, data on costs by store in the Census of Retail Trade is restricted to payroll data which is a noisy measure of fixed cost. A large component of payroll probably reflects variable costs and fails to account for other important fixed costs such as rent or the opportunity cost of funds. Another approach would be to estimate fixed costs that are specific to locations using entry and exit patterns for all stores location by location. Roughly speaking, the estimated fixed cost in a particular location would be the threshold level of profits that induces exit. However, this fails to account for the massive heterogeneity that is likely to exist between stores in their fixed costs. Smaller firms probably have much lower fixed costs per store than large national retailers. A model estimated in this way would overstate the role of exit since low cost/high surplus stores would exit at too fast a rate relative to the data and vice versa for high cost/low surplus stores.

Instead I use a revealed preference approach that identifies store-specific fixed costs through the geography of a firm's network of stores. The main disadvantage of this approach is that it only set identifies fixed costs. Hence I will report results from counterfactuals from the midpoint these sets, with the full set of results in the Appendix. Results computed from the lower bounds of fixed costs are conservative estimates of exit, since it requires a relatively large reduction in sales to induce exit of any store. On the other hand, the results using the upper bounds are less conservative as exit thresholds are passed more easily.

To construct these bounds, I exploit conditions that describe the optimality of each firm's choice of the network of stores to operate. Informally, fixed costs must be low enough that it would not be optimal for firms to close a store in a location (where a store already exists) and high enough that firms do not find it optimal to open another store in the same location. These perturbations are similar in vein to the pairwise deviations in Holmes (2011) and the one-step deviations in Morales et al. (2013) which use minor deviations in optimal policies to generate inequalities that set identify parameters. To identify a firm's lower bound for the

fixed cost in a particular location, define a perturbed set of stores from the firm's optimal set of stores by adding an additional store to that location. Denote the optimal set S_i and the perturbed set $S_i(+,l)$. It must be the case that since we observe the firm operating S_i stores in the location, operating $S_i(+,l)$ stores is suboptimal.

$$\begin{aligned} V(i;S_i) &\geq V(i;S_i(+,l)) \\ \varphi_{l,i} &\geq \pi_l(S_i(+,l)) - \pi_l(S_i) \end{aligned} \tag{3.18}$$

Similarly, denote the perturbation $S_i(-,l)$ which is identical to the optimal set of stores except that a single store is closed in location l. Optimality requires the following hold.

$$\begin{aligned} V(i;S_i) &\geq V(i;S_i(-,l)) \\ \varphi_{l,i} &\leq \pi_l(S_i) - \pi_l(S_i(-,l)) \end{aligned} \tag{3.19}$$

Equations (3.18) and (3.19) define the set of feasible estimates for firm-location specific fixed costs. The lower bound at a location is the increase in gross profit from adding another store in that location, accounting for the effect on price setting by all stores owned by the firm.[9] Another way to interpret these sets are as discrete analogs to the first order conditions that would come from optimization if the number of stores in each location were continuous rather than restricted to integers. Firm-location fixed costs would be identified from a firm's marginal profit with respect to the number of stores at that location if the number of stores were continuous. Because store choice is not continuous, it is not possible to use infinitesimal variations in store policy and hence only upper and lower bounds can be derived.

These bounds reflect the curvature of the firm's gross profits in a location

[9] I assume that firms hold prices at stores owned by other firms fixed.

with respect to the number of stores with the size of the gap between the upper and lower bounds dependent on the concavity of gross profits. Gross profits are concave both because of the structure of the logit probabilities and through cannibalization of profits at stores in other areas. Data on ownership are derived from the Report of Organization which has comprehensive data on the organizational structure of multi-unit firms. Every establishment in the Census of Retail Trade is effectively linked to the highest level firm that has a controlling interest (or chain of controlling interests) in the establishment. Therefore, cannibalization effects incorporate the reduced sales from all stores owned by the firm, not only stores with the same trade name. Failing to account for cannibalization effects for the full set of stores leads to estimates of marginal profits that are too high, biasing both bounds upwards and estimates of fixed costs that are too high, overstating the role of exit.

An advantage of this approach is that the bounds of fixed costs incorporate all factors that firms consider as fixed costs to operating a store, even unmeasured fixed costs. Even if one had complete accounting data for each store, unmeasured costs such as required return on equity vary across firms and affect the exit margins for firms. These unmeasured costs are accounted for here since they are inferred from a firm's choice of the number of stores to operate in a given location. One may be concerned that these estimates of the bounds for fixed costs encompass costs that are not traditionally associated with the fixed cost of operating a store, such as additional financing costs imposed on smaller firms or an owner's personal disutility (in dollar terms) of operating an additional store. However, it is precisely these costs that make this approach more useful since these costs are likely to influence firms' behavior and are unable to be quantified. A firm may decide not to open an additional store because it's unmeasured fixed cost is very large, even though its profit net of measured fixed cost for its sole store may suggest that expansion would be viable.

To understand the estimates of fixed costs that are inferred from store location patterns, it is useful to describe fixed costs from various patterns of store location and geography, assuming only for simplicity that we observe stores operating at max one store in a location. For a given location, larger fixed costs are inferred from stores generating higher gross profits relative to other stores in the same location, holding constant cannibalization effects. If fixed costs were not higher for these stores, it would otherwise be optimal for the corresponding firm to expand in that location. Hence it is not necessarily the smallest stores in an area that are the most sensitive to increases in competition from entrants. For single unit firms (mom and pops), net profit margins are smaller in markets with larger numbers of stores since there is less curvature in gross profits. Marginal gross profits are flat since the market is already saturated with stores, lessening cannibalization effects.

3.4.2.1 Fixed costs estimates

The estimates of fixed costs for each store are summarized in Table 3.8. The median estimate of the midpoint of fixed costs is $196,000 while the 10th and 90th percentiles are $35,000 and $1,062,000. To get a sense of how large fixed costs are relative to gross profits, I report the midpoint of fixed costs relative to gross profits. The 10th, 50th and 90th percentile estimates of fixed costs are 58.1, 94.9 and 99.8 per cent of gross profits respectively. The range of estimates is also relatively small, with the median difference between the upper and lower bounds of estimates for fixed costs representing 3.59 per cent of gross profits. I also report the percentiles of fixed costs relative to the median estimate in the same industry, as a gauge of heterogeneity in fixed costs within narrow retail industries. The heterogeneity within industries is substantial with the 10th and 90th percentile stores exhibiting fixed costs 0.34 and 3.65 times the median store in their industry, indicating that profitability matters for exit.

Table 3.8: Estimates of fixed costs

	Percentile						
	5	**10**	**25**	**50**	**75**	**90**	**95**
Fixed cost ($ '000s) - midpoint	16.1	33.4	96.1	249.5	590.2	1,254.5	1,968.4
Fixed cost (midpoint) - % of gross profits	45.3	64.1	84.3	95.0	98.8	99.8	99.9
Range of estimate - % of gross profits	0.01	0.03	1.55	5.64	16.14	34.21	48.70
Range of estimate ($ '000s)	0.03	0.21	2.47	16.81	80.79	297.73	647.62
Relative to median in industry	0.17	0.28	0.53	1	1.83	3.58	4.94

3.5 Results

To gauge the effect of online retail on the retail sector and the Marketplace Fairness Act (MFA), I use counterfactual exercises that hold other parts of the model constant. The first set of these counterfactuals are based on changes that are likely to have had a role in increasing online sales. These counterfactual exercises are like exogenous shocks which I impose on the model and then measure the resulting effects.

The MFA is assessed by introducing a tax on the online retail good which mimics the proposal to allow states to compel online retailers to collect sales taxes based on the destination of the shipped goods. I apply sales taxes to online retailers based on the state of residence of households and gauge equilibrium responses by both stores and customers. Effects of the rise in sales taxes depend in part on price elasticities, but also on local market conditions. In areas where operating margins are relatively tight, relatively small increases in taxes on online retailers can have a large impact on mitigating store closures and the subsequent reallocation that unfolds with store closures. On the other hand, effects in areas with fat operating margins depend primarily on price elasticities alone.

The counterfactual exercises are implemented on the 2007 data. Therefore, counterfactual exercises represent changes in online retail from 2007 to 2012 and effects from the MFA if it had been implemented in 2007. Since data from the 2012 Economic Census are unavailable, I compare the effects of the counterfactual

exercises relative to the 2007 baseline as a means of assessing the magnitude of the effects. Although I can not directly validate the model's predictions since 2012 data are not unavailable, I show that the model is consistent with reduced form measures.

Computing these counterfactuals all rely on varying some element of the model and assessing responses by firms which requires computing equilibria. Equilibrium consist of a set of prices that satisfy store optimality taking other stores' prices as given and optimal exit decisions by firms. Computing the set of optimal prices is straightforward for a given set of stores as I show in the Appendix that one can write a fixed point algorithm for prices that is a contraction mapping and hence converges to a unique fixed point from any initial guess. The difficulty in computing equilibrium is in determining the set of stores that exist and satisfy firm-level optimality given the set of fixed costs for each firm. I describe the algorithm that finds an equilibrium in detail in the Appendix. Briefly, the algorithm is consistent with the "natural equilibrium" in Abbring et al. (2012), where multiple equilibria are resolved by selecting the equilibrium consistent with the weaker stores exiting first.[10]

3.5.1 Improvements in online retail

It is useful to frame the potential mechanisms that have lead to increases in the online share of sales in relation to the mail order industry, which offered a set of products via physical catalogs available to be shipped direct to customers. Superficially, online retail simply offers an improvement in this catalog technology with better features such as the ability to find products quickly, reviews to support decision making and a broader range of products available. These features can broadly be categorized as improvements in the services offered to customers above

[10]There are many other ways to implement equilibrium store closures. The other extreme is to close stores making the smallest losses first. However, experiments with algorithms that closed these relatively profitable stores first were unstable.

the value of the product itself and are unfortunately hard to measure quantitatively. However, the stages that bookend the online retail experience are measurable: access to the internet (catalogs) and shipping of goods to customers. I measure effects on welfare gains from these two measurable changes, and make up for the residual increase in the online share of sales by an increase that resembles online retailers' expansion into new industries.

