The London School of Economics and Political Science

The Industrial Organisation of Financial Intermediation

Patrick Coen

A thesis submitted to the Department of Economics of the London School of Economics and Political Science for the degree of Doctor of Philosophy, London, June 2020.

Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I declare that my thesis consists of around 31,000 words.

Statement of co-authored work

I confirm that Chapter 1 was jointly co-authored with Jamie Coen and I contributed 50% of this work.

Acknowledgements

I am deeply grateful to my supervisors, Alessandro Gavazza and Christian Julliard, for their guidance and support. I am proud to have been their student.

I would like to thank members of the research community in industrial organisation at LSE, and John Sutton and Mark Schankerman in particular, for many helpful comments. I would also like to thank Jamie Coen, Wolfgang Ridinger, Claudia Robles Garcia, William Matcham, Adrien Bussy and Miguel Bandiera for their help and friendship.

I acknowledge the financial support of the Economics and Social Research Council. I am grateful to the Bank of England for providing data and a visiting position there. The views in this thesis are my own, and not necessarily those of the Bank of England or its committees.

I dedicate this thesis to Peter, Susie, Alex, Macey and, above all, my wife Zita. Without her encouragement, support and hard work none of this would have been possible.

Abstract

This thesis consists of three chapters on the industrial organization of financial intermediation.

The first chapter, which is co-authored with Jamie Coen, considers the interbank market and how it should be regulated. The interbank network, in which banks compete with each other to supply and demand financial products, creates surplus but may also result in risk propagation. We examine this trade-off by setting out a model in which banks form interbank network links endogenously, taking into account the effect of links on default risk. We estimate this model based on novel, granular data on aggregate exposures between banks. We find that the decentralised interbank network is not efficient, primarily because banks do not fully internalise a network externality in which their interbank links affect the default risk of other banks. A social planner would be able to increase surplus on the interbank network by 13% without increasing mean bank default risk or decrease mean bank default risk by 4% without decreasing interbank surplus. We propose two novel regulatory interventions (caps on aggregate exposures and pairwise capital requirements) that result in efficiency gains.

The second chapter considers the effect of the business cycle on outcomes in the mutual fund industry. The business cycle induces turnover in mutual funds: they exit in recessions and enter in recoveries. The effect of this firm turnover on welfare depends on a key trade-off: on the one hand, the business cycle "cleanses" the market of low quality exiting funds and replaces them with entrants that may on average be higher quality. On the other hand, the entrants have no returns history and so investors have less precise beliefs about their ability, where this "information loss" leads to misallocation that harms welfare. I examine this trade-off by estimating a structural model in which rational investors form and update beliefs about competing mutual funds that endogenously choose to enter and exit the market. I estimate this model using data on US mutual funds. I find that the business cycle has material, persistent effects that are negative in the short-term but turn positive as the effect of information loss decays over time.

The third chapter considers local competition between mutual funds. Mutual funds with similar investment strategies compete with each other for investment opportunities. I set out a model of demand for mutual funds in which (i) funds are located within a network depending on similarities in their investment strategies and (ii) funds impose negative spillovers on each other through this network. I structurally estimate this model using data on US equity

mutual funds. I identify these network spillovers based on how investors in a given mutual fund respond to the returns performance of its competitors. I find that local competition has a material impact on fund size, in that absent competition the median fund would be 20% bigger, and on cross-sectional variation in size. I perform counterfactual simulations in which I demonstrate that luck can play an important role even when funds are skilled and investors are rational: I find that luck accounts for 9% of cross-sectional variation in mutual fund size.

Contents

Tables

Figures

Chapter 1:
A structural model of interbank network formation and contagion

with Jamie Coen.[1]

The interbank network, in which banks compete with each other to supply and demand financial products, creates surplus but may also result in risk propagation. We examine this trade-off by setting out a model in which banks form interbank network links endogenously, taking into account the effect of links on default risk. We estimate this model based on novel, granular data on aggregate exposures between banks. We find that the decentralised interbank network is not efficient, primarily because banks do not fully internalise a network externality in which their interbank links affect the default risk of other banks. A social planner would be able to increase surplus on the interbank network by 13% without increasing mean bank default risk or decrease mean bank default risk by 4% without decreasing interbank surplus. We propose two novel regulatory interventions (caps on aggregate exposures and pairwise capital requirements) that result in efficiency gains.

[1]The views in this paper are those of the authors, and not necessarily those of the Bank of England or its committees. We are particularly grateful to Alessandro Gavazza and Christian Julliard for many helpful discussions. We are also grateful for comments by seminar participants at the Bank of England, the Federal Reserve Bank of New York, the London School of Economics, Princeton, Stanford, the Toulouse School of Economics, Universidad Pompeu Fabra, Queen Mary University of London, and conference participants at the RES Junior Symposium 2019 and EARIE 2019. We are grateful to the Bank of England for providing the data. Both authors acknowledge the financial support of the Economic and Social Research Council.

1.1 Introduction

Direct interconnections between banks are important in two ways. First, these interconnections fulfill a function, in that there are gains to trade. The interconnection could, for example, involve providing liquidity or acting as the other party in a hedging transaction, which may result in surplus on both sides of the trade. Second, interconnections can open up at least one side of the transaction to counterparty risk: a lender, for example, runs the risk that the borrowing bank will not pay it back. Both sides of this trade-off were important during the financial crisis and remain important today, and consequently there is significant debate about optimal regulation in this context (Yellen, 2013).

We consider the following fundamental economic questions. How does the network of direct interconnections between banks, which we term the interbank network,[2] affect systemic risk? How do banks form the interbank network, given the effect of such exposures on their risk? What inefficiencies exist in network formation? The answers to these economic questions then lead us to two questions about regulation. Given equilibrium responses by banks, is regulation effective in reducing default risk? If it does reduce default risk, does it do so efficiently in a way that preserves interbank surplus? Understanding the equilibrium effect of prospective regulation on outcomes in this market is of first-order importance, but is a difficult problem because banks respond endogenously to any changes in regulation.

We answer these questions by estimating a structural equilibrium model in which banks form the interbank network endogenously, taking into account the effect of their choices on their default risk. The key mechanism in this model is that when a bank takes on an exposure through the interbank network it earns a return, but it may also become riskier, which endogenously increases its funding costs. We estimate this model based on novel, rich Bank of England data on interbank exposures, and show that the model fits the data well both in and out of sample.

We are the first, to our knowledge, to estimate a structural model of the trade-off between surplus on the interbank network and the causal effect of the network on bank default risk. This allows us to make the following contributions: (1) we show how standard measures of bank systemic importance are biased, (2) we quantify the inefficiency of interbank network formation and (3) we examine the equilibrium effects of regulation, and propose alternative regulation that is more efficient.

[2]The "interbank market" is often used to describe short-term (often overnight) lending between banks. We use the "interbank network" more generally to cover any form of direct interconnection between banks.

The starting point for our work is Bank of England data on interbank exposures. These data are collected by the Bank of England through periodic regulatory surveys of 18 global banks from 2012 to 2018, in which they report the exposures they have to their most important banking counterparties. The data are novel, relative to the data commonly used in this literature, in two ways that are important for our context: (1) the data include a broad range of instruments, making them a reasonable proxy for a bank's *total* exposure to another bank and (2) the data contain rich detail on the types and characteristics of the instruments that make up each exposure. We set out various empirical facts about the network that inform our work, the most important of which is that there is significant variation in the size of exposures between banks, but not much variation in the presence of exposures: in other words, the network is dense but heterogeneous.

The features of our data and the empirical facts we observe guide our modelling choices in the following ways. First, the breadth of the data allows us to specify and estimate an *empirical* model of the effect of exposures on default risk, in a way which would not be feasible if we only observed exposures relating to a single instrument that is only a small subset of total exposures. Second, the fact that we observe a dense, heterogeneous network leads us to consider heterogeneity in *marginal* cost, in contrast to those parts of the empirical networks literature that seek to explain *sparse* network structures using *fixed* costs (Craig and Ma, 2019). Finally, the granularity of our data allows us to specify and then estimate a rich model of network formation, with a focus on allowing for as much observed and unobserved heterogeneity as possible.

With this general guidance in mind, we set out a model consisting of three parts: (1) the default risk process that relates the default risk of a bank to that of other banks and the exposures between them, (2) the demand for interbank financial products and (3) their supply, where demand and supply together determine network formation.

We model the default risk process as being spatially autocorrelated, such that bank i's default risk depends on its fundamentals and on its interbank exposures. These interbank exposures can have a *hedging effect* that reduces default risk, but also a *contagion effect* that increases default risk, where the net effect depends on the characteristics of the exposure and the counterparties involved. We generalise a standard spatially autocorrelated regression by allowing the strength of the contagion effect to vary across pairs: in other words, some links are inherently more risky than others, holding all other things (including exposure size and the default risk of both counterparties) constant. There are various reasons why this could be the case, the most important of which is *risk-sharing*: an exposure held by

bank i to bank j is likely to be particularly risky if the fundamentals of i and j are strongly positively correlated. This *heterogeneous contagion intensity* is an important part of our model. We refer to links with relatively low contagion intensity as "inherently safe" and links with relatively high contagion intensity as "inherently risky". The structure of this spatial autocorrelation is such that in equilibrium a bank's default risk depends on its exposures, but also the exposures of its counterparties and of its counterparties' counterparties (and so on).

Banks demand interbank financial products to maximise profits from heterogeneous technologies that take these differentiated interbank products as inputs. Banks supplying financial products receive a return, but also incur a cost because regulatory capital requirements mandate that they raise a certain amount of capital for the exposure that they take on when they supply. The key mechanism in this part of our model is that the cost of capital a bank incurs is an increasing function of its default risk. This default risk, per the default risk process we describe above, is a function of the bank's exposures, meaning that a bank supplying financial products endogenously changes its cost of capital when it does so. Heterogeneous contagion intensity means that this marginal cost varies across pairs: inherently risky links involve higher marginal cost.

Equilibrium trades and prices depend in an intuitive way on the key parameters of the model: (1) variation in contagion intensity is a key driver of link formation: inherently safe links are less costly and therefore more likely to be large, (2) risky banks pay more to be supplied financial products because contagion means it is more costly to supply them and (3) risky banks supply less, as their funding costs are higher. The most important source of market failure is network externalities, in which banks do not fully internalise the effect that their exposure choices have on the risk (and therefore also the funding cost) of their counterparties. We show that our model is consistent with the key empirical facts in our data, as well as some additional stylised facts from the financial crisis.

We estimate our model by matching two groups of moments: moments related to data on bank default risk and moments related to data on interbank exposures. To represent bank fundamentals we use, amongst other data, variation in regional equity indices: for example, we take a shock to a Japanese equity index as a shock that affects Japanese banks more than European banks. We then use these fundamentals to identify the key parts of our network formation model and the default risk process. The effect of counterparty risk in the default risk process depends on equilibrium exposures, which are endogenous. We address this endogeneity by using insights from the network formation part of our model: the default risk

process is, by assumption, *linear* in the fundamentals of banks, but our network formation game shows that equilibrium network links are *non-linear* functions of bank fundamentals. We therefore use non-linear variation in bank fundamentals as instruments for equilibrium links in the default risk process.

We estimate our model and show that it fits the data well in sample, before testing internal and external consistency in two ways. Our primary motivation for heterogeneous contagion intensity is based on risk-sharing, which implies a relationship between the parameters in our default risk process: links between banks whose fundamentals are closely correlated should be relatively high risk. We do not impose this relationship in estimation, but instead estimate these parameters freely and test the relationship post-estimation. We find evidence for risk-sharing, which we view as evidence of internal consistency. To test external consistency, we run an out of sample test: we use our model to simulate default risk for 2009 to 2011 and compare it to actual bank default risk, and show that (1) our model replicates some key patterns in the data and (2) our model outperforms the out of sample fit of a linear regression of default risk on fundamentals, in a way that the model would predict.

Our results imply that contagion through the interbank network is responsible for, on average, 9.8% of a bank's total default risk. We find significant variation in pairwise contagion intensity: the inherently riskiest links in the network are 50% riskier than the inherently safest links, holding all other things equal.

We then use our estimated results to answer the key questions set out above. We first describe two results relating to how the interbank network affects systemic risk. Our first result is that the overall effect of the interbank network depends on the economic climate: when bank fundamentals are good, then the hedging effect dominates the contagion effect, and the interbank network reduces systemic risk. When bank fundamentals are bad, the opposite is true: the contagion effect dominates the hedging effect and the interbank network increases systemic risk.

Our second result regarding systemic risk is that heterogeneity in contagion intensity has an important implication for the identification of systemically important banks within our network, which in our context means the banks that contribute most to bank default risk. There are various measures of systemic importance, but in general terms a bank is deemed systemically important if it has large exposures to other systemically important banks. Heterogeneous contagion intensity and endogenous network formation together show why this approach is likely to be flawed: *some links are large because they are inherently*

safe. Banks with large links like these would be incorrectly characterised as systemically important using standard network centrality measures based on unweighted network data. We propose an alternative measure of systemic importance based on network data that is weighted by the heterogeneous network effect parameters: an inherently risky (safe) link is scaled up (down). This weighted centrality measure implies materially different centrality rankings among banks: the bank that is most systemically important in our sample based on the unweighted network is only the 5th most systemically important bank based on our alternative risk-weighted centrality measure.

We then consider the efficiency of the decentralised interbank network, which we do by deriving an efficient frontier that shows the optimal trade-off between interbank surplus and bank default risk. We find that the decentralised interbank network is not on the frontier: a social planner would be able to increase interbank surplus by 13.2% without increasing mean bank default risk or decrease mean bank default risk by 4.3% without decreasing interbank surplus. This result is driven by the fact that our empirical results indicate that network externalities are significant. The social planner internalises the externality by considering the effect that a given link has on the risk of other banks, with the result that the social planner would (i) reduce aggregate exposures and (ii) reduce inherently risky exposures by relatively more than inherently safe exposures.

We then use our model to simulate the equilibrium effects of various forms of regulation, including a cap on individual exposures (Basel Committee, 2014b, 2018b) and an increase in regulatory capital requirements (Basel Committee, 2018a). We find that a cap on individual links is relatively ineffective: it has only a small effect on mean bank default risk, as in equilibrium banks shift their supply to uncapped links. Furthermore, a cap on individual links is inefficient, in that it has a large negative effect on interbank surplus, because it penalises large links that in equilibrium are more likely to be inherently safe. We instead propose capping aggregate exposures held by each bank, rather than individual exposures: an aggregate cap is more effective (because it prevents a bank moving capped supply to another bank) and more efficient (because in equilibrium banks respond to a cap on aggregate exposures by reducing relatively risky exposures by more than less risky exposures). Our results suggest that a social planner would strictly prefer our proposed cap on aggregate exposures to a cap on individual exposures.

We find that a general increase in capital requirements that applies equally across exposures to all banks is effective but inefficient: it decreases mean bank default risk, but at the cost of reduced interbank surplus. We instead propose a pairwise adjustment to cap-

ital requirements based on their heterogeneous contagion intensity: we give links that are inherently risky (inherently safe) greater (lower) capital requirements. In other words, we propose directly risk-weighting interbank exposures based on contagion intensity, as this targets regulatory intervention more closely at the network externalities that are the key driver of inefficiency in our model. Our results suggest that a social planner would strictly prefer our proposed pairwise capital requirement to a homogenous capital requirement.

We discuss related literature below. In Section 2, we introduce the institutional setting and describe our data. In Section 3, we set out our model. In Section 4, we describe our approach to estimation. In Section 5, we set out our identification strategy. In Section 6, we set out our results. In Section 7, we undertake counterfactual analyses. In Section 8, we conclude.

1.1.1 Related literature

Our work is related to three strands of literature: (i) the effects of network structure on outcomes in financial markets, (ii) endogenous network formation in financial markets and (iii) optimal regulation in financial markets.

There is an extensive literature on the effect of network structure on outcomes in financial markets, both theoretical (Acemoglu et al., 2015; Ballester et al., 2006; Elliott et al., 2014) and empirical (Denbee et al., 2017; Eisfeldt et al., 2018; Gofman, 2017; Iyer and Peydro, 2011). Our primary innovation is that we connect this empirical literature with the literature on network formation, by estimating a model of the effect of network structure on outcomes (default risk, in our case) simultaneously with a model of network formation. This allows us to make three contributions. First, using insights from our network formation model, we are able to directly address the endogeneity of the network when we estimate network effects, in contrast to large parts of the empirical literature.[3] Second, it allows us to consider equilibrium effects in counterfactual scenarios, taking into account how the network would respond endogenously.[4] Third, by combining a model of network formation with heterogeneous contagion intensity, we are able to show how existing measures of systemic importance are biased.

[3]See De Paula (2017) for a summary.

[4]Various papers (Eisfeldt et al. (2018) and Gofman (2017), for example) adjust the network arbitrarily (usually by simulating a failure) and show the impact on market outcomes holding network structure otherwise fixed. In our model, network structure responds endogenously to a counterfactual change.

There is a growing theoretical literature on network formation in financial markets (Babus, 2016; Farboodi, 2017; Chang and Zhang, 2018; Acharya and Bisin, 2014; Rahi and Zigrand, 2013), but little empirical work (Cohen-Cole et al., 2010; Craig and Ma, 2019; Blasques et al., 2018). Our contribution is that we are the first, to our knowledge, to structurally estimate a model of network formation in which banks trade off gains to interbank trade against contagion. Importantly, this allows us to quantify the extent of inefficiency in the market, and to study the implications of network structure for systemic risk.

We also contribute to the literature regarding optimal regulation in financial markets (Duffie, 2017; Baker and Wurgler, 2015; Greenwood et al., 2017; Batiz-Zuk et al., 2016). Our primary contribution is that by considering bank default risk we are able to evaluate bank regulation comprehensively. Various papers consider the effect of bank regulation on outcomes in specific markets,[5] but without considering bank default risk (which was arguably the primary focus of much recent banking regulation) it is not possible to draw any conclusions about whether regulation is optimal. Furthermore, our network formation model allows us to assess the equilibrium effects of regulation, taking into account the endogenous response of the network.

1.2 Institutional setting and data

We first describe the institutional setting of our work, including the relevant regulation. We then describe our data. We then use this data to set out some empirical facts that will guide our approach to modelling.

1.2.1 Institutional setting

Direct connections between banks fulfill an important function: "*there is little doubt that some degree of interconnectedness is vital to the functioning of our financial system*" (Yellen, 2013). Debt and securities financing transactions between banks are an important part of liquidity management, and derivatives transactions play a role in hedging. There is, however, widespread consensus that direct connections can also increase counterparty risk, with implications for the risk of the system as a whole (see, for example, Acemoglu et al. (2015)). This can be thought of, in loose terms, as a classic risk/reward trade-off. The

[5]Including Kashyap et al. (2010) on bank lending, Kotidis and Van Horen (2018) on the repo market and Bessembinder et al. (2018) and Adrian et al. (2017) on the bond market.

importance of both sides of this trade-off is such that direct interconnections between banks are the subject of extensive regulatory and policy-making scrutiny, whose aim is to: "*preserve the benefits of interconnectedness in financial markets while managing the potentially harmful side effects*" (Yellen, 2013).

After the 2008 financial crisis, a broad range of regulation was imposed on these markets. In this paper, we focus on two in particular: (1) caps on large exposures and (2) increases in capital requirements. We focus on these two because we think they are most relevant to our underlying economic research question, which is to examine the *efficiency* with which this risk/reward trade-off is balanced.

1.2.1.1 Large exposures cap

In 2014 the Basel Committee on Banking Supervision (BCBS) set out new standards for the regulatory treatment of banks' large exposures (Basel Committee, 2014b, 2018b). The new regulation, which came into force in January 2019, introduces a cap on banks' exposures: a bank can have no single bilateral exposure greater than 25% of its capital.[6]. For exposures held between two "globally systemic institutions", as defined in the regulation, this cap is 15%.

These requirements represent a tightening of previous rules, where they existed. For example, in the EU exposures were previously measured relative to a more generous measure of capital and there was no special rule for systemically important banks (AFME, 2017; European Council, 2018).

1.2.1.2 Capital requirements

Banks are subject to capital requirements, which mandate that their equity (where the precise definition of capital, Common Equity Tier 1, is set out in the regulation) exceeds a given proportion of their risk-weighted assets. Additional equity in principle makes the bank more robust to a reduction in the value of its assets, and so less risky. The total amount of capital E_{ij} that bank i is required to raise to cover asset j is the product of the value of the asset A_j, its risk-weighting ρ_{ij} and the capital requirement per unit of risk-weighted asset λ_i:

$$E_{ij} = \rho_{ij}\lambda_i A_j$$

[6]Where the precise definition of capital, in this case "Tier 1 capital", is set out in the regulation (Basel Committee, 2014b, 2018b)

The risk-weights, ρ_{ij}, can be calculated using banks' internal models or based on a standardised approach set out by regulators. Whilst risk-weights from banks' internal models are likely to vary by counterparty, the standardised approach is based on the credit rating relevant to the asset, and for the significant majority of interbank transactions between major banks this will be AAA or AA, the highest credit rating. In other words, for interbank transactions the standardised approach involves very little variation across i or j.[7]

In 2013 all banks in our sample faced the same capital requirement per risk-weighted unit, λ_i, which was 3.5%.[8] Since then, regulators have changed capital requirements in three ways. First, and most importantly, the common minimum requirement that applies to all banks has increased significantly. Second, capital requirements vary across banks, as systemically important banks face slightly higher capital requirements than non-systemically important banks. Third, capital requirements vary countercylically, in that in times of financial distress they are slightly lower (Basel Committee, 2018a). The result of these changes is that mean capital requirements for the banks in our sample has increased significantly, from 3.5% to over 9% in 2019. There have also been changes to the definition of capital and the measurement of risk-weighted assets, with the general effect of making capital requirements more conservative.

1.2.2 Data

1.2.2.1 Exposures

We define in general terms the exposure of bank i to bank j at time t as the immediate loss that i would bear if j were to default, as estimated at time t. The way in which this is calculated varies from instrument to instrument, but in general terms this can be thought of as (1) the value of the instrument, (2) less collateral, (3) less any regulatory adjustments intended to represent counterfactual variations to value or collateral in the event of default (for example, regulation typically requires a "haircut" to collateral when calculating exposures, as in the event of default any financial instruments provided as collateral are likely to be worth less).

[7]Banks are also subject to a leverage ratio requirement (Basel Committee, 2014a) which does not weight exposures according to risk.

[8]We use the minimum capital requirements as published by Basel Committee (2011) as the minimum requirements for banks. National supervisors can add discretionary buffers on top of these requirements, which we do not include in our empirical work.

We use regulatory data on bilateral interbank exposures, collected by the Bank of England. The dataset offers a unique combination of breadth and detail in measuring exposures. Much of the existing literature (such as Denbee et al. (2017)) on empirical banking networks relies on data from payment systems. This is only a small portion of the activities that banks undertake with each other and is unlikely to adequately reflect the extent of interbank activity or the risk this entails.

18 of the largest global banks operating in the UK report their top 20 exposures to banks over the period 2011 to 2018. Banks in our sample report their exposures every six months from 2011 to 2014, and quarterly thereafter. They report exposures across debt instruments, securities financing transactions and derivative contracts. The data are censored: we only see each bank's top 20 exposures, and only if they exceed £5 million. The data include granular breakdowns of each of their exposures: by type (e.g. they break down derivatives into interest rate derivatives, credit derivatives etc.), currency, maturity and, where relevant, collateral type.

We use this dataset to construct a series of snapshots of the interbank network between these 18 banks. We calculate the total exposure of bank i to bank j at time t, which we denote C_{ijt}, as the sum of exposures across all types of instrument in our sample. We winsorize exposures at the 99th percentile. The result is a panel of $N = 18$ banks over $T = 21$ periods from 2011 to 2018 Q2, resulting in $N(N-1)T = 6,426$ observations. For each C_{ijt}, we use the granular breakdowns to calculate underlying "exposure characteristics" that summarise the type of financial instrument that make up the total exposure. These 8 characteristics, which we denote d_{ijt}, relate to exposure type, currency, maturity and collateral type.

Although the dataset includes most of the world's largest banks, it omits banks that do not have a subsidiary in the UK.[9] Furthermore, for the non-UK banks that are included in our dataset, we observe only the exposures of the local sub-unit, and not the group. For non-European banks, this sub-unit is typically the European trading business.

1.2.2.2 Default risk

We follow Hull et al. (2009) and Allen et al. (2011) in calculating the (risk-neutral) probability of bank default implied by the spreads on publicly traded credit default swaps (data obtained from Bloomberg). This represents the market's estimate of bank default risk, as well as wider

[9]This is particularly relevant for some major European investment banks, who operate branches rather than subsidiaries in the UK, and hence do not appear in our dataset.

effects that are unrelated to the default risk of an individual bank (notably variations in the risk premium):

$$Prob(Default_{itT}) = 100(1 - (1 + (CDS_{itT}/10000)(1/rr))^{-T}$$

where rr is the assumed recovery rate, T is the period covered by the swap and CDS_{itT} is the spread.

1.2.2.3 Other data

We supplement our core data with the following:

- Geographic source of revenues for each bank from Bloomberg. Bloomberg summarises information from banks' financial statements about the proportion of their revenues that come from particular geographies, typically by continent, but in some cases by country.
- Macro-economic variables from the World Bank Global Economic Monitor, a panel of 348 macro series from a range of countries.
- Commodity prices from the World Bank "Pink Sheet", which is a panel of 74 commodity prices.
- S&P regional equity indices for US, Canada, UK, Europe, Japan, Asia, Latin America.

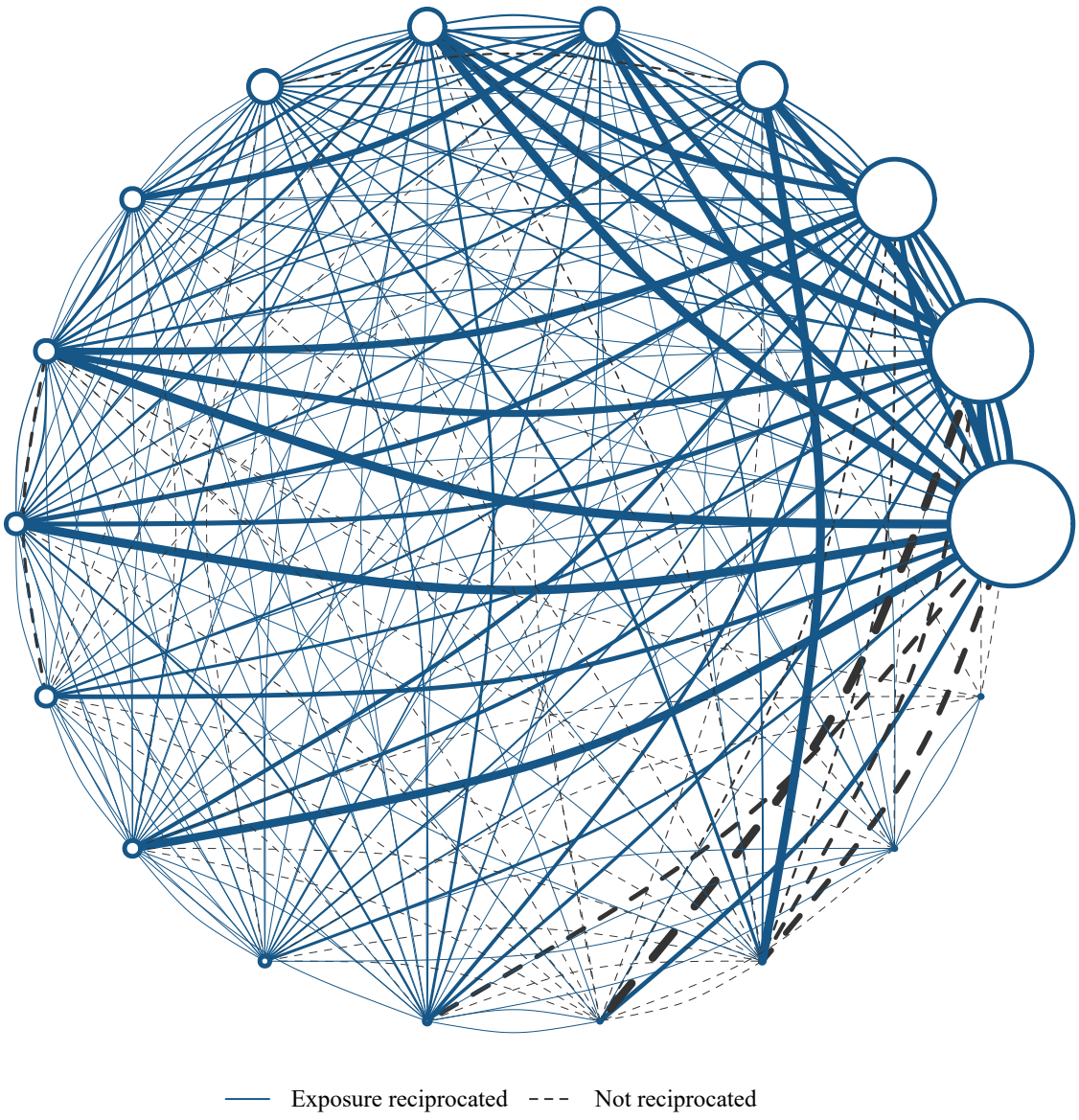
1.2.3 Summary statistics

The data reveal certain empirical observations about exposures and how they vary cross-sectionally and inter-temporally in our sample: (1) exposures in our data are large, (2) our observed network is dense and reciprocal, (3) network links are heterogeneous in intensity and characteristics and (4) the network has become more concentrated over our sample period. We discuss below how we use these empirical observations to guide our modelling.

Empirical fact 1: Exposures are large

The primary advantage of our data, relative to others used in the literature, is that it is intended to capture a bank's *total* exposures. The largest single exposure in our sample is

Figure 1.1: The aggregate network in H1 2015



Note: This is the network of aggregate exposures between banks in H1 2015. Each node is a bank in our sample. A solid line between two nodes shows a reciprocated exposure (each bank has an exposure to the other) and a dashed lines shows an unreciprocated exposure (that goes in one direction only). The line width is proportional to the size of the exposure. The size of the node is proportional to its total outgoings. The network is dense but heterogeneous in the size of individual exposures.

GBP 7,682m, the largest total exposures to other banks in a given period is GBP 26,367m. The mean exposure is GBP 285m and the mean total exposure to other banks in a given period is GBP 4,851m.

In this respect, our data has two important advantages over many of the data used in the literature. First, our dataset is the closest available representation of *total* exposures, when most other empirical assessments of interbank connections rely on a single instrument, such as CDS (Eisfeldt et al., 2018) or overnight loans (Denbee et al., 2017). Second, our data are on exposures, rather than simply market value, in that when banks report their exposures they account for collateral and regulatory adjustments. Data based solely on market value is a representation of bank activity, rather than counterparty risk.

Empirical fact 2: The network is dense and reciprocal

Figure 1.1 shows the network of exposures between banks in 2015 Q2. Our sample is limited to the core of the banking network, and does not include its periphery. Our observed network is, therefore, dense: of the $N(N-1)T$ links we observe in total, only approximately 30% are 0. One implication of the density of the network is that it is reciprocal: of the $N(N-1)T/2$ possible bilateral relationships in our sample, 55% are reciprocal, in that they involve a strictly positive exposure in each direction (that is, bank i has an exposure to bank j and bank j has an exposure to bank i).

Empirical fact 3: The network is heterogeneous in intensity and characteristics

Although the network is dense and so not particularly heterogeneous in terms of the presence of links, it is heterogeneous in the intensity of those links (that is, the size of the exposure), as shown in Figure 1.1. We further demonstrate this in Table 1.1, which contains the results of a regression of our observed exposures C on fixed effects. The R^2 from a regression on it fixed effects is 0.43: if all of bank i's exposures in a given time period were the same, then this would be 1.00. In other words, the low R^2 indicates that there is significant variation in the size of exposures.

There is significant persistence in exposures, as set out in Table 1.1, in which we show that the R^2 for a regression of C_{ijt} on pairwise ij fixed effects is 0.67. In other words, a large proportion of the variation in exposures is between pairs rather than across time.

There is significant variation in product characteristics across banks, in that the average

Table 1.1: Variation and persistence in network

	it	jt	ij
C_{ijt}	$R^2 = 0.43$	0.16	0.67
No. obs	6,426	6,426	6,426

Note: This table shows the R^2 obtained from regressing observed exposures C_{ijt} from bank i to bank j at time t on dummy variables. jt, for example, indicates that the regressors are dummy variables for each combination of j and t.

product supplied by each bank varies according to currency, maturity and type. For example, between 60% and 80% of the exposures held by most banks in our sample relate to derivatives. For one bank, however, this figure is 95%, and for another it is 15%.

Empirical fact 4: The network has increased in concentration over time

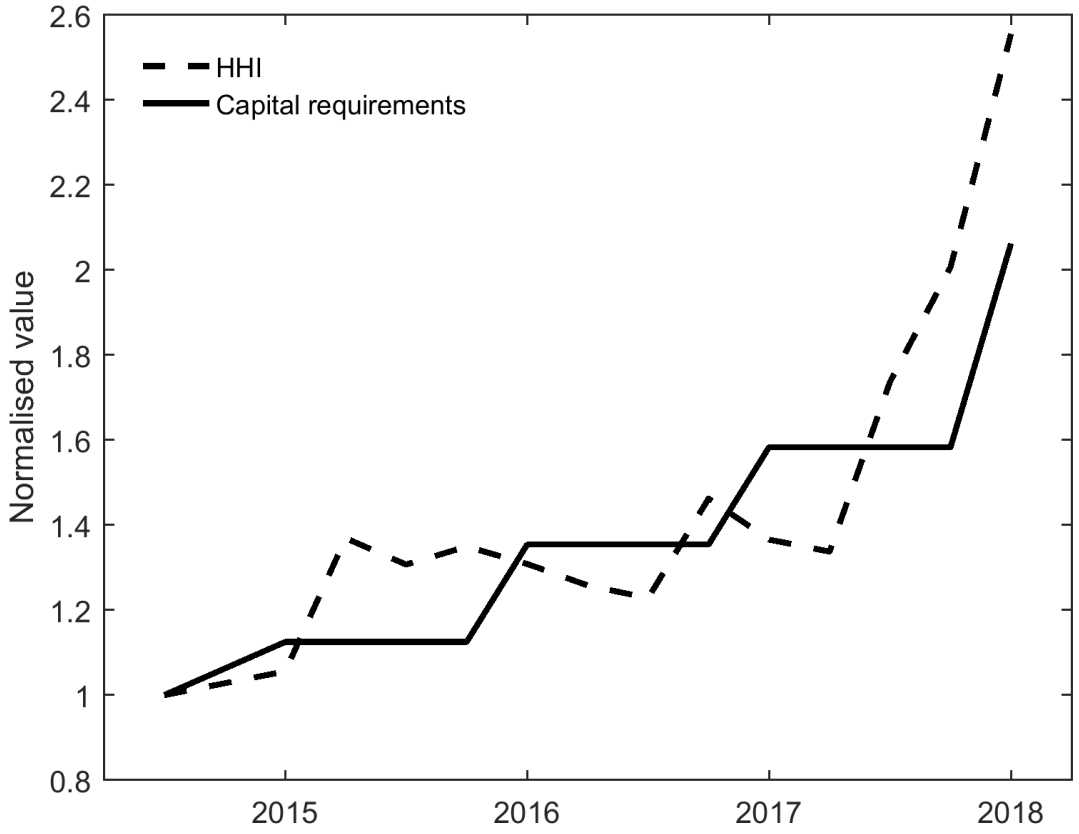
Even though the network is persistent, there is still inter-temporal variation. In particular, concentration in the interbank network has increased over time, in that the Herfindahl-Hirshmann index[10] over exposure supply has increased, as set out in Figure 1.2. In Figure 1.2, we show that the HHI index and regulatory capital requirements are closely correlated. It is obviously not possible to draw any causal conclusions from such a graph, but the relationship between concentration and capital requirements will be an important part of our model and identification.

Empirical fact 5: Bank default risk has decreased

Our sample runs from 2011 to 2018, and therefore earlier periods feature the end of the European debt crisis. Bank default risk has broadly reduced across all banks, as we set out in Figure 1.3. Importantly, though, there is cross-sectional variation across banks, and inter-temporal variation in that cross-sectional variation. We show this in Figure 1.3, in which we highlight the default risk of two specific banks. Bank 1 (Bank 2) was in the top (bottom) quartile by bank default risk in 2011, but the bottom (top) quartile by 2018.

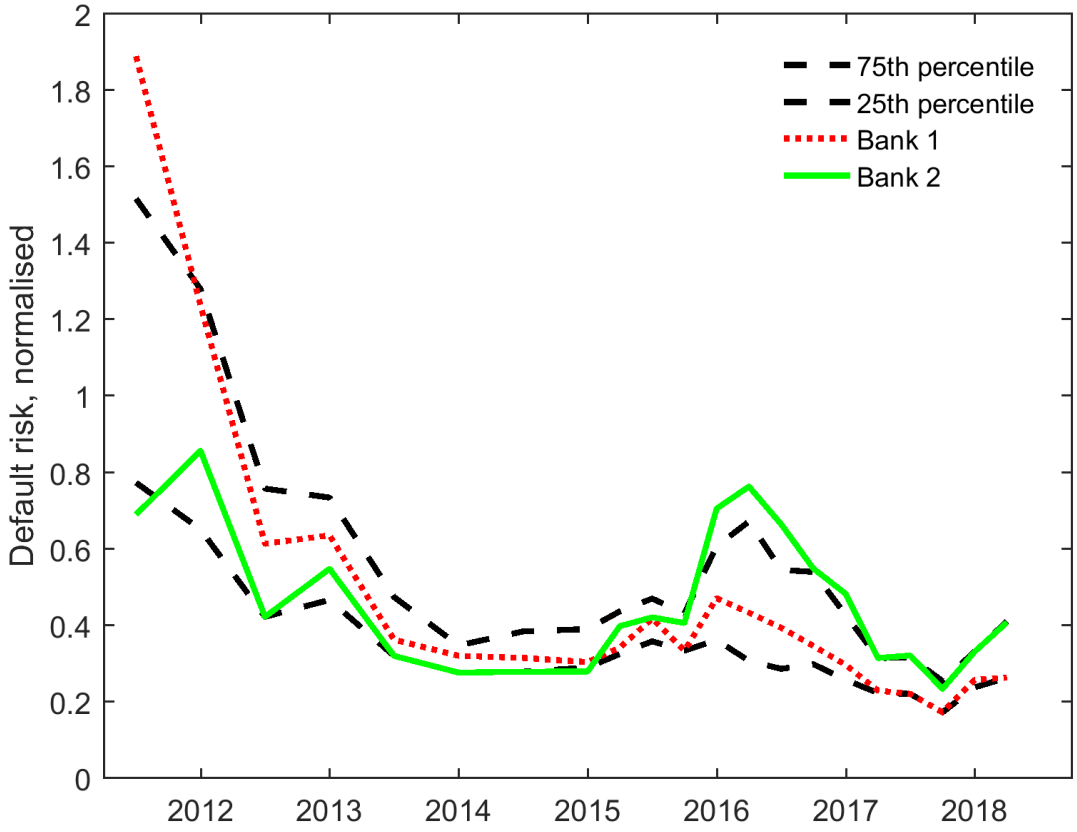
[10] $HHI_t = \frac{1}{N}\sum_j \sum_i s_{ij}^2$, where s_{ij} is the share of bank i in the total supply to bank j: $s_{ij} = \frac{C_{ij}}{\sum_i C_{ij}}$. Larger HHI indicates greater concentration. Because of the group-to-unit measurement issue we describe above, we weight exposures in our calculation of HHI by $(\frac{1}{NT}\sum_t \sum_j C_{ijt})^{-1}$. In this sense our measure of HHI is concentration within the i-bank.

Figure 1.2: Increased concentration



Note: The dashed line is a measure of concentration in exposures. The solid line is the mean capital requirement. There was a change in the way our data was collected that mean comparing concentration before and after 2014 is not meaningful, so we restrict our sample to 2014 onwards.

Figure 1.3: Inter-temporal and cross-sectional variation in default risk



Note: The black dashed lines show the 25th and 75th percentiles of bank default risk over time. The red dotted line and green solid line show how cross-sectional variation changed over time. The red dotted line is a bank that was initially high risk relative to the other banks in our sample, before becoming relatively low risk. The green solid line shows a bank that went from being relatively low risk to relatively high risk.

1.2.4 Stylised facts

Our sample starts in 2011, it does not feature the financial crisis that began in 2008. We note three features that were observed on the interbank network during the 2008 crisis, on the basis that a good model of interbank network formation should be able to replicate what happened during the crisis. First, risky banks were not supplied; in other words, they experienced *lockout* (Welfens, 2011). Second, risky banks did not supply, which we loosely term *liquidity hoarding* (Gale and Yorulmazer, 2013). Third, in the worst periods of the financial crisis there was effectively *market shutdown* in markets for certain instruments, in that very few banks were supplied anything on the interbank network (Allen et al., 2009; Afonso et al., 2011).

1.3 Model

We first introduce the setup of the model and notation. We then describe each of the three parts of the model in turn: the default risk process, demand for financial products and supply. We then set out the equilibrium of our model. Finally, we consider the implications of this model for optimal networks.

1.3.1 Setup and notation

There are N banks. At time t, the interbank network consists of an $N \times N$ directed adjacency matrix of total exposures, $\mathbf{C_t}$. C_{ijt} is the element in row i and column j of $\mathbf{C_t}$, and indicates the total exposure of bank i to bank j at time t. $\mathbf{C_t}$ is directed in that it is not symmetric: bank i can have an exposure to bank j, and bank j can have a (different) exposure to bank i. For each bank i, $\mathbf{d_i}$ is an $L \times 1$ vector of product characteristics for the exposures that it supplies.

$\mathbf{p_t}$ is an $N \times 1$ vector of bank default risks: the element in position i is the probability of default of bank i. $\mathbf{p_t}$ is a function of $\mathbf{C_t}$ and an $N \times K$ matrix of bank fundamentals, which we denote $\mathbf{X_t}$, and which update over time according to some exogenous process. This function is the default risk process, and the effect of $\mathbf{C_t}$ on $\mathbf{p_t}$ represents "contagion", as we will define more formally below.

C_{ijt} results in profits to bank i (we term this supply of exposures) and to bank j (demand

for exposures). These profits depend on bank default risk, in a way we will formalise below. The equilibrium interbank network $\mathbf{C_t}$ is formed endogenously based on the supply- and demand-sides, such that markets clear. Banks choose their supply and demand decisions simultaneously. For simplicity, there is no friction between changes in bank fundamentals and the formation of the network: once fundamentals change, the equilibrium network changes immediately.[11]

1.3.2 Default risk process

Understanding the effect of exposures on default risk is a key part of our research question. In our approach to modelling this default risk process, we are guided by the summary statistics we set out above in three important ways:

- First, in our dataset the *exposures are large and complete* (empirical fact 1), which means that the exposures could reasonably have an impact on the default risk of the banks that hold these exposures, in contrast to papers in the literature that observe exposures relating to a single instrument type (Denbee et al., 2017; Gofman, 2017). In other words, the size of our observed exposures leads us to consider financial contagion on default risk through these exposures.

- Second, there is *cross-sectional variation in exposure characteristics* (empirical fact 3): in other words, firms are trading different financial products. Some financial products may not impact default risk in the same way as others: as a trivial example, holding GBP 100m of senior debt of bank j may have a smaller effect on the default risk of bank i than holding GBP 100m of junior debt. This empirical fact means that we need to take a flexible approach to modelling contagion that accounts for this heterogeneity.

- Third, there is *cross-sectional variation in bank default risk* (empirical fact 5). There is a broad theoretical literature on the importance of such cross-sectional variation for financial contagion: the effect of an exposure to bank j on bank i's default risk is likely to depend on the extent to which their underlying fundamentals are correlated

[11] It is straightforward to introduce some friction in the timing, such that the network does not update immediately once fundamentals change. This would allow more detailed consideration of shock propagation in the *short-run*, which we define as the interval in which the network has not updated. We consider these short-run effects in further work, and consider in this paper only the *long-run* effects of changes in fundamentals.

(Glasserman and Young, 2015; Elliott et al., 2018). Our model of contagion, therefore, needs to be sufficiently flexible to account for this heterogeneity.

We model a bank's default risk process as the sum of two components: a set of fundamentals and a spatially autocorrelated component whereby bank i's default risk depends on its aggregate exposure to bank j, C_{ijt}, and bank j's default risk, p_{jt}. In matrix form:

$$\underbrace{\mathbf{p_t}}_{\text{Default risk}} = \underbrace{\boldsymbol{X_t\beta}}_{\text{Fundamentals}} \underbrace{-\omega\,\mathbf{C_t}\,\boldsymbol{\iota}}_{\text{Hedging}} \underbrace{+\tau_t(\boldsymbol{\Gamma}\circ\mathbf{C_t})\mathbf{p_t}}_{\text{Counterparty risk}} + \mathbf{e_t^p}$$

where $\mathbf{p_t}$ is a $N \times 1$ vector of bank default risks, $\mathbf{C_t}$ is a $N \times N$ directed adjacency matrix of aggregate pairwise exposures, $\boldsymbol{\beta}$ is a $K \times 1$ vector that represents each bank's loadings on a $N \times K$ matrix of fundamentals $\mathbf{X}$, $\boldsymbol{\iota}$ is a $N \times 1$ vector of ones, $\omega > 0$ is a scalar parameter that determines the effect of exposures on default risk through hedging, $\boldsymbol{\Gamma} > \mathbf{0}$ is a $N \times N$ matrix of parameters that determine the effect of exposures on default risk through counterparty risk, τ_t is a scalar that allows the effect of counterparty risk to vary across time and $\circ$ signifies the Hadamard product.

In broad terms, in this model a bank's default risk depends on its fundamentals and on its interbank network exposures. The interbank network can decrease bank default risk through *hedging*: many lending or derivatives transactions between banks are expressly intended to hedge risk. The interbank network can also increase bank default risk through *counterparty risk*: when a banks takes on an exposure to another bank it runs the risk that the other bank will default.

More specifically, this is a spatially autocorrelated regression, as is commonly used in network econometrics (De Paula, 2017), with a generalisation: the parameter governing the size of counterparty risk, Γ_{ij}, is allowed to be heterogeneous across bank pairs. Before we explain the effect of this generalisation, we first define *contagion* from bank j to bank i as the partial equilibrium effect that $\frac{\partial p_{it}}{\partial p_{jt}} > 0$: that is, the default risk of bank j has a causal impact on the default risk of bank i. In our model, $\frac{\partial p_{it}}{\partial p_{jt}} = \tau_t \Gamma_{ij} C_{ijt}$, such that the strength of contagion depends on the size of the exposure and this parameter Γ_{ij}.

Γ can be thought of as *contagion intensity* in that $\Gamma_{ik} > \Gamma_{im}$ implies that $\frac{\partial p_{it}}{\partial p_{kt}} > \frac{\partial p_{it}}{\partial p_{mt}}$ for any common $C_{ikt} = C_{imt}$. That is, bank i's default risk is more sensitive to exposures to bank k than to bank m, holding exposures and fundamentals constant. We refer to links with relatively low contagion intensity as "inherently safe" and links with relatively high

contagion intensity as "inherently risky".

This heterogeneity in contagion intensity could come from three sources. First, it could be a result of correlations in the underlying fundamentals, as described above, whereby if bank i and k (m) have fundamentals that are positively (negatively) correlated then exposure C_{ik} (C_{im}) is particularly harmful (benign). This implies a relationship between the fundamentals processes and Γ_{ij} which we leave open for now, but consider in our empirical analysis. Second, it could be a result of variations in product characteristics, as described above. This difference across products could be modelled using a richer default risk process that separately includes exposures matrices for each instrument type with differing contagion intensities, but this would introduce an infeasible number of parameters to take to data. Third, it could be a result of some other relevant pairwise variation that is unrelated to fundamentals or product, such as geographic location. It could be, for example, that recovery rates in the event of default are lower if bank i and bank j are headquartered in different jurisdictions, making cross-border exposures riskier than within-border exposures.

We allow for contagion intensity to vary across time via τ_t because there are, in principle, things that could affect contagion intensity. One of the purposes of the increase in capital requirement, for example, was to make holding a given exposure C_{ijt} safer, in the sense of Modigliani and Miller (1958) (because it means bank i has a greater equity buffer if bank j defaults). We do not make any assumptions about the relationship between τ_t and capital requirements λ at this stage, but consider it in estimation.

As well as resulting in contagion, the interbank network can reduce default risk by allowing banks to hedge. The partial equilibrium net effect of an exposure C_{ijt} is as follows:

$$\frac{\partial p_{it}}{\partial C_{ijt}} = -\omega + \tau_t \Gamma_{ij} p_{jt}$$

An exposure C_{ijt} is more likely to increase the default risk of bank i if hedging is less important (because ω is small), the counterparty is particularly risky (p_{jt} is large) or the link is particularly risky (Γ_{ij} is large).

To find equilibrium default risk we solve for a fixed point in $\mathbf{p_t}$. Subject to standard regularity conditions on $\mathbf{\Gamma}$ and $\mathbf{C}$ this spatially autocorrelated process can be inverted and expanded as a Neumann series as follows, which we term the Default Risk Process ("DRP"):

$$\mathbf{p_t} = (\mathbf{I} - \tau_t \mathbf{\Gamma} \circ \mathbf{C_t})^{-1}(\mathbf{X_t}\boldsymbol{\beta} - \omega \mathbf{C_t}\boldsymbol{\iota} + \mathbf{e_t^p}) = \sum_{s=0}^{\infty} (\tau_t \mathbf{\Gamma} \circ \mathbf{C_t})^s (\mathbf{X_t}\boldsymbol{\beta} - \omega \mathbf{C_t}\boldsymbol{\iota} + \mathbf{e_t^p})$$

We motivate our approach further in Appendix A, in which we set out a more primitive model of default risk, along the lines of Eisenberg and Noe (2001). In this model, a bank's value is the sum of its fundamentals and its interbank holdings, and it fails if this value falls below some critical value. Once a bank fails, it defaults on its interbank obligations and so reduces the values of its counterparties, resulting in a cascade of bank failures. By drawing repeatedly from the stochastic process that governs bank fundamentals, it is possible to calculate default risk as simply the proportion of draws in which bank i fails. In this model we allow for variations in the correlations of fundamentals across banks, variations in product riskiness and variations in recovery rates at the pairwise level, each of which we describe above.

This model is arguably a more natural model of a default risk process, but it is not suitable for our purposes in two ways: (i) it does not have an analytical solution, which makes it difficult to combine with a model of network formation in which firms take into account the effect of their network choices on their default risk and (ii) it is not easily taken to data, in that we could not separately identify each of these sources of heterogeneity or the critical values at which bank failures occur. Instead, what we do is simulate data from this model, and fit our proposed spatial autoregression. We use the results of this exercise to show that (i) a spatial autoregression fits relatively well and (ii) heterogeneity in contagion intensity Γ_{ij} is important. In this sense, our proposed approach can be thought of as a reduced form representation of this underlying more fundamental model, and Γ_{ij} can be thought of as a reduced form representation of these underlying sources of pairwise heterogeneity.

We run alternative specifications of the default risk process as robustness checks to our results. In particular, we consider an alternative default risk process in which common fundamentals (intended to represent the risk premium) do not propagate through the network.

1.3.3 Demand

In our approach to modelling demand we are guided by one important empirical fact: *product characteristics are heterogeneous* across banks (empirical fact 3). In other words, banks are supplying and demanding different financial products. This has two important implications:

- First, this heterogeneity has implications for the specificity with which we model the payoffs to demanding financial products. For example, if our empirical exposures were uniquely debt, then we would be able to include a standard model of liquidity

management on the demand-side (as in Denbee et al. (2017)). If instead our empirical exposures were uniquely CDS contracts, then we would be able to include a model of credit risk management (as in Eisfeldt et al. (2018)). Instead, we need to model the demand-side in a general way that is applicable across the range of financial products that feature in our data.

- Second, this heterogeneity has implications for how we model competition between banks. In particular, this heterogeneity means we need to consider the extent to which exposures supplied by one bank are substitutable for those supplied by another bank (product differentiation, in other words).

Each j-bank has a technology that maps inputs into gross profit, from which the cost of inputs is subtracted to get net profits. Inputs are funding received from other banks $C_{ij}, \forall i \neq j$ and an outside option C_{0j} designed to capture funding from banks outside our sample and non-bank sources. Net profits are given by:

$$\Pi^D_{jt} = (\zeta_{ij} + \delta_{jt} + e^D_{ijt}) \sum_{i=0}^{N} C_{ijt} - \frac{1}{2}\left(B \sum_{i=0}^{N} C^2_{ijt} + 2 \sum_{i=0}^{N} \sum_{k \neq i}^{N} \theta_{ik} C_{ijt} C_{kjt} \right) - \sum_{i=0}^{N} r_{ijt} C_{ijt}$$

where ζ_{ij} and δ_{jt} represent heterogeneity in the sensitivity of the j-bank's technology to product i, B governs diminishing returns to scale and θ_{ik} governs the substitutability of product i and k. Before we motivate our choices about functional form in more detail, it is helpful to set out what this implies for the j-bank's optimal actions. Bank j chooses C^D_{ijt} to maximise net profit taking interest rates as given, resulting in optimal C^D_{ijt} such that inverse demand is as follows:

$$r^D_{ijt} = \underbrace{\zeta_{ij} + \delta_{jt}}_{\text{Technology}} - \underbrace{BC_{ijt}}_{\text{Own-effect}} - \underbrace{\sum_{k \neq i} \theta_{ik} C_{kjt}}_{\text{Cross-effect}} + \underbrace{\theta_0 C_{0jt}}_{\text{Out.Op.}} + e^D_{ijt}$$

In other words, our functional form assumptions imply that the bank demanding exposures has linear inverse demand.

We assume that the j-bank has an increasing but concave objective function in the funding that it receives. We justify its concavity on the basis that the j-bank undertakes its most profitable projects first (or conversely, if its funding is restricted for whatever reason, it terminates its least profitable projects rather that its most profitable projects). Concavity also means that the returns to receiving funding decrease, in that the j-bank only has a limited number of opportunities for which it needs funding.

The intercept is comprised of three parts: δ_{jt}, ζ_{ij} and e^D_{ijt}. δ_{jt} ensures that the returns that the j-bank gets from funding are time-varying. This time variation is left general, although it could be related to the j-bank's fundamentals. It could be, for example, that when the j-bank's fundamentals are bad then the payoff to receiving funding is greater, in that the projects being funded are more important (if, for example, it needs this funding to undertake non-discretionary, essential projects or to meet margin calls on other funding). This is intended to allow for the importance of the interbank network in times of distress. The technologies possessed by each j-bank vary by ζ_{ij}, which governs the importance of the i-bank's product to the j-bank's technology. We allow this technology to be heterogeneous across pairs. e^D_{ijt} is an iid shock to the returns that bank j gets from receiving funding from bank i.

We also allow for product differentiation, in that the product supplied by bank i may not be a perfect substitute for the product supplied by bank k. We parameterise this product differentiation in parameters we denote θ_{ik}.

1.3.4 Supply

In our approach to modelling the supply side, we are guided by the following empirical observations: the network we are seeking to model is *dense with heterogeneous intensities* (empirical facts 2 and 3). Much of the literature focuses on explaining sparse core-periphery structures, which are often rationalised by *fixed* costs to link formation (Craig and Ma (2019), for example, have a fixed cost of link formation relating to monitoring costs). Variation in fixed cost cannot explain heterogeneity in link intensity, however, so this empirical observations leads us to focus on heterogeneity in *marginal* cost instead.

Bank i has an endowment E_{it} that it can either supply to another bank or to an outside option. When it supplies its product to bank j it receives return r_{ijt} and incurs a per-unit cost puc_{ijt}. We model this per-unit cost as the cost of the equity that the bank has to raise to satisfy its capital requirements; that is, when bank i supplies bank j it pays a certain

rate to raise the necessary equity. We parameterise the cost of equity as a linear function of the bank's default risk: $c^e_{it} = \phi p_{it} + e^c_{it}$, where e^c_{it} is the remaining part of the bank's cost of equity that is unrelated to its default risk. The riskier a bank is, the higher the cost of raising equity:

$$\underbrace{puc_{ijt}}_{\text{Per-unit cost}} = \underbrace{\lambda_{ijt}}_{\text{Reg'n}} \underbrace{c^e_{it}}_{\text{Cost of K}} = \lambda_{ijt}\left(\phi p_{it} + e^c_{it}\right)$$

where λ_{ijt} is the equity bank i needs to raise per-unit of exposure to bank j,[12] c^e_{it} is the cost of raising that equity, p_{it} is the default risk of bank i, e^c_{it} is an error term and ϕ is a parameter governing the relationship between default risk and cost of equity.

This simple parameterisation has three important implications. First, p_{it} is endogenously dependent on bank i's supply decisions, via the default risk process that we define above. In other words, when bank i supplies bank j, it takes into account the fact that doing so makes it riskier and so makes it costlier to raise capital. Second, p_{it} is endogenously dependent on the supply decisions of *other* banks, via the default risk process that we define above. In other words, there are network cost externalities. Third, p_{it} is endogenously dependent on regulation λ_{ijt} through the default risk process described above. In other words, in the spirit of Modigliani and Miller (1958), an increase in λ_{ijt} has two effects on the total cost of capital for firm i: it increases the amount of capital that the i bank needs to raise, but makes the bank safer and so makes the cost of a given unit of capital lower.

Bank i's problem in period t is to choose $\{C_{ijt}\}_j$ to maximise the following, taking $p_{k\neq i,t}$ as given:

$$\begin{aligned}\Pi_{it} &= \Pi^S_{it} + \Pi^D_{it} \\ &= \underbrace{\sum_j C_{ijt}[r_{ijt} - puc_{ijt} + e^S_{ijt}]}_{\text{Interbank supply}} + \underbrace{(E_{it} - \sum_j C_{ijt}) r_{i0t}}_{\text{Supply to Out.Op.}} + \Pi^D_{it}\end{aligned}$$

$$\text{such that } C_{ijt} \geq 0,\ E_{it} - \textstyle\sum_j C_{ijt} \geq 0 \text{ and } puc_{ijt} = \lambda_{ijt}\left(\phi p_{it} + e^c_{it}\right).$$

[12] For ease of exposition we have collapsed the risk-weighting (ρ, using the notation from Section 2) and the capital required per risk-weighted assets (λ) into a single parameter, λ.

For interior solutions the first order condition is as follows:

$$\underbrace{r^S_{ijt} + \frac{\partial r^S_{ijt}}{\partial C_{ijt}} C_{ijt}}_{\text{MB}} = puc_{ijt} + \underbrace{\sum_k \frac{\partial puc_{ikt}}{\partial C_{ijt}} C_{ikt}}_{\Delta \text{ Aggregate K cost}} - \underbrace{\frac{\partial \Pi^D_{it}}{\partial p_{it}} \frac{\partial p_{it}}{\partial C_{ijt}}}_{\Delta \text{ D-side cost}} - \underbrace{r_{i0t}}_{\text{Out.Op.}}$$

The left-hand side is the marginal benefit to i of supplying bank j. The right-hand side is the marginal cost, which consists of four parts (i) the per-unit cost it pays, (ii) the marginal change in the per-unit cost, (ii) the marginal change in i's payoff from demanding interbank products and (iv) the outside option.

Bank i, when choosing to supply C_{ijt}, therefore balances the return it gets from supplying against the effect of its supply on its default risk, via the default risk process described above. Being riskier harms bank i by increasing the price it pays to access capital in two ways. First, it increases the marginal cost bank i pays when supplying interbank exposures (the second term in the preceding equation, labelled 'Δ Aggregate K cost'). Second, being riskier means that bank i pays higher interest rates when demanding exposures (the third term in the preceding equation, labelled 'Δ D-side cost').

1.3.5 Equilibrium

Before considering equilibrium, we summarise what our model implies for the definition of a bank. In our model, bank i is the following tuple: $(E_{it}, d_{i,l}, \beta \mathbf{X_i}, \zeta_\mathbf{i}, \mathbf{\Gamma_i})$: respectively, an endowment, a set of product characteristics, a set of loadings on fundamentals, a technology and a set of contagion intensities. In other words, although the model is heavily parameterised, it allows for rich heterogeneity among banks.

Definition 1 *In this context we define a Nash equilibrium in each period t as: an $N \times N$ matrix of exposures $\mathbf{C^*_t}$ and $N \times 1$ vector of default risks $\mathbf{p^*_t}$ such that markets clear and every bank chooses its links optimally given the equilibrium actions of other banks.*

For interior solutions where $C_{ijt} > 0$, market clearing requires that supply and demand are equal, such that the following equilibrium condition holds, which we term the Equilibrium

Condition ("EQC"):

$$\begin{aligned} 0 \;=\; & \delta_{jt} + \zeta_{ij} + e^D_{ijt} - 2BC_{ijt} - \sum_{k\neq i}^{N} \theta_{ik} C_{kjt} + e^S_{ijt} \\ & -\lambda_{ijt}\phi_1 p_{it}(\mathbf{C_t}) - \phi_1\big[-\omega + \tau_t \Gamma_{ij} p_{jt}(\mathbf{C_t})\big] \sum_{k\neq i}^{N} C_{ikt}\lambda_{ikt} - r_{i0t} \\ & -\phi_1\tau_t\big[-\omega + \tau_t \Gamma_{ij} p_{jt}(\mathbf{C_t})\big] \sum_{k} C_{kit}\Gamma_{ki} \sum_{m} C_{kmt}\lambda_{kmt} \end{aligned}$$

We show our calculations in Appendix C. Note that a bank's default risk is a function of $\mathbf{C_t}$, as we set out in the default risk process, which we repeat here for convenience:

$$\mathbf{p_t} = (\mathbf{I} - \tau_t \mathbf{\Gamma} \circ \mathbf{C_t})^{-1}(\mathbf{X_t}\boldsymbol{\beta} - \omega \mathbf{C_t}\boldsymbol{\iota} + \mathbf{e^p_t}) = \sum_{s=0}^{\infty} (\tau_t \mathbf{\Gamma} \circ \mathbf{C_t})^s (\mathbf{X_t}\boldsymbol{\beta} - \omega \mathbf{C_t}\boldsymbol{\iota} + \mathbf{e^p_t})$$

Substituting $\mathbf{p}$ out of EQC using DRP gives a system of equations in $\mathbf{C}^*$. The form of DRP is such that the EQC become a system of infinite-length series of polynomials, such that in general no analytical solution exists. Instead, we solve these equilibrium conditions numerically. We make no general claims about uniqueness or existence at this stage, but confirm numerically that our estimated results are an equilibrium that is, based on numerical simulations, unique.

We demonstrate how the model works by arguing that our model is consistent with: (1) with the empirical facts we set out above and (2) the stylised facts we set out above regarding how direct interbank connections behaved during the financial crisis.

1.3.5.1 The model is consistent with our empirical facts

We set out certain empirical facts above that we used to guide our modelling. In this subsection, we explain in more detail how exactly the model is consistent with these empirical facts.

First, our empirical network is *heterogeneous* in the intensity of links. There are three main sources of such heterogeneity in our model: (i) firms have heterogeneous technologies ζ_{ij} that require differing inputs from other firms, (ii) contagion intensity Γ_{ij} is heterogeneous, such that some links are intense because they are less risky and (iii) firms have heterogeneous

fundamentals X_{it}, such that some links are intense because the banks involved have good fundamentals.

Second, our empirical network is *persistent* over time. Each of the sources of heterogeneity discussed above is also a source of persistence: ζ_{ij} and Γ_{ij} are by assumption fixed over time, and X_{it} vary over time but may be persistent.

Third, we observe *increased concentration* in our data. In our model this results from the increase in capital requirements across our sample. Consider bank i's decision to supply bank j and/or bank k, where bank k's fundamentals are worse than bank j. For a given level of capital requirement λ, the fact that bank k is riskier means that ceteris paribus bank i supplies more to bank j than bank k. An increase in λ then makes supplying bank k relatively more costly compared to supplying bank j. In other words, an increase in capital requirements penalises risky links that are already likely to be small, resulting in an increase in concentration.

1.3.5.2 The model is consistent with our stylised facts

We also set out above three stylised facts from the crisis. Our model can match each of these stylised facts.

First, risky banks may choose to supply less total exposures, which we loosely term *liquidity hoarding*. All other things being equal, if a bank experiences a negative shock to its fundamentals it supplies less, as it is riskier and so its cost of capital is higher. This is not strictly liquidity hoarding in a structural sense, in that the bank is not lending less because it needs to preserve liquidity for the future, but the effect is the same. In that sense, this mechanism can be thought of as a reduced form for liquidity hoarding.

Second, risky banks may be supplied less, which we term *market lockout*. A shock to the fundamentals of bank j makes supplying it more risky and therefore more costly. This is true holding fixed δ_{jt}, which are fixed effects governing inter-temporal variation in demand. If this is related to X_{jt}, then the effect of variations in fundamentals is more complicated.

Third, when all banks are risky, liquidity hoarding and market lockout combine to result in *market shutdown*, where no bank is supplied anything at all. This follows in our model as the combination of the two previous effects.

1.3.6 Optimal networks

There are three immediate potential sources of inefficiency in our model (plus a fourth one we will define later):

1. Network externalities
2. Market power
3. Inefficient cost allocations

First, there are externalities within the interbank network, as bank k's default risk p_{kt} is affected by C_{ijt} provided that bank k has a chain of strictly positive exposures to i. If $C_{kit} > 0$ then this is trivially true, but it is also true if bank k has a strictly positive exposure to another bank that has a strictly positive exposure to i, and so on. Banks i and j do not fully account for the effect on p_{kt} when they transact bilaterally, such that this negative externality implies that exposures are too large relative to the social optimum. Second, the banks supplying financial products may have market power, such that exposures are too small relative to the social optimum. Third, equilibrium allocations among suppliers may not be efficient, given differing marginal costs. In equilibrium high cost suppliers might supply positive quantities when it would be more efficient for low cost suppliers to increase their supply instead.

These inefficiencies mean that aggregate interbank surplus may not be maximised in equilibrium, where we define aggregate interbank surplus as the sum of aggregate surplus on the demand-side and aggregate surplus on the supply-side across all N banks. In other words, a social planner could specify an exposure network that increased aggregate interbank surplus.

In this context, however, it is insufficient to consider aggregate surplus within the interbank network. A bank's default risk can impact agents outside of the interbank network, such as its depositors, creditors, debtors and various other forms of counterparty. A crisis in the interbank network could, in principle, lead to a wider crisis with implications for the "real" economy. In other words, a social planner would not set exposures and default risk solely to maximise surplus in the interbank network, but instead to maximise total surplus in the economy, including aggregate interbank surplus and real surplus, which we define as follows.

Definition 2 : Real surplus : *We define "real surplus" as surplus outside of the interbank network, and denote it by R_t.*

The relationship between bank default risk and real surplus is important, as if there is such a relationship then it reveals a fourth possible inefficiency:

4. Real externalities: Banks do not take this into account the effect of their network formation decisions on real surplus.

Characterising the relationship between real surplus and default risk, or estimating it empirically, is not straightforward. We do not model or estimate this relationship, but only make the following directional assumption:

Assumption 1 *Suppose real surplus R_t is a function of the mean default risk of banks $\bar{p}_t$: $R_t = r(\bar{p}_t)$. We assume that R_t is strictly decreasing in $\bar{p}_t$.*

This assumption is clearly an approximation of what is likely to be a complex relationship between real surplus and bank default risk. It may not always hold; it may be, for example, that when bank default risk is very low, some additional bank default risk increases real surplus. It could also be that *mean* bank default risk is not the only thing that is important, but also some measure of dispersion or the minimum or maximum. Nevertheless, we think that this assumption reasonably represents the fundamental, local trade-off that regulators face when intervening in these markets: the trade-off between default risk and surplus in the market.

In particular, this assumption allows us to think about optimal default risk and interbank surplus in the sense of Pareto-optimality. That is, denote total surplus in the interbank network by TS_I (where the I subscript emphasises that this is total surplus in the interbank network only) and mean default probability by $\bar{p}$, and suppose $TS_I^H > TS_I^L$ and $\bar{p}^H > \bar{p}^L$. Assumption 1 implies that $(TS_I^H, \bar{p}^L) \succ^{SP} (TS_I^L, \bar{p}^H)$, where $\succ^{SP}$ denotes the social planner's preferences, but it does not allow us to rank $(TS_I^H, \bar{p}^H)$ and $(TS_I^L, \bar{p}^L)$, as we illustrate in Figure 1.4.

It is helpful to think about the trade-off between TS_I and $\bar{p}$ in terms of constrained maximisation of interbank surplus subject to a default risk constraint.

Definition 3 : Efficient frontier : *For an arbitrary, exogenous value of mean default risk, $\bar{p}^F$, define $TS_I^F = max_{\mathbf{C}}\, TS_I(\mathbf{C})$ st $\bar{p}(\mathbf{C}) = \bar{p}^F$. We define the efficient frontier as the locus traced out in $(\bar{p}^F, TS_I^F)$ space as $\bar{p}^F$ is varied.*

In other words, the efficient frontier is agnostic about the scale of externalities outside of the interbank network. It requires only that there is no feasible alternative $(TS_I^A, \bar{p}^A)$ that is a Pareto-improvement in the sense that (i) $TS_I^A > TS_I^F$ and $\bar{p}^F \leq \bar{p}^A$ or (ii) $TS_I^A \geq TS_I^F$ and $\bar{p}^F < \bar{p}^A$. If such a Pareto-improvement existed, we can conclude from Assumption 1 that $(TS_I^A, \bar{p}^A) \succ^{SP} (TS_I^F, \bar{p}^F)$. The extent to which a given point is inefficient can then be loosely characterised by its vertical or horizontal distance from the frontier, as we set out in the definitions below. Figure 1.4 shows the frontier and illustrates what conclusions we can draw using this model about different outcomes.

Definition 4 : p inefficiency : *The default risk inefficiency of some allocation* $(TS_I, \bar{p})$ *is the percentage decrease in* $\bar{p}$ *that could be obtained without decreasing* TS_I*. In other words, it is the vertical distance in percentage terms from the frontier.*

Definition 5 : TS inefficiency : *The total surplus inefficiency of some allocation* $(TS_I, \bar{p})$ *is the percentage increase in* TS_I *that could be obtained without increasing* $\bar{p}$*. In other words, it is the horizontal distance in percentage terms from the frontier.*

Finally, we note that although it is straightforward to consider efficient allocations, it is much more difficult to calculate optimal regulation (in our model, the capital regulations λ_{ijt}^{SP} that a social planner would choose) that fully implements efficient allocations. We consider feasible regulations that are efficiency improvements over the perfectly decentralised market in the section below on counterfactual analysis.

1.4 Estimation

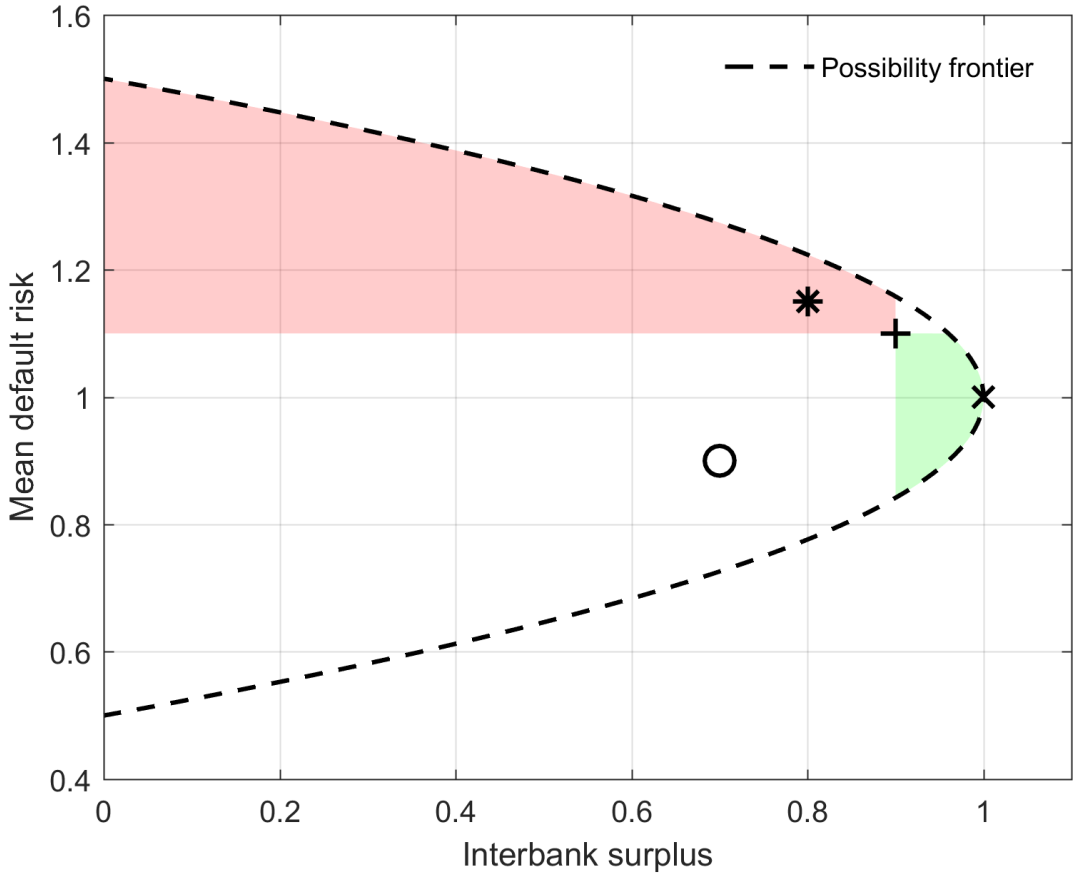
We first describe the data we use to model bank fundamentals and the structure of our estimation approach. We then describe the parameterisations that we make when we take this model to data.

1.4.1 Modelling fundamentals

To represent bank fundamentals $\mathbf{X}$ we use bank-specific and common data.

For bank-specific variation, we take the relevant equity index to be a bank-specific weighted average of global equity indices from S&P, where the weightings are the proportion of the bank's revenues that come from that geography (data provided annually by

Figure 1.4: Stylised example: Interbank surplus and default risk



Note: Point + dominates any point in the red area but is dominated by any point in the green area. For example, $\times \succ^{SP} + \succ^{SP} *$, but we cannot rank $\circ$ relative to the other points. We cannot even rank $\circ$ relative to $\times$ despite $\times$ being on the efficient frontier: the social planner's preferences over $\times$ and $\circ$ depend on the scale of externalities outside of the interbank network, which we leave open. The extent of inefficiency of point $\circ$ can be expressed as the vertical distance south to the efficient frontier and the horizontal distance east to the frontier.

Bloomberg, based on corporate accounts). For example, suppose that at time t bank k obtained 70% of its revenues from the US and the remaining 30% from Japan. In this case, $Z^p_{kt} = 0.7 \times S\&P500_t + 0.3 \times S\&PJapan_t$. Absolute index values are not meaningful, so we normalise each S&P index by its value on 1 June 2019. Although this is clearly an imperfect measure of the bank's fundamentals, we argue it has informative value: this bank k would plausibly be more affected by a slowdown in Japan than some other bank with no Japanese revenues. The S&P indices we use are for the US, Canada, the UK, Europe, Japan, Asia

and Latin America.

In our robustness tests, we test whether our results are sensitive to an alternative measure of bank fundamentals: weighted average consumption growth (where the weighting is bank revenues by jurisdiction, as above).

To capture common variation in bank fundamentals, we use a broad panel of macroeconomic and commodity data from the World Bank. We calculate the first three principal components of this panel, which collectively account for more than 99% of total variation, and include these three variables in $\mathbf{X}$. We also include the Chicago Board Options Exchange Volatility Index, more commonly known as "VIX", which represents expected variation in option prices, and the Morgan Stanley World Index.

1.4.2 Estimation structure

The parameters we seek to estimate are $\boldsymbol{\Theta} = (\tilde{\boldsymbol{\Gamma}}, \tau, \omega\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\zeta}, \tilde{\boldsymbol{\theta}}, \phi)$; respectively, contagion intensities, time-variation in contagion intensities, hedging effect, fundamentals, demand intercept variation, pairwise technology importance, characteristic-based product differentiation, and the cost multiplier. Our estimation process involves two loops. In the inner loop, we solve our model numerically to calculate the network links and default risks implied by a given parameter vector; respectively, $\hat{C}(\Theta)$ and $\hat{p}(\Theta)$. In the outer loop, we search over parameter vectors Θ to minimise two sets of moments, where the relevant instruments are set out in the following section: (1) network formation: $\mathbb{E}[\mathbf{Z}'(\hat{\mathbf{C}}(\boldsymbol{\Theta}) - \mathbf{C})] = 0$ and (2) contagion: $\mathbb{E}[\mathbf{Z}'(\hat{\mathbf{p}}(\boldsymbol{\Theta}) - \mathbf{p})] = 0$. We express $\mathbf{p}$ in logs.

1.4.3 Parameterisations

We impose four parameterisations to feasibly take this model to our data. The first parameterisation we make is with respect to Γ_{ij}. General symmetric Γ_{ij} consists of $N(N-1)/2 = 153$ elements. These are individually identifiable, as we will show below, but because the length of our panel is limited we cannot estimate them with reasonable power. For this reason, our baseline estimation approach imposes the following structure on Γ_{ij}:

$$\Gamma_{ij} = \tilde{\Gamma}_i \tilde{\Gamma}_j$$

where $\tilde{\mathbf{\Gamma}}$ is an $N \times 1$ vector of parameters. This parameterisation is significantly more parsimonious but retains variation at the ij level. It does result in some loss of generality, in that loosely speaking it implies that if Γ_{12} and Γ_{23} are high, then Γ_{13} must also be high. This kind of structure is broadly consistent with each of the three motivations for heterogeneous Γ_{ij} that we introduce above.

The second parameterisation we make relates to τ_t. We include τ_t to allow for time-variation in contagion intensity because higher capital requirements are intended to make a given exposure safer. General τ_t, with a different multiplicative parameter for each time period, is in principle identifiable. In practice, we parameterise τ_t based on capital requirements:

$$\tau_t = e^{-\tau(\lambda_t - \lambda_1)}$$

where λ_t is the mean capital requirement at time t, λ_1 is the mean capital requirement in the first period of our sample, 2011, and τ is a scalar parameter. Thus $\tau_1 = 1$, but $\tau_{t>1}$ can be lower depending on the size of τ. If $\tau = 0$ then $\tau_t = 1$ for $\forall t$ and there is no time-variation in contagion intensity, if τ is large then there is significant time-variation. This is a more parsimonious approach that directly addresses the underlying reason why allowing for time-variation in contagion is important.

The third parameterisation we make relates to θ_{ik}, which governs the extent to which the products supplied by bank i are substitutes for those supplied by bank k. General θ_{ik} cannot be reasonably estimated from our dataset; instead we parameterise it as being a logistic function of certain product characteristics, including maturity, currency and instrument-type.

$$\theta_{ik} = \frac{exp\left(\tilde{\theta} - \sum_l^L \tilde{\theta}_l (d_{i,l} - d_{k,l})^2\right)}{1 + exp\left(\tilde{\theta} - \sum_l^L \tilde{\theta}_l (d_{i,l} - d_{k,l})^2\right)}$$

where $d_{i,l}$ denotes the value for characteristic l of bank i and $\tilde{\theta}_l > 0$ is a parameter that determines the importance of characteristic l to the substitutability of different products. For instrument type, for example, $d_{i,l=type}$ is the proportion of i's product that is derivatives. If banks i and k have very different product characteristics, then θ_{ik} is small and the two are not close substitutes. If, on the other hand, banks i and k have very similar product characteristics then θ_{ik} is large and the two are close substitutes. This parameterisation replaces θ_{ik} (which across all pairs has dimension N^2) with $\tilde{\theta}_l$ (which has dimension $L + 1$).

The fourth parameterisation we make relates to the structure of our data, and in par-

ticular the fact that, as described in Section 1.2, for non-British banks we only observe local-unit-to-group exposures, under-estimating their total exposure. We assume that:

$$C_{ijt} = (1 + a_i)\, \tilde{C}_{ij}$$

where we denote local-unit-to-group exposures by $\tilde{C}_{ijt}$ and group-to-group (that is, total) exposures by C_{ijt}, and a_i are bank-specific parameters that we estimate. These parameters a_i are identified given that (i) some variables, such as X_{jt} and p_{it}, enter the EQC with non-bank-specific coefficients and (ii) for the British banks we know $a = 0$. In principle, a finer disaggregation is identifiable in this way, but we restrict variation to a_i to preserve degrees of freedom.

1.5 Identification

We consider identification of the network formation game and of the default risk process. We then return to our research question, and discuss in intuitive terms the empirical variation that we use to identify each of the key parameters that determine our answer to this research question.

1.5.1 Network formation

The EQC and DRP allow us to solve for equilibrium $\mathbf{C}$ and $\mathbf{p}$ as a function of λ, $\mathbf{X}$ and the jt and it fixed effects described above. In other words, identification is significantly easier when we solve for equilibrium exposures, because the endogenous exposures of other banks and endogenous default risks are substituted out of our empirical specification.

We assume bank fundamentals, as defined above, are exogenous. Treating this as exogenous assumes that a bank's revenue distribution and the equity indices themselves are independent of *pairwise* structural errors in the interbank network. We emphasise that the fact that we are able to include it and jt fixed effects means that the only remaining unobservable variation is pair-specific. We think it is a reasonable assumption that, for example, HSBC, which has deep roots in Asia, would not shift its geographic revenue base in response to pair-specific shocks in the interbank network. Similarly, we think it is a reasonable assumption that the equity indices that form the basis of our bank-specific fundamentals are independent of pair-specific shocks in the interbank network.

We treat product characteristics as exogenous, in keeping with the literature on demand estimation in characteristic space. We treat λ, regulatory capital requirements, as exogenous, in keeping with the literature on the empirical analysis of bank capital requirements (Robles-Garcia, 2018; Benetton, 2018). It is informative to consider how we are able to separately identify the effect of common time variation in capital requirements from the it and jt fixed effects. This relates to Figure 1.2, in which we show the correlation between concentration in the interbank network over time and changes in capital requirements. In our model the effect of the common increases in capital requirements on equilibrium exposures depends on the fundamentals of the banks supplying and demanding the exposures: in other words, although the changes in capital requirements are common across all banks, their effect on exposures is pair-specific.

B is not separately identifiable from the other parameters. We normalise $B = 1$ on the basis that in models of quantity competition what matters for market power is θ/B, not the absolute value of B.