3.5.1.1 Increased internet access

Increased internet access expands the base of potential consumers for online retailers.[11] Figure 3.2 depicts the growth in internet access from 2007-2012 which I derive from CPS data on individuals' access to the internet.[12] While the mean increase in internet access across zip codes is low, the distribution is skewed with some zip codes experiencing a relatively large increase in internet access. To isolate the effect of the rise in internet access, I increase internet access for each location in the model by the estimated value from 2007 to 2012.

Improvements in internet access may have limited effects in terms of generating sales for online retailers because improvements are concentrated in areas with poorer households, who have less spending power. An industry that benefits from wealthier households is unlikely to experience an increase in market access to its desired customer base. On the other hand, improvements in access may lead to relatively large effects if retail stores serving these poorer areas are of relatively lower quality. If that is the case, customers in poorer areas will shift away from traditional retailers more so than customers in less poor areas, compensating for their relatively lower spending power.

[11] Greater internet penetration is similar to greater market penetration as in Arkolakis (2010) except that market penetration costs are paid by consumers.

[12] The derivation of zip code level estimates of internet access are provided in the Appendix.

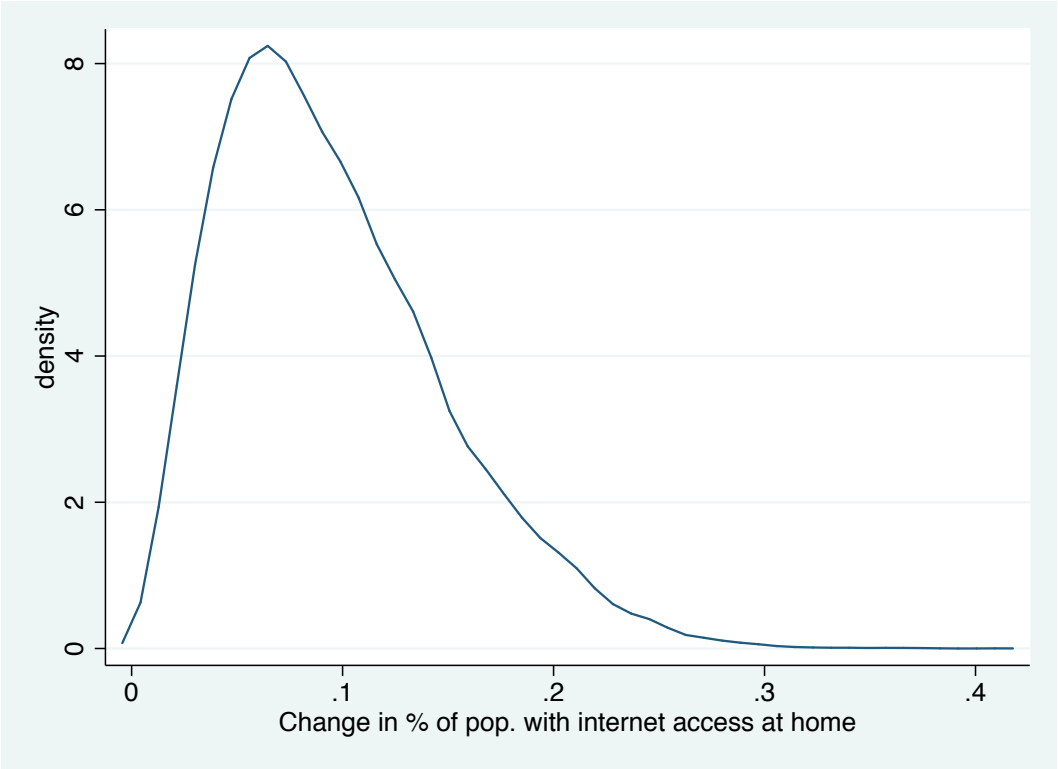


Figure 3.2: Estimated improvements in internet access across zip codes (2007 to 2012)

3.5.1.2 Improved shipping practices

In contrast to increased internet access, there is no obvious way to introduce improvements in shipping into the model. Improvements in shipping are likely to be reflected in increased utility from purchasing online and changes in the end-to-end costs of delivery. The difficulty is in calibrating these changes so that they accurately reflect actual improvements in shipping.

The increased utility from improvements in shipping are based on the notion that delivery is valued since it eliminates the need to travel to stores. Instead of traveling, courier services such as UPS deliver goods from warehouses which in effect, shifts the commute cost to these courier services. Hence the value of delivery to consumers is the value of eliminating of travel to alternate stores,

which varies across consumers. Consumers with access to relatively high quality stores nearby and low costs of commuting are inferred to value delivery less.

To determine the magnitude of the increase in utility, I begin with the assumption that the average disutility per trip experience by a consumer is equivalent to the value of same-day delivery, a shipping experience that replicates traveling to stores. Given this baseline, I calculate the increase in utility from 2007 to 2012 by determining how far away the industry is from same-day delivery in 2012 and measuring the progress made between 2007-12.

Delivery times can be reduced either by reducing the distance from the warehouse to customers or by increasing the speed at which goods travel which includes reducing bottlenecks in the supply chain. I use data on these margins to calculate reductions in shipping time. Specifically, I use the expansion of Amazon.com's shipping and fulfillment warehouses across geographic space over time and, productivity improvements at UPS and FedEx, who handle between 80-88 per cent of parcel deliveries. Given data on these margins, I calculate improvements in shipping by assuming that delivery times are proportional to the average distance between households and warehouses divided by productivity.

Figure 3.3 plots the expansion of Amazon.com's warehouses in terms of square feet and number of states. In 1997, Amazon.com operated warehouses in Delaware and Washington, one on each coast ostensibly to minimize transportation costs. By 2012, there were 14 states with warehouses spread almost uniformly over the US which reflects a large reduction in delivery times for customers. Note that this expansion is not simply maintaining warehouse capacity with the increase in sales since Amazon.com could expand warehouse space intensively in existing states.

To measure productivity, I calculate the total volume of packages delivered per real unit of expenses for UPS and FedEx where nominal expenses are deflated using the CPI.[13] Productivity gains are a reasonable measure of improvements

[13]Using the PPI or implicit price deflator for the parcel delivery industry would result in

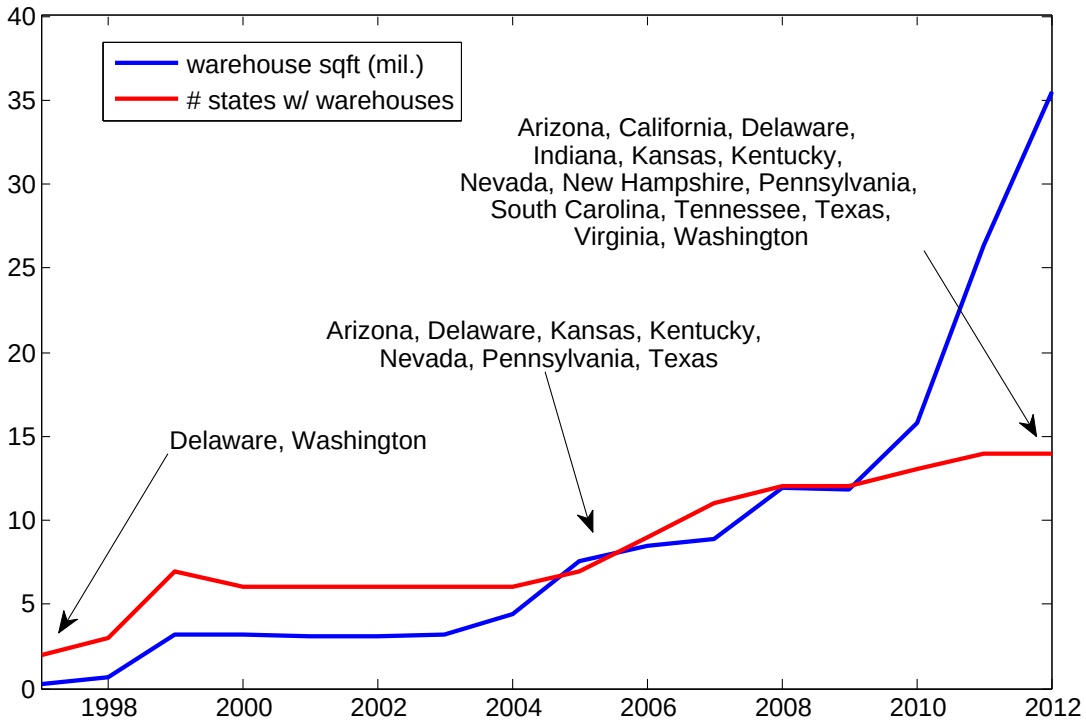


Figure 3.3: Shipping and fulfillment expansion at Amazon.com

experienced by end-users since margins at UPS and FedEx remain flat over the period. Figure 3.4 plots the measure of productivity from 1998 to 2012. From 1998 to 2012, the number of packages delivered per dollar of real expense increased 53.9 per cent with 32.8 percentage points of the improvement coming between 2007 to 2012.

Assuming that warehouses are uniformly distributed, the area covered by each warehouse fell from 0.986 to 0.211 million square miles or roughly 79 per cent between 1998 and 2012. Combined with the estimates of productivity improvements, I estimate that shipping times for the average household in 1997 were 7.14 times longer than in 2012 and 4.43 than in 2007.