1.5.2 Default risk process

We repeat DRP for convenience:

$$\mathbf{p_t} = (\mathbf{I} - \tau_t \mathbf{\Gamma} \circ \mathbf{C_t})^{-1}(\mathbf{X_t}\boldsymbol{\beta} - \omega \mathbf{C_t}\boldsymbol{\iota} + \mathbf{e_t^p}) = \sum_{s=0}^{\infty}(\tau_t \mathbf{\Gamma} \circ \mathbf{C_t})^s(\mathbf{X_t}\boldsymbol{\beta} - \omega \mathbf{C_t}\boldsymbol{\iota} + \mathbf{e_t^p})$$

The advantage of explicitly considering network formation is that we can account for the endogeneity of the network in our spatial DRP model. The key insight to our identification strategy is that DRP is a *linear* function of bank fundamentals $\mathbf{X_t}$, but equilibrium exposures $\mathbf{C_t}$ are a *non-linear* function of $\mathbf{X_t}$. We therefore use non-linear variation in $\mathbf{X_t}$ as pair-specific, time-varying instruments for the network. We motivate this more clearly in three steps. First, we show that equilibrium exposures are indeed non-linear in bank fundamentals. Second, we show that this gives us the pair-specific variation that we need. Third, we set out exactly which variables we use as instruments.

The fact that equilibrium exposures are non-linear in bank fundamentals comes from the non-linearity of the cost function. The key intuition for this is that the cost function is convex in C_{ijt}, provided that ω is small, such that in equilibrium C_{ijt} would never grow linearly with fundamentals as that would lead to marginal cost becoming very large. Consider

a simple example with three banks, 1, 2 and 3, and suppose, for the sake of simplicity, that in equilibrium every network link between those banks is strictly positive. In equilibrium C_{12}^* is such that the marginal cost of supplying exposures is equal to the marginal benefit. The marginal benefit is linear in C_{12}, whereas the marginal cost is convex in C_{12}, as set out in Figure 1.5. Suppose the fundamentals of banks 1, 2 and 3 improve, worsen and remain unchanged, respectively. In these circumstances, we show in Figure 1.5 that C_{12} changes non-linearly relative to the size of these. In Appendix C we show, for a simplified version of our model for which an analytical solution exists, that equilibrium C are a non-linear function of $\mathbf{X}$.

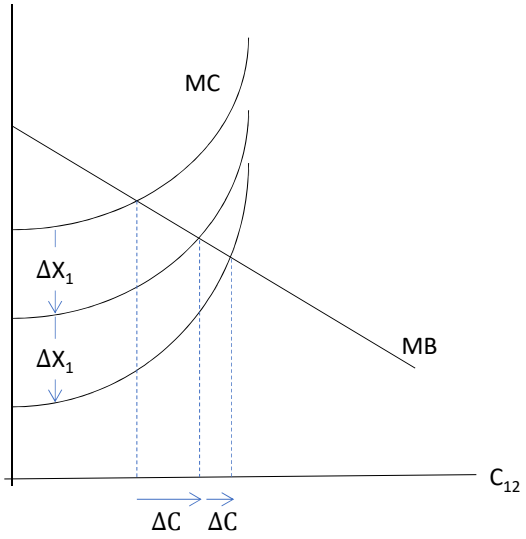
Having shown that exposures are non-linear in fundamentals, it is straightforward, using the same simple example, to show that changes in fundamentals then give us the pair-specific variation that we need for them to be instruments for C_{ijt}. Assume again that the fundamentals of banks 1, 2 and 3 improve, worsen and remain unchanged, respectively. This causes links between banks 1 and 3 to increase (because the improvement in bank 1's fundamentals mean that the marginal cost to bank 1 of supplying bank 3 has gone down, and the marginal cost to bank 3 supplying bank 1 has gone down). For analogous, but opposite, reasons, links between bank 2 and bank 3 decrease. For links between banks 1 and 2 it is not possible to sign the effect, as some elements of marginal cost have gone up and some have gone down. In summary, provided there is reasonable cross-sectional variation in bank fundamentals (which we show in Figure 1.3), then that variation has differing exogenous implications for each of the pairs.

We define $\tilde{X}_{ijt} = \frac{1}{N-2}\sum_{k\neq i,j} X_{kt}$ (that is, average fundamentals of other banks). As instruments for C_{ijt} we use $[X_{it}^2,\ X_{jt}^2,\ \tilde{X}_{ijt}^2,\ X_{it}/X_{jt},\ X_{it}/\tilde{X}_{ijt}...]$, as well as these terms interacted with λ_{ijt} to leverage its time variation. We show the results of first stage regressions in the appendix. Assuming these bank fundamentals are orthogonal to unobserved shocks to bank default risk is more restrictive than in the case of the network formation data, as we have fewer fixed effects available to use. We assume that the equity indices on which we rely are independent of unobserved bank default risk. We justify this on the basis that, although the banks in our sample are large, none are a material proportion of these equity indices.

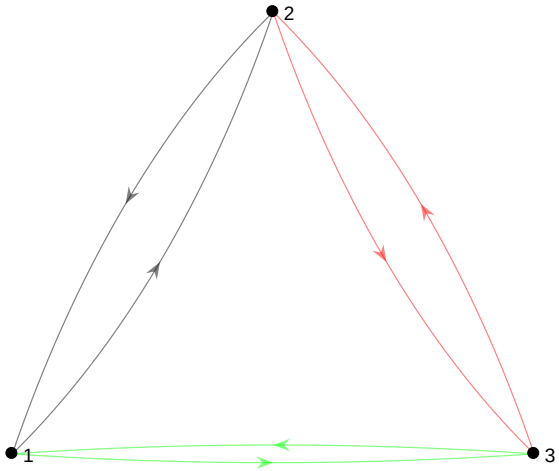
We then use the GMM moments suggested in a spatial context by Kelejian and Prucha (1998) and Kelejian and Prucha (1999).

Figure 1.5: Non-linear bank fundamentals as instruments for C

(a) Non-linear effect of X on C



(b) Pairwise variation



Note to Figure 1.5: Suppose the fundamentals of bank 1, 2 and 3 improve, worse and do not change, respectively. In part (a) we show that equilibrium exposures are non-linear with respect to this variation in fundamentals. In part (b) we show that this this has differing pairwise effects on equilibrium link intensity, where link intensity between 1 and 2 increases, link intensity between 2 and 3 decreases and link intensity between 1 and 2 does not change.

1.5.3 Identification: Back to the research question

Having described our approach to identification, we summarise by considering how identification relates to our core research question regarding the inefficiency of the interbank network. There are three sources of inefficiency in our model, and each is determined by certain parameters in the model:

- Network externalities: The extent of network externalities depends on the size of Γ_{ij}. If these parameters are large, then network effects are large, and so network externalities are large.
- Market power: The extent of market power depends on the size of θ_{ij}. If these are large, then small differences in product characteristics lead to large differences in substitutability, and market power is large.
- Inefficient cost allocations: The extent to which high cost links inefficiently receive equilibrium allocations depends on the dispersion in Γ_{ij}. If these parameters are very dispersed, then cost variations are greater and the resulting inefficiency is greater.

Having argued that these parameters are the key parameters in our model, we summarise the key variation that identifies each of these parameters in Table 1.2. This is important for the robustness with which we answer our research question, as it shows that our answers to these questions are guided by the data rather than by our modelling assumptions.

Table 1.2: Key variation

	Key parameter	Key variation
[1]	$Size\ of\ \theta_{ik}$	$Cov(C_{ijt}, X_{kt} \mid d_i - d_k)$
[2]	$Size\ of\ \Gamma_{ij}$	$Cov(C_{ijt}, X_{jt})$, $Cov(p_{it}, X_{jt} \mid Z^C_{ijt})$
[3]	$Dispersion\ in\ \Gamma_{ij}$	$Cov(s_{ijt}, \lambda_t)$

Note: s_{ijt} denotes proportion of bank i's total supply that is to bank j. All other notation as previously defined.

θ_{ik} determines how closely banks i and banks k compete. We identify the size of θ_{ik} by the covariance between C_{ijt} and X_{kt}, which is an exogenous measure of bank k's cost,

conditional on the extent to which the two banks have similar product characteristics. If this covariance is high, then θ_{ik} is high.

Γ_{ij} determines the contagion intensity from j to i. There are two sources of empirical variation for this: from the network formation data and from the default risk data. On the network formation side, Γ_{ij} is identified by the covariance between C_{ijt} and X_{jt}. If C_{ijt} is sensitive to the fundamentals of bank j, then in the context of our model this means that Γ_{ij} is large. On the default risk side, Γ_{ij} is identified by the covariance between bank i's default risk and the fundamentals of bank j, conditional on the instruments we describe above for the size of C_{ijt}. If this conditional covariance is large, then this means that bank i's default risk is particularly sensitive to bank j's default risk, which in the context of our model means that Γ_{ij} is large.

Finally, we describe a further source of variation that helps identify the dispersion in Γ_{ij}. We set out above how a general increase in capital requirements leads to concentration, as it affects high and low marginal cost links differentially. Γ_{ij} is a key determinant of which links are high and low marginal cost. If, following an increase in capital requirements, bank i supplies relatively less to bank j, then this concentration indicates that Γ_{ij} is high.

1.6 Results

We set out our results in Table 1.3. We find that the model fits the data well, with R^2 of 0.85 and 0.83 for network data and default risk data, respectively. Parameter estimates are of the expected sign and mostly significantly different from zero.

We draw the following immediate implications for contagion intensity from our results:

- **Contagion is material**: on average 9.8% of mean bank default risk is due to interbank contagion, with the remainder due to bank fundamentals.[13] This can be thought of as an aggregate representation of the network effect. We also re-run our estimation taking the network as exogenous in our estimation of the default risk process (that is, without using the instruments for the endogenous network that are implied by our network formation game). This results in parameter estimates that imply 8.0% of mean bank default risk is due to interbank contagion. In other words, incorrectly assuming that the network is exogenous biases our estimation of the network effect downwards.

[13] We calculate this by calculating mean bank default based solely on fundamentals, $p_{it} = X_{it}\beta$, and comparing it to actual bank default risk.

Table 1.3: Results

		[1]	
ϕ		1.84***	
		(2.39)	
τ		9.26***	
		(6.03)	
β_1		-0.02**	
		(1.70)	
ω		0.04***	
		(8.80)	
	Min	*Median*	*Max*
$\tilde{\Gamma}_i$	0.15***	0.24***	0.51***
	(5.59)	(3.43)	(5.07)
$\tilde{\theta}_k$	4.71**	5.21*	27.69***
	(1.97)	(1.70)	(8.46)
a_i	0.01	0.69	5.53**
	(0.06)	(1.01)	(2.03)
Network			
FE		ij, it, jt	
R^2		0.85	
No. obs		6,426	
Default risk			
FE		i	
Controls		Y	
R^2		0.83	
No. obs		378	

Notes: SEs clustered at bank level. Figures in parentheses are t-stats. ***, **, * indicate different from 0 at 1%, 5% and 10% significance, respectively. For the heterogeneous parameters we report estimates and t-stats for the minimum, median and maximum, and plot the full distribution below. **Notation**: ϕ is the sensitivity of cost of equity to default risk, τ is the extent to which contagion intensity varies over time, β_1 is the effect of bank-specific fundamentals, ω is the effect of hedging, $\tilde{\Gamma}_i$ is contagion intensity, $\tilde{\theta}_k$ governs product differentiation based on characteristics and a_i scales exposures for non-UK banks. Controls in the default risk process are VIX, MSWI and macro data.

- **Contagion is heterogeneous**: there is substantial pairwise variation in contagion intensity Γ_{ij}: some links are nearly twice as costly as others, in terms of their effect on default risk. We plot the estimated distribution of Γ_{ij} in Figure 1.6.

- **Contagion is time-varying**: there is evidence that contagion intensity has decreased across our sample, in line with increasing capital requirements. Estimated τ implies that mean contagion intensity decreased by 36% between 2011 and 2018, as we plot in Figure 1.6. This is consistent with a significant improvement in bank default risk in response to the banks becoming better capitalised.

- **The effect of the network on default risk is time-varying**: in our model interbank exposures can decrease default risk through hedging or increase it through counterparty risk. When bank fundamentals are bad earlier in our sample, then the effect of counterparty risk dominates the effect of hedging, as set out in Figure 1.7. When bank fundamentals are good later in our sample, then the reverse is true.

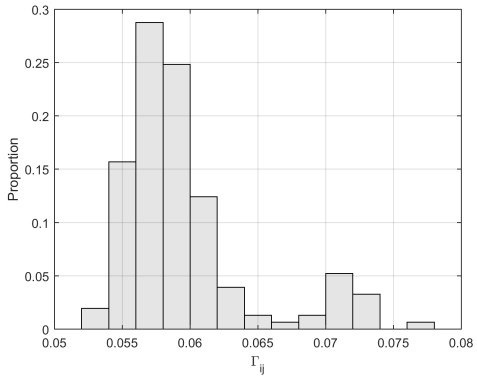
Our results also have implications for the form of competition between banks. We plot our estimated $\widehat{\theta}_{ij}$ in Figure 1.6, and show that there is significant product differentiation based on product characteristics. Generally, most $\widehat{\theta}_{ij}$ are small, indicating that only pairs producing very similar products are substitutes. The most important product characteristics in determining substitutability are (i) the proportion of total exposures that is denominated in EUR and (ii) the proportion of exposures with maturity greater than 1 year.

1.6.1 Robustness

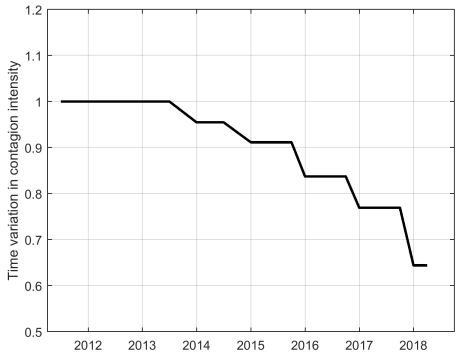
We run two alternative specifications as robustness tests, both of which test how sensitive our results are to how we treat time-variation in risk premium. In the first robustness test, we use alternative measures of bank default risk and bank-specific fundamentals that exclude the risk premium, but otherwise estimate our baseline specification. In the second robustness test, we use the same data as in our baseline results but amend the default risk process so that common time-variation in the risk premium does not propagate through the interbank network. We describe these tests in more detail and set out the results in Appendix D. In both cases, the results are quantitatively and qualitatively similar to our baseline results.

Figure 1.6: Distributions of parameter estimates

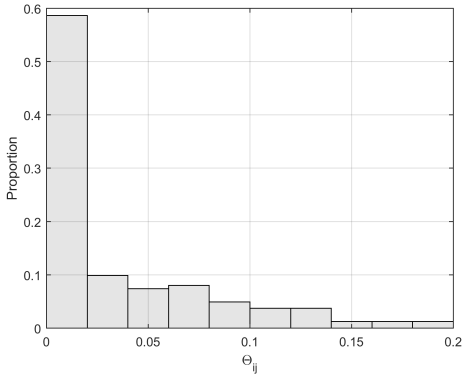
(a) Heterogeneous contagion intensity



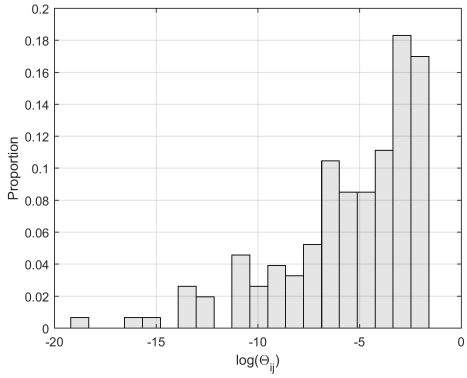
(b) Time variation in contagion intensity



(c) Product differentiation

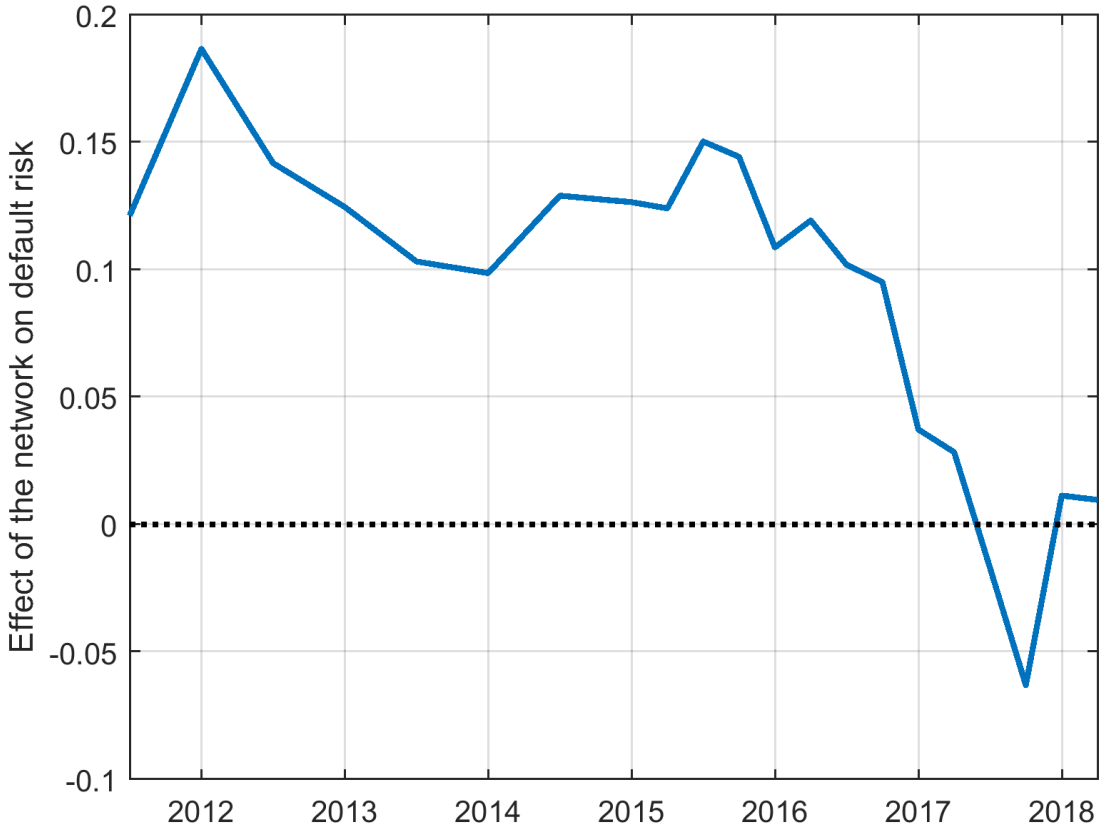


(d) Product differentiation (in logs)



Note: These figures show the distribution of our estimated parameters. Panel (a) shows that there is material variation in the intensity of contagion. Panel (b) shows that contagion intensity has decreased over time. Panels (c) and (d) show that there is variation in product differentiation, based on exposure characteristics.

Figure 1.7: Time-varying effect of the network



Note: We define the effect of the network as the difference between actual mean bank default risk and simulated mean bank default risk in which every interbank exposure is set to 0. A value of 0.1 means that actual mean bank default risk is 10% higher than if there were no interbank exposures. In our model interbank exposures can decrease default risk through hedging or increase it through counterparty risk. When bank fundamentals are bad earlier in our sample, then the effect of counterparty risk dominates the effect of hedging. When bank fundamentals are good later in our sample, then the reverse is true.

1.6.2 Cross-checks of our results

We run two cross-checks of our results, to test the extent to which they are reasonable. First, we show that the heterogeneity in contagion intensity that we estimate is consistent with risk-sharing. Second, we show that the model fits well out of sample.

1.6.2.1 Cross-check 1: Contagion is related to risk sharing

The first cross-check is a test of internal consistency: we set out above various motivations for why contagion intensity Γ_{ij} could be heterogeneous. One of these motivations is heterogeneity in the extent to which bank fundamentals are correlated; risk sharing, in other words. This implies a relationship between fundamentals, which we estimate as $X\beta$, and contagion intensity Γ_{ij}. We do not impose this relationship in estimation, but estimate general Γ_{ij} and test the existence of such a relationship post-estimation. These post-estimation tests, which we describe in Appendix E, support risk-sharing: where banks i and j are in the same jurisdiction, Γ_{ij} is higher when the fundamentals of banks i and j are more closely positive correlated. We view this as an important test of the consistency of our model and empirical approach.

1.6.2.2 Cross-check 2: The model fits well out of sample

The second cross-check we run relates to external consistency, in that we test the fit of our model out of sample. We do this in two ways: (1) using publicly available historical data on default risk data and (2) using stylised facts about what happens to interbank exposures in times of financial stress.

We do not have access to historical data on interbank exposures. We do, however, have access to historical CDS premia (bank default risk $\mathbf{p}$) and macro-economic variables (bank fundamentals $\mathbf{X}$), meaning that we can simulate interbank exposures and model-implied default risk backwards. We do this for 2009 to 2011, and compare the predicted default risk values with actual observed default risk. As set out in Figure 1.8, we find that the model fits out of sample variation in the mean and dispersion in bank default risk reasonably well. Some of this fit is driven by our choice of fundamentals, rather than our network formation model per se. We test the extent of this by also showing the out of sample fit of a linear model solely on bank fundamentals (that is, $\mathbf{p_t} = \mathbf{X_t B}$). We find that (1) the out of sample fit of the linear model is materially worse than the full model (the mean square error out of sample of the linear model is 18% greater than that of the full model and (2) the linear model is biased upwards relative to the full model, particularly when bank fundamentals are relatively good (as in 2010 in Figure 1.8).

We cannot compare simulated interbank exposures to actual historical interbank exposures, because we do not have the data. We do, however, have certain stylised facts about

how interbank exposures behaved during the financial crisis, as we describe in Section 1.2 above: we know that during the financial crisis some parts of the interbank network froze, in that no transactions occured. We forward simulate a generic recession by arbitrarily varying bank fundamentals and show the implications for bank default risk and network exposures in Figure 1.9 below. We find that the simulated interbank network dries up at a level of bank fundamentals that is broadly consistent with what we know about what happened during the financial crisis. This is in this sense a pseudo out of sample test in which we match a stylised fact rather than data.

1.6.3 Implications of our results

Having described our results and the cross-checks we run, we now discuss two important implications of our results regarding (1) forward simulation of recessions using our model and (2) the identification of systemically important banks.

1.6.3.1 Forward simulation

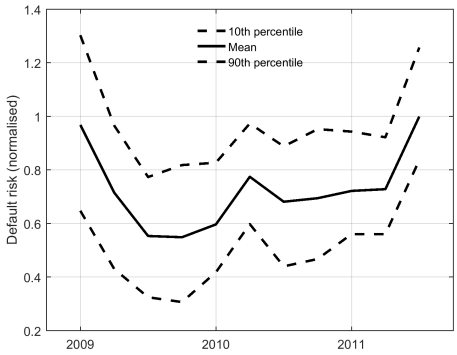
In Figure 1.9 below we simulate the effect of a recession on the interbank network and default risk. We do this by simulating an arbitrary increase (deterioration) in bank fundamentals. As the shock increases in severity the network shrinks and, when the recession is sufficiently severe, dries up. This is an important cross-check of our work, as we describe above. One implication of this is that bank default risk is convex with respect to bank fundamentals: as fundamentals deteriorate, the endogenously declining network dampens the effect of the change on fundamentals on default risk. There is, however, a zero lower bound, such that once the network has dried up then it cannot dampen the response to fundamentals. In other words, bank default risk is more sensitive to fundamentals in severe recessions.

This fact also has implications for predicting the impact of recessions. Suppose, for example, that when modelling the response of default risk p to fundamentals X the endogenous network was ignored, and instead p was simply regressed on X. Because severe recessions are very infrequently observed, a regression of p on X in normal times would *understate* the extent to which p would respond to X in a severe recession. We show true simulated default risk (the black solid line) and such a naively estimated default risk (the red dashed line) in Figure 1.9, and show that this bias can be material.

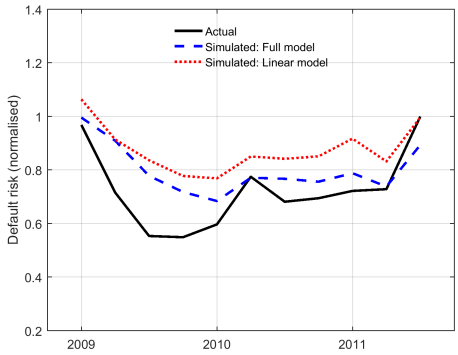
In Figure 1.8 above we show out of sample fit, and show that during periods in which

Figure 1.8: Out of sample fit: Bank default risk

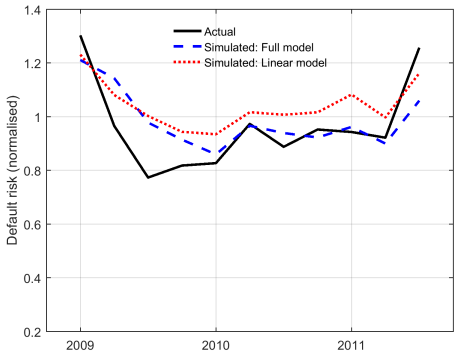
(a) Actual variation in default risk



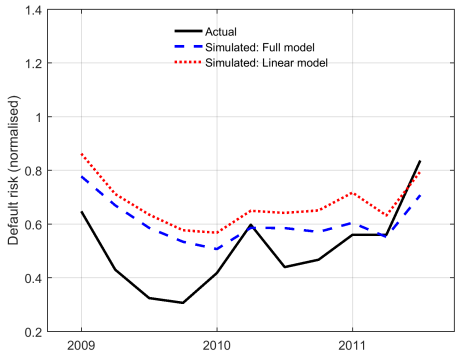
(b) Default risk (mean)



(c) Default risk (90th percentile)



(d) Default risk (10th percentile)



Note: These figures show the out of sample fit of our model. The black lines show the 10th percentile, mean and 90th percentile of actual historical default risk. The dashed blue line shows the out of sample fit of our estimated network formation and contagion model. The blue dotted line shows the out of sample fit of a linear model that ignores the interbank network and simply regresses default risk on fundamentals. This test shows that our model is robust in three ways: (1) our model fits well out of sample, (2) our model outperforms the simple linear model and (3) the performance of the simple linear model (and notably the fact that the linear model performs badly in the middle of the out-of-sample period when fundamentals were relatively good) is consistent with the predictions of our model, as we show below in Figure 1.9.

bank fundamentals were moderate (as in 2010), the bias goes the other way: estimated bank default risk using this linear model *overstates* true bank default risk. We explain this feature using the simulated recession set out in Figure 1.9, in which estimated default risk also overstates simulated true default risk in moderate fundamentals (such as in period 5): the bias arises from the difficulties a linear model has fitting an inherently non-linear process. In other words, our model predicts the shape of how a linear model should perform out of sample, and this indeed the shape we observe in the data. This is, therefore, an additional element to our robustness test.

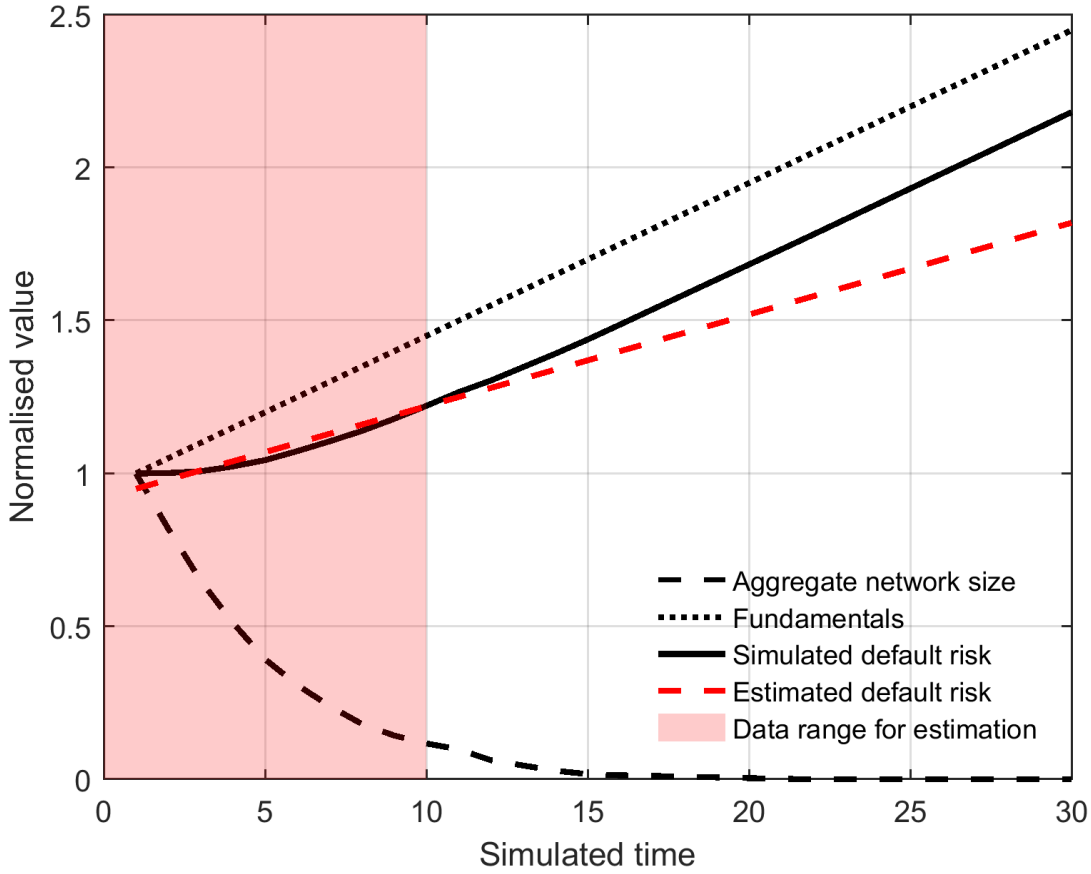
1.6.3.2 Systemic importance

The second implication of our results relates to systemic importance. A recurring issue in the network literature is the identification of "important" nodes. We have an equilibrium process that relates an outcome (bank default risk, in our case) to a network, and it is reasonable to ask which node in the network contributes most to the outcome in which we are interested. Understanding this communicates important information about this equilibrium process, but may also have implications for regulation (as we describe above, large parts of the banking regulatory framework are stricter for banks that are judged to be "systemically important" (Basel Committee, 2014b)). Various measures of systemic importance, or centrality, exist, where the most appropriate measure depends on the context and on the way in which nodes interact with each other (Bloch et al., 2017). Our contribution to this literature is not about the most appropriate measure, but instead about how any such measure should be calculated: it must account for the heterogeneity in contagion intensity Γ_{ij}.

We illustrate this by reference to one of the simplest measures of centrality: Eigenvector Centrality. Broadly speaking, node n's centrality score is the n'th entry in the eigenvector associated with the maximal eigenvalue of the adjacency matrix $\boldsymbol{C_t}$. A central node using this measure is close to other nodes that are central: this measure of centrality is in this sense self-referential. Nodes that have many large links to other nodes that have many large links are more central.

Applying this centrality measure to the network $\boldsymbol{C_t}$ therefore gives a ranking of which banks are most systemically important in driving bank default risk. If contagion intensity is homogenous, $\Gamma_{ij} = \Gamma$, then the level of Γ has no impact on this relative ranking. If, however, contagion intensity is heterogeneous, then accounting for this heterogeneity is important when assessing centrality: a more reasonable measure of centrality would be based on the

Figure 1.9: Simulated recession



Note: We simulate a recession by arbitrarily inflating (where an increase is a deterioration) bank fundamentals by an increasing factor (the dotted black line). As fundamentals deteriorate, the interbank network (the black dashed line) contracts and eventually dries up. Mean bank default risk (the solid black line) increases, but is convex because the network contraction dampens the effect of fundamentals. The red dashed line shows the results of observing a limited set of data (the red shared area) and fitting a linear regression of default risk on bank fundamentals: ignoring endogenous network formation understates how bank default risk changes with fundamentals in (infrequently observed) recessions. This is consistent with the findings of our out of sample test, as set out in Figure 1.8.

weighted adjacency matrix $\mathbf{\Gamma} \circ \boldsymbol{C_t}$. Importantly, the effect of this weighting on the ranking of systemic importance is not random noise, because the equilibrium network depends on this weighting. More specifically, links C_{ij} where Γ_{ij} is low (high) are inherently safe (uinherently risky) and so are more likely to be large (small), all other things being equal. In other words,

assessing centrality based on the raw, unweighted exposures matrix is biased and likely to overstate the centrality of more central nodes and understate the centrality of less central nodes. This holds only when holding other things equal: in our model of network formation, links can be large even if they are not safe (if they are technologically important through ζ_{ij}, for example).

In Figure 1.10, we show that calculating Eigenvector Centrality based on unweighted $\boldsymbol{C_t}$ and weighted $\boldsymbol{\Gamma} \circ \boldsymbol{C_t}$ lead to quite different rankings of systemic importance. Bank 18, for example, would be identified as the most systemically important node based on the unweighted network. Based on the weighted network, however, 4 other banks are most systemically important than Bank 18: in other words, Bank 18's links are large because its links are relatively safe. Bank 5's centrality, on the other hand, is significantly understated when looking solely at the unweighted network: in other words, Bank 5's links are small because its links are relatively unsafe. We do this for Eigenvector Centrality, but the same point applies to other measures (including, for example, Katz-Bonacich centrality).

1.7 Counterfactual Analysis

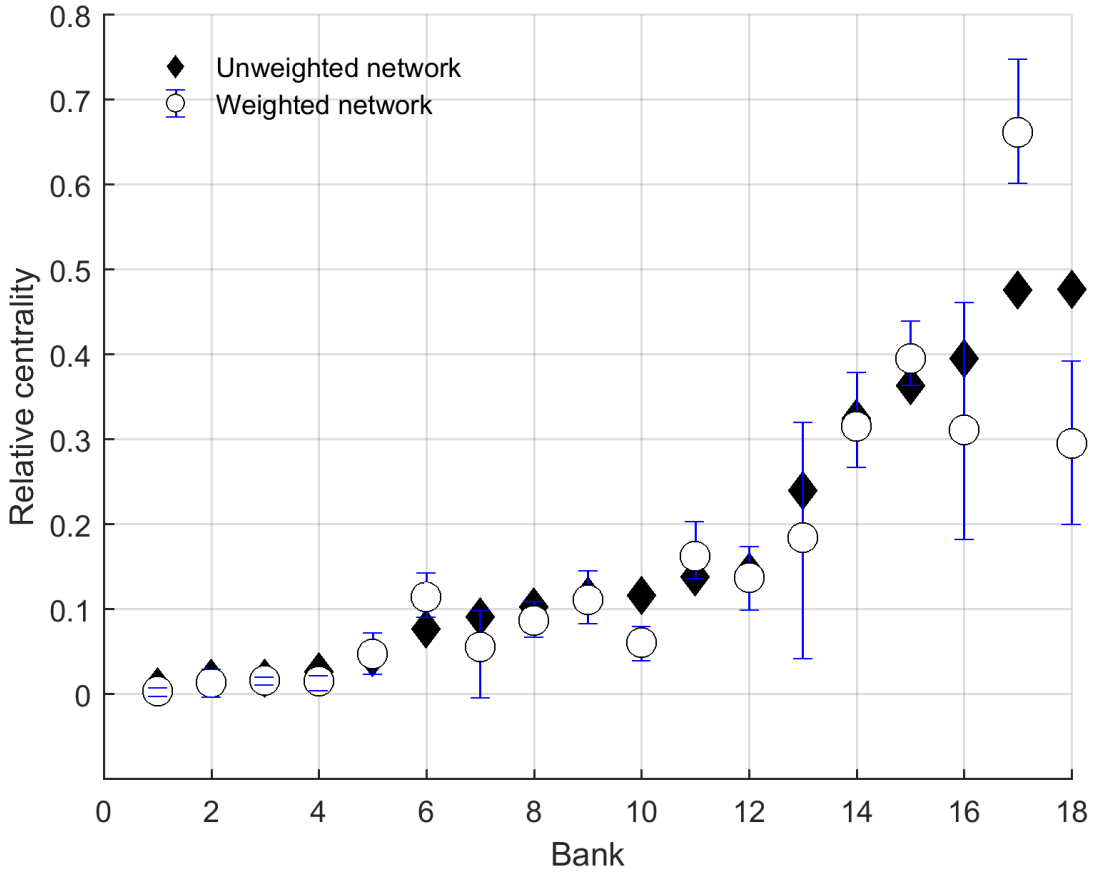
In our counterfactual analyses, we first consider the social planner's solution, and show what that implies for efficiency. We then consider two broad forms of regulation: caps on exposures and capital ratios.

Before we describe the counterfactual analyses in detail, we describe two uses of our model that play an important role in each of these counterfactual analyses. Our model, together with the parameters we have estimated, allow us to do two things. First, the estimated model provides a mapping from any arbitrary network of exposures $\mathbf{C_t}$ to (i) bank default risk and (ii) interbank surplus. Second, the estimated model provides a mapping from the exogenous parts of the model (fundamentals, regulation, etc) to decentralised equilibrium exposures $\mathbf{C_t}$. Together, these two uses of our model and results allow us to quantify surplus and default risk in counterfactual equilibria.

1.7.1 Efficiency

We describe above how our model implies a trade-off between mean bank default risk and interbank surplus, and how there is an efficient frontier on which this trade-off is optimised.

Figure 1.10: Identifying systemic nodes



Note: This figure plots the relative centrality of each of the 18 banks in our sample using Eigenvector Centrality. The black diamonds show relative centrality based on the unweighted network of observed exposures: banks with large exposures are more central. The white circles show relative centrality based on observed exposures weighted by their relative contagion intensities: relatively risky links are given a higher weighting. The blue lines show a 95% confidence interval around this weighted measure. Taking into account heterogeneous contagion intensity materially changes the relative systemic importance of banks: bank 18 is the most central bank based on the unweighted network, but only the 5th most central bank based on the weighted network. This is because in our network formation model banks endogenously choose large (small) exposures where those links inherently safe (inherently risky).

We use our estimated model to derive this frontier, by choosing $\mathbf{C_t}$ to maximise interbank surplus, subject to mean bank default risk being less than some critical value. We then vary

this critical value to trace out the efficient frontier. As described above, we do not know what allocations a social planner that was maximising aggregate surplus would choose, as we do not directly model the relationship between bank default risk and real surplus. We do know that this optimal allocation would be somewhere along the efficient frontier. The distance to the frontier in either direction is in this sense an estimate of inefficiency, as we describe above when we define p inefficiency and TS inefficiency.

We find that the decentralised interbank network is not on the efficient frontier: a social planner would be able to increase interbank surplus by 13.2% without increasing mean bank default risk or decrease mean bank default risk by 4.3% without decreasing interbank surplus, as set out in Figure 1.11. This result comes primarily from the fact that contagion (and thus network externalities) is significant. Exposure allocations on the frontier are more concentrated in favour of inherently safe links than actual observed exposures.

1.7.1.1 Comparative statics for efficiency

We emphasise that our conclusions on efficiency are driven by the data, rather than our modelling choices. We demonstrate this by undertaking comparative statics and showing how the extent of inefficiency varies according to the parameters chosen. We set out the results of these simulations in Table 1.4.

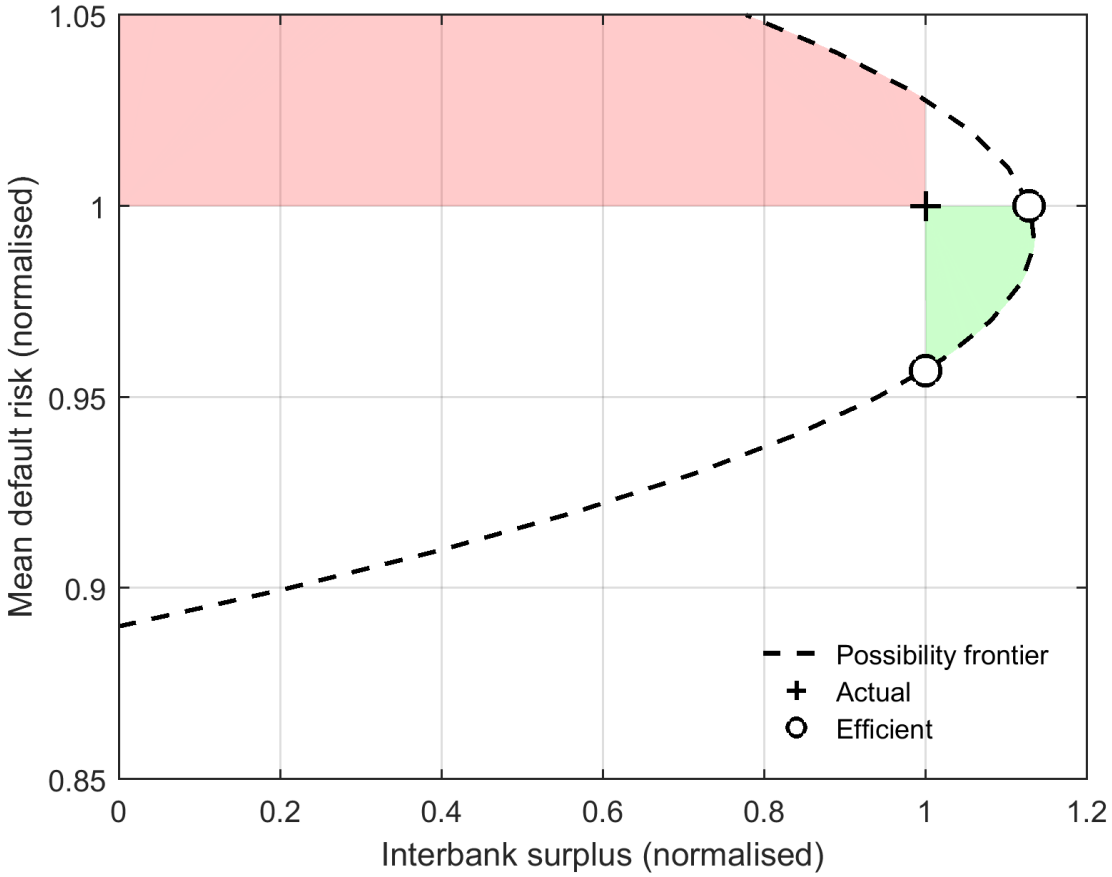
Table 1.4: Comparative statics

	[A]	[B]	[C]	[D]
	Baseline	$\downarrow mean(\theta_{ij})$	$\uparrow var(\Gamma_{ij})$	$\uparrow mean(\Gamma_{ij})$
p inefficiency	4.3%	5.4%	6.0%	8.7%
TS inefficiency	13.2%	15.6%	14.6%	14.2%

Note: [A] is our baseline results set out above; [B] is the baseline, with every θ_{ij} multiplied by a factor of 0.8; [C] is the baseline, with a mean-preserving spread of Γ_{ij} such that its variance increases by a factor of 1.5; [D] is the baseline, with every Γ_{ij} multiplied by a factor of 1.5.

First, market power is determined by θ_{ij}, which governs the extent of product differentiation. If θ_{ij} is large (small), then products i and j are close substitutes and market power is low (high). We illustrate the impact of increased market power by multiplying every θ_{ij} by a factor of 0.5 (Column B in Table 1.4). As set out in Table 1.4, this increases the distance

Figure 1.11: Decentralised inefficiency



Note: This figure shows that the decentralised outcome in the interbank network is inefficient. The + sign shows the mean bank default risk and interbank surplus that our model implies for actual exposures, both normalised to 1. The white circles show what a social planner who chooses the entire interbank network could achieve by (1) minimising mean default risk without decreasing interbank surplus and (2) maximising interbank surplus without increasing mean default risk. The dotted line shows the efficient possibility frontier of combinations of surplus and risk.

between the decentralised outcome and the efficient frontier.

Second, the efficiency of decentralised cost allocations is driven by the extent of variation in marginal cost across banks. If marginal cost is the same for all banks, then decentralised cost allocations are not inefficient. If marginal cost is highly variable, then the decentralised equilibrium will inefficiently involve some high cost links being positive. The extent of

variation in marginal cost across banks is driven primarily by the extent of variation in contagion intensity Γ_{ij}. We illustrate this by applying a mean-preserving spread to Γ_{ij} such that its variance increases by a factor of 2 (Column C in Table 1.4). This increases the distance between the decentralised outcome and the efficient frontier.

Third, the extent of externalities depends on the scale of network effects, which in our model is the size of Γ_{ij}. If these are large, then there are significant externalities and the decentralised equilibrium is more likely to be inefficient. We illustrate this by increasing every Γ_{ij} by a factor of 2. This also increases the distance between the decentralised outcome and the efficient frontier.

1.7.2 Caps on exposures

As discussed in Section 1.2, in 2019 a cap on individual exposures came into force: a bank can have no single bilateral exposure greater than 25% of its capital.[14] For exposures held between two "globally systemic institutions"[15] this cap is 15%.

We evaluate the effects of a cap on individual exposures by simulating new equilibrium exposures C_{ij}^C under a generic cap, using our estimated parameters and assuming that fundamentals are unchanged. We consider a generic, binding cap at the i-bank level:

$$C_{ij}^C \leq 0.9 \cdot max_j\{C_{ij}\}$$

In other words, we assume that any exposure held by bank i has to be less than or equal to 90% of its largest exposure. This cap is stylised, in that it is defined relative to observed exposures, rather than relative to its capital. This avoids issues about measuring capital appropriately and measuring total exposures (our exposures do not include every possible financial instrument), while still showing the economic effect of a cap in general. We simulate the effect of this cap in Figure 1.12 below, and find that such a cap has a very small impact on default risk, for two reasons. First, a cap on individual exposures binds on the bank's largest exposures, which are more likely to be relatively safe (that is, they have low Γ_{ij}). Second, a cap on individual links creates excess supply and unmet demand that causes other uncapped links in the network to increase. That is, the network topology changes endogenously.

We propose an alternative form of regulation in which total exposures held by bank i are

[14] Where the precise definition of capital, "Tier 1 capital", is set out in the regulation.
[15] As defined in the regulation.

capped, rather than individual exposures:

$$\sum_j C^C_{ijt} \leq 0.9 \sum_j C_{ijt}$$

A cap on total exposures held by bank i prevents other parts of the network from increasing in response to a capped link. A cap on total exposures also causes bank i to inherently risky (high Γ_{ij}) exposures by relatively more than inherently safe (low Γ_{ij}) exposures. In other words, a cap on individual exposures targets inherently safe exposures, whereas a cap on total exposures targets inherently risky exposures. We simulate the effect of this cap in Figure 1.12, and find that it reduces mean default risk by significantly more than an individual cap and actually *increases* interbank surplus. Our results suggest a social planner therefore would strictly prefer a cap on total exposures to a cap on individual exposures.

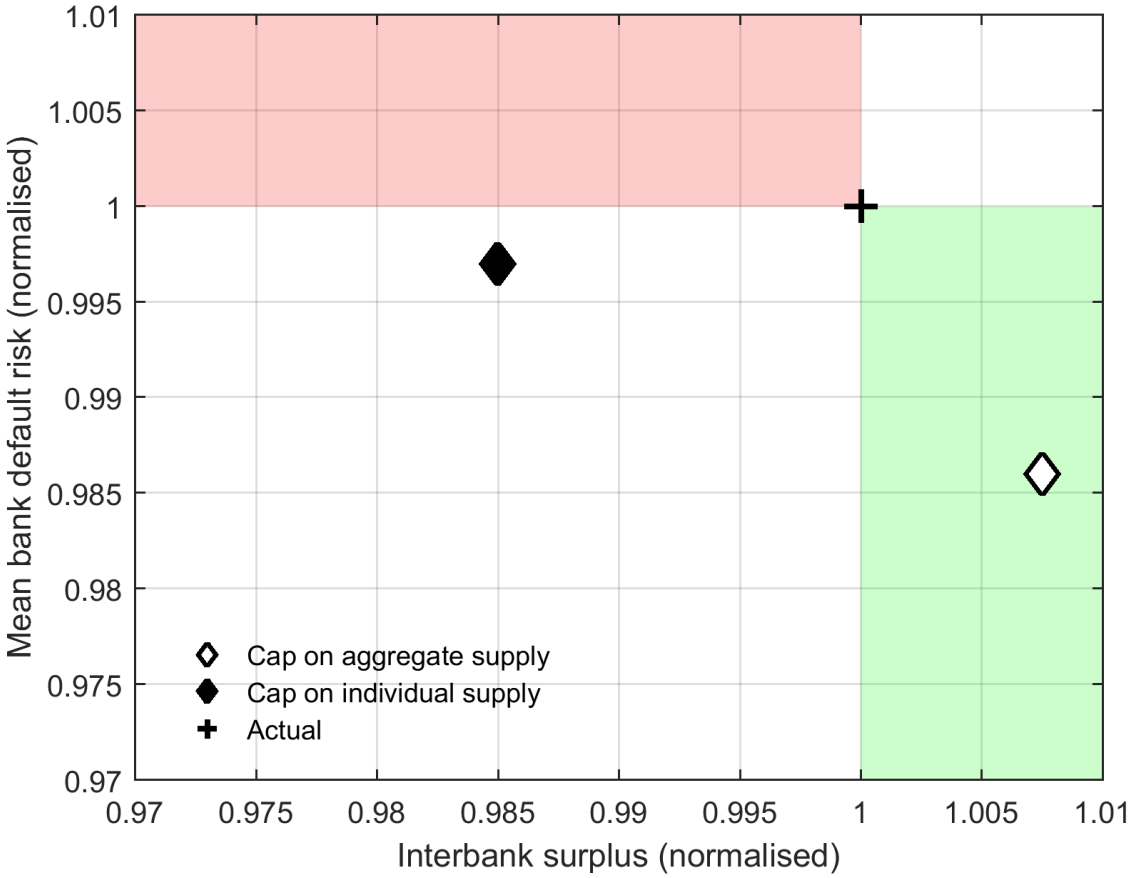
1.7.3 Capital ratios

The second form of regulation we consider is a minimum capital requirement, as applied by regulators since the crisis. As described in Section 1.2, there is very little variation in risk-weights for exposures to banks under the standardised approach to risk-weighting. To assess the effect of a stylised risk-insensitive capital requirements, we simulate a further increase in λ_{it} by up to 2% holding bank fundamentals constant, as set out in Figure 1.13.

We propose a pairwise adjustment (that is, we allow λ_{ijt} to vary at the pair level) to capital ratios that is more closely targeted at network externalities. The key parameter in our model is Γ_{ij}, contagion intensity: links where this is high are particularly costly in terms of their effect on default risk. We propose increasing the capital requirements for any link with $\Gamma_{ij} > median(\mathbf{\Gamma})$ ("high risk links") by some value b (where we increase the value b from 0% to 10% in Figure 1.13). For any link where Γ_{ij} is less than the 20th percentile of the distribution ("low risk links"), we propose *decreasing* the associated capital requirements by $b + 1.5\%$.[16] Our results suggest a social planner would strictly prefer this targeted change in capital ratios to a risk-insensitive increase in capital ratios.

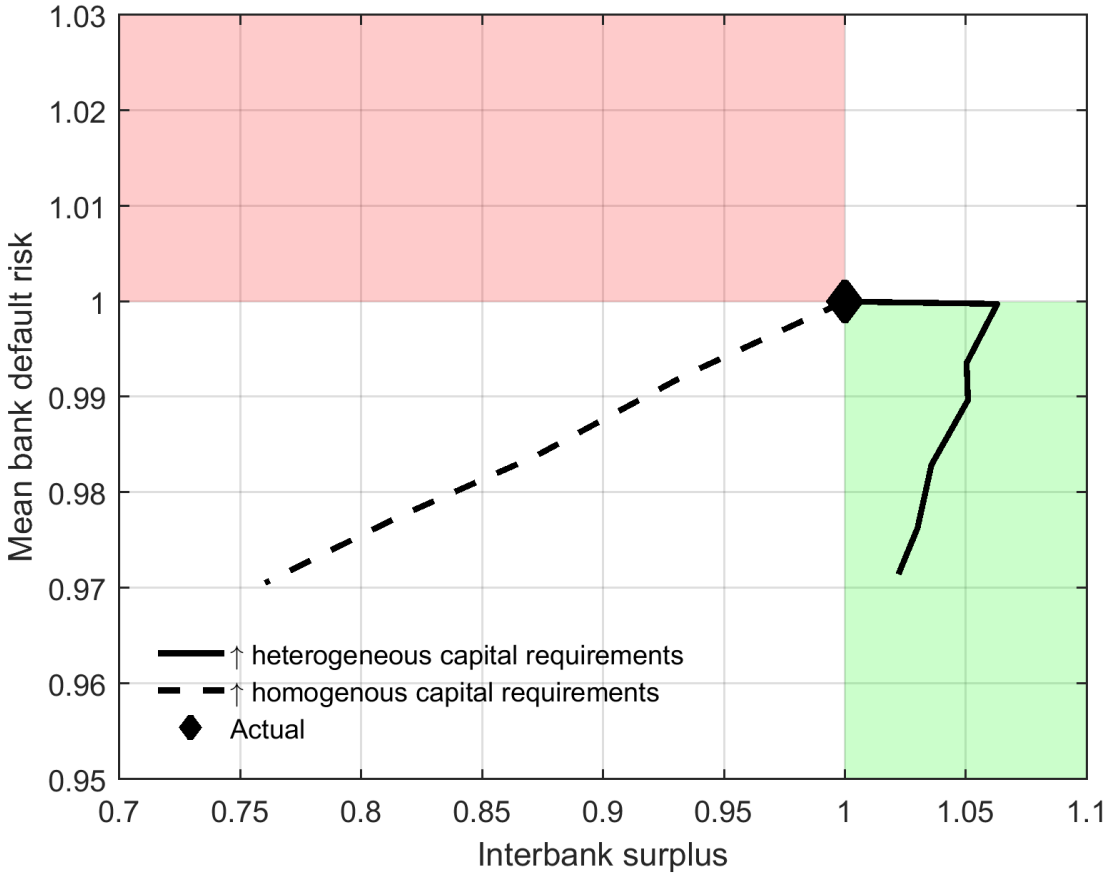
[16] Any spread like this is an improvement over homogeneous capital requirements, this particular spread is one we have chosen arbitrarily as one that produces good results.

Figure 1.12: Counterfactual analysis of caps



Note: The + sign indicates actual normalised default risk and interbank surplus. The black diamond simulates a cap on individual exposures, C_{ij}. The white diamond simulates a cap on each bank's aggregate exposures, $\sum_j C_{ij}$. The social planner would strictly prefer a cap on aggregate exposures to a cap on individual exposures.

Figure 1.13: Counterfactual analysis of capital requirements



Note: This figure starts with actual normalised default risk and interbank surplus (the black diamond). We then plot the effect of (i) homogenous increases in capital requirements for all banks up to an additional 2% (the dashed line) and (ii) heterogeneous adjustments to capital requirements, as we describe in the text (the solid line). Heterogeneous capital requirements can reduce bank default risk by the same amount as homogeneous capital requirements, whilst materially increasing interbank surplus.

1.8 Conclusion

In this paper we structurally estimate a model of network formation and contagion. In contrast to much of the literature on financial networks, our model of network formation

is in the spirit of the wider industrial organisation literature in two ways. First, we model network formation as the interaction of demand for financial products and their supply, with a focus on identifying the relevant underlying cost function. Second, in specifying our model and taking it to data we pay particular attention to the role of unobserved firm- and pair-level heterogeneity. In particular, the core of this paper is heterogeneity in contagion intensity, including (i) why one might reasonably expect contagion intensity to be heterogeneous, (ii) how this heterogeneity can be identified empirically and (iii) what implications this heterogeneity has for strategic interactions between firms and their regulation. The primary message of this paper is that this heterogeneity in contagion intensity has material implications for systemic importance, efficiency and optimal regulation.

References

Acemoglu, D., Ozdaglar, A., and Tahbaz-Salehi, A. (2015). Systemic risk and stability in financial networks. *American Economic Review*, 105(2):564–608.

Acharya, V. and Bisin, A. (2014). Counterparty risk externality: Centralized versus over-the-counter markets. *Journal of Economic Theory*, 149:153–182.

Adrian, T., Boyarchenko, N., and Shachar, O. (2017). Dealer balance sheets and bond liquidity provision. *Journal of Monetary Economics*, 89:92–109.

AFME (2017). Crd 5: The new large exposures framework. Technical report, Association for Financial Markets in Europe.

Afonso, G., Kovner, A., and Schoar, A. (2011). Stressed, not frozen: The federal funds market in the financial crisis. *The Journal of Finance*, 66(4):1109–1139.

Allen, F., Babus, A., and Carletti, E. (2009). Financial crises: theory and evidence. *Annu. Rev. Financ. Econ.*, 1(1):97–116.

Allen, J., Hortaçsu, A., and Kastl, J. (2011). Analyzing default risk and liquidity demand during a financial crisis: The case of Canada. Technical report, Bank of Canada Working Paper.

Babus, A. (2016). The formation of financial networks. *The RAND Journal of Economics*, 47(2):239–272.

Baker, M. and Wurgler, J. (2015). Do strict capital requirements raise the cost of capital? bank regulation, capital structure, and the low-risk anomaly. *American Economic Review*, 105(5):315–20.

Ballester, C., Calvó-Armengol, A., and Zenou, Y. (2006). Who's who in networks. wanted:

The key player. *Econometrica*, 74(5):1403–1417.

Basel Committee (2011). Basel III: a global regulatory framework for more resilient banks and banking systems. Technical report, Bank for International Settlements.

Basel Committee (2014a). Basel III leverage ratio framework and disclosure requirements. Technical report, Bank for International Settlements.

Basel Committee (2014b). Supervisory framework for measuring and controlling large exposures. Technical report, Bank for International Settlements.

Basel Committee (2018a). Countercyclical capital buffer. Technical report, Bank for International Settlements.

Basel Committee (2018b). The treatment of large exposures in the Basel capital standards. Technical report, Bank for International Settlements.

Batiz-Zuk, E., López-Gallo, F., Martínez-Jaramillo, S., and Solórzano-Margain, J. P. (2016). Calibrating limits for large interbank exposures from a system-wide perspective. *Journal of Financial Stability*, 27:198–216.

Benetton, M. (2018). Leverage regulation and market structure: An empirical model of the UK mortgage market.

Bessembinder, H., Jacobsen, S., Maxwell, W., and Venkataraman, K. (2018). Capital commitment and illiquidity in corporate bonds. *The Journal of Finance*, 73(4):1615–1661.

Blasques, F., Bräuning, F., and Van Lelyveld, I. (2018). A dynamic network model of the unsecured interbank lending market. *Journal of Economic Dynamics and Control*, 90:310–342.

Bloch, F., Jackson, M. O., and Tebaldi, P. (2017). Centrality measures in networks.

Chang, B. and Zhang, S. (2018). Endogenous market making and network formation.

Cohen-Cole, E., Patacchini, E., and Zenou, Y. (2010). Systemic risk and network formation in the interbank market.

Craig, B. and Ma, Y. (2019). Intermediation in the interbank lending market. *Manuscript, Stanford Graduate School of Business*.

De Paula, A. (2017). Econometrics of network models. In *Advances in Economics and Econometrics: Theory and Applications, Eleventh World Congress*, pages 268–323. Cambridge University Press Cambridge.

Denbee, E., Julliard, C., Li, Y., and Yuan, K. (2017). Network risk and key players: A structural analysis of interbank liquidity.

Duffie, D. (2017). Financial regulatory reform after the crisis: An assessment. *Management Science*, 64(10):4835–4857.

Eisenberg, L. and Noe, T. H. (2001). Systemic risk in financial systems. *Management*

Science, 47(2):236–249.

Eisfeldt, A. L., Herskovic, B., Rajan, S., and Siriwardane, E. (2018). OTC intermediaries.

Elliott, M., Golub, B., and Jackson, M. O. (2014). Financial networks and contagion. *American Economic Review*, 104(10):3115–53.

Elliott, M., Hazell, J., and Georg, C.-P. (2018). Systemic risk-shifting in financial networks.

European Council (2018). Banking: Council agreement on measures to reduce risk. Technical report, European Council.

Farboodi, M. (2017). Intermediation and voluntary exposure to counterparty risk.

Gale, D. and Yorulmazer, T. (2013). Liquidity hoarding. *Theoretical Economics*, 8(2):291–324.

Glasserman, P. and Young, H. P. (2015). How likely is contagion in financial networks? *Journal of Banking & Finance*, 50:383–399.

Gofman, M. (2017). Efficiency and stability of a financial architecture with too-interconnected-to-fail institutions. *Journal of Financial Economics*, 124(1):113–146.

Greenwood, R., Stein, J. C., Hanson, S. G., and Sunderam, A. (2017). Strengthening and streamlining bank capital regulation. *Brookings Papers on Economic Activity*, 2017(2):479–565.

Hull, J. et al. (2009). *Options, futures and other derivatives/John C. Hull.* Upper Saddle River, NJ: Prentice Hall,.

Iyer, R. and Peydro, J.-L. (2011). Interbank contagion at work: Evidence from a natural experiment. *The Review of Financial Studies*, 24(4):1337–1377.

Kashyap, A. K., Stein, J. C., and Hanson, S. (2010). An analysis of the impact of 'substantially heightened' capital requirements on large financial institutions. *Booth School of Business, University of Chicago, mimeo*, 2.

Kelejian, H. H. and Prucha, I. R. (1998). A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *The Journal of Real Estate Finance and Economics*, 17(1):99–121.

Kelejian, H. H. and Prucha, I. R. (1999). A generalized moments estimator for the autoregressive parameter in a spatial model. *International economic review*, 40(2):509–533.

Kotidis, A. and Van Horen, N. (2018). Repo market functioning: The role of capital regulation.

Modigliani, F. and Miller, M. H. (1958). The cost of capital, corporation finance and the theory of investment. *The American*, 1:3.

Rahi, R. and Zigrand, J.-P. (2013). Arbitrage networks. *Available at SSRN 1430560*.

Robles-Garcia, C. (2018). Competition and incentives in mortgage markets: The role of

brokers.

Welfens, P. J. (2011). The transatlantic banking crisis: lessons, EU reforms and G20 issues. In *Financial Market Integration and Growth*, pages 49–126. Springer.

Yellen, J. (2013). Interconnectedness and systemic risk: Lessons from the financial crisis and policy implications. *Board of Governors of the Federal Reserve System, Washington, DC*.

A An underlying model of default risk

We set out above a default risk process, which we repeat here for convenience:

$$\underbrace{\mathbf{p_t}}_{\substack{\text{Default}\\\text{risk}}} = \underbrace{\boldsymbol{X_t\beta}}_{\substack{\text{Funda-}\\\text{mentals}}} - \underbrace{\omega\,\mathbf{C_t}\,\boldsymbol{\iota}}_{\text{Hedging}} + \underbrace{\tau_t(\boldsymbol{\Gamma}\circ\mathbf{C_t})\mathbf{p_t}}_{\substack{\text{Counterparty}\\\text{risk}}} + \mathbf{e_t^p}$$

Our proposed default risk process is a reduced form for an underlying, more fundamental, default risk process. This underlying model is more fundamental in that the relationship between default risk and fundamentals is structurally grounded in a balance-sheet based model of contagion. It is not, however, feasible to take this underlying model to data.

In this appendix, we first describe this underlying model. We then use this underlying model to simulate data, and estimate our default risk process using this simulated data. We show that (i) our proposed default risk process fits the simulated data well and (ii) contagion intensity Γ_{ij} is heterogeneous.

A.1 Underlying model

This model builds on Eisenberg and Noe (2001), where the variation is to allow for some heterogeneity at the bank- and pair-level. There are N banks. These banks have fundamentals $\mathbf{F}$, an $N \times 1$ vector whose i'th element denotes the fundamentals of bank i. Banks trade P products, resulting in N by N directed adjacency matrices $\mathbf{C^p}$ for $p \in \{1, ..., P\}$, where C_{ij}^p is the exposure of i to j relating to product p. For the purposes of this example, these matrices are exogenous; we are interested in their effect on bank default risk, not their formation.

Fundamentals update according to a random walk, $\mathbf{F}' = \mathbf{F} + \mathbf{e}$, where $\mathbf{e} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. Each bank has value V_i, which is the sum of its fundamentals and its interbank holdings:

$$V_i = F_i^{'} + \sum_p \sum_j C_{ij}^p[\delta_j + (1 - \delta_j) rr_{ij}^p]$$

where $\delta_i = 1$ if bank i is solvent, and $\delta_i = 0$ otherwise, and rr_{ij}^p is the recovery rate of C_{ij}^p in the event of default. In other words, bank i's value is the sum of its fundamentals and its impaired interbank holdings, where the impairment relates to losses on any interbank exposures to insolvent banks.

The recovery rate is exogenous, and varies according to the product and the banks involved. There are G groups, and each bank is a member of a single group. Recovery rate varies as follows:

$$rr_{ij}^{p} = 1 - r_{ij}^{\tilde{g}} r^{p}$$

where $r^p \in [0,1]$ for $\forall\ p$ and $r^1 < r^2 < ... < r^p$. In other words, recovery rate varies by product: this is a simple representation of some products being riskier than others. $r^{\tilde{g}} \in [0,1]$ for $\forall\ \tilde{g}$, where $\tilde{g} = 1$ if i and j are in the same group and 0 otherwise, and $r^{\tilde{g}=1} < r^{\tilde{g}=0}$. This is a simple representation of pairwise variation in riskiness; for example, banks with headquarters/histories in different countries may involve a lower recovery rate. G generally covers any relevant pairwise variation that is not directly related to the banks' risk profile or exposures matrix.

A bank defaults if its value V_i falls below some critical value $\bar{V}_i$. In that sense, δ is a function of $\mathbf{V}$), such that the problem is simply finding a fixed point in $\mathbf{V}$. Eisenberg and Noe (2001) propose the following iterative algorithm, for a given draw of $\mathbf{e}$:

1. Set initial $\delta_j^{m=0} = 1$ for $\forall j$.
2. Calculate $\mathbf{V^{m=0}} = f(\delta^{\mathbf{m=0}}, \mathbf{F}')$.
3. For any bank where $V_j^{m=0} < \bar{V}_j$, set $\delta_j^{m=1} = 0$.
4. Calculate $\mathbf{V^{m=1}} = f(\delta^{\mathbf{m=1}}, \mathbf{F}')$.
5. Iterate until $\delta^{\mathbf{m}}$ converges.

In other words, we propose embedding three sources of pair-wise heterogeneity (apart from obvious heterogeneity in aggregate exposures) in a simple model of bank default risk. The first source of heterogeneity is in $\mathbf{\Sigma}$: the fundamentals of some banks may be positively correlated, some may be negatively correlated. The second source of heterogeneity is in the product-type of the exposures matrix: some pairs may have larger exposures in riskier types. The third source of heterogeneity is in groups: some banks belong to the same group, and so are less risky.

A.2 Simulated fit

We use this underlying model to simulate data. We specify distributions for each of the primitives, including an exogenous network, and randomly draw realisations. We then sample repeatedly from the distribution of $\mathbf{e}$, and calculate default risk of bank i as simply the proportion of the draws in which each bank fails. This gives us a panel of exogenous bank fundamentals and network exposures and endogenous bank default risk.

We then fit our proposed default risk model, as described above, to this simulated data. We set out the results in Table 1.5.

Table 1.5: Results

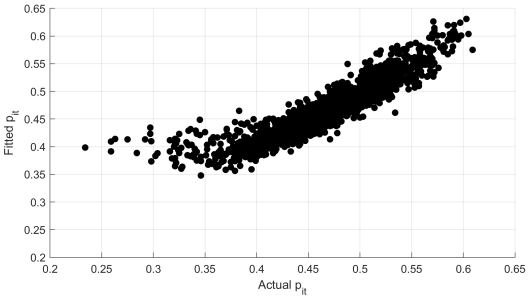
		[1]	[2]	[3]
	Min			0.149
				(-)
Γ_i	$Median$		0.144***	0.178
			(12.96)	(-)
	Max			0.205
				(-)
β		-0.0261***	0.0395***	0.0468
		(-45.73)	(7.76)	(93.96)
FE		t	t	t
R^2		0.77	0.79	0.83
No. obs		2,000	2,000	2,000

Note: Figures in parentheses are t-statistics. ***, **, * indicate different from 0 at 1%, 5% and 10% significance, respectively.

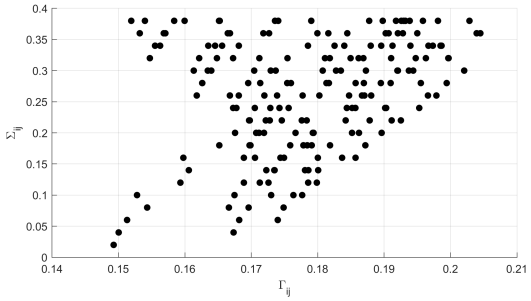
We find that (1) a spatial autoregression fits the simulated data well, (2) heterogeneity in contagion intensity is important (the fit is materially better in column [3] than in column [2]) and (3) contagion intensity is related to pairwise covariance in fundamentals, as set out in Figure 1.14. In this sense, our proposed default risk process can be thought of as a reduced form representation of this underlying more fundamental model, and heterogeneous contagion intensity can be thought of as a reduced form representation of underlying sources of pairwise heterogeneity related to risk sharing, jurisdiction effects and exposure type.

Figure 1.14: Estimation results

(a) Fit



(b) Risk sharing



Note: Panel (a) shows simulated true default risk and fitted default risk. Panel (b) shows a positive relationshiop between contagion intensity Γ_{ij} and the covariance in the fundamentals of i and j.

B First stage regression results

Table 1.6: First stage: Default risk

	p_{it}
X_{it}^1	-0.82***
	(-2.61)
FE	i
Other X	Y
R^2	0.82
No. obs	378

Note: Figures in parentheses are t-statistics. ***, **, * indicate different from 0 at 1%, 5% and 10% significance, respectively. X_{it}^1 is a revenue-weighted average of stock market indices and the other fundamentals include the Morgan Stanley World Index, VIX and the first two principal components of World Bank macroeconomic data, as we describe in the text.

Table 1.7: First-stage: Network formation results

	Estimate	t statistic
X_{it}	-0.57***	-3.73
X_{jt}	0.22	1.51
X_{kt}	0.35***	11.90
X_{it}^2	0.01	0.18
X_{jt}^2	0.28*	1.88
X_{kt}^2	-0.39***	-10.05
X_{jt}/X_{it}	0.01	1.42
X_{jt}/X_{kt}	-0.46	-0.55
X_{it}/X_{kt}	-1.44*	-1.69
$\lambda_{it}X_{it}$	13.86***	7.31
$\lambda_{it}X_{jt}$	-2.94	-1.57
$\lambda_{it}X_{kt}$	-10.69***	-13.64
$\lambda_{it}X_{it}^2$	-0.27	-0.35
$\lambda_{it}X_{jt}^2$	-9.24***	-5.38
$\lambda_{it}X_{kt}^2$	11.70***	8.84
$\lambda_{it}X_{jt}/X_{it}$	-0.192**	-2.05
$\lambda_{it}X_{jt}/X_{kt}$	10.32	0.93
$\lambda_{it}X_{it}/X_{kt}$	-21.27*	-1.89
FE	ij	
R^2	0.70	
No. obs	6,426	

C Mathematical appendix

C.1 EQC

In this appendix, we derive the equilibrium quantity condition, EQC. The first order supply condition is:

$$r_{ijt} = -\frac{\partial r_{ijt}}{\partial C_{ijt}} C_{ijt} + puc_{ijt} + \frac{\partial p_{it}}{\partial C_{ijt}} \sum_k \frac{\partial puc_{ikt}}{\partial p_{it}} C_{ikt} - \frac{\partial \Pi^D_{it}}{\partial p_{it}} \frac{\partial p_{it}}{\partial C_{ijt}} + r_{i0t} + e^S_{ijt}$$

It follows immediately from DRP that $\frac{\partial p_{it}}{\partial C_{ijt}} = \tau_t \Gamma_{ij} p_{jt}$, from our assumed cost function that $\frac{\partial puc_{kjt}}{\partial p_{it}} = \phi_1 \lambda_{kjt}$ and from our demand model that $\frac{\partial r_{ijt}}{\partial C_{ijt}} = -B$ and $\frac{\partial \Pi^D_{it}}{\partial p_{it}} = -\sum_k \frac{\partial r_{kit}}{\partial p_{it}} C_{kit}$:

$$r_{ijt} = BC_{ijt} + \phi_1 \lambda_{ijt} p_{it} + \phi_1 \big[-\omega + \tau_t \Gamma_{ij} p_{jt} \big] \sum_m \lambda_{imt} C_{imt} + \frac{\partial p_{it}}{\partial C_{ijt}} \sum_m \frac{\partial r_{mit}}{\partial p_{it}} C_{mit} + r_{i0t} + e^S_{ijt}$$

For ease of exposition we then repeat the same equation for supply from bank k to bank i:

$$r_{kit} = BC_{kit} + \phi_1 \lambda_{kit} p_{kt} + \phi_1 \big[-\omega + \tau_t \Gamma_{ki} p_{it} \big] \sum_m \lambda_{kmt} C_{kmt} + \frac{\partial p_{kt}}{\partial C_{kit}} \sum_m \frac{\partial r_{mkt}}{\partial p_{kt}} C_{mkt} + r_{k0t} + e^S_{kit}$$

When bank i considers how much to supply to bank j, it takes into account the impact of the resulting increase in p_{it} on its profits from being supplied exposures. That is, it takes into account the effect of its supply on r_{kit}. We assume that bank i takes the interest rates of transactions involving other parties as given, such that:

$$\frac{\partial r_{kit}}{\partial p_{it}} = \phi_1 \tau_t \Gamma_{ki} \sum_m \lambda_{kmt} C_{kmt}$$

Substitute this and the equation for demand into supply, and we obtain the EQC:

$$\begin{aligned} 0 \;=\; & \delta_{jt} + \zeta_{ij} + e^D_{ijt} - 2BC_{ijt} - \sum_{k \neq i}^{N} \theta_{ik} C_{kjt} + e^S_{ijt} \\ & -\lambda_{ijt} \phi_1 p_{it} - \phi_1 \big[-\omega + \tau_t \Gamma_{ij} p_{jt} \big] \sum_{k \neq i}^{N} C_{ikt} \lambda_{ikt} - r_{i0t} \\ & -\phi_1 \tau_t \big[-\omega + \tau_t \Gamma_{ij} p_{jt} \big] \sum_k C_{kit} \Gamma_{ki} \sum_m C_{kmt} \lambda_{kmt} \end{aligned}$$

C.2 Equilibrium links are non-linear in fundamentals

Consider a simplified version of the model in which banks do not consider the impact of their supply decisions on Π^D; that is, they consider the impact on their funding costs when supplying on the interbank network, but not on their funding costs when demanding from the interbank network. This means that the EQC is linear in C. Furthermore, for simplicity of exposition (and without loss of generality regarding the form of equilibrium C) suppose $\zeta = \omega = e^D = e^S = r_0 = 0$, $2B = \phi_1 = \lambda = 1$, $\theta_{ij} = \theta$, $\Gamma_{ij} = \Gamma$ for all banks and parameters are such that all equilibrium exposures are strictly positive. The EQC is then as follows:

$$0 = \delta_{jt} - C_{ijt} - \theta \sum_{k \neq i}^{N} C_{kjt} - p_{it} - \Gamma p_{jt} \sum_{k \neq i}^{N} C_{ikt}$$

In this case an analytical expression for equilibrium exposures exists, where $\boldsymbol{C}$ is a $N(N-1) \times 1$ vector of endogenous exposures, $\boldsymbol{p}$ is a $N \times 1$ vector of default probabilities, $\boldsymbol{X}$ is a $N \times 1$ vector of fundamentals, $\boldsymbol{M_i}$, $\boldsymbol{M_j}$, $\boldsymbol{M_{\sum i}}$ and $\boldsymbol{M_{\sum j}}$ are matrices that select and sum the appropriate elements in $\boldsymbol{C}$ and $\boldsymbol{p}$ and . and $\circ$ signify matrix multiplication and the Hadamard product, respectively:

$$\boldsymbol{C} = \Big[\boldsymbol{I} + \theta \boldsymbol{M_{\sum j}} + (\boldsymbol{M_j}.\boldsymbol{p}) \circ \boldsymbol{M_{\sum i}}\Big]^{-1} \Big[\boldsymbol{M_j}.\boldsymbol{\delta} - \boldsymbol{M_i}.\boldsymbol{p}\Big]$$

Given that p is a linear function of X, as set out in the DRP, it follows that equilibrium C is a non-linear function of X.

D Robustness tests

We run two alternative specifications as robustness tests, both of which test how sensitive our results are to how we treat time-variation in the risk premium. In the first robustness test, we use alternative measures of bank default risk and bank-specific fundamentals that exclude the risk premium, but otherwise estimate our baseline specification. In the second robustness test, we use the same data as in our baseline results but amend the default risk process so that common time-variation in the risk premium does not propagate through the interbank network. In both cases, the results are quantitatively and qualitatively similar to our baseline results.

D.1 Robustness: Removing the effect of the risk premium

We attempt to remove the effect of the risk premium by using different data. For bank default risk, we use a proprietary Bloomberg estimate of bank default risk (DRISK), excluding the risk premium, based on market data about the bank. For bank-specific fundamentals, we calculate the weighted average consumption growth in various geographic regions, where the weighting is the proportion of a bank's revenues that came from that region. We plot some summary statistics in Figure 1.15. Our estimation procedure is otherwise the same as our baseline specification. In Table 1.8 we set out our results, which are quantitatively and qualitatively similar to our baseline results.

D.2 Robustness: Preventing the risk premium from propagating through the network

In this robustness test, we amend the default risk process. As in our baseline specification, let $\boldsymbol{p_t}$ signify the default risk implied by Credit Default swap premia, $\boldsymbol{X_{1,t}}$ signify the matrix of bank-specific equity indices and $X_{2,t}$ signify the Morgan Stanley World Index, which we use to control for common variation in the risk premium. In the following specification, we amend the default risk process so that the risk premium does not propagate through the interbank network.

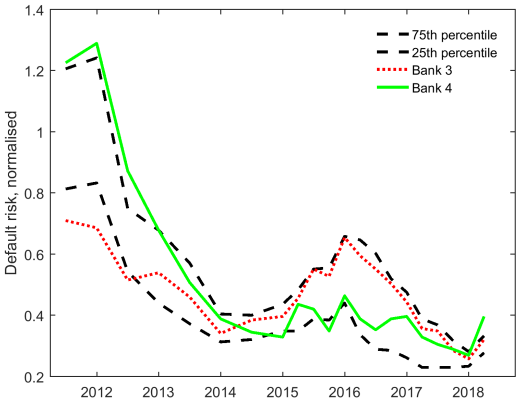
Table 1.8: Results: Robustness check 1

		[1]	
ϕ		2.16	
		(1.04)	
τ		7.64***	
		(4.32)	
β_1		-0.04***	
		(-3.07)	
ω		0.03***	
		(7.91)	
	Min	*Median*	*Max*
$\tilde{\Gamma}_i$	0.15***	0.19***	0.74***
	(12.23)	(8.10)	(18.65)
$\tilde{\theta}_k$	5.24	7.07*	31.22***
	(0.37)	(1.80)	(11.96)
a_i	0.04	0.71	5.53***
	(0.21)	(1.47)	(2.55)
Network			
FE		ij, it, jt	
R^2		0.85	
No. obs		6,426	
Default risk			
FE		i	
Controls		Y	
R^2		0.85	
No. obs		378	

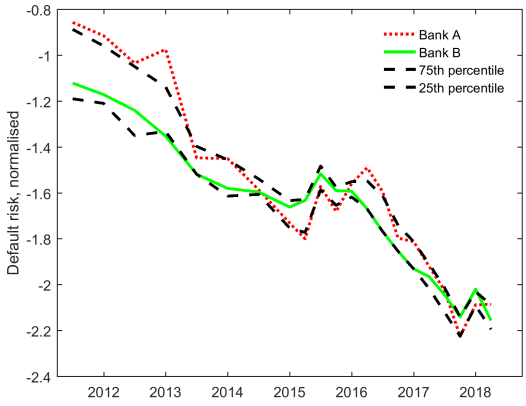
Notes: SEs clustered at bank level. Figures in parentheses are t-stats. ***, **, * indicate different from 0 at 1%, 5% and 10% significance, respectively. For the heterogeneous parameters we report estimates and t-stats for the minimum, median and maximum, and plot the full distribution below. **Notation**: ϕ is the sensitivity of cost of equity to default risk, τ is the extent to which contagion intensity varies over time, β_1 is the effect of bank-specific fundamentals, ω is the effect of hedging, $\tilde{\Gamma}_i$ is contagion intensity, $\tilde{\theta}_k$ governs product differentiation based on characteristics and a_i scales exposures for non-UK banks. Controls in the default risk process are VIX, MSWI and macro data.

Figure 1.15: Removing the effect of the risk premium

(a) Default risk excluding risk premium



(b) Revenue-weighted consumption growth



Note: Panel (a) shows a Bloomberg measure of default risk that excludes the risk premium. Panel (b) shows a bank-specific fundamental measure that is the weighted average consumption growth in various geographic regions, where the weighting is the proportion of a bank's revenues that came from that region (normalized by a negative number for ease of comparison with panel (a)).

$$\underbrace{\mathbf{p_t}}_{\substack{\text{Default}\\ \text{risk}}} = \underbrace{\boldsymbol{X_t\beta}}_{\substack{\text{Funda-}\\ \text{mentals}}} + \underbrace{(\mathbf{I} - \tau_t \boldsymbol{\Gamma} \circ \mathbf{C_t}) X_{2,t}\beta_2}_{\substack{\text{Risk}\\ \text{premium}}} - \underbrace{\omega\, \mathbf{C_t}\, \boldsymbol{\iota}}_{\text{Hedging}} + \underbrace{\tau_t(\boldsymbol{\Gamma} \circ \mathbf{C_t})\mathbf{p_t}}_{\substack{\text{Counterparty}\\ \text{risk}}} + \mathbf{e_t^p}$$

Re-arranging for equilibrium $\boldsymbol{p_t}$:

$$\mathbf{p_t} = X_{2,t}\beta_2 + \sum_{s=0}^{\infty} (\tau_t \boldsymbol{\Gamma} \circ \mathbf{C_t})^s (\mathbf{X_t}\boldsymbol{\beta} - \omega \mathbf{C_t}\boldsymbol{\iota} + \mathbf{e_t^p})$$

This allows bank default risk and therefore their cost of equity to vary with the risk premium, but the effect of the risk premium on bank default risk does not depend on the interbank network. We set out our results below in Table 1.8, and find that our results are quantitatively and qualitatively similar to our baseline results.