The remaining step is to determine the the improvements necessary from 2012

an even slower increase in real expenses, implying that the estimates of productivity gains are conservative.

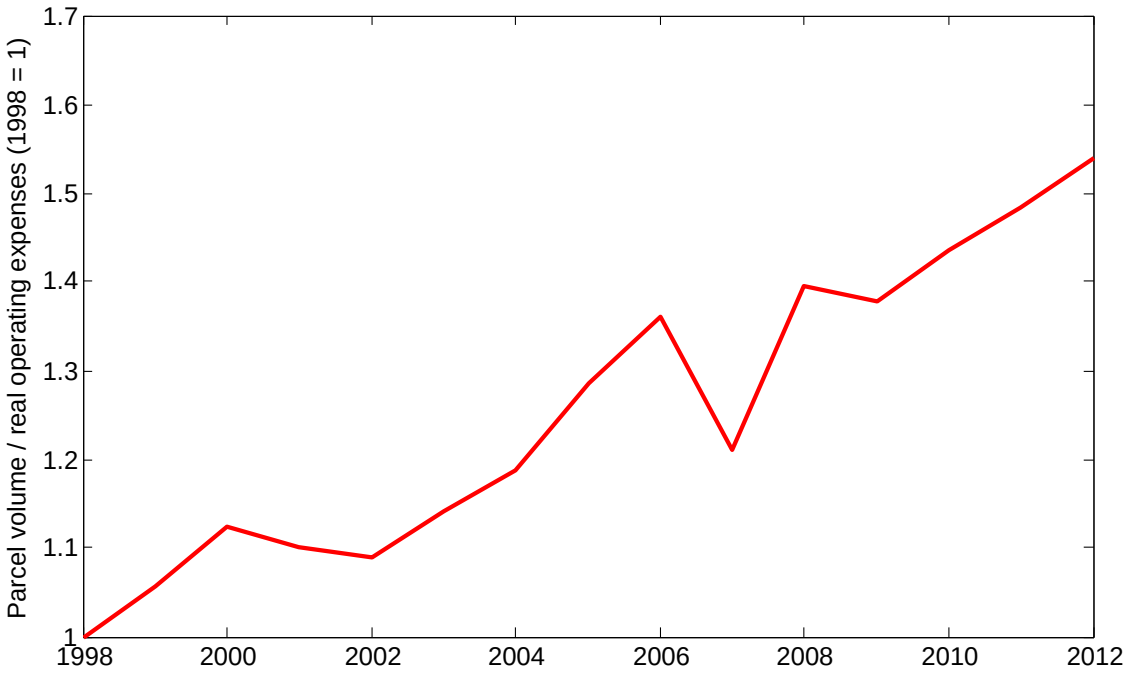


Figure 3.4: Productivity improvements at FedEx and UPS

to make same day shipping possible. Amazon.com's financial statements and annual reports suggest that same-day shipping will be available to the majority of the US population by the end of 2014.[14] By 2014, Amazon will have expanded its warehouses into Connecticut, Florida, New Jersey and Wisconsin, and productivity is estimated to rise by 7.1 per cent implying a 26.4 per cent reduction in shipping times from 2012. Therefore, I infer that a 55.1 per cent reduction in shipping times from 2007 is required to deliver same-day shipping and that 38.2 percentage points of that reduction occurred from 2007 to 2012.

Of course, this massive geographic expansion of warehouses was not without cost. To measure the corresponding increase in cost, I use data on fulfillment expenses as a fraction of the cost of goods sold from Amazon.com's annual reports in 2007 and 2012. Fulfillment expenses incorporate both direct shipping costs paid

[14]As of April 2014, same-day delivery is currently available in Baltimore, Boston, Chicago, Indianapolis, Las Vegas, New York City (and parts of New Jersey), Philadelphia, Phoenix, San Bernardino Area, Seattle, and Washington, D.C.

to courier services as well as amortized investments required for fulfillment and hence are a good measure of the total cost of sending a good to a final customer. From 2007 to 2012, fulfillment expenses increased from 11.3 to 13.9 per cent of Cost of Goods Sold, roughly 5.1 billion dollars (unadjusted for inflation). When calculating welfare gains from reductions in shipping, I include the corresponding 2.6 percentage point increase in fulfillment expenses per purchase.

3.5.1.3 Expansion into new industries

Figure 3.5 shows that the distribution of the online share of sales is concentrated in a few industries. Less than 10 per cent of the 6-digit NAICS industries have an online share of sales greater than 10 per cent. Moreover, very large industries such as supermarkets and grocery stores have virtually non existent online sales. These facts point to potentially large gains for online retailers of expanding into these untapped markets. As anecdotal evidence, consider Amazon.com's entry into the grocery market with Amazon Fresh. Amazon.com seems to exploiting its large shipping and fulfillment network to enter the grocery business, which offers a potentially large source of growth.

As evidence of expansion into new industries, I plot mean annual online sales growth by products between 2007-11 and the log of online sales of products in 2007. Although the relationship is not perfect, there is a negative relationship suggesting that the source of growth in online sales post 2007 has been in the industries with relatively few online sales. The relationship is even stronger if one excludes food, beer and wine which will likely become one of the largest sources of growth for online retailers.

To capture this in the model, I improve online retailers' relative utility, where relative utility is defined as online retailers utility relative to the unweighted mean

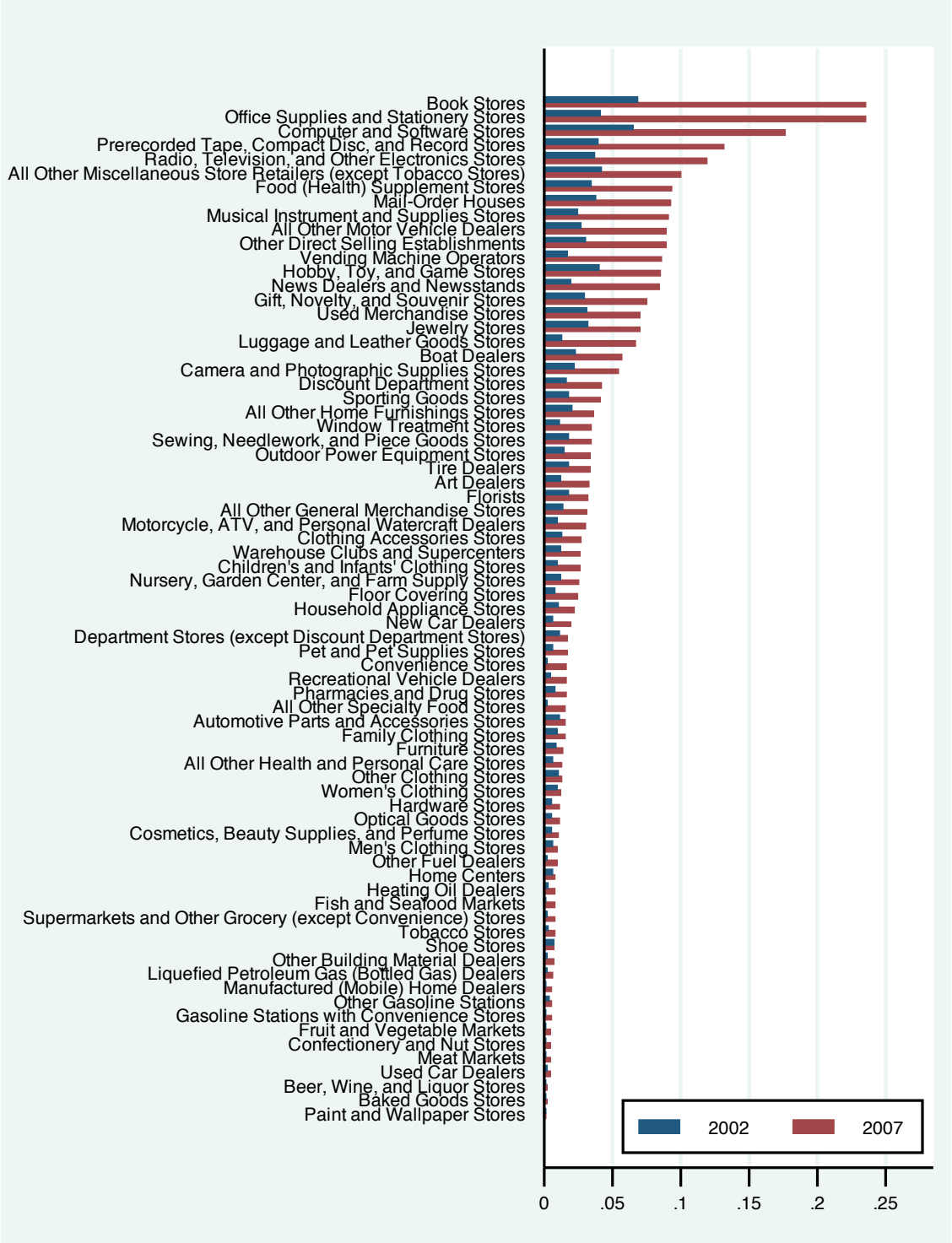


Figure 3.5: Online market shares by industry

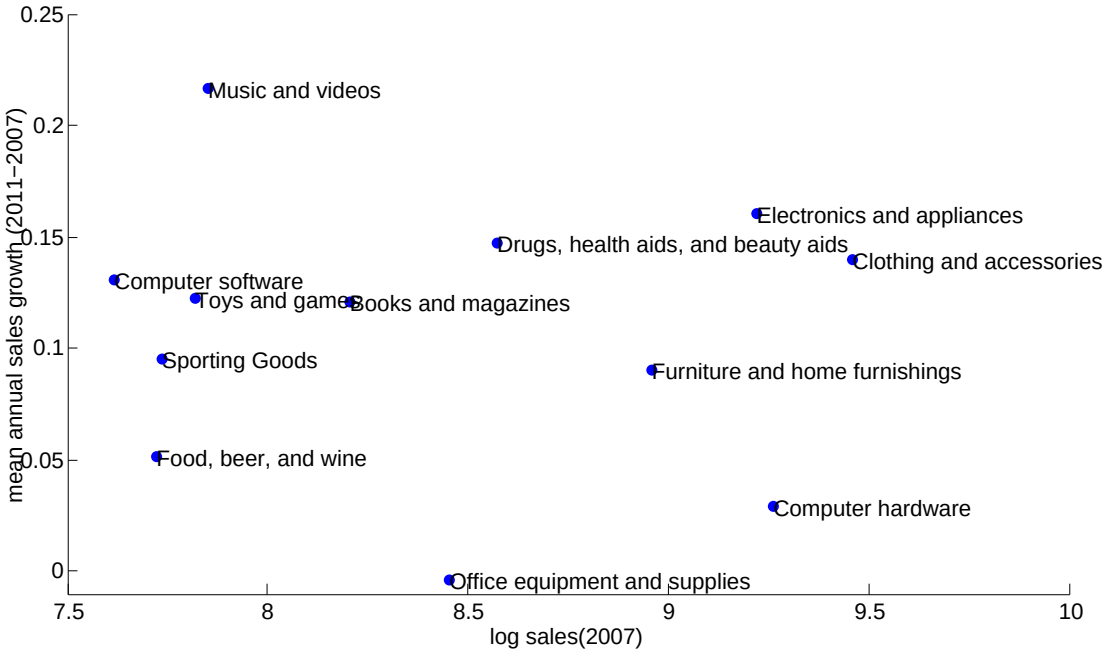


Figure 3.6: E-commerce sales growth across products (2007 to 2011)

of utility at brick and mortar stores.

$$\bar{u}_o - \bar{u}_{bm} = \mathbb{E}\left(u_{l,o}\right) - \mathbb{E}\left(u_{l,i}\right) \tag{3.20}$$

For industries with relatively few online sales, relative utility is small as each visit online by a household generates relatively little surplus compared with a visit at a brick and mortar store. Without a straightforward way to quantitatively measure these improvements, I treat online retailers' expansion into new industries as a residual that soaks up the part of the rise in the online share of sales unaccounted for by increased internet access and reduced shipping times. In particular, I increase the relative utility in industries with below average relative utilities to the value that is necessary for the model to explain the total rise in the online share of sales.

3.5.1.4 Effects on aggregates

Table 3.9 summarizes the results from these counterfactuals. I calculate equilibrium outcomes for scenarios where store closure is and is not allowed, and where the size of the shocks mirror those observed in the data and the maximum possible change. Regardless of the counterfactual, effects are largest when one allows stores closures. The rise in the online share of sales increases by between 0.3-0.5 percentage points with store closures. Regardless of exit, improvements in online retail lead some consumers to substitute away from stores to online retailers. However, there is a second bout of reallocation with store closures as the consumers of stores that subsequently close, who did not initially shop online, are forced to find an alternative. From this point on, I focus on the scenario with exit unless stated otherwise.

The results indicate that reduced delivery times and online retailers' expansion into new industries are more important than increased internet access in driving the rise in online sales. More widespread internet access increased the online share of sales by 1.4 percentage points, just over half the impact from reduced shipping times and less than one third relative to the general expansion into new industries. Combined, increased internet access and reduced shipping times are capable of explaining just under half of the observed increase in the online share of sales.

Going forward in time, a similar ordering applies with the maximum feasible effects. Moving to complete internet penetration raises online sales by 32.9 per cent from 2007 which represents an increase of 14.9 percent from the value estimated for 2012. Shipping improvements are much more effective, increasing online sales by 30 percent from 2012 on. The larger impact of shipping mirrors the online retail industry's focus on ever-faster shipping.

With all three shocks present, the online share of sales increases by 5.2 per-

centage points. Bear in mind that this exactly matches the observed increase in the online share of sales as the expansion into new industries counterfactual is calibrated to match the observed increase. The aggregate impact on the retail sector is large. Sales at brick and mortar stores fall 15.5 percent with 78,000 stores closing.

$$OX(i) = \sum_{p \in P} \frac{sales_{p,2002}(i)}{sales_{2002}(i)} \left(\frac{sales_{p,2007}(online)}{sales_{p,2002}(online)} - 1 \right) \tag{3.21}$$

In order to validate these results, I compare the predicted outcomes from the model to data. However, I do not compare them directly to observables since the great recession occurred during the same period. Instead, I compare them to the predicted values from a reduced form model of the relationship between brick and mortar sales and the rise in the online share of sale.

The obvious strategy would be to use variation in online shares of sales and retail sales across industries. The issue with this strategy is that industry based models of the effect of the internet on retail store sales suffer from the problem of confounding effects from the internet with unobservable industry level effects. For instance, sales at book stores may appear to decline because of a corresponding rise in the online share of sales but instead are driven by changing consumer preferences that are independent of online retail. To overcome this problem, I use within-industry variation in the product mix comprising each store's sales.

As an example, consider book stores. The bookstores that support their sales with a relatively large amount of food and beverage sales are relatively insulated from the rise in online sales of books, relative to those that sell only books since food and beverage sales are not subject to online competition whereas books are. The caveat for this identification strategy to be valid is that industry level shocks are not correlated with the product mix of stores. If it were the case that book stores who sell disproportionately more books bore the brunt of changing

Table 3.9: Results from counterfactual exercises

	Observed		**Maximum**	
	No exit	Exit	No exit	Exit
Increased internet access				
Δ online share of sales (ppt)	0.5	0.8	1.1	1.6
Δ online sales (%)	10.8	15.6	23.5	32.9
Δ brick and mortar sales (%)	-1.3	-3.7	-2.8	-6.5
Δ brick and mortar stores (%)	-	-5.1	-	-8.6
Δ brick and mortar stores ('000s)	-	-26	-	-44
Reduced shipping times				
Δ online share of sales (ppt)	1.1	1.4	2.6	3.1
Δ online sales (%)	25.2	29.9	60.0	68.9
Δ brick and mortar sales (%)	-1.3	-3.3	-2.7	-6.1
Δ brick and mortar stores (%)	-	-7.9	-	-12.7
Δ brick and mortar stores ('000s)	-	-40	-	-64
Expansion into new industries[1]				
Δ online share of sales (ppt)	2.4	2.9		
Δ online sales (%)	46.7	57.2		
Δ brick and mortar sales (%)	-6.2	-10.2		
Δ brick and mortar stores (%)	-	-5.3		
Δ brick and mortar stores ('000s)	-	-26		
Expansion into new industries				
+ Reduced shipping times				
Δ online share of sales (ppt)	1.8	2.4	4.3	5.4
Δ online sales (%)	39.1	47.8	96.4	115.1
Δ brick and mortar sales (%)	-2.7	-6.3	-6.0	-7.8
Δ brick and mortar stores (%)	-	-10.9	-	-18.7
Δ brick and mortar stores ('000s)	-	-55-	-95	
Increased internet access				
+ Reduced shipping times				
+ Expansion into new industries				
Δ online share of sales (ppt)	5.2	5.2	8.1	10.7
Δ online sales (%)	95.9	162.2	166.0	208.9
Δ brick and mortar sales (%)	-9.6	-15.5	-13.8	-21.4
Δ brick and mortar stores (%)	-	-15.4	-	-23.2
Δ brick and mortar stores ('000s)	-	-78	-	-118

[1] Expansion into new industries increase online retailers' relative-utility-to-stores in industries with below mean online shares of sales to the mean relative-utility-to-stores.

consumer preferences, then the identification strategy would not be valid.

To measure each store's exposure to online competition, I create an establishment specific online exposure index which is the weighted average of each product's exposure to online retail where the weights come from the composition

of the establishment's sales. A product's exposure to online retail is defined as the change in online sales for that product relative to the product's total sales in the initial period. For example, a bookstore with 20 per cent of sales from coffee and 80 per cent books has an online exposure index weighted 20 and 80 percent towards the growth of online sales relative to initial sales of coffee and books respectively. More precisely, let subscript p denote a product and i denote an establishment. The online exposure index is defined as follows.

$$OX(i) = \sum_{p \in P} \frac{sales_{p,2002}(i)}{sales_{2002}(i)} \left(\frac{sales_{p,2007}(online)}{sales_{p,2002}(online)} - 1 \right) \tag{3.22}$$

To determine the relationship between online retail and brick and mortar sales, I estimate a linear regression model of the log change in a store's sales on the store's online exposure index that takes into account 6 digit industry effects, firm size, population density at the store location, establishment size and payroll share of sales. Note that since I include industry dummies, any industry level effects are accounted for. The results are provided in Table 3.10 and as expected, indicate that stores with greater exposure to online competition experienced less sales growth.

Table 3.10: Regression on store sales

Variable[1]	**Log sales**
constant	0.414 (0.018)
age	-0.006 (0.000)
log(firm size)	0.018 (0.001)
log(pop. density)	-0.012 (0.001)
log(firm payroll / firm sales)	0.110 (0.005)
$OX(i)$	-1.848 (0.343)
$OX(i)\times$log(pop. density)	0.098 (0.027)
$OX(i)\times$ log(firm size)	-0.111 (0.015)

[1] N = 266,000 establishments; Industry dummy variables at the 6 digit level included.

It is straightforward to use the reduced form model to generate a prediction for the reduction in total retail store sales due to online retail. Between 2007

to 2011, average firm size in retail was on 26 employees with average population density in the US 34.2 people per square kilometer. The online exposure index for online retail as a whole in 2007 is 0.064. With these figures, the model predicts that online retail would be responsible for 10.4 per cent decline in retail store sales between 2007 and 2012. In comparison, the structural model's prediction is for a reduction of between 10.5 to 20.5 per cent (lower and upper bounds) in brick and mortar sales which suggests that the structural and reduced form models offer similar predictions.