Table 1.9: Results: Robustness check 2

		[1]	
ϕ		1.89***	
		(4.43)	
τ		7.60***	
		(4.15)	
β_1		-0.98***	
		(-3.28)	
ω		0.04***	
		(7.51)	
	Min	*Median*	*Max*
$\tilde{\Gamma}_i$	0.15***	0.21***	0.52***
	(9.86)	(10.16)	(3.86)
$\tilde{\theta}_k$	5.20	6.98***	31.06***
	(0.57)	(5.43)	(8.70)
a_i	0.04	1.28**	5.53***
	(0.21)	(1.93)	(2.54)
Network			
FE		ij, it, jt	
R^2		0.85	
No. obs		6,426	
Default risk			
FE		i	
Controls		Y	
R^2		0.82	
No. obs		378	

Notes: SEs clustered at bank level. Figures in parentheses are t-stats. ***, **, * indicate different from 0 at 1%, 5% and 10% significance, respectively. For the heterogeneous parameters we report estimates and t-stats for the minimum, median and maximum, and plot the full distribution below. **Notation**: ϕ is the sensitivity of cost of equity to default risk, τ is the extent to which contagion intensity varies over time, β_1 is the effect of bank-specific fundamentals, ω is the effect of hedging, $\tilde{\Gamma}_i$ is contagion intensity, $\tilde{\theta}_k$ governs product differentiation based on characteristics and a_i scales exposures for non-UK banks. Controls in the default risk process are VIX, MSWI and macro data.

E Additional post-estimation tests

E.1 Default risk and cost of equity

In this sub-section, we show test our parameterisation of a bank's cost of equity as a function of its default risk is reasonable. We run a linear regression of a bank's cost of equity, taken from Bloomberg and based on a simple CAPM model, on its default risk.

$$c^e_{it} = \phi p_{it} + FE_i + FE_t + e^e_{it}$$

As we set out below, we find that the relationship between the two is positive and significant, as expected. Riskier banks face a higher cost of capital, even when controlling for time fixed effects.

Table 1.10: Cost of equity and default risk

	c^e_{it}
p_{it}	1.31***
	(2.94)
FE	i,t
R^2	0.69
No. obs	346

Note: Figures in parentheses are t-statistics. ***, **, * indicate different from 0 at 1%, 5% and 10% significance, respectively.

E.2 Testing heterogeneous contagion intensity

We set out above three motivations for heterogeneous contagion intensity Γ_{ij}: (1) correlations in fundamentals (risk sharing, in other words), (2) variations in product and (3) other pairwise variations, including common jurisdiction. We estimate general Γ_{ij} without imposing any of these motivations in estimation, meaning we can test them post-estimation. In particular, risk sharing implies a relationship between $\mathbf{X}\boldsymbol{\beta}$ and Γ_{ij}, which we test in the following way.

As bank-specific fundamentals we use equity indices weighted by the geographic revenues of each bank, as we describe above. This implies that banks that get their revenues from the same geographic areas will have positively correlated fundamentals, and banks that have differing geographic revenue profiles will have less correlated fundamentals. For each pair of banks we calculate the empirical correlation coefficient as $\widehat{\rho}_{ijt} = Corr(\mathbf{X_{it}}\widehat{\boldsymbol{\beta}}, \mathbf{X_{jt}}\widehat{\boldsymbol{\beta}})$.

We then divide our bank pairs into two groups, "more correlated" and "less correlated", by defining the dummy variable $1_{\rho_{ij}} = 1$ if $\widehat{\rho}_{ij} > median(\widehat{\rho}_{ij})$ and $1_{\rho_{ij}} = 0$ otherwise. We divide bank pairs similarly regarding Γ_{ij}, into "safe links" and "risky links", by defining the dummy variable $1_{\Gamma_{ij}} = 1$ if $\widehat{\Gamma}_{ij} > median(\widehat{\Gamma}_{ij})$ and $1_{\Gamma_{ij}} = 0$ otherwise. Risk sharing implies that safe links should be less correlated, and risky links should be more correlated. Risk sharing is, however, difficult to separately identify from other motivations for heterogeneous contagion intensity. In particular, less correlated links are more likely to go across jurisdictions than more correlated links, where going across jurisdictions may make links less safe. We test this by identifying the home jurisdiction of each of the $N = 18$ banks in our sample and classifying each as being in the UK, North America, Europe or Asia. We then define the dummy variable $1_G = 1$ if they share the same home jurisdiction, and 0 otherwise. We do not attempt to test the effect of product variations, as there are many product characteristics and we do not have a clear ranking of their relative riskiness.

We run the following linear regression:

$$1_{\Gamma_{ij}} = \alpha_0 + \alpha_1 1_{\rho_{ij}} + \alpha_2 1_G + \alpha_3 1_G 1_{\rho_{ij}} + e_t^{\alpha}$$

The coefficient on the interaction term is positive and significant: where banks are in the same jurisdiction, then more correlated links are less safe. We interpret this as evidence in support of a risk sharing motivation for heterogeneous contagion intensity. The coefficient on 1_G is the right sign (indicating that links within the same jurisdiction are safer), but insignificant. The coefficient on $1_{\rho_{ij}}$ is negative and significant: this suggests that when links go across jurisdictions, less correlated links are actually less safe. This could still be because of confounding jurisdictional effects: within the set of links that cross jurisdictions, more distant links will be riskier but also less correlated.

Table 1.11: Drivers of heterogeneous contagion intensity

	[1]
$1_{\rho_{ij}}$	-0.280***
	(-4.44)
1_G	-0.204
	(-1.61)
$1_G 1_{\rho_{ij}}$	0.600***
	(3.96)
R^2	0.09
No. obs	153

Note: Figures in parentheses are t-statistics. ***, **, * indicate different from 0 at 1%, 5% and 10% significance, respectively.

Chapter 2:
Information loss over the business cycle

The business cycle induces turnover in mutual funds: they exit in recessions and enter in recoveries. The effect of this firm turnover on welfare depends on a key trade-off: on the one hand, the business cycle "cleanses" the market of low quality exiting funds and replaces them with entrants that may on average be higher quality. On the other hand, the entrants have no returns history and so investors have less precise beliefs about their ability, where this "information loss" leads to misallocation that harms welfare. I examine this trade-off by estimating a structural model in which rational investors form and update beliefs about competing mutual funds that endogenously choose to enter and exit the market. I estimate this model using data on US mutual funds. I find that the business cycle has material, persistent effects that are negative in the short-term but turn positive as the effect of information loss decays over time.

2.1 Introduction

The business cycle induces firm turnover: firms exit in recessions, and enter in recoveries. What impact does this firm turnover have on outcomes post-recovery? How persistent is this impact? How does the impact vary with the characteristics of the business cycle?

I seek to answer these questions by exploring a key trade-off that underpins them. On the one hand, the business cycle can improve outcomes by "cleansing" the market of low quality firms: replacing low quality firms that exit during the recession with higher quality firms that enter during the subsequent recovery. On the other hand, the firms that exit have a track record of performance, whereas the entrants that replace them do not. To the extent that this information was valuable and had an impact on outcomes, this "information loss" could harm outcomes.

This trade-off between cleansing and information loss is important in the wide class of markets in which unobserved quality is important for outcomes and past performance is informative about quality. The mutual fund industry is such a market, and is a natural setting in which to study this trade-off for the following reasons. First, there is a broad literature exploring whether quality or ability is important for mutual funds. Second, there is clear evidence that investors in mutual funds respond to past returns.

I evaluate this trade-off by estimating a structural equilibrium model of investor and mutual fund behaviour. I estimate this model, and I use the results to run counterfactuals in which I simulate business cycles of varying types and quantify the impact of the resulting firm turnover. This allows me to draw novel conclusions about the size and persistence of business cycle shocks. This paper is the first, to my knowledge, to structurally estimate the impact of cleansing and information loss over the business cycle.

The model consists of two parts. On the demand side, rational investors invest in mutual funds based on their beliefs about the heterogeneous abilities of funds to generate excess returns, and update those beliefs over time as they observe fund performance, following Berk and Green (2004). The ability of a given mutual fund to generate excess returns is decreasing in the total size of the mutual fund industry (which in the spirit of Pástor and Stambaugh (2012) is the way in which I model competition between mutual funds) and also varies with a macro-economic factor. The aggregate surplus generated by a fund is the total payout to the fund managers and to investors. This aggregate surplus is increasing and convex in fund ability, and is also increasing in the precision of investor beliefs: if these

beliefs are imprecise, then there is misallocation (over-investment in low ability funds or under-investment in high ability funds) that harms surplus.

On the supply side, funds make dynamic decisions to exit and enter. Funds take the size of the mutual fund industry as given, and form beliefs about how its size varies with the macro-economic factor. If a fund exits it receives a scrap value that represents the use of its human capital elsewhere. If a fund enters it incurs a fixed entry cost and randomly draws ability from the population distribution.

I take the size of the mutual fund market as given, and instead focus on *compositional inefficiencies* regarding the types of funds that make up the market: incumbent funds do not take into account that if they exit then new funds could enter, which may improve aggregate surplus depending on their relative characteristics. A business cycle (which I model as a negative shock to the macro-economic factor, followed by a recovery) results in exactly this exchange of funds: during the recession funds exit, which reduces competition and allows funds to enter during the subsequent recovery.

The impact of this firm turnover depends on two countervailing effects. Low ability funds are smaller and are more likely to exit during the recession, whereas the firms that replace them are of average, and therefore higher, ability. Surplus is increasing and convex in ability, whereas fund size is increasing and linear in ability. This means that although the aggregate size of the exiting and entering funds is the same, the surplus generated by the higher ability entering funds is higher. This is the cleansing effect. The entering funds, however, have no returns history, meaning that investors have less precise beliefs about their ability. This results in more misallocation in equilibrium, which is bad for aggregate surplus. This is the information loss effect. Cleansing is about the *first moment* in ability (entrants are higher ability on average), whereas information loss is about the *second moment* (there is greater uncertainty about the ability of entrants).

The model allows me to formalise the key parameters that determine the relative strength of these two countervailing effects. The strength of the cleansing effect depends on the dispersion in the distribution of fund abilities, the differing extents to which low and high ability funds exit and the convexity of a fund's surplus with respect to its ability. The strength of the information loss effect depends on the informational content of returns and the age of exiting funds.

I estimate both parts of this model using data on US Equity mutual funds. I fit the demand-side to data on mutual fund size, taking into account fund returns. I do not identify

the ability of funds directly, but I do identify the beliefs of investors about ability from the size of the fund: the model implies that bigger funds, all else being equal, are believed to be higher ability. I identify the value of information from the rate at which investors adjust their holdings in response to past performance: a returns history is valuable if investors are responsive to returns.

I fit the supply-side to data on fund entry and exit. In doing so, I allow the scrap value to vary according to the state and the type of the fund for two reasons. First, it allows me to more accurately capture exit dynamics: it stands to reason that funds of different types have differing outside options. Second, it ensures model consistency: in my analysis I observe and hold fixed the equilibrium relationship between the total size of the mutual fund industry and the macroeconomic factor. State-type variation in scrap values means I can ensure in my estimation that the equilibrium fund-specific exit dynamics implied by my model are consistent with these aggregate dynamics. In the extreme case in which I estimate a different scrap value for each state-type combination, I am able to perfectly match observed exit rates.

I find that the model fits well on both the demand- and supply-side. I find that investors are relatively slow to respond to past returns: the estimated signal-to-noise ratio implies that investors consider the informational content in their priors to be roughly equivalent to 4 years of returns data. I also find that scrap values vary in intuitive ways with the state and type of the fund: funds have better outside options when the macro-economic factor is good and when they are believed to be high ability.

I use my results to counterfactually simulate a business cycle of varying depths, where deeper business cycles result in more firm turnover. I then compare the surplus generated by the exiting funds and the entering funds at various points post-recovery to reach two main conclusions.

First, I find that the business cycle harms surplus in the short-term and improves surplus in the long-term. The information loss effect dominates the cleansing effect in the short-term, such that the firm turnover harms aggregate surplus. Post-recovery, both the exiting and entering funds age and so benefit from additional returns information: the benefit of this extra information is greater for the entering funds who started with no information, and so over time the information loss effect decays. There is a "switching point" at 27 months, by which time information loss effect decays to the point where it is dominated by the cleansing effect. From this point onward, aggregate surplus is higher due to the cleansing effect.

Second, I find that deeper business cycle have bigger persistent effects in the short-term and the long-term. For the deepest business cycle I model (which is roughly equivalent to the financial crisis), the aggregate surplus of entering funds is 20% less than the aggregate surplus of the exiting funds in the first month after the recovery. By month 80, the information loss has decayed to the point where the aggregate surplus of entering funds is 30% greater than that of exiting funds. The impact on total surplus in the market is small but material: the short-term harm from the information loss is 0.5% of total surplus in the market (including funds that neither exited nor entered) and the long-term benefit from the cleansing effect is 0.9%. The switching point at which the cleansing effect dominates the information loss effect is around 27 months regardless of the depth of the business cycle. I also model the *cumulative* impact of the firm turnover over time,[1] taking into account the one-off costs incurred by entering firms. I find that it takes 75 months for the cumulative impact to become positive.

The persistent effects of the business cycle have been extensively studied in macroeconomic contexts, but less so in market-specific contexts. The main contribution of this paper is to develop an under-explored implication of business cycles: the information loss that results from firm turnover. I explore the conditions under which this information loss dominates the cleansing effect, and I quantify how this trade-off changes over time.

I review the literature below. In Section 2, I introduce the data and set out some guiding empirical facts. In Section 3, I set out my model. In Section 4, I describe my empirical approach. In Section 5, I report my results. In Section 6, I undertake counterfactual analyses. In Section 7, I conclude.

2.1.1 Related literature

This paper is related to three broad strands of literature.

First, this paper is related to the literature on cleansing that goes back to Schumpeter et al. (1939), and is featured more recently in Caballero and Hammour (1996) and Castillo-Martınez (2018). In this paper, I document and measure cleansing in the context of mutual funds. I also show how cleansing may bring first-moment benefits but second-moment costs in the form of information loss. This loss of information over the business cycle has not been studied extensively. Relatedly, Hale (2012) set out reduced form evidence that recessions

[1] Where the cumulative impact at time t is the sum of the impacts in all previous periods.

affect connections between firms and banks which, in a relationship banking context, could have implications for the extent of information asymmetry.

Second, this paper is related to the literature on mutual funds generally (Berk and Green, 2004; Berk and Van Binsbergen, 2015; Pástor and Stambaugh, 2012; Pollet and Wilson, 2008; Fama and French, 2010) and more specifically the effect of the business cycle on mutual fund outcomes (Kosowski, 2011; Glode, 2011; Kacperczyk et al., 2014, 2016). There is a more limited literature that estimates structural models related to mutual funds, including Roussanov et al. (2018) and Gavazza (2011). I introduce information loss over the business cycle as a new consideration within this literature, and quantify its importance in a structural econometric context.

Third, this paper is related to the literature on exit and entry, and in particular the estimation of such models (Hotz and Miller, 1993; Rust, 1987). I estimate such a model in the context of mutual funds in which I argue that funds take as given the aggregate size of the industry and form beliefs over its future dynamics. I then show how estimating state-type-specific opportunity costs allows me to consistently match those beliefs in equilibrium.

2.2 Data

I first describe how I select funds and calculate excess returns. I then describe the key empirical facts that motivate my research question and guide my modelling.

2.2.1 Sample selection

I obtain data on mutual fund characteristics and their monthly returns and assets from the database maintained by the Center for Research in Security Prices (CRSP), The University of Chicago Booth School of Business. I select data from January 1990 to December 2016. I limit my sample to actively managed US Equity funds that (i) are never smaller than USD 1m in size, (ii) have at least 12 months of returns data and (iii) have data on their expense ratio. This is the standard approach in the literature (see for example Berk and Van Binsbergen (2015) for an overview of mutual fund selection), but with slightly looser size and history thresholds: this is important because propensity to exit is likely to be correlated with data availability. In other words, the standard thresholds exclude many of the funds I am seeking to study. I am left with a sample of 3,420 funds and a total of 452,222 month-fund

observations.

2.2.2 Calculating excess returns

I calculate excess returns following Berk and Van Binsbergen (2015). I regress returns in excess of the risk-free rate (R_{it}) on a set of 11 common factors ($\mathbf{F_t}$) which are the returns to the main index funds operated by Vanguard, which I list in the table below. The fund's excess return, α_{it} is the residual in this regression:

$$R_{it} = \boldsymbol{\beta}_i \mathbf{F_t} + \alpha_{it} \tag{1}$$

This is a more reasonable benchmark for mutual funds than, for example, a benchmark involving momentum investing returns that would be prohibitively costly to implement in practice. See Berk and Van Binsbergen (2015) for a fuller discussion

Table 2.1: Benchmark

Fund Name	Ticker	Asset Class
S&P 500 Index	VFINX	Large-Cap Blend
Extended Market Index	VEXMX	Mid-Cap Blend
Small-Cap Index	NAESX	Small-Cap Blend
European Stock Index	VEURX	International
Pacific Stock Index	VPACX	International
Value Index	VVIAX	Large-Cap Value
Balanced Index	VBINX	Balanced
Emerging Markets Stock Index	VEIEX	International
Mid-Cap Index	VIMSX	Mid-Cap Blend
Small-Cap Growth Index	VISGX	Small-Cap Growth
Small-Cap Value Index	VISVX	Small-Cap Value

2.2.3 Empirical facts

I set out four empirical facts:

1. **Heterogeneity in fund size**. Funds vary significantly in size at the point of entry and over their lifetime, as I show in Figure 2.1. This is true even controlling for the

macro-economic conditions at the time of entry: in other words, this is cross-sectional variation not inter-temporal variation.

2. **Exit is correlated with size**. Smaller funds are significantly more likely to exit in any given period than bigger funds.

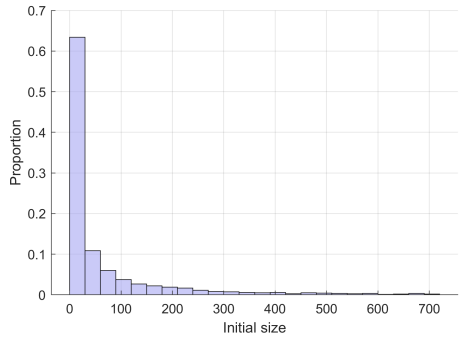
3. **Exit is counter-cyclical**. Funds are more likely to exit when the S&P500 (which I denote macro-economic factor M_t) is low than when it is high, as I show in Figure 2.2.

4. **The size of the mutual fund industry is pro-cyclical**. There is, unsurprisingly, a close relationship between the S&P500 and the aggregate size of the mutual fund industry, which I denote Q_t. I show this graphically in Figure 2.3 and in the regression results in Table 2.2. The R^2 of a regression of Q_t on M_t is 0.75, rising to 0.9 if I include a structural break in the financial crisis.

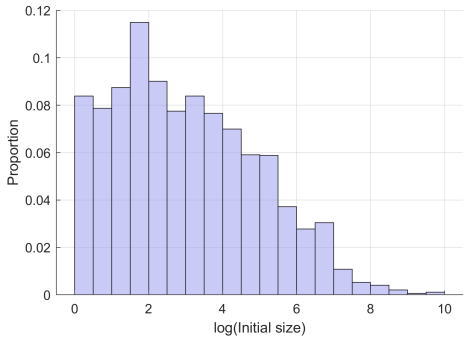
To these empirical facts I add that investors respond to past returns, on which there is a large literature (see, for example, Chevalier and Ellison (1997)). These facts combined naturally give rise to my research question: given that exiting funds are observably different from the average fund, what impact does this exit have on aggregate outcomes? Given that investors clearly attach some value to past returns, what impact does the absence of past returns have on entrants? The macro-economic factor clearly has an impact on aggregate trends in the mutual fund industry, but what about on its composition?

Figure 2.1: Heterogeneity in fund size

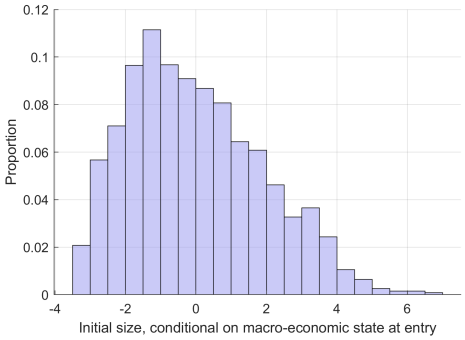
(a) Initial size



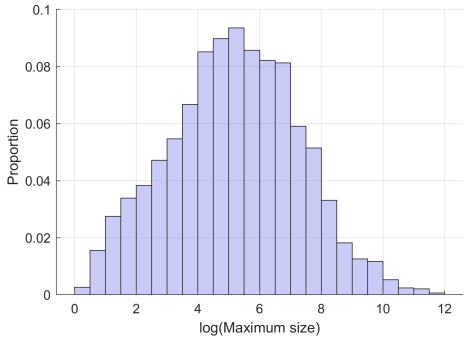
(b) Log initial size



(c) Log initial size conditional on M_1



(d) Log maximum size



Note: Panel (a) shows the distribution of fund size in the first period of its life, excluding the top 5% of funds by size. Panel (b) shows the distribution of the natural log of initial size. Panel (c) conditions on M_1, the level of the SP500 in the period in which the fund entered. Panel (d) shows the log of the maximum size the fund attains during my sample.

Figure 2.2: Exiting funds and the S&P500

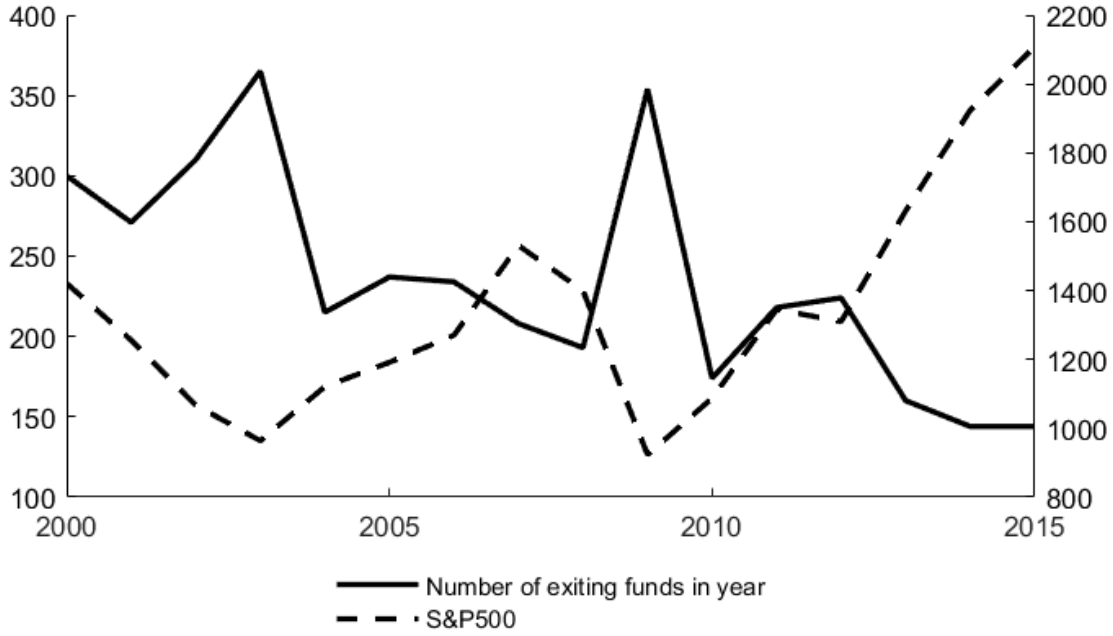
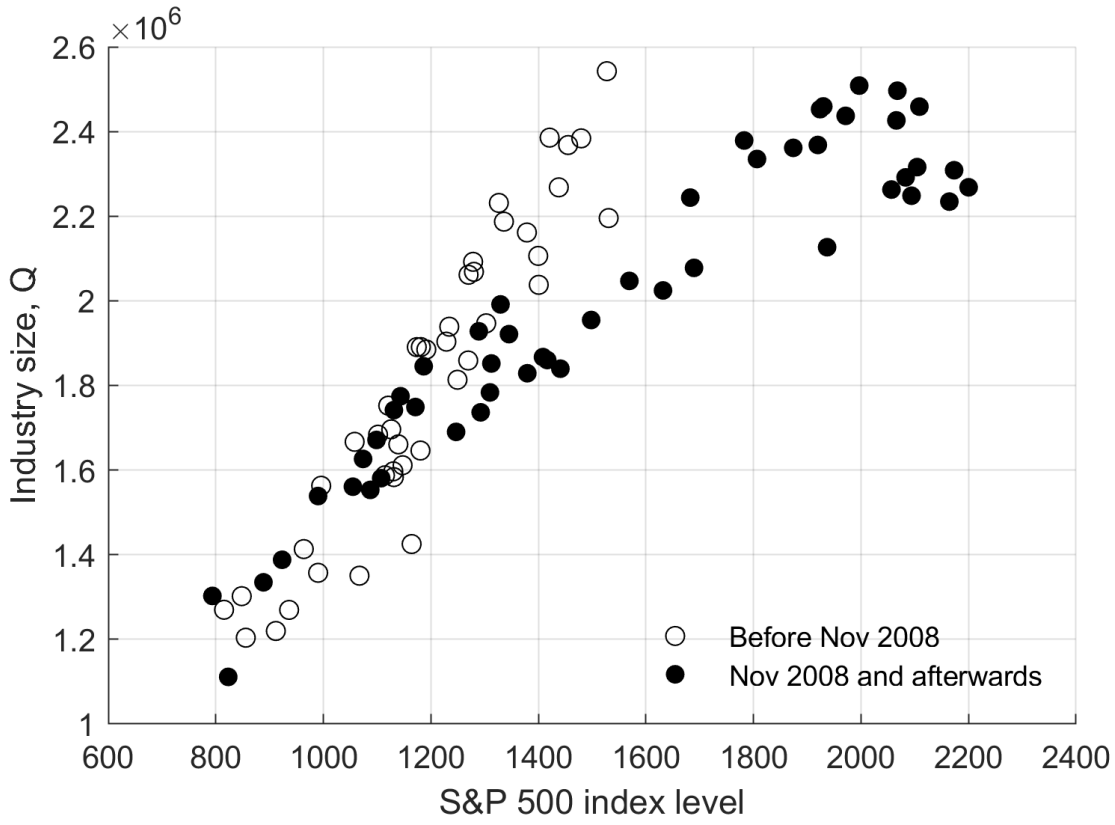


Table 2.2: Relationship between Q_t and M_t

	[1]	[2]	[3]
	ΔQ_t	Q_t	Q_t
Intercept	0.001	7.38 x 10^{5***}	-3.62 x 10^{5***}
	(0.007)	(7.36 x 10^4)	(1.22 x 10^5)
$\mathbf{1}_{Post2008}$			1.10 x 10^{6***}
			(1.38 x 10^5)
M_t		843.82***	1826.8***
		(51.35)	(100.94)
M_t $\mathbf{1}_{Post2008}$			-1022.1***
			(108.6)
ΔM_t	0.466***		
	(0.117)		
R^2	0.15	0.75	0.90
No. obs	90	91	91

Note: Figures in parentheses are standard errors. ***, **, * indicate different from 0 at 1%, 5% and 10% significance, respectively. Q_t is the size of the mutual fund industry, M_t is the SP500 index and $\mathbf{1}_{Post2008}$ is a dummy variable that is one after 2008. The dataset is from 2001 to 2016, at a frequency of 2 months.

Figure 2.3: The relationship between Q and S&P500



Note: Q_t is aggregate assets under management across funds in my sample in period t.

2.3 Model

The model consists of two parts: (1) a model of demand by rational investors for mutual funds and (2) a model of supply by mutual funds. I describe each part of the model, before considering the implications of the model for aggregate surplus, efficiency and the role of the business cycle.

2.3.1 Demand

The model of demand is based on Berk and Green (2004), in that it shares the following two core components. First, there are *decreasing returns to scale* in the ability of funds to earn

excess returns. Bigger funds, all else being equal, earn lower returns because their ability to gather and exploit information is diluted or because of price effects or execution costs. Second, *ability is unobserved*, but *investors learn* as they observe past returns. These two core components in combination mean that rational investors form beliefs about the ability of funds and invest up to the point where, given decreasing returns to scale, those returns are competed away. As investors observe past returns of the fund, they update their beliefs about the ability of the fund and adjust their holdings.

To these core components I add the following to suit my research question and to allow the model to be reasonably taken to data. First, following Pástor and Stambaugh (2012), I model *competition between mutual funds* by allowing the returns earned by funds to be decreasing in the total size of the mutual fund industry: a mutual fund earns lower excess returns, all other things being equal, if there are many other mutual funds trying to earn excess returns from the same set of investment opportunities. Second, I include *a role for the business cycle* by allowing the ability of funds to earn excess returns to vary according to a macro-economic factor that varies exogenously over time.

More formally, I follow Berk and Green (2004) and draw a distinction between the *net* excess return that investors actually earn, and the *gross* excess return the fund would have earned on a single dollar of investment (that is, before the effect of decreasing returns to scale). The total risk-adjusted payout in dollar terms to investors from investing q_{it} in mutual fund i with gross return α_{it}^g and fee rate f_i is:

$$TP_{it} = q_{it}\alpha_{it}^g - C(q_{it}) - q_{it}f_i$$

where $C(q_{it})$ is a cost function representing the decreasing returns to scale in the ability to earn excess returns. I parameterise the cost function as $C(q_{it}) = \phi_i q_{it}^2$ where $\phi_i > 0$, such that when $q > 0$: $C(q) > 0$, $C'(q) > 0$, $C''(q) > 0$, $C(0) = 0$ and $\lim_{q\to\infty} C(q) = \infty$. The *net* α_i^n excess return is what investors actually earn, and is simply this payout divided by the size of the investment:

$$\alpha_{it}^n = \frac{TP_{t+1}}{q_{it}} = \alpha_{it}^g - \frac{C(q_{it})}{q_{it}} - f_i = \alpha_{it}^g - \phi_i q_{it} - f_i \tag{2}$$

I disaggregate the fund's gross excess return into three components. First, the fund's true ability to generate excess returns α_i, where $\alpha_i \sim N(\mu_i, \tau_{i,\alpha}^{-1})$. Second, a fund-specific iid shock ϵ_{it}, where $\epsilon_{it} \sim N(0, \tau_{i,e}^{-1})$ and $\alpha_i \perp\!\!\!\perp \epsilon_{it}$. Third, common variation in ability across

funds δ_t:

$$\alpha_{it}^g = \alpha_i + \epsilon_{it} + \delta_t \tag{3}$$

I then disaggregate the common variation into a further three components: an age effect $\delta_{a(it)}$, the effect of macro-economic factor M_t and the effect of industry size Q_t:

$$\delta_t = \delta_{a(it)} + \beta M_t + \theta Q_t \tag{4}$$

I estimate unrestricted age effects $\delta_{a(it)}$, macro-effects β and industry-size-effects θ in my empirical analysis, as described below. A natural interpretation at this stage, however, is that $\beta > 0$ and $\theta < 0$. $\beta > 0$ implies that funds are more able to earn excess returns when the macro-economic factor is good. $\theta < 0$ represents competition, in that a larger mutual fund industry means more competition for the same investment opportunities, reducing excess returns.

Investors choose q_{it} before ϵ_{it} is realised. Furthermore, investors do not know the true ability of the fund α_i, but form expectations based on the information available to them at the point of investment, which I denote I_{t-1}. I define these expectations as $e_{it} \equiv \mathbb{E}[\alpha_i \mid I_{t-1}]$. All other components of the return are known to the investor, including ϕ_i and δ_t.

Investors supply capital with infinite elasticity to any fund with positive expected *net* returns α_{it}^n, taking aggregate investment q_t in the fund as given. In equilibrium, q_t is then such that $\mathbb{E}[\alpha_{it}^n \mid I_{t-1}] = 0$. Substituting in Equations 2 and 4, this means that:

$$q_{it} = \frac{e_{it} + \delta_t - f_i}{\phi_i} \tag{5}$$

Investor demand for mutual fund i is therefore increasing in its expected ability e_{it}, increasing in its scalability ϕ_i, decreasing in its fee rate f_i and subject to common variation δ_t. Note that for ease of reference I refer to e_{it} as "ability" and ϕ_i as "scalability", but in some sense both are fund-specific measures of the ability to generate excess returns on q_{it}.

To complete the model of demand, I need to characterise the expectations formation process behind e_{it}. Investors observe past net excess returns, $\alpha_{is<t}^n$ and from this can infer gross returns α_{is}^g. Investors cannot separately identify α_i from ϵ_{is}, but can extract a signal about α_i given their relative distributions.

Given these distributional assumptions, there are simple closed-form expressions for how investors form and update their posterior beliefs about α_i in responses to these signals.

Defining the signal-to-noise ratio $\lambda = \frac{\tau_{i,e}}{\tau_{i,\alpha}}$ and $s(\lambda, t) = 1 + (t-1)\lambda$:

$$q_{it} = \frac{1}{\phi_i}\left[\delta_t - f_i + \frac{\mu_i}{s(\lambda,t)} + \frac{\lambda}{s(\lambda,t)}\sum_{m=1}^{t-1}\alpha_{im}^{g}\right] \tag{6}$$

I leave implicit the lower bound of zero. I repeatedly substitute in Equation 2 to solve forward for optimal q_{it} in terms of *net* returns (which are observed by the econometrician), instead of *gross* returns (which are not directly observed by the econometrician):

$$q_{it} = \frac{1}{\phi_i}\left[\mu_i - f_i + \delta_t + \lambda\sum_{m=1}^{t-1}\frac{\alpha_{im}^{n} - f_i}{s(\lambda, m+1)}\right] + u_{it}^{q} \tag{7}$$

I add an error term, u_{it}^{q}, that represents shocks to q_{it} beyond this expectations formation process. This could include, for example, noise traders. I leave further discussion of this error term and its distribution to the section below on my empirical analysis. This Equation 7 characterises equilibrium investor demand for fund i. In what follows I define the "observable type" of mutual fund i as $\mathbf{\Theta_i} = (\mu_i, \phi_i, \sigma_i^a, \sigma_i^e, f_i)$ and its "unobservable type" as α_i.

2.3.2 Supply

On the supply-side, firms make three decisions: (1) they choose to enter or not to enter, (2) they set a single fee at the start of their life and (3) they choose to exit or not to exit. Before modelling these three choices, I describe firm beliefs about the evolution of industry size, which will be important for each choice.

2.3.2.1 Firm beliefs about industry size

The payoff to a mutual fund depends on macroeconomic factor M_t and competition through the size of the mutual fund industry Q_t, as I set out in equation 4. I set out in Figure 2.3 and in Table 2.2 how closely M_t and Q_t co-move, with a R^2 value of 0.75 in a linear regression of Q_t on M_t.

The key assumption on the supply-side is that funds take aggregate industry size Q_t as given, and form beliefs about its dynamics based on its co-movement with M_t:

$$Q_t = g(M_t) = 738 + 0.844\ M_t \tag{8}$$

I argue that this assumption is reasonable given that there are a very large number of funds, the vast majority of which are a very small proportion of total Q_t. There are admittedly a small number of very large mutual funds for which this assumption may not be reasonable: these, however, are mostly established, older funds that are very unlikely to exit. That is, this is a reasonable assumption to make when studying, as I am, the entry and exit of mutual funds.

I assume that M_t develops according to some exogenous Markov process. This means that Q_t does too, such that firms have expectations about how industry size will develop over time, regardless of their own decisions or those of their competitors. This assumption has obvious computational benefits: the modelling environment is not a game, but a series of individual decisions by each mutual fund. The remaining complication, which I consider below, is ensuring that the individual decisions result in aggregate dynamics that are consistent with the firm beliefs set out in Equation 8.

2.3.2.2 Exit

In each period, a mutual fund can choose to exit and obtain a scrap value, which is intended to capture the use of its human capital elsewhere. This decision is dynamic, and depends on (1) the type of the mutual fund and (2) the state, including investor beliefs about the mutual fund e_{it}, the age of the mutual fund a_{it} and the macro-economic factor M_t:

- Type: $\mathbf{\Theta_i} = (\mu_i, \phi_i, \sigma_i^a, \sigma_i^e, f_i)$
- State: $\mathbf{S_t} = (e_{it}, a_{it}, M_t)$
- Action: $z_{it} = 0$ if exit, $z_{it} = 1$ if do not exit.

Firms take expectations over (1) the development of beliefs about their ability e_{it} and (2) changes in M_t. Their age and the precision of investor beliefs about their ability update deterministically. As is standard in the literature on exit (see for example Hotz and Miller (1993)) funds receive an action-specific shock $\eta(z_{it})$ that is distributed Type-1 extreme value. In recursive Bellman form:

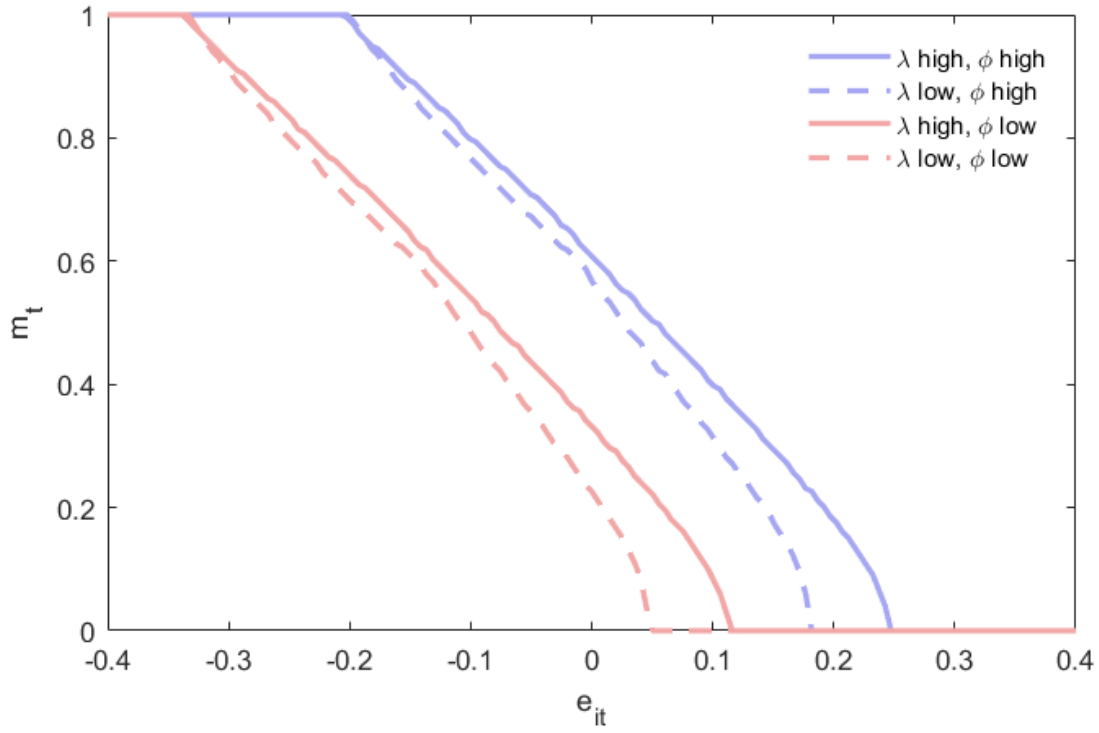
$$V_{it}(\mathbf{S_t}; \mathbf{\Theta_i}) = \max_{z_{it}} \; z_{it} f_i q_{it}(\mathbf{S_t}; \mathbf{\Theta_i}) + (1 - z_{it}) W(\mathbf{S_t}; \mathbf{\Theta_i}) + \eta(z_{it}) + z_{it} \beta \mathbb{E}[V_{it+1}(\mathbf{S_{t+1}}; \mathbf{\Theta_i})] \quad (9)$$

I allow the scrap value $W(\mathbf{S_t}; \mathbf{\Theta_i})$ to be state- and type-specific. This assumption has an

intuitive justification: mutual funds are more likely to have good outside options if they are good funds and/ or if the macro-economic state is good.

To illustrate the equilibrium exit decisions resulting from this model, I solve for $z^*(\mathbf{S_t}; \mathbf{\Theta_i})$ under various combinations of ability beliefs e_{it}, macroeconomic states m_t and parameter values in Figure 2.4 below, ignoring the action-specific shock $\eta(z_{it})$. These numerical results indicate a cutoff rule: funds exit when they are perceived to be bad or when the macroeconomic state is bad, or some convex combination thereof.

Figure 2.4: Exit decisions



Note: The area under the curve shows the combinations of ability belief (e_{it}) and business cycle state (m_t) in which a fund exits: funds exit when they are perceived to be bad or when the macroeconomic state is bad, or some convex combination thereof. Funds are less likely to exit when returns are less informative (λ is low) and/or when their ability scales up easily (ϕ is low).

2.3.2.3 Entry

Firms decide to enter without knowing their observable type $\mathbf{\Theta_i}$. Once they enter, they are randomly allocated a type (excluding the fee rate f_i, which they choose) from some distribution $h_{\mathbf{\Theta_i}}$. Firms choose to enter if the expected value of entry, taking expectations over $\mathbf{\Theta_i}$, exceeds an entry cost that is constant across firms but can vary over time:

$$\int V_{it}(\mathbf{S_t}; \mathbf{\Theta_i})\, d\Theta_i \geq F_t^{entry} \tag{10}$$

After deciding to enter, firms learn their observable type $\mathbf{\Theta_i}$. Firms then choose f_i to maximise their initial value, given their observable type $\mathbf{\Theta_i}$ and the prevailing state M_t:

$$f_i^* = \arg \max_{f_i} V((\mathbf{S_t}; \mathbf{\Theta_i})) \tag{11}$$

This generates cross-sectional variation in fee rates through the random allocation of types: funds that are given a better random type charge a higher fee. This also generates inter-temporal variation in fee rates through variation in the macro-economic factor M_t: funds that enter in good times charge a higher fee.

2.3.3 Equilibrium

In equilibrium, (1) investors invest in any fund with positive expected excess returns, as per Equation 7; (2) mutual funds choose to enter, set fees and exit optimally, given their type, the state, investor behaviour and their beliefs about future competition, as per Equations 9, 10 and 11; and (3) mutual fund beliefs about the dynamics of future competition are consistent with optimal mutual fund behaviour.

Expanding on the third of these equilibrium requirements: the entry and exit rules, conditional on $g(.)$, induce dynamics in Q_t, which we call $h(., g(.))$. Equilibrium is a fixed point such that $g^*(.) = h(., g^*(.))$. In other words, in equilibrium the entry and exit rules induce fund behaviour that is consistent with the overall dynamics in Q_t.

I do not solve for this equilibrium function. Instead, in the empirical analysis below, I observe and estimate this equilibrium function and hold it constant in the counterfactuals I run. This clearly places restrictions on the counterfactuals in which this equilibrium function could plausibly be held constant, which I discuss below.

2.3.4 Aggregate surplus

I follow Berk and Van Binsbergen (2015) in defining the surplus (or value-added, in the words of Berk and Van Binsbergen (2015)) generated by a given fund i as the dollar return to fund and investors:

$$s_{it} = f_i q_{it} + \alpha_{it}^n q_{it} \tag{12}$$

Aggregate surplus is then the sum of individual fund surplus: $AS_t = \sum_i s_{it}$. s_{it} depends on unknown true ability α_i: taking expectations gives $\mathbb{E}[s_{it} \mid I_{t-1}] = f_i q_{it}$. That is, in expectation mutual funds are the only ones to receive positive payoff because investors compete away their payoff. The model I set out above has two important implications for how s_{it} varies across funds.

First, the surplus generated by a given fund i is increasing and convex in true unknown ability α_i, as set out in Figure 2.5. The convexity comes from the fact that both equilibrium fee rate f_i^* and mutual fund size q_{it} are increasing in α_i. Intuitively, the market power of fund i is increasing in α_i, and thus so is surplus because investor payoff is competed away in any case.

Second, conditional on true α_i, the surplus generated by a given fund is typically increasing in the precision of investor beliefs. The general intuition for this is straightforward: investor beliefs are correct in expectation, but in particular realisations investors can think a particular fund is good when it is bad, and vice versa. This uncertainty results in mis-allocation (investing too much (little) in bad (good) funds) that harms surplus. More formally, substitute Equations 2 and 5 into Equation 12 for surplus and assume for ease of exposition that $f_i = 0$:

$$s_{it} = \frac{1}{\phi_i}(\alpha_{it}^g - e_{it})e_{it} \tag{13}$$

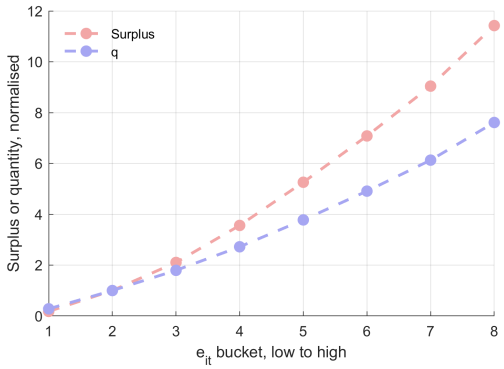
Let $e_{it} = \alpha_i + \epsilon_{it}^e$, where ϵ_{it}^e denotes the error in investor beliefs for fund i. It follows that s_{it} is decreasing in this error: mis-allocation is harmful to surplus.

The primary determinant of the precision of investor beliefs is the age of the mutual fund. Older mutual funds have a returns history that is a signal of their ability, and so allows investors to form more precise beliefs. As the age of a fund goes to infinity, the error term ϵ_{it}^e goes to zero. In Figure 2.5, I set out an example of how the surplus of a given fund is typically increasing in its age.

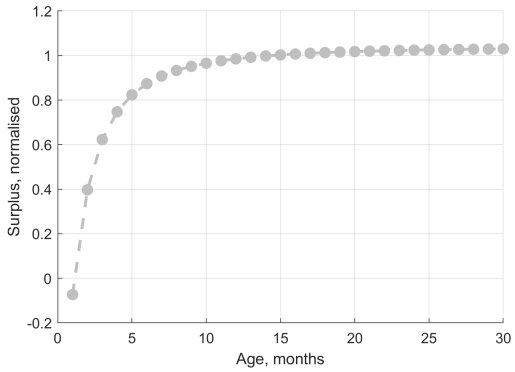
I caveat above that additional information as a fund ages *typically* increases surplus.

Figure 2.5: The effect of age and ability on surplus

(a) The effect of ability on surplus



(b) The effect of age on surplus



Note: Panel (a) shows how expected surplus varies with investor beliefs about ability e_{it}: it is increasing and convex. The size of the mutual fund q_{it} is also increasing in ability e_{it} and convex, but to a lesser degree. Panel (b) shows how surplus is typically increasing in age, because as funds age investor beliefs become more precise as they observe returns.

Whether this is always the case depends on the age of the fund and the fee rate f_i: when f_i is set too low or too high relative to the fund's true ability, then this introduces a distortion. This distortion can interact with the effect of aging in a way that means that, beyond a certain age, surplus is no longer increasing in age. I discuss this in more detail in Figure 2.16 in the appendix. For the purposes of my research question, it suffices to say that funds with no returns history have lower surplus than those that have a returns history, all other things being equal.

2.3.4.1 Efficiency

I consider the choices of a social planner without additional information: that is, the social planner does not know true fund ability α_i or have any more information than investors or funds.

There are various inefficiencies on the supply-side: mutual fund i's choices over entry, fee-setting and exit all affect Q_t and, through the effect of competition, the payoffs of other mutual funds. Mutual fund i does not account for any of these externalities.

My research question, however, is about how the business cycle affects the types of mutual fund in the industry. In other words, I am interested in the *composition* of the mutual fund industry, rather than its size. That is, I consider the second-best problem of optimising the

composition of mutual funds, whilst taking as given the aggregate size of the mutual fund industry Q_t and its dynamics.

I illustrate a *compositional inefficiency* by considering the mutual fund industry in equilibrium, in which no incumbent fund wishes to exit and and no potential entrant wishes to enter, given the prevailing macro-economic state M_t and the size of the industry Q_t. I ask whether the social planner would be willing to swap some number n incumbent funds of a particular type and size q^{exiter} for a single randomly drawn new entrant of expected size $q^{entrant}$. To focus on composition only, n, the size of the exiting fund and the expected size of the new entrant must be such that the overall mutual fund industry size Q_t does not change. That is, $nq^{exiter} = q^{entrant}$.

To see that such a compositional inefficiency is feasible in practice, suppose that the social planner chooses exiting funds of the worst expected ability type, and replaces them with a new entrant of average expected ability. Both the surplus generated by a given fund and its size are increasing and convex in true ability, as described above in Figure 2.5. Importantly, surplus is significantly more convex than fund size, meaning that $s^{entrant} > n\ s^{exiter}$: the new entrant generates greater aggregate surplus than the exiting funds. This surplus-improving swap need not occur in the decentralised equilibrium because the incumbent funds have no incentive to exit so that the better fund can enter.

I consider the social planner's preferences to illustrate the inefficiencies in the model. I do not, however, formally model the social planner's choices over composition, but instead consider the social planner's preferences over business cycles.

2.3.5 The role of the business cycle

I discuss above whether the social planner would be willing to swap some bad funds for an average entrant. This type of swap is exactly what results from a business cycle in my model.

Suppose the mutual fund industry is currently in equilibrium at time 0 at (Q_0, M_0). Potential new entrants are indifferent between entering or not. At the start of period 1, there is a shock to the macro-economic factor, in that $M_1 < M_0$. This downward shock causes existing funds to shrink (as per Equation 7) and causes some funds to exit (as per the exit rule set out in Figure 2.4), such that $Q_1 < Q_0$. At the start of period 2, there is a recovery and $M_2 = M_0$. The existing funds increase in size and new entrants face an incentive to enter, up to the point that $Q_2 = Q_0$.

I emphasise, though, that a business cycle is not necessary for exit and replacement to occur in my model. Even without any change in macro-economic factor M_t funds would still exit when they get a negative shock to their expected ability e_{it} or when they draw a negative shock to their profits η_{it}. In that sense, a business cycle accelerates firm turnover, but is not necessary for firm turnover.

The effect of this firm turnover on aggregate surplus depends on two countervailing effects:

1. **The cleansing effect**. Exiting funds are more likely to be low expected ability, as illustrated in Figure 2.4. Entrants are randomly allocated true ability from the population distribution and so are, on average, better than exiting funds. Higher ability funds result in more surplus, as set out in Figure 2.5, giving rise to a cleansing effect that increases surplus. The strength of this effect depends on the size of the ability differential between exiters and entrants, which in turn depends on: (i) the dispersion of the distribution in abilities and (ii) the extent to which exit rates are greater for low ability funds than for high ability funds.

2. **The information loss effect**. Exiting funds have a returns history, whereas entrants do not: it is in this sense that the business cycle results in information loss. Investors therefore have more precise beliefs about the ability of the exiting fund, which holding all other things equal results in greater surplus, as set out in Figure 2.5. The strength of this effect depends on the value of the information contained in past returns, as measured by the signal-to-noise ratio λ.

Cleansing, then, is about the first moment in ability (entrants are higher ability on average), whereas information loss is about the second moment (there is significantly greater uncertainty about the ability of entrants).

As I set out above, the strength of each of the effects depends on the parameters of the model: if, for example, returns are not particularly informative about ability, fund abilities are highly dispersed and low ability funds are significantly more likely to exit, then the cleansing effect is more likely to dominate the information loss effect. The model, therefore, cannot provide a general answer about the effect of the business cycle on outcomes: it is an empirical question.

2.4 Empirical approach

There are three aspects to my empirical approach: (1) I estimate some exogenous processes that are outside the model, (2) I calibrate some parameters and (3) I estimate the remaining parameters by matching observed quantities, entry and exit decisions. I discuss each of these in turn, before considering identification.

2.4.1 Exogenous processes

I model two exogenous processes outside the model. The first is the dynamics of macro-economic factor M_t, which in my empirical analysis is the S&P500 index. I assume that the index follows a random walk, with an iid error term:

$$M_t = M_{t-1} + e_t^M \tag{14}$$

where $e_t^M \sim N(0, \sigma^M)$. I recover an estimate of σ^M from the time-series of M_t. In my simulations I impose an upper and lower bound on M_t, $\overline{M}$ and $\underline{M}$, respectively.

The second exogenous process is the relationship between Q_t and M_t. I use the results set out in column 3 of Table 2.2.

2.4.2 Calibration

On the supply-side, I set the discount factor to 0.99. On the demand-side, all of the parameters in Equation 7 are separately identifiable, including ϕ_i and μ_i. In practice, to keep the number of parameters to be estimated down, I calibrate ϕ_i and μ_i based on how q_{it} evolves over time.

I set ϕ_i to be the inverse of the maximum size that fund i reaches in my sample: $\phi = \frac{1}{q_{i,max}}$, where $q_{i,max} = \max_t q_{it}$. This is effectively a fund-specific normalisation such that the product $q_{it}\phi_i \in [0, 1]$ for any i. This means that I do not use the cross-sectional variation in the size of the funds to identify the other parameters, but only the variation over time. In other words, I assume that Vanguard's largest funds are not large relative to other funds because they earned very large returns early in their life, they are large for fund-specific reasons that I effectively encode and leave fixed in ϕ_i.

I infer μ_i from the size of fund i in the first period of its life. Setting $t = 1$ in Equation

7 and re-arranging: $\mu_i = q_{i1} - \delta_{i1}$. This results in computational benefits, relative to simply estimating μ_i as a fixed effect, as it can be done outside of the main estimation loop. It also better matches the interpretation of μ_i as an initial prior belief about fund ability at the start of its life.

Implementing these calibrations in Equation 7 for demand, re-arranging and defining the within-style transformation $\tilde{\delta}_t = \delta_t - \delta_{i1}$:

$$\frac{q_{it} - q_{i1}}{q_{i,max}} = \tilde{\delta}_{it} + \lambda \sum_{m=1}^{t-1} \frac{\alpha_{im}^n - f_i}{s(\lambda, m+1)} + u_{it}^q \tag{15}$$

2.4.3 Estimation

I estimate the demand-side and the supply-side separately for tractability. From the supply-side I need to estimate the entry cost F_t^{entry} and the scrap values $W(\mathbf{S_t}; \mathbf{\Theta_i})$. I estimate all remaining parameters from the demand side.

On the demand-side, I run non-linear least squares on Equation 15, where the only non-linear parameter is λ. Given estimates of the parameters in Equation 15, it is then straightforward to infer estimates of α_{it}^g from Equation 2, and from that $\tau_{i,e}^{-1} = var(\alpha_{it}^g)$ and $\tau_{i,\alpha}^{-1} = \lambda \tau_{i,e}^{-1}$.

On the supply-side, I undertake a nested-fixed point estimation in which I match observed probabilities of exit with model-implied probabilities. I discretise the state-space into 8 buckets for e_{it}, 8 buckets for M_t and 4 buckets for the fund's age. I do this for 3 types of ϕ_i, meaning I have a total of 768 state-type combinations. The estimates of the demand-side and the first-stage estimates relating to the evolution of M_t allow me to model transition probabilities between each of these buckets. I show in Figure 2.14 in the appendix the exit rules implied by this coarser state space: it matches the key characteristics of the exit rules implied by the finer state space in Figure 9. For each state-type bucket, I calculate the observed exit probabilities over 8 years between 2005 and 2012, $\hat{Pr}(z = 1 \mid \mathbf{S_t}; \mathbf{\Theta_i})$.

To calculate model-implied probabilities, I first set out the following mean choice-specific utilities, averaging across funds in the same state-type buckets:

$$v_t(z = 1, \mathbf{S_t}; \mathbf{\Theta_i}) = f_i q_{it}(\mathbf{S_t}; \mathbf{\Theta_i}) + \beta \mathbb{E}[V_{it+1}(\mathbf{S_{t+1}}; \mathbf{\Theta_i})]$$
$$v_t(z = 0, \mathbf{S_t}; \mathbf{\Theta_i}) = W(\mathbf{S_t}; \mathbf{\Theta_i})$$

Given the assumed distribution of $\eta(z_{it})$, the probability of exit is then a function of the scrap values:

$$Pr(z = 1 \mid \mathbf{S_t}; \mathbf{\Theta_i}) = \frac{exp(W(\mathbf{S_t}; \mathbf{\Theta_i}))}{exp(v_t(\mathbf{S_t}; \mathbf{\Theta_i})) + exp(W(\mathbf{S_t}; \mathbf{\Theta_i}))}$$

I then choose $\hat{W}(\mathbf{S_t}; \mathbf{\Theta_i})$ to minimise the difference between observed $\hat{Pr}(z = 1 \mid \mathbf{S_t}; \mathbf{\Theta_i})$ and model-implied $Pr(z = 1 \mid \mathbf{S_t}; \mathbf{\Theta_i})$, solving the model for each iteration. This nested fixed point iteration is more efficient than the methodology proposed by Hotz and Miller (1993) that avoids solving the model. With state-type-specific scrap values $W(\mathbf{S_t}; \mathbf{\Theta_i})$ the number of unknowns and the number of observations is the same. The scrap values are just identified, and fit the observed exit probabilities exactly.

As well as estimating individual scrap values for each state-type combination, I estimate two more parsimonious variants. First, I set the scrap value to be the same for all state-type combinations: $W(\mathbf{S_t}; \mathbf{\Theta_i}) = W$. Second, I parameterise $W(\mathbf{S_t}; \mathbf{\Theta_i})$ as a function of states and types: that is, as a function of e_{it}, ϕ_i, age_{it} and M_t. I describe the exact parameterisation of this function in the results section below, chosen to imitate my results for state-type specific $\hat{W}(\mathbf{S_t}; \mathbf{\Theta_i})$.

To estimate F_t^{entry}, I sample observed funds randomly to take expectations over $\mathbf{\Theta_i}$, and use the demand-side parameter estimates to calculate $\mathbb{E}[V((\mathbf{S_t}; \mathbf{\Theta_i}))] = \hat{F}_t^{entry}$ in equation 10. I do this for each year: in other words, F_t^{entry} is effectively a residual that ensures that entrants are indifferent between entering and staying out in any given year.

2.4.4 Identification

The primary challenge in identification is the role of unobserved shocks to mutual fund size. In the context of the model, the error term u_{it}^q represents investment in the fund that is unrelated to beliefs of investors about the ability of the fund: noise traders, in other words. Correlations in noise trading across funds and across time create challenges in identification in two ways.

First, Q_t is endogenous in the presence of unobserved shocks that are common across funds. If, for example, the mutual fund industry is popular with noise traders in time t, then both q and Q would be large: this would bias our estimate of the effect of Q on q away from zero. Second, α_{it-1}^n is a function of q_{it-1} and so of u_{it-1}^q: this means that historical returns are endogenous in the presence of serially correlated unobserved noise trading. If,

for example, firm i is popular among noise traders for two consecutive periods, then returns are low and the fund is big: this would bias our estimate of the responsiveness of rational investors to past returns λ downwards.

We control for unobserved noise trading by controlling for the size of "similar" funds, where we define a similar fund as one that has similar β, as in Equation 1. If funds follow similar investment strategies, then it is likely that $\boldsymbol{\beta_i} \approx \boldsymbol{\beta_j}$. We define the following distance measure on $K \times 1$ vectors $\boldsymbol{\beta}$:

$$d_{ij} = || \boldsymbol{\beta}_i - \boldsymbol{\beta}_j ||$$

We define group $g(i)t$ as the 10 closest funds to i in terms of d_{ij} at time t (where time variation in the group comes from the composition of funds, not the constant distance measure d_{ij}). We then define $\overline{q}_{g(i)t}$ as the mean size within this group, and include this as a control within our estimation.

To demonstrate the role of this control more formally, I disaggregate u^q_{it}, the unobserved shocks to fund i, into three parts:

$$u^q_{it} = u_t + u_{g(i)t} + u_{it}$$

u_t is common to every mutual fund, $u_{g(i)t}$ is common to every fund in group $g(i)$ and u_{it} is idiosyncratic to fund i. $\overline{q}_{g(i)t}$ controls for u_t and $u_{g(i)t}$, such that the remaining identifying assumption is that (1) Q_t is independent of idiosyncratic, fund-specific shocks e_{it}, which requires that u_{it} and u_{jt} are not correlated, and (2) α^n_{it-1} is independent of u_{it}, which requires that u_{it} and u_{it-1} are not serially correlated. In other words, identification requires that the unobserved component of q_{it} is iid across i and t: including $\overline{q}_{g(i)t}$ as a control makes this assumption more reasonable, as it limits the unobserved component to fund-specific unobservables.

2.5 Results

I set out the results of my estimation in Tables 2.3 and 2.4 and Figures 2.6 to 2.9.

On the demand-side, the results have the following implications:

1. **The role of past returns**: I estimate λ to be 0.0193. This means that investors respond to returns, but relatively slowly. It implies, for example, that the investor's

priors about a fund are as important to the investor as 52 months of returns history.

2. **The role of competition**. The coefficient on Q_t is negative and significant, indicating that competition between mutual funds plays a role. Furthermore, this parameter estimate is sensitive to the inclusion of the control $\overline{q}_{g(i)t}$ in the way one would expect: failing to control for common shocks understates the importance of competition.

3. **The role of the business cycle**. The coefficient on M_t is positive and significant: funds are larger when the macro-economic factor is good. As well as this direct effect, M_t has an indirect effect on q_{it} via Q_t. The net effect of M_t, taking into account both the direct and indirect effect, is positive: when the macro-economic factor is good, Q_t is higher (which has a negative impact on q_{it} because of the impact of competition), but not to the extent that it dominates the direct effect.

On the supply-side, the key implications of the results are as follows:

1. **Variation in exit rates**. Based on the results from the demand-side, I allocate each fund to the state-type buckets described above, and calculate the exit rates in those buckets. As set out in Figure 2.6, I find that funds are more likely to exit when my model indicates that they are low expected ability (e_{it} is low) or do not scale up well (ϕ_i is high).

2. **Variation in scrap value**. I estimate state-type specific scrap values, and show their estimated distribution in Figure 2.7. In Figure 2.8 I show that these scrap values vary across types and states in an intuitive way. Scrap value co-moves closely with the expected ability of the fund e_{it} and with its scalability ϕ_i: funds have better outside options external to the mutual fund industry if they are higher ability and/ or are able to scale that ability up easily. Scrap value also co-moves weakly with the macro-economic factor in the way one would expect, in that outside options are slightly better when the macro-economic factor M_t is good.

   I emphasise two important benefits to estimating state-type specific scrap values. First, it allows me to more accurately model exit dynamics: it stands to reason that better funds have better outside options, and imposing a single homogeneous scrap value would miss this. Second, it helps ensure model consistency. In my model, funds assume a particular equilibrium relationship between aggregate Q_t and M_t (Equation 8) when they decide to exit. State-type-specific scrap values allows my model to perfectly match

observed exit rates, meaning that the behaviour of individual funds is consistent with this equilibrium relationship between Q_t and M_t.

3. **Variation in entry cost**. The expected value of entering is greater when the macro-economic factor M_t is good. Given that I assume that new entrants are always indifferent between entering or not, this means that the fixed cost of entry F_t^{entry} is also increasing in M_t, as I set out in Figure 2.9. The effect is limited as F_t^{entry} does not vary by more than 6%.

2.5.1 Alternative specifications

As well as estimating individual scrap values for each state-type space, I estimate two more parsimonious variants. First, I set the scrap value to be the same for all state-type combinations: $W(\mathbf{S_t}; \mathbf{\Theta_i}) = W$. I set out the results in Table 2.4, and show that this specification does not perform well: the R^2 is only 0.19, meaning that there is significant unexplained variation in observed exit rates.

Second, I parameterise $W(\mathbf{S_t}; \mathbf{\Theta_i})$ as a function of states and types. As described above and set out in Figure 2.8, when I estimate state-type specific scrap values I find that they are sensitive to expected ability e_{it} and scalability ϕ_i, but less sensitive to macro-economic factor M_t and age. I therefore parameterise scrap values as follows:

$$W^P(\mathbf{S_t}; \mathbf{\Theta_i}) = w_0 + w_1 e_{it} + w_2 e_{it}^2 + w_3 \phi_i + w_4 \phi_i^2 + w_5 e_{it} \phi_i + e^w \tag{16}$$

I choose parameters $w_0, w_1, w_2, w_3, w_4, w_5$ to minimise the distance between model-implied and observed exit probabilities. I set out the results of this parameterisation in Table 2.4 and show that it performs better than constant scrap values, in that it explains 44% of the variation in observed exit probabilities. I use this parameterised scrap value in my counterfactuals below instead of fully state-type specific scrap values for two reasons. First, it is less sensitive to noise in observed state-type exit rates. Second, it allows me to extend my analysis to *counterfactual* states and types that are not observed in the data.

Table 2.3: Demand-side results

	[1] q_{it}	[2] q_{it}	[3] q_{it}
λ	0.019***		0.019***
	(0.003)		(0.003)
M_t	2.03***	2.22***	2.29***
	(3.52 x 10^{-2})	(3.59 x 10^{-2})	(3.53 x 10^{-2})
Q_t	-3.08***	-6.57***	-6.82***
	(2.07 x 10^{-1})	(2.22 x 10^{-1})	(2.18 x 10^{-1})
$\overline{q}_{it}$		1.76***	1.72***
		(3.51 x 10^{-2})	(3.45 x 10^{-2})
μ_{it}	0.311	0.311	0.311
ϕ_{it}	0.037	0.037	0.037
Age FE	Y	Y	Y
R^2	0.69	0.75	0.77
No. obs	226111	226,111	226,111

Note: Figures in parentheses are standard errors. ***, **, * indicate different from 0 at 1%, 5% and 10% significance, respectively. q_{it} is the size of mutual fund i at time t, λ is sensitivity to past returns, Q_t is the size of the mutual fund industry, $\overline{q}_{it}$ is the mean size of local funds to fund i and M_t is the SP500 index. Specification [3] is my baseline specification, specifications [2] and [3] show the role of λ and $\overline{q}_{it}$, respectively. I calibrate fund-specific priors μ_{it} and scalability ϕ_i and report the mean across funds here.

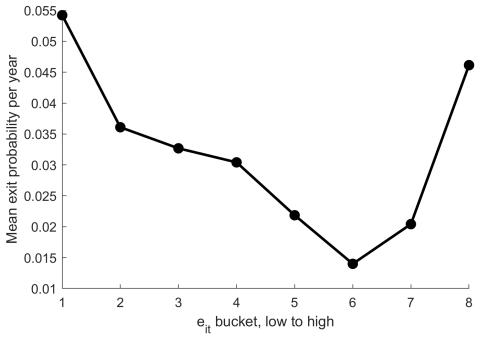
Table 2.4: Supply-side results

	[1] $\hat{Pr}(Exit \mid \mathbf{S_t}; \mathbf{\Theta_i})$	[2] $\hat{Pr}(Exit \mid \mathbf{S_t}; \mathbf{\Theta_i})$
Intercept	6.44	11.77
e_{it}		9.50
e_{it}^2		-457.33
ϕ_i		4.08
ϕ_i^2		-6.41
$\phi_{it} e_{it}$		-763.56
R^2	0.19	0.44
No. obs	768	768

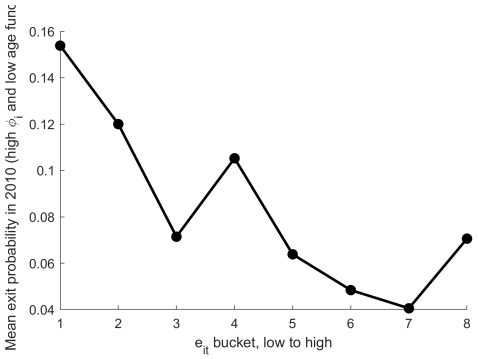
Note: I parameterise state-type scrap costs according to Equation 16 and choose the coefficients to fit the implied exit probabilities to observed exit probabilities.

Figure 2.6: Observed firm exit by state and type

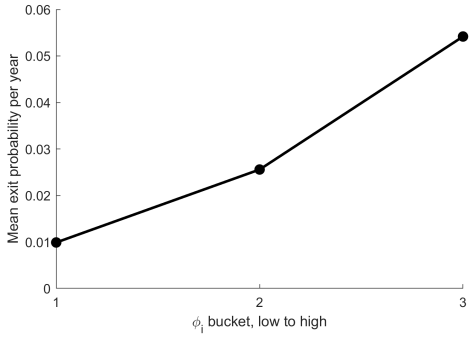
(a) Ability, e_{it}



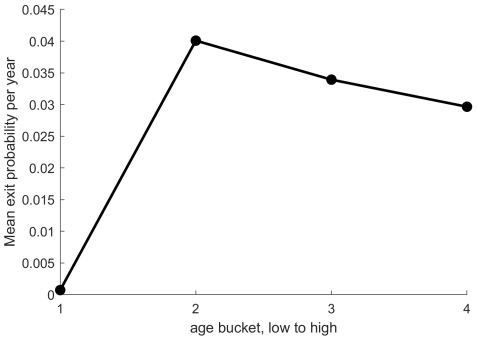
(b) Conditional ability, $e_{it} \mid \phi_i, M_t, age_{it}$



(c) Fund scalability, ϕ_i

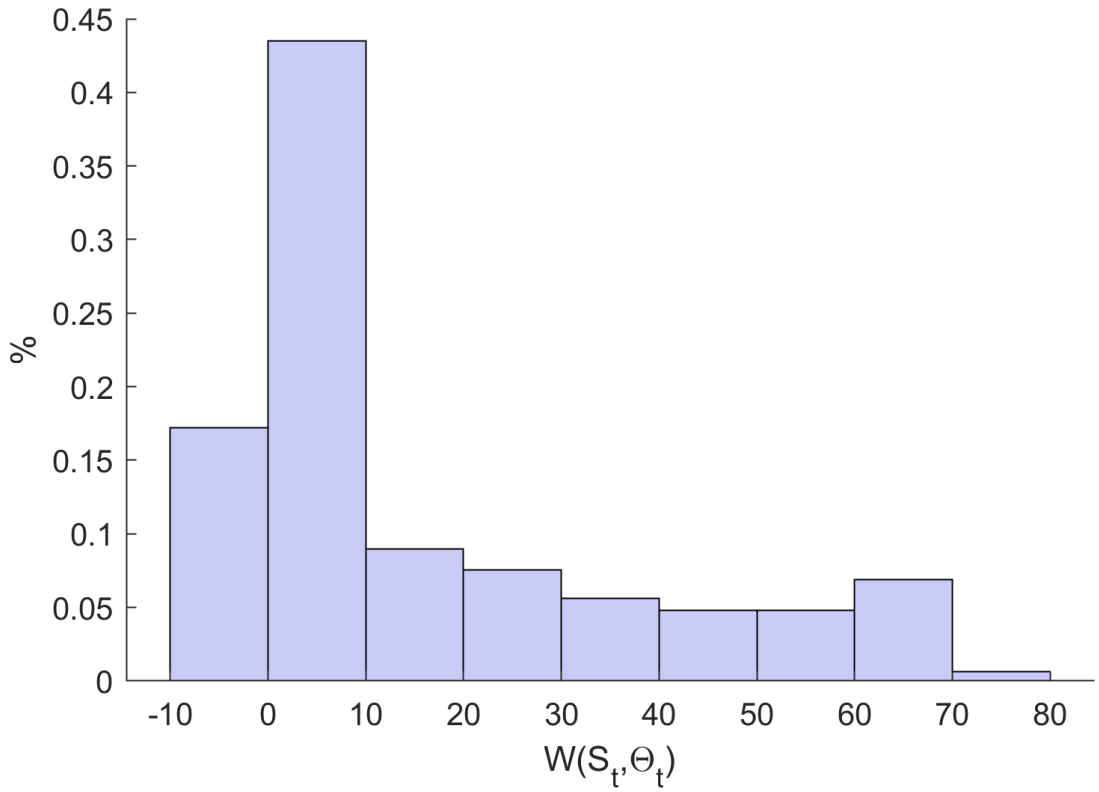


(d) Age, age_{it}



Note: I calculate the observed probability of exit in a year for each state-type bucket. In this figure I show how these observed probabilities vary on average with states and types. Note that the ability of a fund to scale up in size is decreasing in ϕ_i.

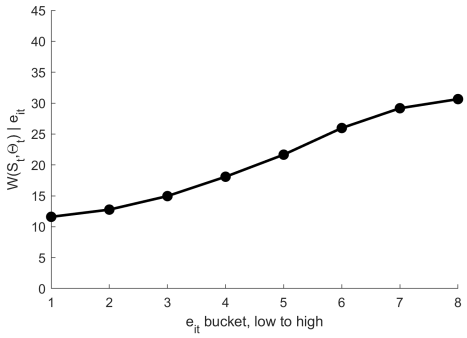
Figure 2.7: The distribution of state-type-specific scrap values



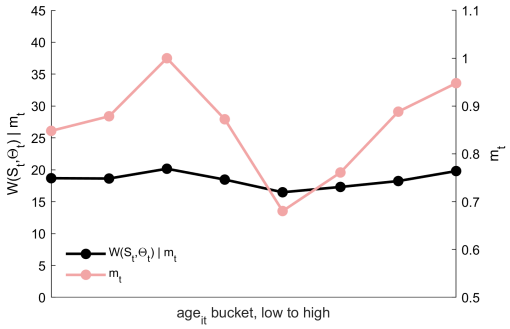
Note: In this figure I plot the distribution of estimated state-type-specific scrap values.

Figure 2.8: State-type-specific scrap values

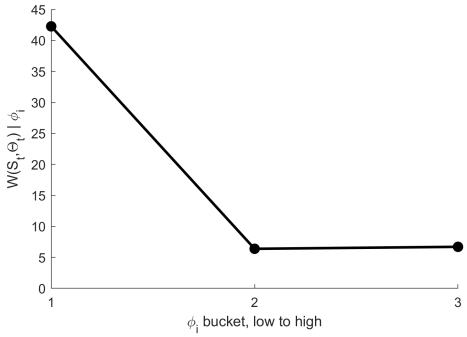
(a) Fund ability, e_{it}



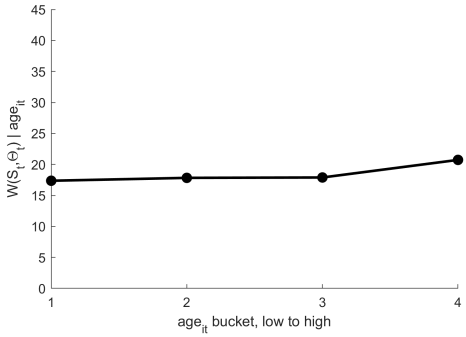
(b) Macro-economic factor, m_t



(c) Fund scalability, ϕ_i

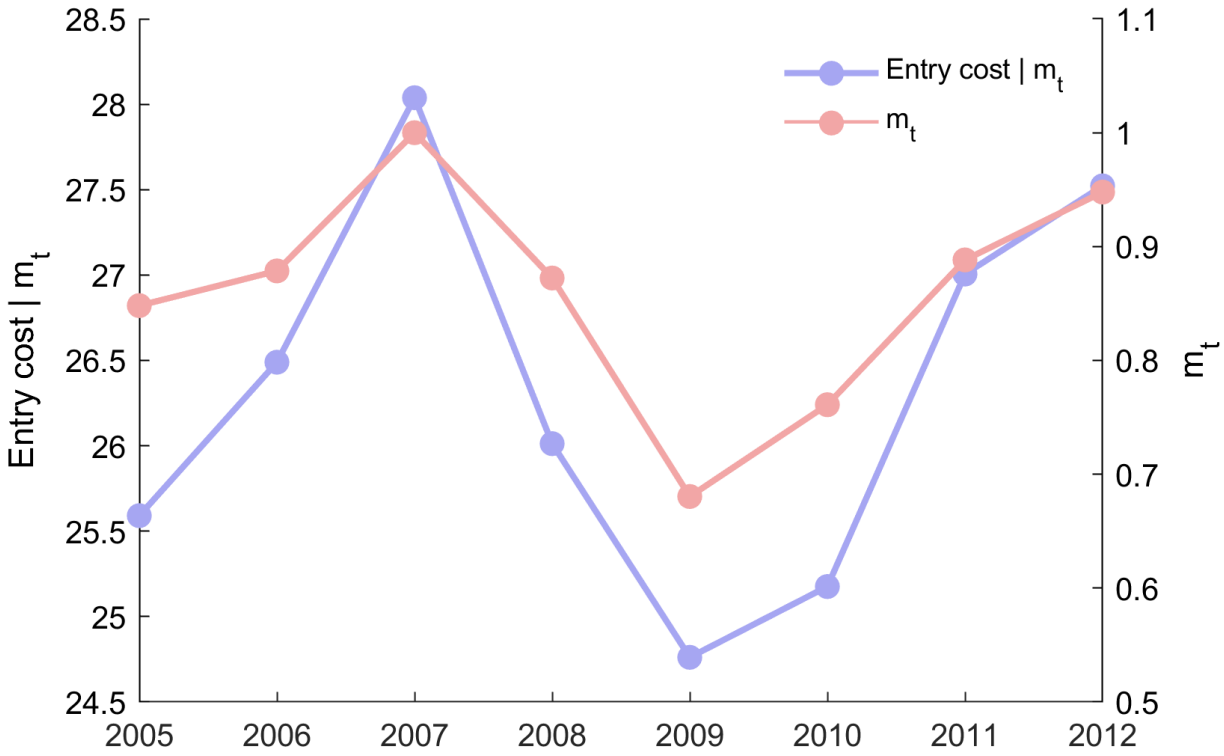


(d) Fund age, age_{it}



Note: I estimate scrap values for each state-type combination. In this figure I show how these scrap values vary according to the state and type of the fund. Funds have better scrap values when they have higher expected ability (panel (a)), when the macro-economic factor is good (panel (b)), when they scale well (panel (c)) and are relatively younger (panel (d)).

Figure 2.9: Variation in the fixed cost of entry over time



Note: In this figure I plot how the estimated fixed cost of entry F_t^{entry} varies over time (black line). The cost of entry is correlated with the macro-economic factor M_t (the red line), but the effect is relatively weak: the maximum entry cost is only 6% greater than the minimum entry cost.

2.6 Counterfactual analysis

I am interested in the effect of the *depth* of the business cycle on outcomes *post-recovery*. To assess this, I simulate a business cycle (that is, a recession, followed by a recovery) in the macro-economic factor M_t of varying depths, model the resulting counterfactual firm turnover, and then set out the effect on aggregate surplus.

Based on this counterfactual analysis, I draw two main conclusions:

1. **The business cycle harms surplus in the short-term and improves surplus in the long-term**.

2. **Deeper business cycles have bigger, persistent effects in the short-term and long-term**.

I set out the impact on firm turnover in Figure 2.10. The number of exiting firms, the number of entering firms and the ratio of entering firms to exiting firms are all increasing in the depth of the recession. The ratio responds in this way because the mean size of exiting funds is bigger in deeper recessions (in which medium-sized firms to exit as well as small firms).

In assessing the impact of this firm turnover, I hold the set of entering and exiting funds fixed: that is, I do not simulate further entry or exit, but only compare these two sets of funds. I compare the annual surplus of these two sets of funds immediately after the recovery, and then at various points in time subsequent to the recovery. As funds age post-recovery, they obtain a returns history and the precision of investor beliefs improves.

In Figure 2.11, I show the net effect of firm turnover on aggregate surplus per-period. It is initially negative, indicating that the information loss effect dominates the cleansing effect. Over time, as the funds age, the information loss effect decays, such that there is a "switching point" in month 27 when the effect of the firm turnover is reversed: the cleansing effect dominates the information loss effect, and per-period aggregate surplus is higher. Deeper business cycles have larger short-term and long-term effects, but the same switching point. In other words, the strength of the information loss effect and the strength of the cleansing effect are both increasing in the depth of the business cycle, but their *relative* strength is not.

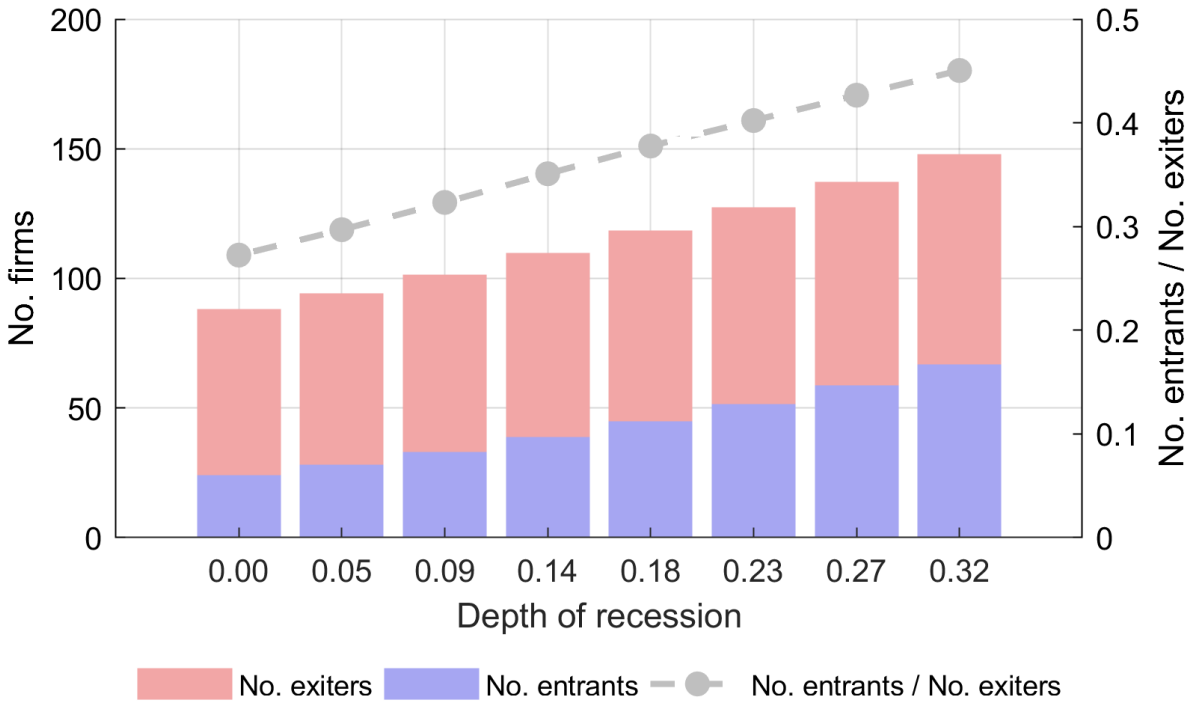
The magnitudes of both the short-term and long-term effects are material and are increasing in the depth of the business cycle, and are material. For the deepest business cycle I model (which is roughly equivalent to the financial crisis), the aggregate surplus of entering funds is 20% less than the aggregate surplus of the exiting funds in the first month after the recovery. By month 80, the information loss has decayed to the point where the aggregate surplus of entering funds is 30% greater than that of exiting funds. The effects on total surplus in the market (including funds that did not exit) are small but material, ranging between -0.5% and 0.9% of total mutual fund surplus for the deepest business cycles.