3.5.1.5 Effects on Welfare

Although the reduced form and structural model predict similar outcomes, the structural model has the advantage that it is informative about the heterogeneity in effects from online competition and the associated effects on welfare. Given the expression for welfare, it is straightforward to decompose the change in aggregate welfare into direct effects and indirect effects. Letting $\frac{u_{l,i}}{\eta} + p_i - c \equiv z_{l,i}$, the change in aggregate welfare is comprised of: direct changes in welfare, indirect effects from net reallocation and changes in store operating costs.

$$W' - W = \sum_l \sum_i (\Delta y_{l,i} z_{l,i} + y_{l,i} \Delta z_{l,i}) + \sum_i \Delta \varphi_i \tag{3.23}$$

Net reallocation measures the effect of reallocating households from one store to another and occurs in the model for two reasons. The first is the initial switching by some households to online retailers as utility per visit improves at online retailers. This initial bout of reallocation reduces sales at some stores, leading to store closures and a second bout of reallocation as households who previously shopped at closed stores have to find alternatives. Net reallocation is not necessarily a positive force on welfare as store closures force some households to switch to less preferred alternatives. A more subtle reason why net reallocation may

reduce aggregate welfare is due to measurement and the random component in utility. When online retailers improve, some households shopping at stores offering relatively large consumer surplus receive a large idiosyncratic shock and decide to shop online. Because my measure of aggregate welfare does not capture he random component of utility, this may appear as a household switching to a lower surplus alternative.[15]

Reallocation also leads to store closures and reduces the total cost of operating stores, partially offsetting the negative effect on consumer surplus of store closures.[16] However, note that the reallocation towards online retailers is not without cost. Increased marginal costs from the online retailer are taken from financial statements at Amazon.com and accounted for in the calculation of welfare gains. Marginal costs as measured by operating costs over cost of goods sold are slightly higher at Amazon.com relative to the average retail store.

The final component of aggregate welfare gains come from the direct increase in welfare from improvements by online retailers, holding fixed the allocation of consumers. The total surplus per purchase increases at online retailers as consumer surplus increases more than costs do, ignoring both product and shipping prices.

Table 3.11 shows the effects on aggregate welfare from these counterfactuals. Consider the results from the counterfactual with both the observed increase in internet access and reductions in shipping time. These observed changes can account for roughly half of the increase in the online share of sales and lead to welfare gains of 8 per cent. Results including online retailers' expansion into new industries are much larger with welfare estimated to increase by 13.4 per cent, with the bulk of these gains coming the direct component of welfare gains.

[15]This is analogous to entry by a McDonalds store offering terrible service stealing customers from all restaurants in a location, rather than just the poorly performing ones.

[16]Of course, this represents a transfer of surplus if ownership of firms is unequal across households.

Table 3.11: Welfare decomposition

	Observed			
	No exit		Exit	
Increased internet access				
+ Reduced shipping time				
direct welfare gains		5.4		5.7
net reallocation				
non exiters	4.3		3.2	
exiters	-	4.3	-13.1	-9.9
store operating costs		-		12.1
Δ welfare (%)		9.7		8.0
Increased internet access				
+ Reduced shipping time				
+ Expansion into new industries[a]				
direct welfare gains		10.3		8.2
net reallocation				
non exiters	9.2		13.5	
exiters	-	9.2	-21.9	-8.4
store operating costs		-		13.5
Δ welfare (%)		19.5		13.4

[a] Expansion into new industries such that all counterfactuals with all three improvements matches observed increase in online share.

These results highlight the substantial dampening effect store closures can have on aggregate welfare. Exiters subtract between 13.1 to 21.9 percentage points off welfare gains with the subsequent reallocation to second best alternatives recuperating only 3.2 and 13.5 percentage points. The substantial negative effect of net reallocation indicates that there are many consumers who actually lose out from improvements by online retailers. Overall, these store closures and reallocation leads to positive effects on welfare as there are savings in the costs of operating stores. However, these savings are captured by owners of firms which are unlikely to compensate consumers unless they are owners of retail firms.

3.5.1.6 Effects on firms

Table 3.12 decomposes the aggregate change in retail profits caused by improvements in online retail. The counterfactuals suggest that aggregate profits decline

by 8.2 per cent. The aggregate decline masks divergent outcomes for brick and mortar firms compared to online retailers. Online retailers add 8.8 percentage points to aggregate profits which is considerable given their size while brick and mortar profits subtract 17 percentage points. The fall in brick and mortar profits is driven by losses at both exiters and non exiters, who subtract 3 and 14 percentage points off aggregate growth respectively.

Table 3.12: Decomposing aggregate profits

	Observed			
	No exit		Exit	
Increased internet access				
+ Reduced shipping time				
+ Expansion into new industries				
online retailers		7.8		8.8
brick and mortar stores				
non exiters	-23.2		-14.0	
exiters	-	-23.2	-3.0	-17.0
Δ total profits (%)		-15.4		-8.2

Although improvements by online retailers unequivocally lead to less sales at traditional retailers and store closures, effects on retail firms are heterogeneous since competitive pressure varies across areas. Stores that manage to survive can often experience an increase in profits if they soak up enough market share from exiting stores to offset the loss of customers to online retailers. When increased competition from online retail induces a store to close, former customers of the closed store are reallocated to a variety of stores, not only to online retailers. Stores that have relatively fat operating margins are less likely to exit from an initial bout of increased competition, and stand to gain if there are a large number of store closures nearby.

To gauge the extent to which some firms are better off, I calculate percentiles of the percentage change in firm profits across surviving firms. The results are displayed in Table 3.13. The median surviving firm suffers a 12.2 reduction in profits with the top 10 percent of surviving firms actually experiencing an increase

Table 3.13: Distribution of profit changes at surviving firms

	Percentiles - change in profits (%)						
Average across industries	**5**	**10**	**25**	**50**	**75**	**90**	**95**
Increased internet access							
+ Reduced shipping time							
+ Expansion into new industries							
No exit	-100.0	-100.0	-100.0	-54.2	-17.1	-5.2	-2.6
Exit	-69.6	-54.3	-30.2	-12.2	-3.0	0.1	20.2

in profits. Because 17.9 percent of firms exit with all three shocks, the model suggests that 8.2 percent of firms increase profits in response to improvements by online retails. Moreover, these figures are likely to understate the increase in profits at some firms as the model does not account for increased online sales by previously brick and mortar-only firms.

3.5.2 Effects of the Marketplace Fairness Act (2013)

Because of a 1967 Supreme Court case, out-of-state retailers have been exempt from collecting state sales taxes from customers due to the complexity of collecting taxes for multiple states, with the many variations in tax rates, exemptions and record keeping requirements. The Marketplace Fairness Act, before Congress at the time of publication, seeks to overcome this ruling by allowing states to compel online retailers to collect state taxes in compensation for the simplification of states sales tax laws.

Since state sales taxes are already collected by traditional retailers, the passing of the act would represent an increase in the relative price of goods purchased online. Supporters therefore claim that the act would level the playing field between online retailers and traditional retailers. To simulate the implementation of the act, I simply increase the effective markup that online retailers charge over marginal costs by the sales tax rate based on the purchasing customer's location. States are required to establish a uniform tax rate before they can compel online retailers to collect sales taxes. For this reason, I use sales tax rates based on the

state tax rate and the average combined city and county tax rates as of August 2013. A summary of these tax rates are listed in Table 3.14.

Table 3.14: Average state tax rates inclusive of city and county rates

Tax rate (%)	**States**
≤ 5	Alaska, Delaware, Hawaii, Maine, Montana, New Hampshire, Oregon
>5 & ≤ 6	District of Columbia, Kentucky, Maryland, Michigan, Nebraska, North Dakota, South Dakota, Virginia, Wisconsin, Wyoming
>6 & ≤ 7	Colorado, Connecticut, Florida, Georgia, Idaho, Iowa, Indiana, Massachusetts, New Jersey, New Mexico, North Carolina, Pennsylvania, Rhode Island, Utah, Vermont, West Virginia
>7 & ≤ 8	Arizona, Ohio, Kansas, South Carolina, Texas
> 8	Arkansas, California, Oklahoma, Illinois, Louisiana, Tennessee, Washington

The estimates for the utility from the online option do not include the contribution of markups since it was not necessary to estimate the components separately. To get the price of the online good from which to implement the sales taxes, I use data from the financial statements of Amazon.com to calculate markups inclusive of shipping to get measures of net markups for the online sector in 2012.[17] Gross margins net of shipping discounts at Amazon.com were 20.1 per cent in 2012 which implies a value for p_o of 1.252. Given this estimate of the price of the online good, the utility from shopping online becomes $u_{o,tax}$ with the tax rate, τ, dependent on the location of the purchasing customer.

$$u_{o,tax} = u_{o,t} - \eta p_o \tau$$

Before imposing the online sales tax, I extrapolate the model to 2012 by using the counterfactual where internet access improves, delivery times fall and there is a general expansion of online retail across newer retail categories. Therefore results should be interpreted as the predicted outcome had the MFA been implemented sometime during 2007-12.

The results from are displayed in Table 3.15. Absent implementation of the

[17]Domestic sales at Amazon.com represented roughly 13 per cent of total online sales in 2011.

Table 3.15: Estimated potential effects of Marketplace Fairness Act (2007-12)

	No exit		Exit	
	Baseline	**Tax**	**Baseline**	**Tax**
Δ online share of sales (ppt)	5.2	2.6	5.2	3.4
Δ online sales (%)	95.9	58.0	162.2	69.9
Δ brick and mortar sales (%)	-9.6	-4.8	-15.5	-9.3
Δ brick and mortar stores (%)	-	-	-15.4	-11.9
Δ brick and mortar stores ('000s)	-	-	-78	-60

tax, online sales increase by between 162.2 per cent with 78,000 store closures. The tax has a large effect, paring increases in online sales by roughly 90 percentage points leading to online sales in 2012 that are 0.54 times online sales without the tax. These numbers are larger than estimated from a price comparison website in Ellison and Ellison (2009), which calculates a decline in online sales of 30 per cent if offline sales taxes were eliminated. It is not surprising that the impact here is larger since the data in Ellison and Ellison (2009) come from a price comparison website for memory modules, where price sensitivity is likely to be relatively high.

3.6 Conclusion

In this paper, I estimated that online retail had a large positive effect on welfare. The estimated model suggests that the 5.2% increase in the online share of sales from 2007-12 is associated with a 13.4% increase in welfare, which incorporates both producer and consumer surplus. This magnitude of these effects have broader implications because the Retail and Wholesale Trade sectors are large, and also because they are indicative of further gains in other sectors being transformed by online services.[18] The welfare gains from this internet technology are in addition to the prior productivity improvements in logistics and inventory management due to information technology. In all, these technological advances and their effects on welfare are strong evidence of the role of recent advances in computer-based

[18]For example, consider the rapid growth of Uber, Lyft and Sidecar in transportation and Square in financial services.

technology in spurring economic growth.

Key to quantifying the effects on welfare is measuring each store's contribution to welfare, akin to store-level retail productivity, accurately. Instead of using standard measures of retail productivity such as value added per worker to measure a store's contribution to welfare, I use a measure that reflects the surplus (consumer surplus and profits) that each visit by a consumer generates. This new measure is robust to variation in competitive pressure, unlike standard measures of productivity which are distorted. Of course, the tradeoff is that a measure of competition is required for every store, which I obtain by using a structural model.

I focus on two margins by which online retail affects aggregate welfare in the retail sector. The first are the gains in consumer surplus from improvements in online retail holding fixed market shares. The next is the change in the market structure of retail induced by reallocation. The changing market structure is complicated as reallocation occurs locally rather than at an aggregate level and the bulk of reallocation is achieved by store closures. Reallocation incorporates the optimal switching of consumers from stores to online retail, the negative effect on consumer surplus from store closures and the savings in store operating costs from closing stores.

To account for these rich local-level selection dynamics and to measure store-level economic surplus, I develop a model of retail that is estimated on store-level data spanning the universe of retail stores in the US. The model takes the geography of the US seriously with stores existing in locations that reflect their real-world location, with competition for customers occurring at a local level. Crucially, the model is one of industry equilibrium with stores setting prices based on local market power and store closures based on firm profit maximization. Therefore, counterfactuals can be used to identify the effects on aggregate welfare, holding all else constant.

Estimates from the model exhibit substantial within-industry heterogeneity in

stores' fixed costs and distributions of store quality (economic surplus per store) across locations. The extent of heterogeneity in fixed costs implies that selection is based on store profitability rather than on store quality alone. Moreover, since distributions of store quality vary across locations, selection occurs between stores operating in local markets rather than between retail firms competing in an aggregate market. Selection dynamics are markedly different from selection effects in a typical heterogeneous firms model, where the lowest productivity establishments are the first to exit in response to an adverse shock. Stores offering consumers relatively little rather than high quality stores may survive if they are more profitable. In counterfactual exercises, a large fraction of firms actually experience an increase in profits from improvements by online retailers as they soak up enough market share from exiting stores to offset lost market share to online retailers.

The counterfactual exercises assess the effect on aggregate welfare from 3 potential drivers of the increase in online sales. The observed increase in internet access from 2007 to 2012 is estimated to increase the online share of sales by 0.8 percentage points. Reduced shipping times that reflect Amazon.com's geographic expansion of its shipping and fulfillment warehouses and productivity improvements at UPS and FedEx from 2007 to 2012 are estimated to have a much larger effect, increasing the online share by 1.4 percentage points. The largest effect comes from expansion into new industries, which increases the online share by 2.9 percentage points. Results from a reduced form model confirm the aggregate decline in brick and mortar sales suggested by the structural model. The cumulative effect of these 3 changes increase welfare by 13.4%. Net reallocation subtracts 8.4 percentage points off the aggregate figure, while direct welfare gains and savings in store operating costs add 8.2 and 13.5 points respectively.

Finally, the paper considers the effect of the Marketplace Fairness Act, a bill before Congress that would allow states to compel online retailers to collect state sales taxes. I introduce that the tax as a state specific distortion to relative prices.

The model estimates a large effect from the tax if implemented between 2007-12. Sales at brick and mortar stores would have decreased by one-third less, saving 18,000 stores in the process. The results suggest that online sales would fall by 46 percent, slightly higher than the figure of 30 percent estimated in Ellison and Ellison (2009) which uses data from a price comparison site for memory modules.

3.7 Appendix

3.7.1 Estimates of internet access by zip code

To obtain estimates of internet access, I use microdata from the CPS. Unfortunately, the finest geographic level at which the data are available is at the MSA level which is relatively coarse and covers only metropolitan areas. However, the CPS provides data on each respondent to the survey, with weights that aggregate up to the population level. This demographic data allows me to generate zip code level estimates of internet access by applying estimates from individual level models to more aggregated data. First, I use demographic and geographic data on each respondent to estimate a model of individual level internet access. Given these estimates, I construct estimates for the fraction of the population with access to the internet by zip code, by applying the model to demographic data for each zip code. This is equivalent to assuming that representative individuals populate each zip code, each characterized by zip code-level demographic data.

Table 3.16: Internet access and demographics

Variables[1]	**1997**	**2002**	**2007**	**2012**
constant	-14.77	-14.15	-11.82	-9.40
age	0.0022	-0.0008	0.0012	0.0029
age^2	-0.0003	-0.0003	-0.0003	-0.0003
log(household income)	1.30	1.33	1.22	1.06
Dummy Variables				
white	0.28	0.74	0.39	0.33
black	-0.66	-0.11	-0.26	-0.21
asian	0.32	1.01	0.49	0.49
hispanic	-0.68	-0.29	-0.60	-0.36
high school	-0.05	0.08	0.14	0.27
some college	0.48	0.58	0.74	0.72
associates degree	0.46	0.57	0.78	0.81
bachelors degree or higher	0.85	0.98	1.15	1.25

[1] N $\equiv$ min. 270 miliion; includes state dummy variables. Race and education groups are mutually exclusive.

The results are broadly consistent with priors that internet access increases with income and education, decreases with age with whites and asians having an

advantage over hispanics, blacks and others. As expected, the model suggests that the fraction of the population with internet access has increased over time which is consistent with aggregate CPS data which indicate that the percentage of the population with internet access at home was 29.9, 56.6, 75.6 and 85.8 in 1998, 2003, 2007 and 2012 respectively. Increased access to the internet country wide is reflected in the estimate of the constant parameter which is increasing monotonically with time. While internet access has increased overall, the model also indicates that the increase has been relatively larger for some groups, likely reflecting saturation. In 1997, a white individual had 2.55 and 2.61 times the odds of having access to the internet compared to black and hispanic individuals. By 2012, the differences in the odds of internet access for whites compared to blacks and hispanics had dropped to 1.71 and 1.99 respectively as relatively more blacks and hispanics obtained access to the internet. Similarly, household income becomes less important over time as a predictor of internet access. Note that this is not driven by inflation since I deflate household income using the CPI to 2007 dollars.

To extrapolate these individual level estimates to the zip code level, I require zip code level data on all the variables that the individual level model was estimated on. All the variables map to variables available by zip code in the 2000 Decennial Census and the 5 year American Community Survey (ACS) 2007-2011. I use data from the 2000 Decennial Census for years 1997 and 2002 and data from the ACS for the years 2007 and 2012. Results have also been generated using linear interpolation/extrapolation but do not generate any meaningful differences and hence are not reported. To obtain estimates of internet access by zip code for a given year, I apply the individual level model with coefficients from the given year to demographic data by zip code for that year. The key difference is that rather that having dummy variables, I have proportions of a zip code's population that fit each education and race group. Hence the estimates are akin to estimating

the probability that a representative individual has access to the internet for each location.