As well as these per-period effects, the business cycle also triggers one-off entry costs. Given these one-off costs, and the fact that the per-period effect of the business cycle starts negative and turns positive over time, it is natural to ask at what point the *cumulative* impact of the business cycle becomes positive.[2] I show this in Figure 2.12 and 2.13: the cumulative effect of the business cycle is negative and downward sloping until month 27. At this point it turns upwards, and becomes positive in month 75. As with the per-period effects, deeper business cycles have stronger persistent impacts but the same switching point.

In Figure 2.12 I also demonstrate the importance of allowing the exiting funds to age counterfactually absent the firm turnover. In other words, had the funds not exited they too would have extended their returns history and improved the precision of investor beliefs. Because the exiting funds are older, however, the marginal improvement in investor precision over time is much smaller than for the new entrant funds. An extra datapoint is more valuable for funds with few datapoints. In other words, the decay of the information loss effect over time is not about the change in the *absolute* precision of investor beliefs about entrants, but instead about the change in their precision *relative* to the precision about exiting funds.

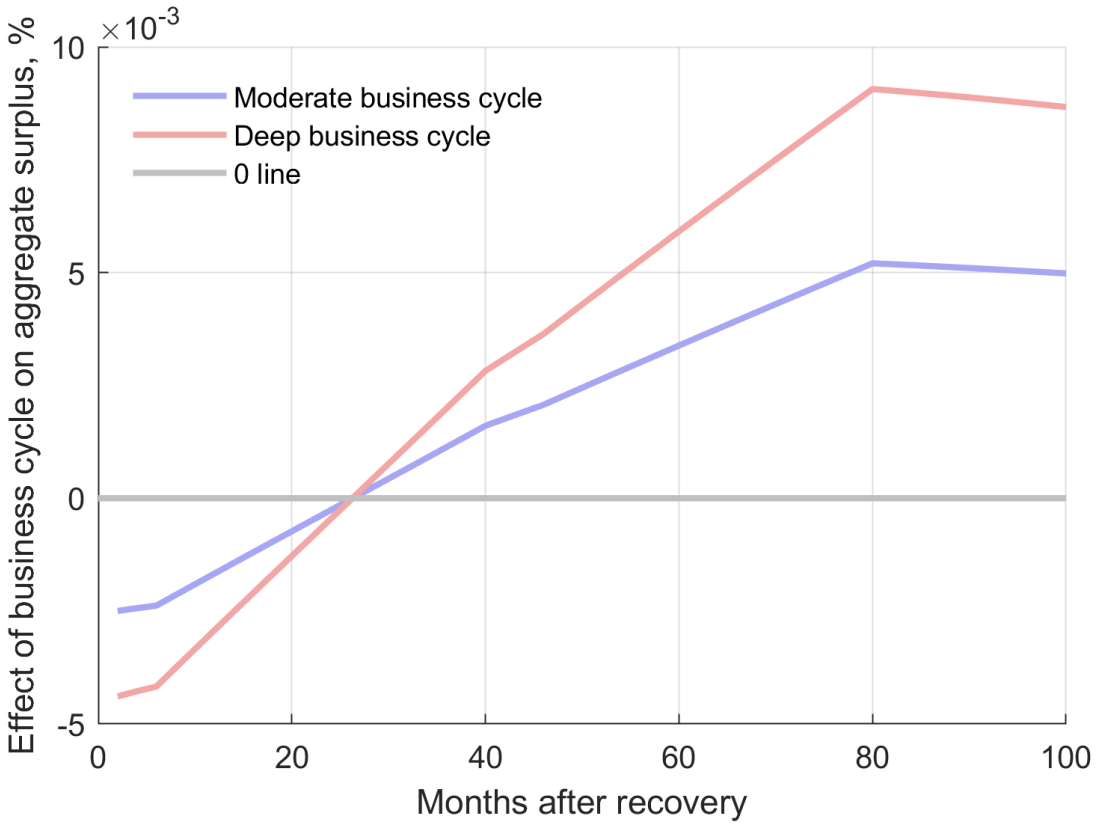
[2]The cumulative impact of the business cycle in period j is the sum of the impacts in all previous periods, including the fixed entry cost: $cum_t = \sum_{t=1}^{j} \left[\sum_{k=1}^{N^{entrants}} s_{kt}^{entrants} - \sum_{l=1}^{N^{exiters}} s_{lt}^{exiters} \right] - N^{entrants} F^{entry}$.

Figure 2.10: Firm turnover over the business cycle



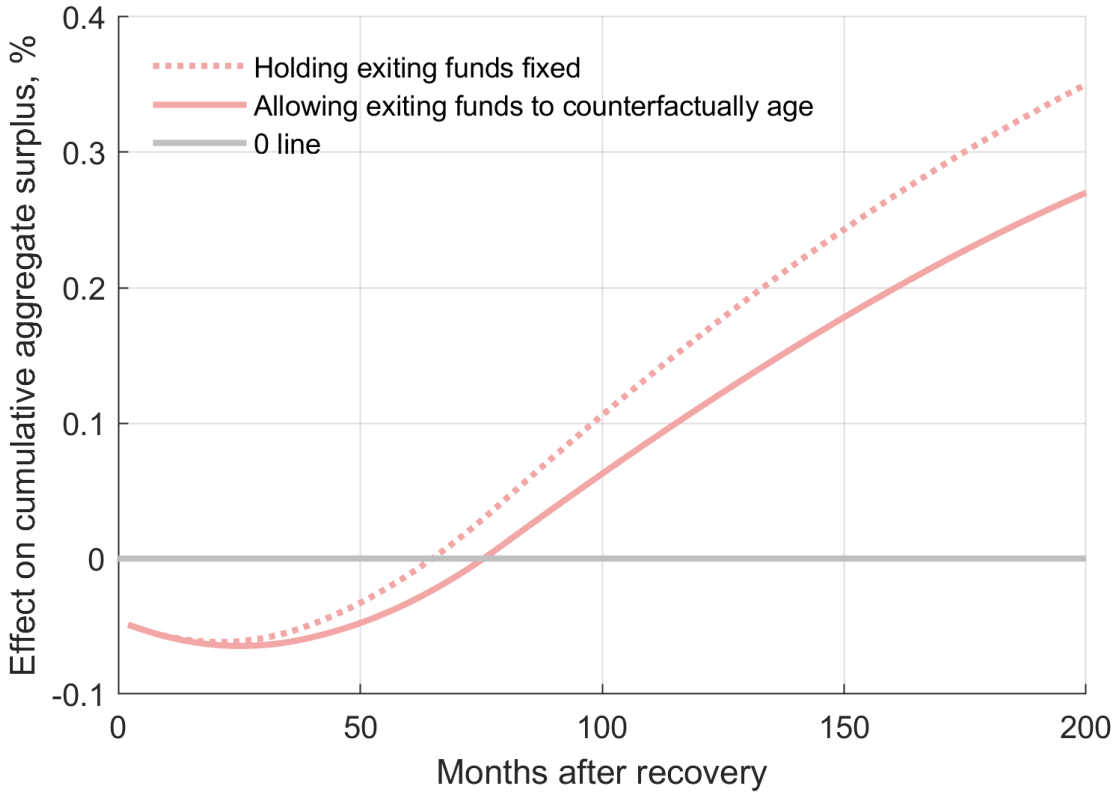
Note: I simulate a business cycle of various depths, and show the number of exiting funds during the recession (red bars), the number of entering funds during the recovery (blue bars) and the ratio of entrants to exiters (grey line). A deeper business cycle results in more firm turnover but also a larger marginal exiting firm, meaning that the ratio of entrants to exiters is greater than in a shallow business cycle.

Figure 2.11: The effect of the business cycle on aggregate surplus



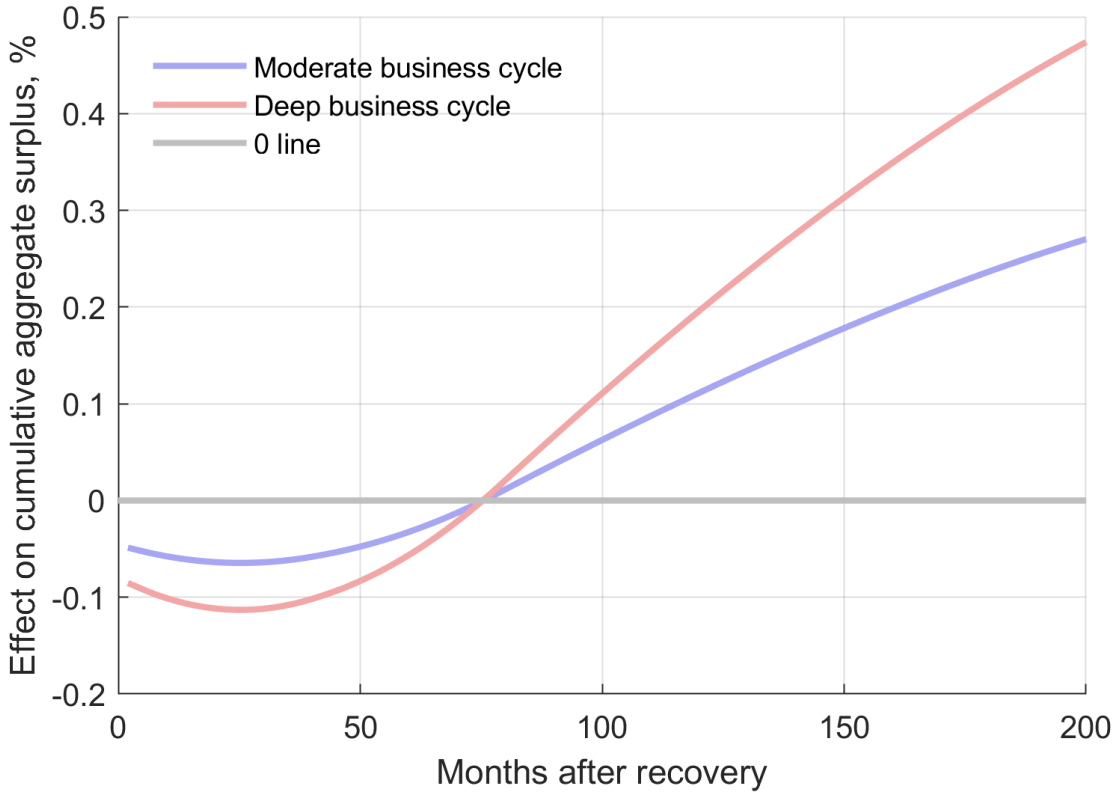
Note: Deeper recessions and subsequent recoveries result in greater firm turnover. In this figure I plot the net effect of this firm turnover on per-period aggregate surplus over time. Immediately after the recession, aggregate surplus in the mutual fund industry is up to 0.5% lower: the information loss dominates the cleansing effect. As the entrants age, investors obtain a returns history and the information loss effect decays: 27 months after the recovery the cleansing effect dominates the information loss effect and the firm turnover improves aggregate surplus. By 80 months the firm turnover improves aggregate surplus by up to 0.9%. The depth of the recession affects the magnitude of both the information loss effect and the cleansing effect, but not the point at which their net effect switches.

Figure 2.12: The effect of the business cycle on cumulative aggregate surplus



Note: As set in previous figure, firm turnover harms per-period aggregate surplus in the short-term and improves it in the long-term. In this figure I plot the effect of the firm turnover on *cumulative* aggregate surplus over time, expressed as a proportion of per-period aggregate surplus. In the first month after the recovery the impact of the firm turnover is negative because of the firm turnover costs and the role of the information loss effect. Over time the information loss effect decays and the per-period effect becomes positive in month 27 (when this graph turns upwards) and the cumulative effect becomes positive in month 75 (when it crosses zero). Holding the exiting funds fixed (the dashed line) results in the net effect becoming positive faster than if the exiting funds are allowed to counterfactually age absent the firm turnover (the solid line).

Figure 2.13: The effect of the business cycle on cumulative aggregate surplus



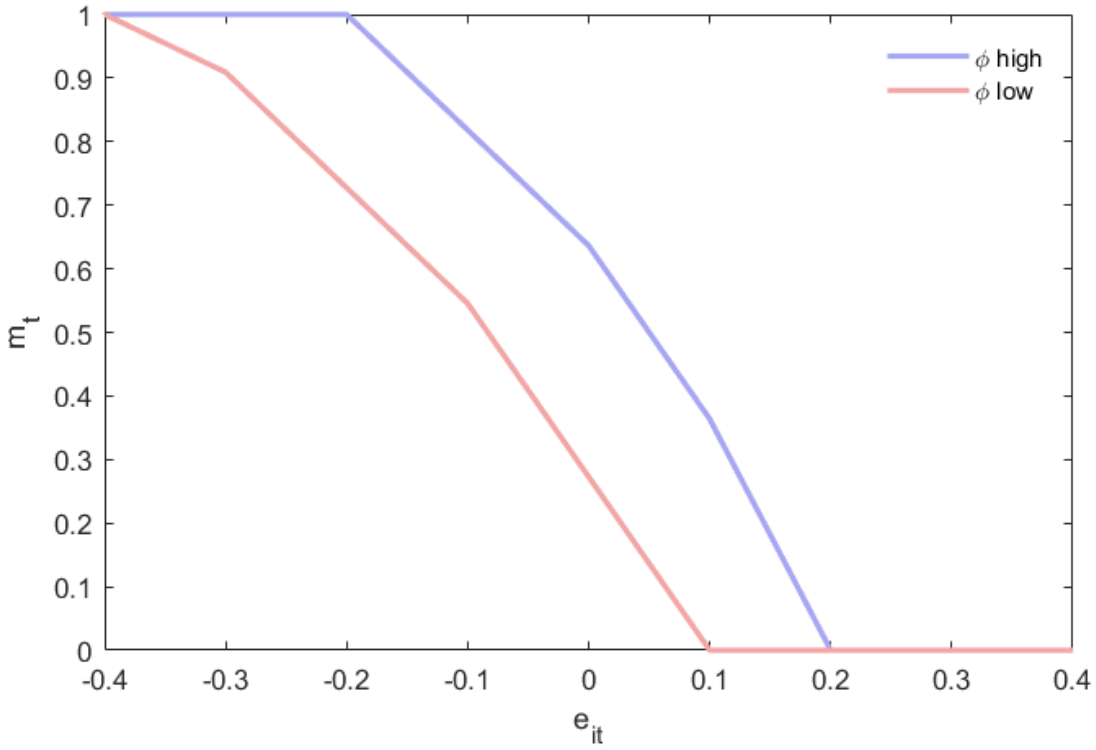
Note: In this figure I show how the depth of the business cycle affects the cumulative impact of firm turnover over time. Deeper business cycles (the red line) have stronger short-term and long-term impacts than moderate business cycles (the blue line), but the point at which the net effect becomes positive does not depend on the depth of the cycle.

2.7 Conclusion

The persistent effects of the business cycle have been extensively studied in macroeconomic contexts, but less so in market-specific contexts. The main contribution of this paper is to develop an under-explored implication of business cycles: the information loss that results from firm turnover. I explore the conditions under which this information loss dominates the cleansing effect, and I quantify how this trade-off changes over time.

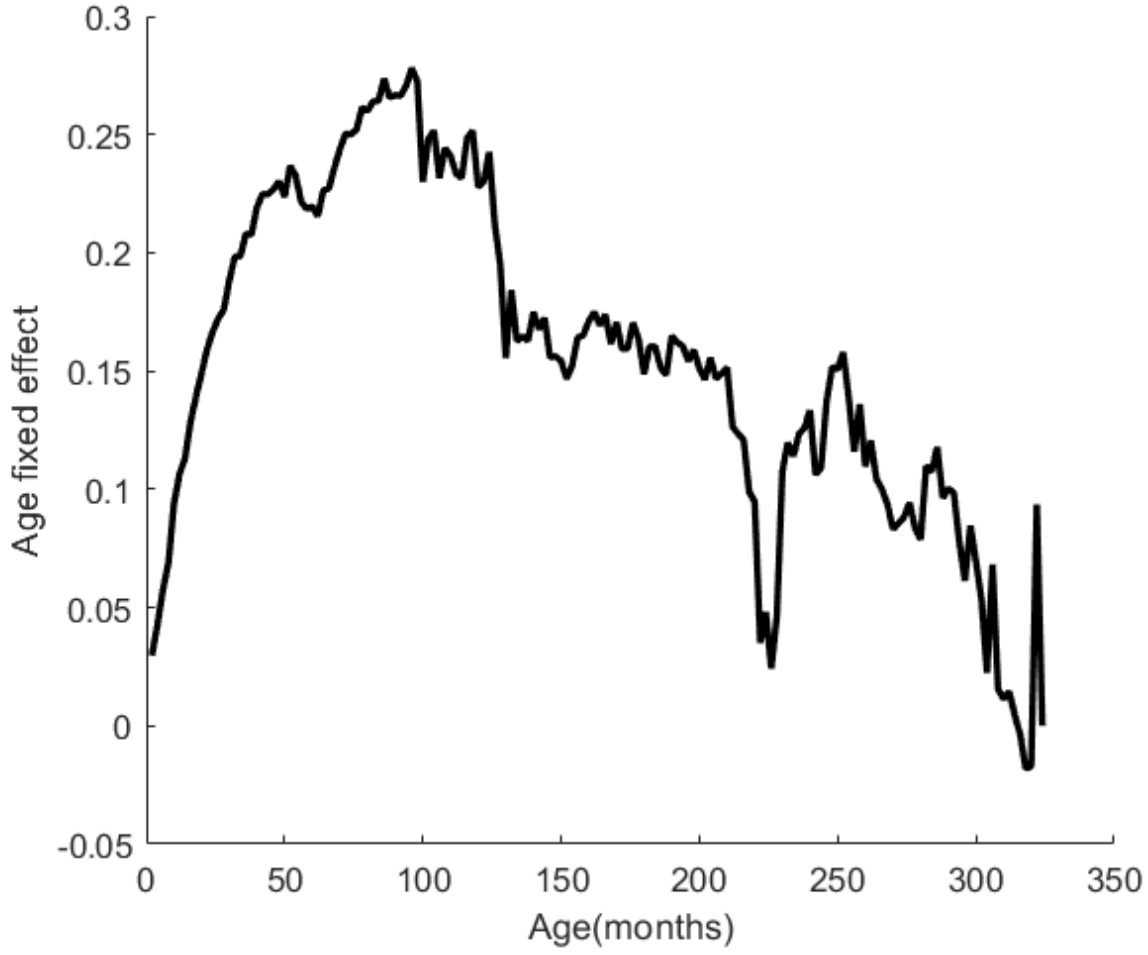
A Additional figures

Figure 2.14: Exit decisions



Note: The area under the curve shows the combinations of ability belief (e_{it}) and business cycle state (m_t) in which a fund exits: funds exit when they are perceived to be bad or when the macroeconomic state is bad, or some convex combination thereof. This figure is the same as figure 2.4, but with a coarser state space.

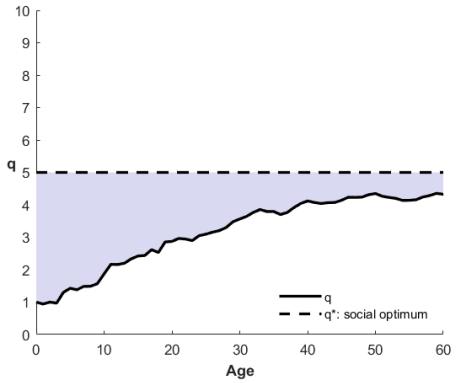
Figure 2.15: The effect of fund age on fund size



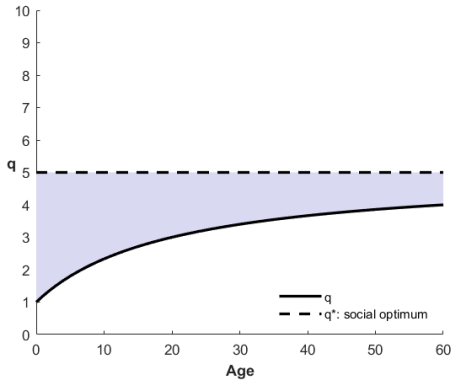
Note: This figure plots the age dummy that I estimate on the demand-side. On average, young funds grow quickly, peak at age 100 months, and then decline as they age further.

Figure 2.16: The effect of age on value-added

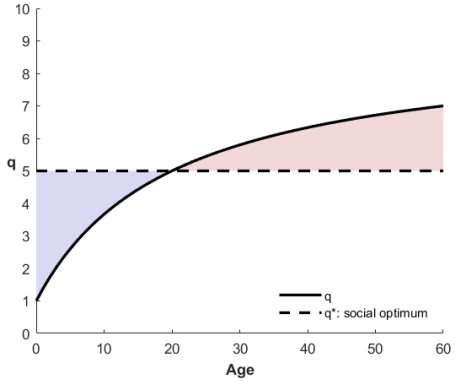
(a) Noisy convergence to optimum



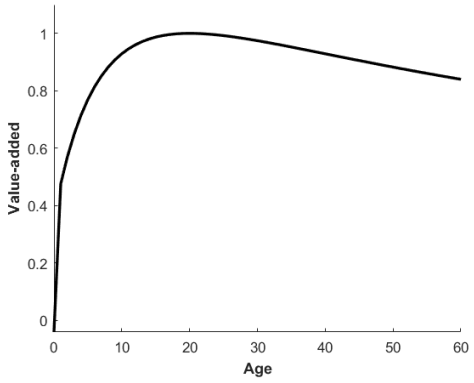
(b) Convergence to optimum



(c) Convergence towards and then away from optimum when fee too low



(d) Convergence towards and then away from optimum when fee too low



Note: Suppose that for a given fund true $\alpha = 0.1$ and other parameters are such that optimal fund size $q^* = 5$. If investors priors are incorrect, $\mu \neq \alpha$ and $q \neq q^*$. q is a function of investor beliefs about ability (which converge to true α as the funds ages and investors observe returns) and the fee rate f (which is fixed). This means that q converges to q^* only if f is ex-post optimal, which in this example means $f^* = \alpha/2$. In panels (a) and (b) $\mu < \alpha$ and $f = \alpha/2 = f^*$ such that q converges to q^*, with noise in the signal (panel (a)) and with the noise in the signal turned off (panel (b)). In panel (c) $\mu < \alpha$ and $f = \mu/2 < f^*$, which means that q initially converges to q^* (the blue area), but then overshoots and moves away from q^* (the red area). In other words, fund value-added does not increase monotonically with age, but is n-shaped, as in panel (d). The same is true if $\mu > \alpha$ and so $f > f^*$.

References

Berk, J. B. and Green, R. C. (2004). Mutual fund flows and performance in rational markets. *Journal of Political Economy*, 112(6):1269–1295.

Berk, J. B. and Van Binsbergen, J. H. (2015). Measuring skill in the mutual fund industry. *Journal of Financial Economics*, 118(1):1–20.

Caballero, R. J. and Hammour, M. L. (1996). On the timing and efficiency of creative destruction. *The Quarterly Journal of Economics*, 111(3):805–852.

Castillo-Martınez, L. (2018). Sudden stops, productivity, and the exchange rate.

Chevalier, J. and Ellison, G. (1997). Risk taking by mutual funds as a response to incentives. *Journal of Political Economy*, 105(6):1167–1200.

Fama, E. F. and French, K. R. (2010). Luck versus skill in the cross-section of mutual fund returns. *The journal of finance*, 65(5):1915–1947.

Gavazza, A. (2011). Demand spillovers and market outcomes in the mutual fund industry. *The RAND Journal of Economics*, 42(4):776–804.

Glode, V. (2011). Why mutual funds "underperform". *Journal of Financial Economics*, 99(3):546–559.

Hale, G. (2012). Bank relationships, business cycles, and financial crises. *Journal of International Economics*, 88(2):312–325.

Hotz, V. J. and Miller, R. A. (1993). Conditional choice probabilities and the estimation of dynamic models. *The Review of Economic Studies*, 60(3):497–529.

Kacperczyk, M., Nieuwerburgh, S. V., and Veldkamp, L. (2014). Time-varying fund manager skill. *The Journal of Finance*, 69(4):1455–1484.

Kacperczyk, M., Van Nieuwerburgh, S., and Veldkamp, L. (2016). A rational theory of mutual funds' attention allocation. *Econometrica*, 84(2):571–626.

Kosowski, R. (2011). Do mutual funds perform when it matters most to investors? us mutual fund performance and risk in recessions and expansions. *The Quarterly Journal of Finance*, 1(03):607–664.

Pástor, L. and Stambaugh, R. F. (2012). On the size of the active management industry. *Journal of Political Economy*, 120(4):740–781.

Pollet, J. M. and Wilson, M. (2008). How does size affect mutual fund behavior? *The Journal of Finance*, 63(6):2941–2969.

Roussanov, N., Ruan, H., and Wei, Y. (2018). Marketing mutual funds. Technical report, National Bureau of Economic Research.

Rust, J. (1987). Optimal replacement of gmc bus engines: An empirical model of harold

zurcher. *Econometrica: Journal of the Econometric Society*, pages 999–1033.

Schumpeter, J. A. et al. (1939). *Business cycles*, volume 1. McGraw-Hill New York.

Chapter 3:
A structural model of local competition between mutual funds

Mutual funds with similar investment strategies compete with each other for investment opportunities. I set out a model of demand for mutual funds in which (i) funds are located within a network depending on similarities in their investment strategies and (ii) funds impose negative spillovers on each other through this network. I structurally estimate this model using data on US equity mutual funds. I identify these network spillovers based on how investors in a given mutual fund respond to the returns performance of its competitors. I find that local competition has a material impact on fund size, in that absent competition the median fund would be 20% bigger, and on cross-sectional variation in size. I perform counterfactual simulations in which I demonstrate that luck can play an important role even when funds are skilled and investors are rational: I find that luck accounts for 9% of cross-sectional variation in mutual fund size.

3.1 Introduction

There is significant cross-sectional variation in size amongst mutual funds. I examine two of the many potential drivers of this variation that have been put forward in the literature: local competition[1] and luck.[2] My contribution is that by building and structurally estimating a model with a role for local competition and luck I am able to *quantify* their impact on mutual fund outcomes in a way that, to my knowledge, has not been done before.

The starting point for my assessment of local competition is the extent of investment strategy overlap, in that funds with similar strategies are competing for the same investment opportunities. I proxy for investment strategy overlap between a given pair of funds by estimating the distance between their respective betas: two funds with very similar (different) investment strategies will have similar (different) betas. I combine these pairwise distance measures into a network summarising the relative locations of all funds. I then show that this network has a core-periphery structure, in that most funds either have many closely located funds (the core) or very few (the periphery). This empirical observation motivates my primary research question: how does a fund's location within this network affect its outcomes?

To answer this question, I set out a model based on Berk and Green (2004) in which funds draw individual unknown ability to generate excess returns ("ability" henceforth). I incorporate a role for local competition by assuming that there are spillovers across funds along this network: all other things being equal, a large, closely located competitor makes it harder for a given fund to earn excess returns. Demand for mutual funds is a spatially auto-correlated process in which the effect of competition depends on a fund's location within the network and a parameter that governs the intensity of spillovers along that network. In Berk and Green (2004) a fund's size changes over time as investors observe its returns and update their beliefs about its ability. In my model, these network spillovers mean that the size of a given fund depends on investor beliefs about that fund, but also on investor beliefs about that fund's competitors. Consequently the size of a given fund changes in response to the returns of that fund and the returns of its competitors.

I estimate this spatially auto-correlated demand model using data on US Equity funds between 1990 and 2016. The challenge with identification is that the spatial structure implies that the size of a given fund and the size of its competitors are endogenously co-determined

[1] Wahal and Wang (2011), Hoberg et al. (2018).
[2] Berk and Van Binsbergen (2015), Fama and French (2010).

in equilibrium. This means that estimating the spillover parameter based on the extent to which fund sizes co-move is likely to under-estimate the true competitive effect. Instead I instrument for the size of a competing fund using the past excess returns earned by that fund. Relevance follows from the long-established empirical observation that investors respond to past performance (see, for example, Chevalier and Ellison (1997)). Validity is implied by investor rationality in the model: past returns must be uncorrelated with contemporaneous shocks to fund size, otherwise investors would have an incentive to change their holdings in mutual funds. I allocate each fund to one of 10 clusters, and include time fixed effects for each cluster to capture local variation in mutual fund outcomes. I include age fixed effects to account for the fact that on average mutual funds grow over their lifetime irrespective of their returns.

I find that the model fits the data well. The network spillover parameter is significant and negative, indicating that there is a role for local competition in mutual fund outcomes. This estimated spatially autocorrelated process gives me an intuitive, tractable model of demand for use in counterfactuals.

I run two sets of counterfactual simulations. These are partial equilibrium only, in that I model how demand changes in response to a counterfactual change, but hold fixed supply-side choices by funds regarding entry, exit and fees. The first simulation quantifies the role of competition: I turn off competition across funds by setting the network spillover parameter to zero. I find that the median fund would be 20% bigger in this counterfactual scenario. I also find that local competition is an important determinant of cross-sectional variation, in that there is significant heterogeneity across funds depending on their location within the network. Funds in the core are much more sensitive to local competition that funds in the periphery. In other words, local competition has material effects on mutual fund outcomes.

The second set of simulations I run relates to the role of luck. There is an extensive literature on whether successful mutual funds are skilled at producing excess returns or just lucky and, relatedly, whether the investors in these mutual funds are rational or not. I do not address this question directly, but instead I use these counterfactual simulations to demonstrate that there is a role for luck even in a model in which funds are skilled and investors are rational.

Specifically, by "luck" I mean two stochastic aspects of my model: the error in investor priors (for example, when investors believe a fund is high ability when it is in fact low ability) and the noise in fund returns (returns are only a noisy signal of true fund ability). These

forms of luck must even out across funds (in other words, on average half the funds are lucky and half are unlucky) and decay over time (in the limit investors observe a long set of fund returns and develop precise beliefs about fund ability regardless of these particular stochastic realisations). Nevertheless, the way in which investors form beliefs means that these stochastic realisations can have persistent effects across time: the impact of a positive prior draw or a positive return shock on investor beliefs only decays to zero as investors observe an infinitely long returns history. Luck can have permanent effects even in the limit if being unlucky results in a fund exiting.

To understand the role of luck in prior formation, I simulate investor demand replacing their actual *prior* about that fund with their *posterior* having observed the fund's returns. A fund that is unlucky in this sense is one that investors initially thought was low ability, but subsequently revised their beliefs upwards over time. Absent this error in prior formation, the fund would have been bigger. To understand the role of luck in returns shocks, I simulate investor demand turning off all inter-temporal variation in the signal that the investors extract from excess returns.

This allows me to quantify the impact of luck for each fund. It averages out to zero across funds, but can have material effects on individual funds: the median *absolute* impact of luck on fund size is 9%, of which about 5% is due to priors and about 4% is due to return shocks. In other words, even in a model with rational investors and skilled funds, luck is responsible for a material proportion of observed cross-sectional variation in funds.

I also find that the impact of luck varies between funds that exited during my sample period and funds that did not exit. Exiting funds were (i) more likely to experience unlucky returns shocks towards the end of their life, (ii) more likely to experience lucky prior draws which they subsquently under-performed (indicating that the *trajectory* of investor beliefs is important for exit, as well as simply the level of those beliefs) and (iii) more likely to experience extreme good or bad luck (indicating that the extent or *volatility* of luck is important for exit, as well the particular realisation of luck).

In this paper I show that local competition between funds can be captured in a tractable, estimable network model of demand. I use this estimated model to make two primary contributions, in that I am able to quantify the impact of competition and the impact of luck in a way that, to my knowledge, has not been done before.

I discuss the related literature below. In Section 2, I introduce the data and set out some guiding empirical facts. In Section 3, I set out my model. In Section 4, I describe my

empirical approach. In Section 5, I report my results. In Section 6, I undertake counterfactual analyses. In Section 7, I conclude.

3.1.1 Related literature

This paper is related to two strands of literature regarding (i) competition amongst mutual funds and (ii) the role of luck in mutual fund outcomes.

Pástor and Stambaugh (2012) set out a model in which there is homogeneous competition amongst funds depending on the aggregate size of the industry. Wahal and Wang (2011) and Hoberg et al. (2018) set out reduced form evidence that there is a local component to competition that depends on the extent of investment overlap. There are a small number of papers that analyse mutual fund competition in a structural econometric setting, including Gavazza (2011) (which focuses on the role of the broader fund family) and Roussanov et al. (2018) (which focuses on the role of marketing). The contribution of this paper is that it considers local competition in a structural econometric setting, which ultimately allows me to quantify the effect of local competition through counterfactual analysis.

There is a very large literature on mutual fund outcomes (Elton et al., 1993; Carhart, 1997; Busse et al., 2010; Bollen and Busse, 2005; Kosowski et al., 2006; Cremers and Petajisto, 2009; Kacperczyk et al., 2014; Chen et al., 2004; Pástor et al., 2015; Pollet and Wilson, 2008; Kacperczyk et al., 2016; Kacperczyk and Seru, 2007; Huang et al., 2011) and in particular whether these outcomes are the result of luck or skill (see, for example, Berk and Green (2004), Berk and Van Binsbergen (2015), Fama and French (2010)). I structurally estimate a model based on Berk and Green (2004) and use it to show that there is a role for luck even in a model in which funds are skilled and investors are rational. Furthermore, I am able to use my estimated model in counterfactual analysis to quantify the role of luck in a novel way.

3.2 Data

I first describe how I select funds and calculate excess returns. I then describe the key empirical facts that motivate my research question and guide my modelling.

3.2.1 Sample selection

I obtain data on mutual fund characteristics and their monthly returns and assets from the database maintained by the Center for Research in Security Prices (CRSP), The University of Chicago Booth School of Business. I select data from January 1990 to December 2016. I limit my sample to actively managed US Equity funds that (i) are never smaller than USD 1m in size, (ii) have at least 12 months of returns data and (iii) have data on their expense ratio. This is broadly the standard approach in the literature (see for example Berk and Van Binsbergen (2015) for an overview of mutual fund selection). I am left with a sample of 3,420 funds and a total of 452,222 month-fund observations.

3.2.2 Calculating excess returns

I calculate excess returns following Berk and Van Binsbergen (2015). I regress returns in excess of the risk-free rate (R_{it}) on a set of 11 common factors ($\mathbf{F_t}$) which are the returns to the main index funds operated by Vanguard (listed in the table below).[3] The fund's excess return, α_{it} is the residual in this regression:

$$R_{it} = \boldsymbol{\beta}_i \mathbf{F_t} + \alpha_{it} \tag{1}$$

[3]This is a more reasonable benchmark for mutual funds than, for example, a benchmark involving momentum investing returns that would be prohibitively costly to implement in practice. I refer to Berk and Van Binsbergen (2015) for a fuller discussion.

Table 3.1: Benchmark

Fund Name	Ticker	Asset Class
S&P 500 Index	VFINX	Large-Cap Blend
Extended Market Index	VEXMX	Mid-Cap Blend
Small-Cap Index	NAESX	Small-Cap Blend
European Stock Index	VEURX	International
Pacific Stock Index	VPACX	International
Value Index	VVIAX	Large-Cap Value
Balanced Index	VBINX	Balanced
Emerging Markets Stock Index	VEIEX	International
Mid-Cap Index	VIMSX	Mid-Cap Blend
Small-Cap Growth Index	VISGX	Small-Cap Growth
Small-Cap Value Index	VISVX	Small-Cap Value

3.2.3 Empirical facts

I set out four empirical facts:

1. **Heterogeneity across funds in size at start of life**: There is significant variation in the initial size of a fund at the start of its life, as I show in Figure 3.1, even controlling for the state of the economy at the time of entry.

2. **Variation over time in relative fund size**: The cross-sectional heterogeneity in fund size is not fixed over time, in that the relative ranking of funds changes over time. I show this for 5 representative funds in Figure 3.2: the biggest of these funds at the start of their lives is only the 4th biggest 6 years later.

3. **Heterogeneity across funds in excess return variability**: Excess returns are more volatile for some funds than for other funds. I show this heterogeneity in Figure 3.3.

4. **Heterogeneity across funds in their location in β-space**: I summarise heterogeneity in fund investment strategies by calculating the distance between the betas of each pair of funds:

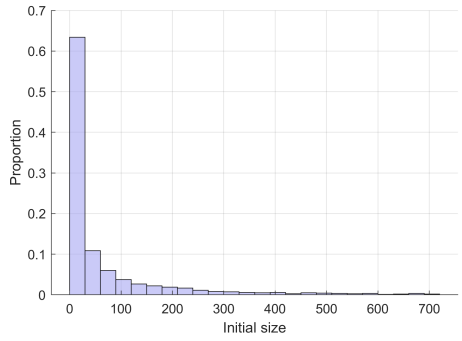
$$d_{ij} = || \boldsymbol{\beta}_i - \boldsymbol{\beta}_j || \tag{2}$$

This measure d_{ij} is a proxy for the similarity in the investment strategies of funds i and j: if they have similar investment strategies, then it is likely that they also have similar betas. I show the distribution of d_{ij} in Figure 3.4, and show that there is significant variation in this figure across pairs. This distance measure can also be thought of as representing a fund's location within the network. I summarise the number of close connections each fund has in in Figure 3.5, and show that there is evidence of a core-periphery structure in β-space: some funds have lots of close connections (the core), and some funds have very few (the periphery).

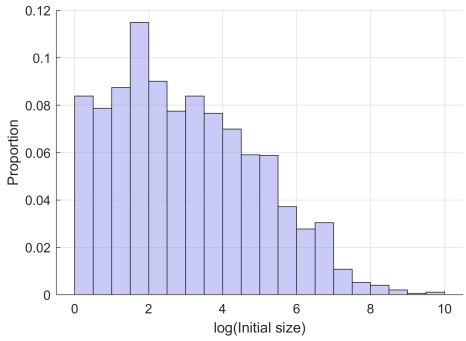
These empirical facts are the basis for my research question. Figure 3.5 shows that there are observable differences in where funds are located in the network: what impact does this have on competition between funds? To the extent that size represents investor beliefs about fund skill, then Figure 3.1 shows that there is significant variation in investor priors about funds at the start of their lives: how persistent are the effects of these priors? Figure 3.2 shows that investors updated these priors over time, in that some funds turned out to be better or worse than initially believed: what is the impact of this error in prior formation? These are the questions I seek to answer in this paper.

Figure 3.1: Heterogeneity in fund size

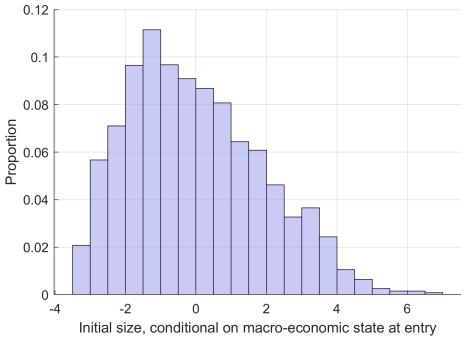
(a) Initial size



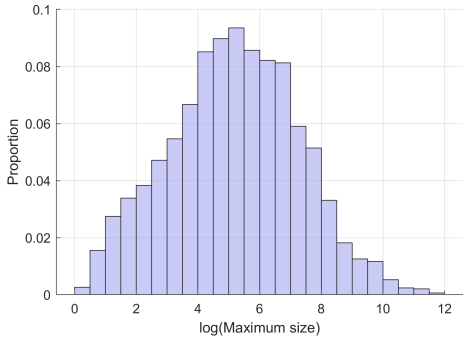
(b) Log initial size



(c) Log initial size conditional on M_1

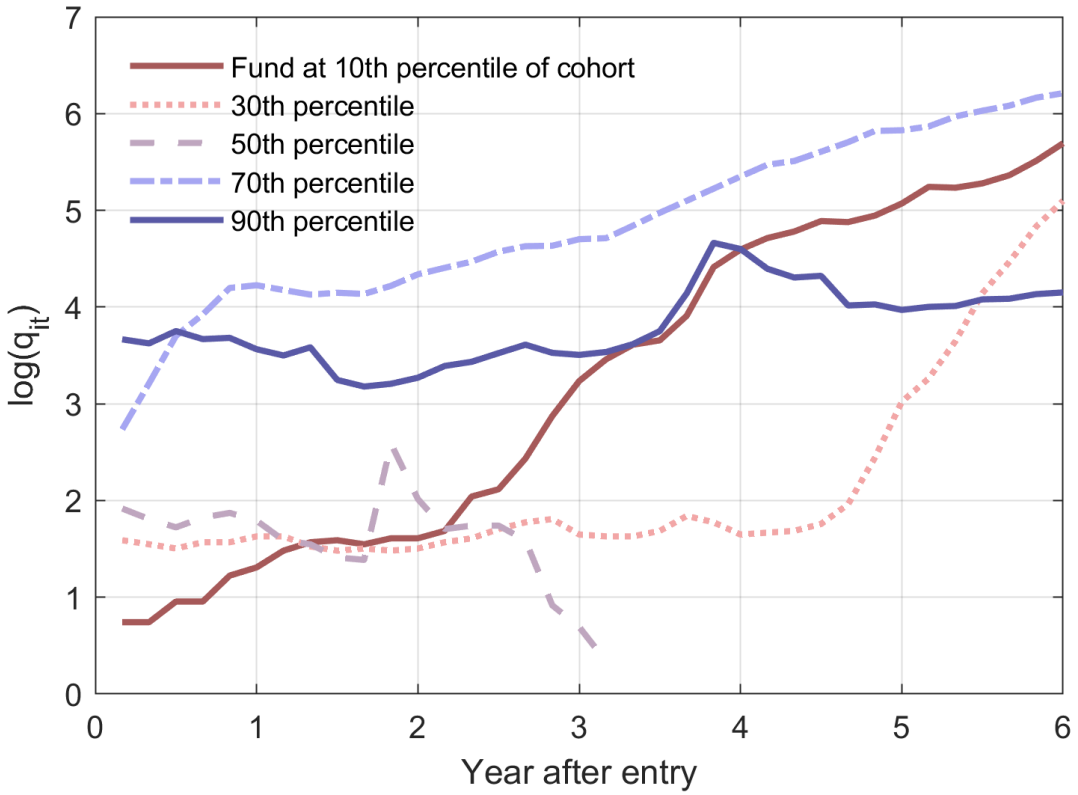


(d) Log maximum size



Note: Panel (a) shows the distribution of fund size in the first period of its life, excluding the top 5% of funds by size. Panel (b) shows the distribution of the natural log of initial size. Panel (c) conditions on M_1, the level of the SP500 in the period in which the fund entered. Panel (d) shows the log of the maximum size the fund attains during my sample.