3.7.2 Proof of uniqueness and existence

For clarity, I prove uniqueness and existence for the model with complete internet access. The proof of the model with heterogeneous internet access is a straightforward extension where weights of market shares need to be accounted for. The proof that equilibrium prices exist and are unique relies on proving that the following operator is a contraction. Define the operator $T : \mathbb{R}^k \to \mathbb{R}^k$, element by element as follows

$$T_j(p) = p_j + \beta \left(\log \left(c + \frac{1}{\eta \left(1 - \bar{v}_i(p)\right)} \right) - \log(p_j) \right) \tag{3.24}$$

for some $\beta \in (0, 1)$ and $\bar{v}_i(p)$ is calculated using p.

$$\bar{v}_i(p) = \sum_{l \in \mathbb{H}(i)} \omega_l v_{l,i}(p)$$

where

$$\omega_l = \frac{m(l)n(l)}{\sum_{l \in \mathbb{H}(i)} m(l)n(l)}$$

Lemma 1. *$\exists\ \beta$ such that $\forall j, k$, $\frac{\partial T_j(p)}{\partial p_k} \geq 0$ and $\sum_k \frac{\partial T_j(p)}{\partial p_k} < 1$.*

Proof. Let i be an arbitrary store and denote $v_{l,max}(i)$ as the market share in location l for store i when $p_i = c$ and $\forall j \neq i, p_j \to \infty$. Define the maximum market share as follows $\bar{v} = \max_i \max_l v_{l,max}(i)$. Note that $\bar{v} < 1$ due to the presence of the outside options, which always have positive market share due to

the logit error. Taking derivatives of equation (3.24),

$$\frac{\partial T_j(p)}{\partial p_j} = 1 - \beta \left(\frac{\sum_l \frac{\omega_l(1-v_{j,l})}{1-\bar{v}_j} v_{j,l}}{c(1-\bar{v}_j) + \frac{1}{\eta}} + \frac{1}{p_j} \right)$$

and $\forall k \neq j$,

$$\sum_k \frac{\partial T_j(p)}{\partial p_k} = \beta \frac{\sum_l \sum_k \frac{\omega_l v_{j,l}}{1-\bar{v}_j} v_{k,l}}{c(1-\bar{v}_j) + \frac{1}{\eta}}$$

It suffices to show that $\frac{\partial T_j(p)}{\partial p_j} \geq 0$ and that $\sum_k \frac{\partial T_j(p)}{\partial p_k} < 1$ which in turn reduce to showing that the following hold.

$$\beta \left(\frac{\sum_l \frac{\omega_l(1-v_{j,l})}{1-\bar{v}_j} v_{j,l}}{c(1-\bar{v}_j) + \frac{1}{\eta}} + \frac{1}{p_j} \right) \leq 1 \tag{3.25}$$

$$\frac{\sum_l \sum_k \frac{\omega_l v_{j,l}}{1-\bar{v}_j} v_{k,l}}{c(1-\bar{v}_j) + \frac{1}{\eta}} < \frac{\sum_l \frac{\omega_l(1-v_{j,l})}{1-\bar{v}_j} v_{j,l}}{c(1-\bar{v}_j) + \frac{1}{\eta}} + \frac{1}{p_j} \tag{3.26}$$

To ensure the equation (3.25) holds, one can simply set β low enough such that equation (3.25) binds for a value for the term inside the brackets guaranteed to be greater than for any store. Setting $\beta = \beta_{bound}$ ensures equation (3.25) is satisfied for all stores as $\forall j, p_j > c, v_{j,l} < 1$ and $\bar{v}_j < \bar{v}$.

$$\frac{1}{\beta_{bound}} = \frac{\sum_l \frac{\omega_l}{1-\bar{v}}}{c(1-\bar{v}) + \frac{1}{\eta}} + \frac{1}{c}$$

Finally, equation (3.26) holds since prices are positive, $p_j > 0$, and the presence of the outside options guarantee that $\sum_k v_{k,l} < 1 - v_{j,l}$.

□

Lemma 2. *There is a value, $\bar{p}$, such that if for any j, $p_j > \bar{p}$, then for some k, $T_k(p) < \bar{p}$.*

Proof. Set $\bar{p} = c + \frac{1}{\eta(1-\bar{v})} + \epsilon$ for any $\epsilon > 0$ where $\bar{v}$ is defined in the previous

lemma. Let j be an arbitrary store and assume $p_j > \bar{p}$. It suffices to show that $\log\left(c + \frac{1}{\eta(1-\bar{v}(i;p))}\right) - \log(p_j) < 0$.

$$\begin{aligned}\log\left(c + \frac{1}{\eta\left(1-\bar{v}(i;p)\right)}\right) &\leq \log\left(c + \frac{1}{\eta\left(1-\bar{v}\right)}\right) \\ &< \log\left(p_j\right)\end{aligned}$$

□

Proof of existence and uniqueness

Proof. The operator T satisfies the conditions of the contraction proof in Berry et al. (1995) as Lemmas 1 and 2 hold. Since T is a contraction, there exists a unique fixed point to the operator T. This implies that the equilibrium defined by (3.9) exists and is unique since the existence of another equilibrium would contradict the uniqueness of the fixed point to the operator T. □

3.7.3 BLP algorithm with endogenous prices

Guess initial values for each store, $\left\{\hat{z}_j^0\right\}$. For every store and all its customers, calculate the probability that a customer in a location chooses that store assuming that for all brick and mortar stores prices are zero and the $\left\{\hat{z}_j^0\right\}$ are the unobserved store effects.

$$\hat{v}_l(i) = \frac{e^{\theta_{f(i)} + \mathbf{x_i}'\theta_i - \left(\zeta_0 + \zeta_1 \log popden_{i,l}\right)d_l(i) + \hat{z}_i}}{1 + e^{u_{o,t}} + \sum_{j \in \mathbb{S}(l)} e^{u_l(j)}} \tag{3.27}$$

Calculate prices according to the following equation

$$p(i) = c + \frac{1}{\eta\left(1 - \hat{v}(i)\right)} \tag{3.28}$$

where $\hat{v}(i)$ represents a store's sales-weighted average market share using equation (3.27).

$$\hat{v}(i) = \sum_{l \in \mathbb{H}(i)} \frac{y_l(i)}{\sum_{l' \in \mathbb{H}(i)} y_{l'}(i)} v_l(i)$$

Use equations (3.27) and (3.28) to calculate a store's nominal sales to customers in a given location and aggregate over locations to get total sales for each store. For each location, calculate total sales for stores in that location and each store's share of the total. Update the initial guess of $\left\{\hat{z}_j^0\right\}$ with the following updating equation.

$$z_i^k = z_i^{k-1} + \log(s_{i,data}) - \log(s_{i,model}) \tag{3.29}$$

Repeat the above until the algorithm converges to a fixed point.

Lemma 3. *A store's share of nominal sales calculated according to the algorithm is decreasing in its own $\hat{z}_i$.*

Proof. By using equation (3.27), it is straightforward to show that the probability of a customer visiting the store falls relative to all other stores in the same location, implying that the store's share of customers fall. Equation (3.28) shows that the store's price relative to other stores' prices in the same location falls as the store's relative share of customers fall. Finally, stores in the same location compete across a common set of markets implying that the store's relative prices and quantities fall in all markets. □

Theorem 4. *The operator defined by the BLP algorithm with endogenous prices is (i) a contraction, (ii), generates prices that are optimal and (iii) finds store level fixed effects that match observed market shares.*

Proof. Because Lemma 3 holds, the standard proof that the BLP algorithm is a contraction is applicable (Berry et al., 1995). To show that the algorithm generates optimal prices, calculate the store-level fixed effect by transforming the solution

to the fixed point problem with the algorithm-generated prices for each store.

$$z_j = \hat{z}_j + \eta p_j \tag{3.30}$$

By definition, the fixed point $\{\hat{z}_j\}$ and associated prices $\{p_j\}$ satisfies equations (3.27) and (3.28). Use (3.30) to substitute out $\{\hat{z}_j\}$ in equations (3.27) and (3.28). Define $\{z_j\}$ as the store level fixed effect. With this definition, it is clear that prices are optimal as the modified pricing equation, (3.28), resembles the model pricing equation, (3.9). Finally, $\{z_j\}$ and $\{p_j\}$ generate market shares that match observed market shares since $\{\hat{z}_j\}$ is a fixed point of the operator. □

3.7.4 Entry and exit algorithm

It is likely that multiple equilibria exist given that variations in the order of exit lead to different equilibria.[19] As an example, suppose that Store A and Store B (owned by separate firms) exist in some location with no other competitors and that a sudden increase in internet access amongst customers in that location reallocates sales away from Stores A and B. If it is the case that both would want to close if the other remained open and both would want to remain open if the other closed, then two equilibria exist, characterized by either store remaining open. I focus on the equilibrium generated by the following iterative procedure.

Algorithm:

1. Exit step.

   (a) For each store in operation, calculate the value to the firm of closing that store assuming that all other open stores remain open while

[19]Note that this problem does not arise in estimation since I assume that fixed costs are such that the equilibrium generates the observed set of stores. When performing counterfactuals, there is no such observed set of stores.

accounting for equilibrium price changes.

   (b) Rank those stores wishing to close in order of the absolute size of losses.

   (c) Close the store making the largest loss. If no stores wish to close, stop.

   (d) Return to step 1a.

2. Entry step.

   (a) For each store in operation, calculate the value to the firm of opening another store in the same location holding fixed the set of other stores while accounting for equilibrium price changes.

   (b) Rank those stores wishing to expand/re-open in order of the absolute size of gross profits.

   (c) Open the store with the largest potential profit. If no stores wish to open, stop.

   (d) Return to step 2a.

3. If both exit step and entry step conclude without changes in set of stores, stop otherwise return to step 1.

The algorithm has a natural interpretation where the industry alternates between an exit stage and an entry stage and concludes when no store wishes to exit or enter.. In the exit stage, stores exit sequentially in order of those making the largest losses. Stores exit sequentially with the ordering recalculated with each exit. When the exit step concludes, no store in operation wishes to exit. However, this situation is not necessarily an equilibrium as exit of some stores may make entry by other firms optimal and moreover, subsequent exits in the exit step may make previous exits suboptimal. Hence the algorithm needs to account for new store openings and re-entry by stores that exited prematurely in the exit step. The entry step ranks potential firm profits from opening a new store/re-opening

a previously closed store and opens stores sequentially until no more firms wish to enter/re-enter. By definition, if both the entry and exit steps conclude without changes in the set of stores, the current set of stores is an equilibrium since no firm wishes to adjust on the extensive margin.

As an example, it is illustrative to see how this resolves the dilemma above. Calculate the losess for each store assuming both stores remain open. Close the store with the largest loss, say, Store B. By definition, the remaining store, Store A, is now profitable. In the entry step, the firm owning Store B by definition does not wish to re-open. For simplicity, assume that the firm which owns Store A does not wish to open a new store. Redoing the exit and entry steps leads to no changes and hence we are left with the equilibrium where Store A remains open, the natural equilibrium where the store making the largest losses closes first or the store making the largest profits remains open.[20]

3.7.5 Results at bounds of fixed costs

[20]The equilibrium is similar in spirit to the "natural equilibrium" in Abbring et al. (2012).