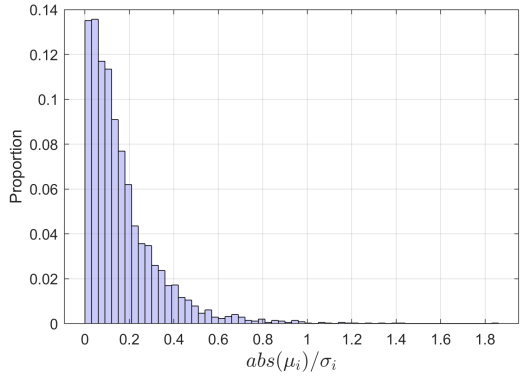
Figure 3.2: Relative variation in fund size



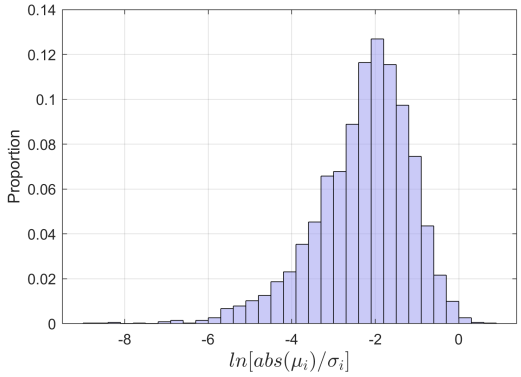
Note: This graph tracks the growth of 5 individual representative mutual funds that entered in 2002 over the 6 years after their entry. The funds chosen are the 10th, 30th, 50th, 70th and 90th percentiles by size at the time of entry. The relative ranking of these funds changes materially over the course of this period: the largest of these funds, for example, is only the 4th largest after 6 years. Note that the median fund exits in its third year.

Figure 3.3: Heterogeneity in excess return variability

(a) Coefficient of variation



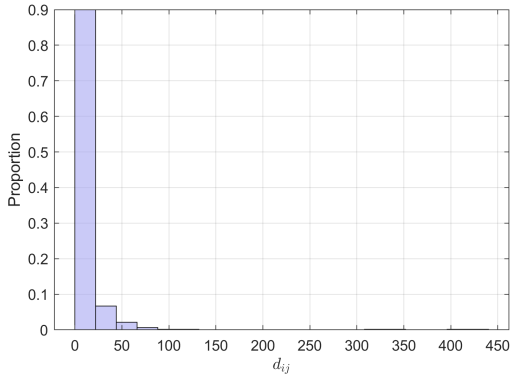
(b) Log coefficient of variation



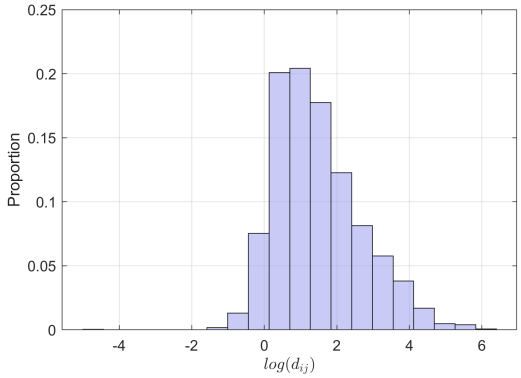
Note: For each fund i, I calculate the mean excess return over the fund's life μ_i and the standard deviation in excess return σ_i. I report their ratio, which is the inverse of the fund-specific coefficient of variation, in order to show that (1) there is significant variation in excess returns and (2) there is significant heterogeneity across funds in the extent of this variation.

Figure 3.4: Heterogeneity in β-space

(a) d_{ij}

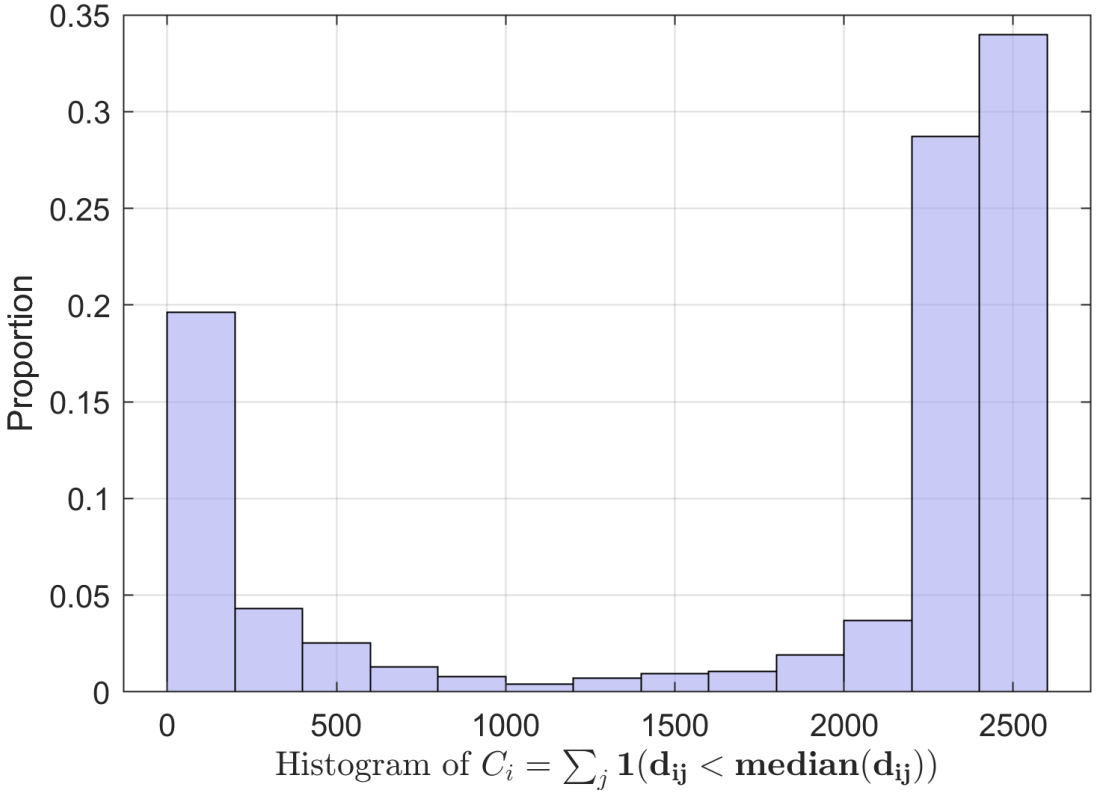


(b) $log(d_{ij})$



Note: d_{ij} is a measure of distance between funds i and j in β-space. That is, d_{ij} is low when β_i and β_j are similar.

Figure 3.5: Core-periphery structure in β-space



Note: d_{ij} is the distance in β-space between funds i and j, as per Equation 2. For every fund, I count the number of funds within a certain distance (the median of all d_{ij}) as a measure of local competition. The distribution is bi-modal: some funds have lots of close competitors (the core), some funds have very few (the periphery).

3.3 Model

I first set out a model of demand for mutual funds. I then discuss the role for luck in this model.

3.3.1 Demand

The model of demand is based on Berk and Green (2004), in that there are decreasing returns to scale in the ability of a mutual fund to earn excess returns. To this model I add an element of *local competition*, in that the ability of a fund is decreasing in the size of other mutual funds as well as in its own size.

Mutual fund i earns *gross* excess return α_{it}^g: this is not the actual return that investors receive, but is instead a notional return that the fund would earn on the first dollar of investment before the impact of decreasing returns to scale. I define the total risk-adjusted payout in dollar terms to investors from investing q_{it} in mutual fund i with gross return α_{it}^g and fee rate f_i as:

$$TP_{it} = q_{it}\alpha_{it}^g - C(q_{it}) - q_{it}f_i$$

where $C(q_{it})$ is a cost function representing the decreasing returns to scale in the ability to earn excess returns. I parameterise the cost function as $C(q_{it}) = \phi_i q_{it}^2$ where $\phi_i > 0$, such that when $q > 0$: $C(q) > 0$, $C'(q) > 0$, $C''(q) > 0$, $C(0) = 0$ and $\lim_{q\to\infty} C(q) = \infty$. The *net* α_i^n excess return is what investors actually earn after the impact of decreasing returns to scale, and is simply this payout divided by the size of the investment:

$$\alpha_{it}^n = \frac{TP_{t+1}}{q_{it}} = \alpha_{it}^g - \frac{C(q_{it})}{q_{it}} - f_i = \alpha_{it}^g - \phi_i q_{it} - f_i \tag{3}$$

To meaningfully take this model to data, I need to capture some of the ways in which the ability of a fund to earn excess returns can vary intertemporally and in the cross-section. To that end, I disaggregate the fund's gross excess return into five components:

$$\alpha_{it}^g = \alpha_i + \epsilon_{it} + \delta_{a(it)} + \delta_t + \sum_j \theta_{ij} q_{jt} \tag{4}$$

where:

- α_i represents the fund's true ability to generate returns. I allow it to vary across funds but keep it fixed across time. This is a simple way of allowing some funds to be higher ability than others.
- ϵ_{it} represents a fund-specific shock to ability at time t.
- $\delta_{a(it)}$ represents an age effect. I denote the age of fund i at time t as $a(it)$, and I allow

ability to vary with age. This captures, in a simple way, the possibility of learning by doing. More practically, this variation will allow me to account for the empirical observation in Figure 3.2 that funds tend to grow, at least initially, regardless of their net returns.

- δ_t is a common time effect across all funds. In Coen (2020), for example, I show that the size of funds varies with the business cycle.

- $\theta_{ij}q_{jt}$ represents the effect on mutual fund i of competition from mutual fund j. If $\theta_{ij} < 0$ and mutual fund j is large then, all other things being equal, fund i finds it harder to earn excess returns because there is more competition for the same investment opportunities. θ_{ij} varies across pairs and captures the intensity of this effect.

This specification nests that of Coen (2020). In that paper, I model specific types of intertemporal variation in ability, whereas in this paper I leave the intertemporal variation in ability as some general δ_t. In that paper I impose homogeneous competition between funds ($\theta_{ij} = \theta$ for $\forall i, j$), whereas here I allow for the effects of competition to be heterogeneous across pairs according to θ_{ij}.

Investors choose q_{it} before ϵ_{it} is realised. Investors do not know the true ability of the fund α_i, but form expectations based on the information available to them at the point of investment, which I denote I_{t-1}. I define these expectations as $e_{it} \equiv \mathbb{E}[\alpha_i \mid I_{t-1}]$. All other components of the return are known to the investor.

Investors supply capital with infinite elasticity to any fund with positive expected *net* returns α^n_{it}, taking aggregate investment q_t in the fund as given. In equilibrium, q_t is then such that $\mathbb{E}[\alpha^n_{it} \mid I_{t-1}] = 0$. Substituting in Equations 3, this means that:

$$q_{it} = \frac{e_{it} + \delta_{a(it)} + \delta_t + \sum_j \theta_{ij} q_{jt} - f_i}{\phi_i} \tag{5}$$

Investor demand for mutual fund i is therefore increasing in its expected ability e_{it}, increasing in its scalability ϕ_i, decreasing in its fee rate f_i and decreasing in the extent of local competition.

To complete the model of demand, I need to characterise the expectations formation process behind investor beliefs e_{it}. To do this, I make the following assumptions about the

distribution of fund abilities:

- **Ability draw**: Funds draw ability from a normal distribution: $\alpha_i \sim N(\overline{\mu}, \overline{\tau}_{\alpha}^{-1})$. This type is not observed by investors or the funds themselves.
- **Prior formation**: At the start of the fund's life, funds and investors observe an initial signal α_{i0} about the true ability of fund i: $\alpha_{i0} = \alpha_i + v_i^{\mu}$, where v_i^{μ} is the error in the prior. I assume that $v_i^{\mu} \sim N(0, \sigma_i^v)$ and $v_i^{\mu} \perp\!\!\!\perp \alpha_i$. Given this signal, investors form fund-specific prior beliefs $\alpha_i \mid \alpha_{i0} \sim N(\mu_i, \tau_{i\alpha}^{-1})$. Note that I allow the precision of the prior error σ_i^v (and thus the precision of the updated beliefs $\tau_{i\alpha}^{-1}$) to be heterogeneous across funds: investors are more uncertain about some funds than others.
- **Return shocks**: Return shocks are independent of true ability and also normally distributed: $\epsilon_{it} \sim N(0, \tau_{i,e}^{-1})$. The precision of these return shocks is also fund-specific, but the relationship between the precision of return shocks and of the prior formation is constant across funds: I define the homogeneous signal-to-noise ratio as $\lambda = \frac{\tau_{i,e}}{\tau_{i,\alpha}}$. In other words, investors are more uncertain about some funds than others, but this greater uncertainty is equally true of both the funds' priors and the funds' return signals.

Investors observe past net excess returns, $\alpha_{is<t}^n$ and from this can infer gross returns α_{is}^g. Investors cannot separately identify α_i from ϵ_{is}, but can extract a signal about α_i given their relative distributions.

Given these distributional assumptions, there are simple closed-form expressions for how investors form and update their posterior beliefs about α_i in responses to these signals. I define the function $g(A_{it}; \lambda) = \sum_{s=1}^{t-1} \frac{\lambda\, \alpha_{is}^n}{1+(s-1)\lambda}$ and express mutual fund demand as follows:

$$q_{it} = \frac{1}{\phi_i}\left[\mu_i - f_i + \delta_{a(it)} + \delta_t + \sum_j \theta_{ij} q_{jt} + g(A_{it}; \lambda)\right] + u_{it}^q \tag{6}$$

I add an error term, u_{it}^q, that represents shocks to q_{it} beyond this expectations formation process. This could include, for example, noise traders. I leave further discussion of this error term and its distribution to the section below on my empirical analysis.

This equation contains endogenous mutual fund sizes on both sides. To solve for equilibrium fund size, I express the same equation in matrix notation. A bold variable indicates an $N \times 1$ vector stacking the non-bold variable (such that, for example, $\boldsymbol{\mu}$ is an $N \times 1$ vector

stacking μ_i) and $\mathbf{\Gamma}$ is an $N \times N$ matrix with ϕ_i in position (i,i) on the diagonal and θ_{ij} in position (i,j) off-diagonal. It is then straightforward to invert Equation 6:

$$\boldsymbol{q_t} = \mathbf{\Gamma}^{-1}\big[\boldsymbol{\mu} - \boldsymbol{f} + \boldsymbol{\delta_a} + \delta_t + \boldsymbol{g}(\boldsymbol{A};\lambda) + \boldsymbol{u_t}\big] \tag{7}$$

This allows me to characterise the equilibrium effects of local competition. It implies that the size of mutual fund i, q_{it} is increasing in beliefs about the ability of fund i and decreasing in beliefs about the ability of fund j, where the intensity of the competitive cross-effects depends on θ_{ij} and the relative positions in the network of i and j. If fund j receives a positive shock to beliefs about its ability, $\alpha^n_{jt} > 0$, then all other things being equal this causes fund i to shrink in equilibrium.

3.3.2 The role of luck

I examine four stochastic elements of mutual fund size which I shall call "luck".

- First, the size of a given mutual fund is sensitive to its random draw of an investor prior. Two otherwise identical (including in true ability) mutual funds can draw different priors and, as per Equation 7 this has a persistent impact on their size.
- Second, the size of a given mutual fund is sensitive to the random draw of investor priors for its close competitors. Two otherwise identical funds can have their nearest competitor draw differing priors and this would also have a persistent impact on their size.
- Third, the size of a given mutual fund is sensitive to return shocks. These return shocks impact the expectations formation process of investors and so have persistent effects. Two otherwise identical funds that received a positive and a negative shock, respectively, would have persistent differences in size. The timing of shocks matters as well as their sign: consider a mutual fund A (mutual fund B) that receives a positive (negative) returns shock at time t and a negative (positive) returns shock at time $t' > t$. Mutual fund A will be bigger than mutual fund B between t and t'.
- Fourth, the size of a given mutual fund is sensitive to return shocks of local competitors, in an analogous way to above. A fund is unlucky if its closest competitor is lucky.

I exclude quantity shocks: they are random and a form of luck, but they do not have persistent effects.

Conditional on the fund surviving, these four forms of luck have persistent but not permanent effects: in the limit, the investor observes enough fund returns to learn their true ability. In other words, differences between funds brought about luck decay to zero over time. The speed of this decay depends critically on λ, the signal-to-noise ratio of observed returns. If this ratio is high, then returns are informative and random variation in investor priors decay in importance quickly. If returns are informative, then the immediate impact of returns shocks is bigger but that impact decays quickly.

Luck can, however, have permanent effects through its impact on firm failure. I do not model the decision of mutual funds to exit, but the zero lower bound implicit in Equation 6 implies exit if investor beliefs are too low. If investor beliefs are such that fund i is not expected to produce positive net returns even if q_{it} is arbitrarily small (such that decreasing returns to scale have no impact), then investors will invest nothing and the fund exits permanently. Luck can in this way have a permanent effect if a fund exits because it drew a poor prior or a negative returns shock early in its life.

The timing of returns shocks is particularly important in this context: a fund is more likely to exit if it draws a negative returns shock in a 'bad' period in which δ_t is low. For example, consider if fund A and fund B are identical and each draw a negative return shock, but fund A draws a negative shock in bad times (δ_t is low) and exits, whereas fund B draws it in good times and so does not exit.

The extent of this random variation varies across funds, because the variance of the returns shock and the prior error varies across funds. Empirically, I set out in Figure 3.3 that excess returns are more variable for some funds than others. In other words, there is a bigger role for luck for some funds than other funds.

The effect of luck therefore depends on (i) the specific realisations and timing of these fund-specific random draws, (ii) the fund-specific volatility of these random draws and (iii) the signal-to-noise ratio λ. The effect of luck depends on the values of these parameters and is, therefore, an empirical question.

Finally, I emphasise that these elements are random within the context of the model. In taking this model to the data, it is worth considering the impact of potential model mis-specification on this definition of luck. If, for example, the fund can affect its prior or its excess return volatility then, to a certain extent, what I am calling luck reflects these choices.

3.4 Empirical approach

There are three aspects to my empirical approach: (1) I calibrate some parameters, (2) I impose parametric restrictions on parameters relating to competition and (3) I estimate the remaining parameters by matching observed quantities. I discuss each of these in turn.

3.4.1 Calibration

I follow Coen (2020) and calibrate ϕ_i and μ_i based on how q_{it} evolves over time. I set ϕ_i to be the inverse of the maximum size that fund i reaches in my sample: $\phi = \frac{1}{q_{i,max}}$, where $q_{i,max} = \max_t q_{it}$. This is effectively a fund-specific normalisation such that the product $q_{it}\phi_i \in [0, 1]$ for any i. This means that I do not use the cross-sectional variation in the size of the funds to identify the other parameters, but only the variation over time. In other words, I assume that Vanguard's largest funds are not large relative to other funds because they earned very large returns early in their life, they are large for fund-specific reasons that I effectively encode and leave fixed in ϕ_i.

I infer μ_i from the size of fund i in the first period of its life. Setting $t = 1$ in Equation 6 and re-arranging: $\mu_i = q_{i1} - \delta_{i1}$. This results in computational benefits, relative to simply estimating μ_i as a fixed effect, as it can be done outside of the main estimation loop. It also better matches the interpretation of μ_i as an initial prior belief about fund ability at the start of its life.

3.4.2 Parameterisation

Competition is heterogeneous according to θ_{ij}. It would not be feasible to estimate all of these parameters, so I follow the industrial organization literature by parameterising these cross-effects by reference to *characteristics* rather than by reference to funds. The variation in θ_{ij} is intended to capture local variation in the extent to which funds are competing for the same investment opportunities. The key characteristic I am seeking to measure therefore is what Wahal and Wang (2011) refer to as *overlap*: the extent to which mutual funds have the same holdings.

I use the distance in β-space, d_{ij} in Equation x, as a proxy for overlap, on the basis that funds with similar holdings will have similar betas. If d_{ij} is large, then funds i and j do not have similar holdings and θ_{ij} is likely to be low. I define $\tilde{d}_{ij} = ln(1/d_{ij})$ and parameterise

θ_{ij} as follows:

$$\theta_{ij} = \theta \tilde{d}_{ij} \tag{8}$$

Let $\boldsymbol{D}$ be the matrix with row-normalized $\tilde{d}_{ij}$ at coordinate (i, j) and 0 on the diagonal and let $\boldsymbol{\Phi}$ be a diagonal matrix with ϕ_i in position (i, i) on the diagonal and 0 off-diagonal. $\boldsymbol{D}$ is in effect a network, and the location of a fund within this network determines the extent of local competition it faces. Equation 6 for fund size can then be expressed as follows:

$$\boldsymbol{\Phi}\, \boldsymbol{q_t} = \boldsymbol{\mu} - \boldsymbol{f} + \boldsymbol{\delta_a} + \delta_t \, + \, \boldsymbol{g}(\boldsymbol{A}; \lambda) + \theta \boldsymbol{D} \boldsymbol{q_t} \, + \, \boldsymbol{e_t} \tag{9}$$

Rearranging for equilibrium $\boldsymbol{q_t}$:

$$\boldsymbol{q_t} = [\boldsymbol{\Phi} - \theta \boldsymbol{D}]^{-1} \big[\boldsymbol{\mu} - \boldsymbol{f} + \boldsymbol{\delta_a} + \delta_t + \boldsymbol{g}(\boldsymbol{A}; \lambda) \, + \, \boldsymbol{e_t}\big] \tag{10}$$

In other words, mutual fund size is a spatially autocorrelated process where the measure of spatial proximity between funds is in β-space. The effect of competition on fund i depends on its location within the network relative to other funds and on the intensity of the spillovers governed by θ.

3.4.3 Estimation

I estimate this spatially autocorrelated process by GMM. I calculate $\mathbf{D}$ by estimating a fund's β over its entire life and calculating distance as per Equation 2. I assume that the fund's location in β-space is exogenous to contemporaneous shocks to fund size. The spatial structure implies that q_{jt} is endogenous in Equation 9: an unobserved positive size shock to fund i means that fund j is small all else being equal.

I instrument for q_{jt} using the returns of fund j and its initial size, in the following first stage:

$$q_{it} = \omega_1 \, g(A_{it}; \lambda) + \omega_2 \, q_{i1} + \eta_{it} \tag{11}$$

$g(A_{it}; \lambda)$, as defined above, is a weighted average of past returns α^n_{jt-1}. In other words, I identify the competitive effect of fund j on fund i by looking at how q_{it} responds to the returns of fund j, α^n_{jt-1}, conditional on the distance between them in β-space, $\tilde{d}_{ij}$. I include q_{i1} as a pre-determined proxy for fund size. I use these instruments to construct moments

and estimate the parameters in Equation 9 by GMM.

The identifying assumption is that α^n_{jt-1} is independent of the unobserved size shock u^q_{it}. Within the context of the model u^q_{it} represents noise investors, and investor rationality requires that this cannot be serially correlated: if noise trader error u^q_{it} were predictable using time $t-1$ information then rational investors would alter their holdings to account for this.

In this sense identification comes from the model. It is worth, however, examining the ways in which identification would fail if the model is mis-specified and so the residual comprises more than just the impact of noise traders. In particular, the model implies that α^n_{jt-1} has an impact on q_{it} only through q_{jt}. Suppose instead there was a local, unobserved component to excess returns that affected both fund i and fund j. In this case, α^n_{jt-1} would be a signal of this local, unobserved component and so q_{it} would respond to this signal directly. This would result in me underestimating the true impact of competition θ_{ij}.

I try and account for the effect of any such mis-specification on identification by including additional time dummy variables. I use statistical clustering tools to allocate each fund to one of 10 clusters in β-space, where the kmeans++ algorithm that I use chooses the boundaries of each cluster to minimise the total distance of each fund from the cluster centre. I then include a separate time dummy for each cluster to account for local shocks within that cluster.

Once I have estimated the parameters in Equation 9, it is then straightforward to infer estimates of α^g_{it} from Equation 3, and from that $\tau^{-1}_{i,e} = std(\widehat{\alpha}^g_{it})$. That is, I calculate the return uncertainty for each fund from the observed variation in the fund's excess returns. From this, I can infer the fund-specific uncertainty in the prior: $\tau^{-1}_{i,\alpha} = \widehat{\lambda}\tau^{-1}_{i,e}$. That is, I am able to observe the fund-specific noise in returns and also how quickly investors respond to those returns: given that investors respond to returns based on their signal-to-noise ratio $\widehat{\lambda}$, this tells me the fund-specific uncertainty in the investors' prior about the fund.

3.5 Results

I set out the results of my estimation in Table 3.2. I find that the model fits the data well. In particular, allowing for local competition results in materially improved fit over the nested model in which there is no local competition. I find that the parameter governing network spillovers, θ, is significant and negative as expected.

Table 3.2: Estimation results

	[1] q_{it}	[2] q_{it}
λ		0.088***
		(0.002)
θ	-0.150***	-0.067***
	(0.014)	(0.016)
μ_{it}	0.311	0.311
ϕ_{it}	0.037	0.037
Age FE	Y	Y
Time×Cluster FE	Y	Y
R^2	0.70	0.74
No. obs	226,111	226,111

Note: Figures in parentheses are standard errors. ***, **, * indicate different from 0 at 1%, 5% and 10% significance, respectively. q_{it} is the size of mutual fund i at time t, λ is sensitivity to past returns and θ governs the impact of local competition. I calibrate fund-specific priors μ_{it} and scalability ϕ_i and report the mean across funds here.

3.6 Counterfactual analysis

I run four counterfactual simulations. In the first counterfactual I quantify the role of competition by comparing actual outcomes with counterfactual outcomes in which there is no local competition.

The remaining counterfactuals relate to the role of luck, which as defined above I use to mean errors in investor prior formation and return shocks. I do not observe true fund ability and so cannot entirely remove the error in the investors' prior about ability. My model does, however, allow me to infer investor beliefs about a fund's ability from the size of the fund. This means that I can observe investors' *posterior* beliefs taking into account its

lifetime returns performance. To understand the role of luck in prior formation, for example, I simulate investor demand replacing their actual *prior* about that fund with their *posterior* having observed the fund's returns. A fund that is unlucky in this sense is one that investors initially thought was low ability, but subsequently revised their beliefs upwards over time. Absent this error in prior formation, the fund would have been bigger. To understand the role of luck in returns shocks, I simulate investor demand turning off all inter-temporal variation in the signal that the investors extract from excess returns.

To describe these simulations more formally it is helpful to characterise my model for q_{it} as a function of, amongst other things, three things: (1) the effect of local competition governed by θ, (2) investor priors about that fund μ_i, (3) fund returns $\{\alpha_{is}^g\}_{s=1}^{T_i}$. In each counterfactual I compare actual $q_{it}(\theta, \mu_i, \{\alpha_{is}^g\}_{s=1}^{T_i})$ with counterfactual q_{it} in which I vary one of these three elements.

1. **The effect of competition**: I assess the effect of competition by comparing actual fund sizes with counterfactual fund sizes absent competition, which I simulate by setting $\theta = 0$. That is, I calculate $q_{it}(\theta = 0, \mu_i, \{\alpha_{is}^g\}_{s=1}^{T_i})$.

2. **The effect of incorrect priors**: Over time, investors observe fund returns and update their initial priors to the following:

$$e_{it} = \frac{\mu_i}{1 + (t-1)\lambda} + \frac{\lambda \sum_s^{t-1} \alpha_{is}^g}{1 + (t-1)\lambda} \tag{12}$$

   I correct priors by setting the prior investor belief μ_i equal to the posterior given the fund's returns over its lifetime. That is, I calculate $q_{it}(\theta, \mu_i = e_{iT_i}, \{\alpha_{is}^g\}_{s=1}^{T_i})$.

3. **The effect of random return shocks**: α_{it}^g consists of true fund ability α_i and an idiosyncratic return shock. I switch off these return shocks by setting $\alpha_{it}^g = e_{iT_i}$ for $\forall t$. That is, I calculate $q_{it}(\theta, \mu_i, \{e_{iT_i}\}_{s=1}^{T_i})$.

4. **The effect of incorrect priors and random return shocks**: I correct priors and remove random return shocks simultaneously, as I do individually in the previous two counterfactuals. That is, I calculate $q_{it}(\theta, \mu_i = e_{iT_i}, \{e_{iT_i}\}_{s=1}^{T_i})$.

In the figures and table that follow, I summarise along various dimensions the percentage difference between this counterfactual quantity and actual quantity. If, for example, a fund's prior μ_i is lower than its posterior e_{iT_i}, then this number is positive and the fund was "unlucky" with the draw of its prior.

From these counterfactual simulations I draw three main conclusions. First, **the effect of competition is significant and heterogeneous**. I summarise the effect of competition in panel (a) of Figure 3.6. The median increase in fund size absent competition is 20.0%, but there is significant variation across funds: for funds in the periphery this number is close to zero, for funds in the core it can be closer to 50%.

Second, **the effect of luck is significant and heterogeneous**: I summarise the impact of prior formation and return shocks in panels (b), (c) and (d) of Figure 3.6. Luck averages out across funds, but not within funds: the average size of some funds over their lifetime is materially affected by their priors and returns shocks. The median *absolute* impact of luck across funds is 9%, of which about 5% is due to priors and 4% is due to return shocks.

Third, **luck is related to exit**. In Table 3.3 and Figure 3.7 I show the impact of luck conditioning on whether a fund exited during my sample period or survived.

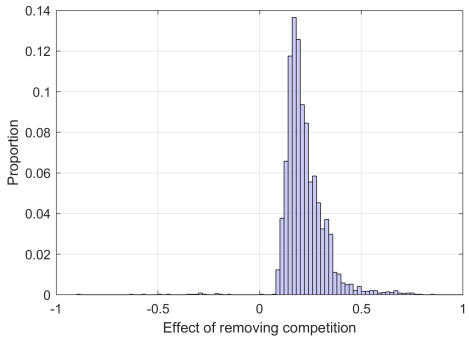
- I find that exit is related to return shocks in an intuitive way, in that exiting funds were unluckier in the sense that they were more likely to experience negative return shocks in the last few months of their lives than surviving funds.

- Exiting funds were luckier than surviving funds, however, in their prior draw. In other words, exiting funds were more likely to under-perform an initial overoptimistic prior. This suggests that the *trajectory of investor beliefs* is important for exit, as well as simply the absolute level of those beliefs. Consider, for example, if investors had the same posterior beliefs about fund A and fund B, but the investors' prior beliefs at fund entry were higher for A than for B. This counterfactual simulation indicates that A would be more likely to exit than B.

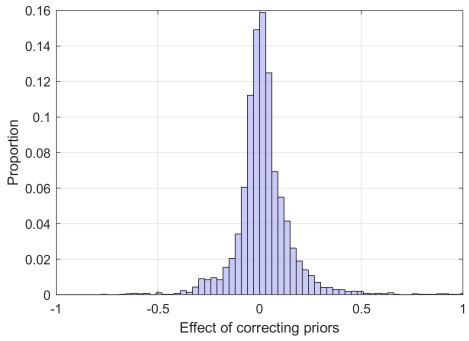
- Exiting funds experienced more extreme good luck and bad luck than surviving funds. In other words, the fund-specific volatility of luck seems to matter for exit as well as the specific realisation of luck.

Figure 3.6: The effect of competition and luck on mutual fund size

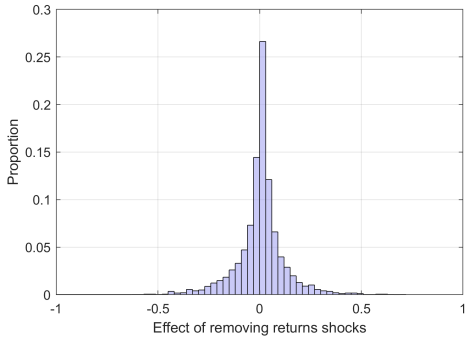
(a) The effect of competition



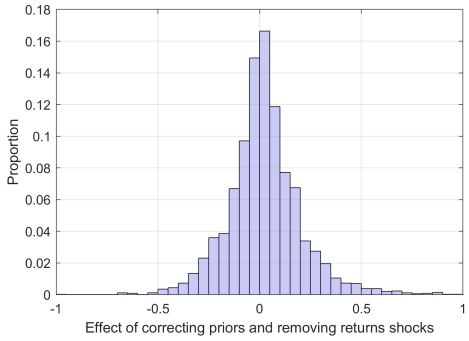
(b) The effect of incorrect priors



(c) The effect of return shocks



(d) The effect of incorrect priors and returns shocks



Note: I show the fund-specific percentage change in fund size resulting from a counterfactual change. In panel (a), I remove the effect of local competition and show that funds would be significantly bigger, to varying degrees, without local competition. In panels (b), (c) and (d), I show the impact of correcting for various forms of luck, where a positive (negative) number indicates the fund was unlucky (lucky) because correcting for luck makes the fund bigger (smaller). In panel (b) I correct investor priors about funds, in panel (c) I remove return shocks and in panel (d) I correct priors and remove return shocks.

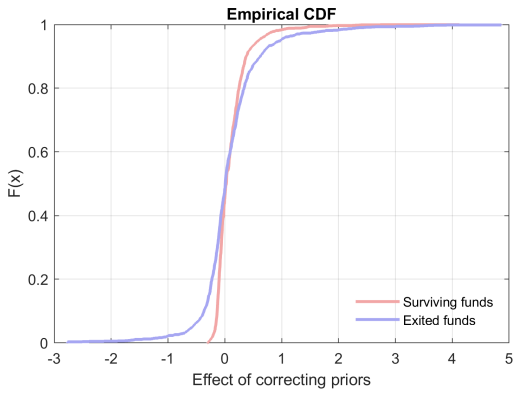
Table 3.3: Differences between exiting and surviving funds

	[1]	[2]
$\%\Delta q$	Exiting funds	Surviving funds
Correcting priors		
Mean value	0.083	0.102
Median value	0.008	0.026
Mean absolute value	0.357	0.189
Median abs. value	0.213	0.123
Removing return shocks		
Mean value	0.011	0.001
Median value	0.011	0.004
Mean absolute value	0.119	0.033
Median abs. value	0.056	0.015

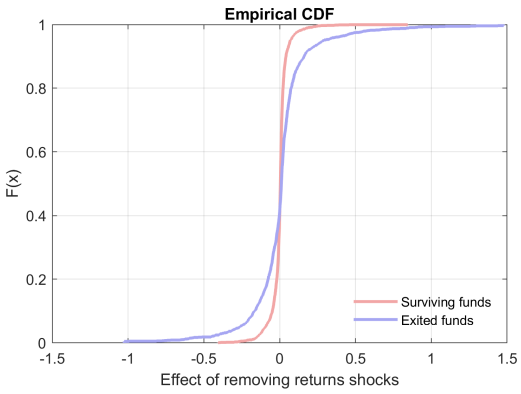
Note: I summarise the percentage change in mutual fund size resulting from (1) correcting priors and (2) removing return shocks, where the bigger the number the more unlucky the fund. I do this for funds that exited during my sample period and funds that survived. I find that on average exiting funds were luckier than surviving funds in their draw of investor prior beliefs, but unluckier in their return shocks. Exiting funds were more affected by luck in absolute terms than surviving funds, indicating their priors and returns were more volatile.

Figure 3.7: Differences between surviving and exiting funds

(a) The effect of incorrect priors



(b) The effect of return shocks



Note: This figure shows the cumulative distribution functions corresponding to panels (b) and (c) of Figure 3.6, but conditioning on whether the fund exited during my sample period or survived. As described in Figure 3.6, a positive (negative) number on the x-axis indicates the fund was unlucky (lucky). Exiting funds had more extreme lucky and unlucky outcomes.

3.7 Conclusion

I estimate a network of investment strategy overlap and show that a given fund's location within this network has a material impact on its size. I then build and use structural model of demand to show quantitatively how luck can have persistent and in some cases permanent effects on mutual fund outcomes even when funds are skilled and investors are rational.

To more fully understand local competition between mutual funds it is necessary to consider two further issues. First, I consider only the demand-side behaviour of investors, not the supply-side behaviour of funds when they decide to enter, exit or set fees. Modelling the supply-side is challenging in this context in which funds are heterogeneously located in the network, as it involves forming expectations over the dynamics of every other fund. It would, however, permit analysis of the equilibrium effects of counterfactual changes. Second, I take a fund's location within the network as given, but a natural starting point for further work would be to endogenise a fund's location choice (or, in other words, to consider network formation as well as network spillovers).

References

Berk, J. B. and Green, R. C. (2004). Mutual fund flows and performance in rational markets. *Journal of Political Economy*, 112(6):1269–1295.

Berk, J. B. and Van Binsbergen, J. H. (2015). Measuring skill in the mutual fund industry. *Journal of Financial Economics*, 118(1):1–20.

Bollen, N. P. and Busse, J. A. (2005). Short-term persistence in mutual fund performance. *The Review of Financial Studies*, 18(2):569–597.

Busse, J. A., Goyal, A., and Wahal, S. (2010). Performance and persistence in institutional investment management. *The Journal of Finance*, 65(2):765–790.

Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of finance*, 52(1):57–82.

Chen, J., Hong, H., Huang, M., and Kubik, J. D. (2004). Does fund size erode mutual fund performance? the role of liquidity and organization. *American Economic Review*, 94(5):1276–1302.

Chevalier, J. and Ellison, G. (1997). Risk taking by mutual funds as a response to incentives. *Journal of Political Economy*, 105(6):1167–1200.

Coen, P. (2020). Information loss over the business cycle. *Working paper*.

Cremers, K. M. and Petajisto, A. (2009). How active is your fund manager? a new measure that predicts performance. *The review of financial studies*, 22(9):3329–3365.

Elton, E. J., Gruber, M. J., Das, S., and Hlavka, M. (1993). Efficiency with costly information: A reinterpretation of evidence from managed portfolios. *The review of financial studies*, 6(1):1–22.

Fama, E. F. and French, K. R. (2010). Luck versus skill in the cross-section of mutual fund returns. *The journal of finance*, 65(5):1915–1947.

Gavazza, A. (2011). Demand spillovers and market outcomes in the mutual fund industry. *The RAND Journal of Economics*, 42(4):776–804.

Hoberg, G., Kumar, N., and Prabhala, N. (2018). Mutual fund competition, managerial skill, and alpha persistence. *The Review of Financial Studies*, 31(5):1896–1929.

Huang, J., Sialm, C., and Zhang, H. (2011). Risk shifting and mutual fund performance. *The Review of Financial Studies*, 24(8):2575–2616.

Kacperczyk, M., Nieuwerburgh, S. V., and Veldkamp, L. (2014). Time-varying fund manager skill. *The Journal of Finance*, 69(4):1455–1484.

Kacperczyk, M. and Seru, A. (2007). Fund manager use of public information: New evidence on managerial skills. *The Journal of Finance*, 62(2):485–528.

Kacperczyk, M., Van Nieuwerburgh, S., and Veldkamp, L. (2016). A rational theory of mutual funds' attention allocation. *Econometrica*, 84(2):571–626.

Kosowski, R., Timmermann, A., Wermers, R., and White, H. (2006). Can mutual fund "stars" really pick stocks? new evidence from a bootstrap analysis. *The Journal of finance*, 61(6):2551–2595.

Pástor, L. and Stambaugh, R. F. (2012). On the size of the active management industry. *Journal of Political Economy*, 120(4):740–781.

Pástor, L., Stambaugh, R. F., and Taylor, L. A. (2015). Scale and skill in active management. *Journal of Financial Economics*, 116(1):23–45.

Pollet, J. M. and Wilson, M. (2008). How does size affect mutual fund behavior? *The Journal of Finance*, 63(6):2941–2969.

Roussanov, N., Ruan, H., and Wei, Y. (2018). Marketing mutual funds. Technical report, National Bureau of Economic Research.

Wahal, S. and Wang, A. Y. (2011). Competition among mutual funds. *Journal of Financial Economics*, 99(1):40–59.