Table 3.17: Results from counterfactual exercises

	Observed		**Maximum**	
	Low FC	High FC	Low FC	High FC
Increased internet access				
Δ online share of sales (ppt)	0.5	1.2	1.1	2.1
Δ online sales (%)	11.2	20.0	25.1	40.8
Δ brick and mortar sales (%)	-1.3	-6.1	-3.1	-10.0
Δ brick and mortar stores (%)	-4.0	-6.1	-7.4	-9.9
Δ brick and mortar stores ('000s)	-20	-31	-38	-50
Reduced shipping times				
Δ online share of sales (ppt)	1.2	1.7	2.7	3.5
Δ online sales (%)	26.5	33.4	62.5	74.3
Δ brick and mortar sales (%)	-1.4	- 5.3	-3.1	-9.1
Δ brick and mortar stores (%)	-6.8	-9.1	-12.4	-13.0
Δ brick and mortar stores ('000s)	-35	-46	-62	- 67
Expansion into new industries[1]				
Δ online share of sales (ppt)	2.5	3.8		
Δ online sales (%)	49.3	65.1		
Δ brick and mortar sales (%)	-6.5	-14.0		
Δ brick and mortar stores (%)	-4.9	-5.6		
Δ brick and mortar stores ('000s)	-25	-28		
Expansion into new industries				
+ Reduced shipping times				
Δ online share of sales (ppt)	1.9	3.0	4.5	6.4
Δ online sales (%)	40.8	54.9	102.4	127.8
Δ brick and mortar sales (%)	-2.9	-9.8	-6.9	-16.4
Δ brick and mortar stores (%)	-9.9	-12.0	-18.2	-19.2
Δ brick and mortar stores ('000s)	-50	-61	-93	-97
Increased internet access				
+ Reduced shipping times				
+ Expansion into new industries				
Δ online share of sales (ppt)	5.2	5.2	8.9	12.6
Δ online sales (%)	102.5	129.4	180.2	219.7
Δ brick and mortar sales (%)	-10.5	-20.5	-16.0	-26.9
Δ brick and mortar stores (%)	-14.6	-16.2	-23.2	-23.3
Δ brick and mortar stores ('000s)	-74	-82	-118	-118

[1] Expansion into new industries increase online retailers' relative-utility-to-stores in industries with below mean online shares of sales to the mean relative-utility-to-stores.

Bibliography

Abbring, J. H., Campbell, J. R., and Yang, N. (2012). Simple Markov-Perfect Industry Dynamics. 21, Federal Reserve Bank of Chicago.

Alvarez, F., Lippi, F., and Paciello, L. (2011). Optimal Price Setting with Observation and Menu Costs. *Quarterly Journal of Economics*, 126(4).

Angeletos, G.-M. and La'O, J. (2009). Incomplete Information, Higher-order Beliefs and Price Inertia. *Journal of Monetary Economics*, 56(S1):S19–S37.

Arkolakis, C. (2010). Market Penetration Costs and the New Consumers Margin in International Trade. *Journal of Political Economy*, 118(6).

Atkeson, A. and Burstein, A. T. (2010). Innovation, Firm Dynamics, and International Trade. *Journal of Political Economy*, 118(3):433–484.

Atkeson, A. and Kehoe, P. J. (2005). Modeling and Measuring Organization Capital. *Journal of Political Economy*, 113(5):1026–1053.

Bartelsman, E. J., Haltiwanger, J. C., and Scarpetta, S. (2009). Cross-Country Differences in Productivity: The Role of Allocation and Selection. Technical Report 15490, National Bureau of Economic Research.

Berry, S., Levinsohn, J., and Pakes, A. (1995). Automobile Prices in Market Equilibrium. *Econometrica*, 63(4):841–890.

Bonomo, M., Carvalho, C., and Garcia, R. (2010). State-Dependent Pricing Under Infrequent Information: A Unified Framework. Staff Reports 455, Federal Reserve Bank of New York.

Bresnahan, T. F. and Reiss, P. C. (1991). Entry and Competition in Concentrated Markets. *Journal of Political Economy*, 99(5):977–1009.

Cabral, L. M. B. and Mata, J. (2003). On the Evolution of the Firm Size Distribution: Facts and Theory. *American Economic Review*, 93(4):1075–1090.

Campbell, J. R. and Hopenhayn, H. A. (2005). Market Size Matters. *The Journal of Industrial Economics*, 53(1):1–25.

Caplin, A. S. and Spulber, D. F. (1987). Menu Costs and the Neutrality of Money. *The Quarterly Journal of Economics*, 102(4):703–25.

Chiou, L. (2009). Empirical Analysis of Competition between Wal-Mart and Other Retail Channels. *Journal of Economics & Management Strategy*, 18(2):285–322.

Collard-Wexler, A. and Loecker, J. D. (2013). Reallocation and Technology: Evidence from the U.S. Steel Industry. Technical Report 18739, National Bureau of Economic Research.

Davis, P. (2006). Spatial Competition in Retail Markets: Movie Theaters. *The RAND Journal of Economics*, 37(4):964–982.

Dinlersoz, E. M. and Yorukoglu, M. (2012). Information and Industry Dynamics. *American Economic Review*, 102(2):884–913.

Ellison, G. and Ellison, S. F. (2009). Tax Sensitivity and Home State Preferences in Internet Purchasing. *American Economic Journal: Economic Policy*, 1(2):53–71.

Fajgelbaum, P. (2013). Labor Market Frictions, Firm Growth, and International Trade. Technical Report 19492, National Bureau of Economic Research.

Fishman, A. and Rob, R. (2003). Consumer Inertia, Firm Growth And Industry Dynamics. *Journal of Economic Theory*, 109(1):24–38.

Foster, L., Haltiwanger, J., and Krizan, C. J. (2006). Market Selection, Reallocation, and Restructuring in the U.S. Retail Trade Sector in the 1990s. *Review of Economics and Statistics*, 88(4):748–758.

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

Foster, L., Haltiwanger, J., and Syverson, C. (2012). The Slow Growth of New Plants: Learning about Demand? Technical Report 12-06, Center for Economic Studies, U.S. Census Bureau.

Golosov, M. and Lucas, R. E. (2007). Menu costs and phillips curves. *Journal of Political Economy*, 115(2):171–199.

Gourio, F. and Rudanko, L. (2011). Customer Capital. Technical Report 17191, National Bureau of Economic Research.

Haltiwanger, J., Jarmin, R. S., and Miranda, J. (2012). Who Creates Jobs? Small versus Large versus Young. *Review of Economics and Statistics*, 95(2):347–361.

Holmes, T. J. (2011). The Diffusion of Wal-Mart and Economies of Density. *Econometrica*, 79(1):253–302.

Hopenhayn, H. A. (1992). Entry, Exit, and Firm Dynamics in Long Run Equilibrium. *Econometrica*, 60(5):1127–50.

Hsieh, C.-T. and Klenow, P. J. (2012). The Life Cycle of Plants in India and Mexico. Technical Report 18133, National Bureau of Economic Research.

Jarmin, R. S. and Miranda, J. (2002). The Longitudinal Business Database. Working Papers 02-17, Center for Economic Studies, U.S. Census Bureau.

Kaplan, G. and Menzio, G. (2014). The Morphology of Price Dispersion. Technical Report 19877, National Bureau of Economic Research.

Kinney, S. K., Reiter, J. P., Reznek, A. P., Miranda, J., Jarmin, R. S., and Abowd, J. M. (2011). Towards Unrestricted Public Use Business Microdata:

The Synthetic Longitudinal Business Database. Working Papers 11-04, Center for Economic Studies, U.S. Census Bureau.

Klenow, P. J. and Kryvtsov, O. (2008). State-Dependent or Time-Dependent Pricing: Does It Matter for Recent U.S. Inflation? *The Quarterly Journal of Economics*, 123(3):863–904.

Krusell, P. and Smith, A. A. (1998). Income and Wealth Heterogeneity in the Macroeconomy. *Journal of Political Economy*, 106(5):867–896.

Lagakos, D. (2009). Superstores or Mom and Pops? Technology Adoption and Productivity Differences in Retail Trade. Staff Report 428, Federal Reserve Bank of Minnesota.

Levy, D., Mark Bergen, S. D., and Venable, R. (1997). The Magnitude of Menu Costs: Direct Evidence from Large U.S. Supermarket Chains. *The Quarterly Journal of Economics*, 112(3):791–825.

Luttmer, E. G. J. (2011). On the Mechanics of Firm Growth. *Review of Economic Studies*, 78(3):1042–1068.

Mackowiak, B. and Wiederholt, M. (2009). Optimal Sticky Prices under Rational Inattention. *American Economic Review*, 99(3):769–803.

Mankiw, N. G. and Reis, R. (2002). Sticky Information Versus Sticky Prices: A Proposal To Replace The New Keynesian Phillips Curve. *The Quarterly Journal of Economics*, 117(4):1295–1328.

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

Morales, E., Sheu, G., and Zahler, A. (2013). Gravity and Extended Gravity: Estimating a Structural Model of Export Entry.

Nakamura, E. and Steinsson, J. (2008). Five Facts about Prices: A Reevaluation of Menu Cost Models. *The Quarterly Journal of Economics*, 123(4):1415–1464.

Nakamura, E. and Steinsson, J. (2010). Monetary Non-Neutrality in a Multi-Sector Menu Cost Model. *Quarterly Journal of Economics*, 125(3).

Woodford, M. (2001). Imperfect Common Knowledge and the Effects of Monetary Policy. Technical Report 8673, National Bureau of Economic Research.