The London School of Economics and Political Science

Essays on Over-the-Counter Markets

Jamie 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, May 2022.

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 35,000 words.

Statement of co-authored work

I confirm that Chapters 2 and 3 were jointly co-authored with Patrick Coen and I contributed 50% of the work.

Acknowledgements

I am deeply grateful to my supervisor, Alessandro Gavazza, and to Patrick Coen for their advice and support throughout the PhD. I am grateful to the industrial organisation group at LSE for their support and feedback. I am also grateful to Daniel Albuquerque, Lu Liu, Will Matcham, Heidi Thysen and Di Song Tan for their help and friendship throughout the PhD.

I acknowledge the financial support of the Economics and Social Research Council and the Paul Woolley Centre. I am grateful to the Bank of England—and, in particular, Anne-Caroline Hüser, Caterina Lepore, Jack McKeown, Amar Radia, Laura Silvestri, Rhiannon Sowerbutts and Nick Vause—for providing coauthorship and support. The views in this thesis are my own, and not necessarily those of the Bank of England or its committees.

Abstract

This thesis consists of three chapters. The first chapter critically reviews the literature on over-the-counter financial markets, with a particular focus on studies that structurally estimate models of these markets. I consider the two dominant approaches—those based on network models and those based on models of random search—and discuss the merits and applicability of each in different contexts.

The second chapter, co-authored with Patrick Coen, studies how firm heterogeneity determines liquidity in over-the-counter markets. We build a model of search and trading in which firms have heterogeneous search costs, and estimate this model using granular data on trading in the secondary market for sterling corporate bonds. We show that the supply of liquidity is highly concentrated, with the 8% most active traders supplying as much liquidity as the remaining 92%. We draw implications for the resilience of these markets to shocks to these active traders, the impact of banking regulation, and why new trading technologies struggle to gain traction in these markets.

The third chapter, co-authored with Patrick Coen, studies the network of interconnections between banks. This network involves a fundamental trade-off: these interconnections enable banks to realise gains from trade, but expose them to risk. We build a network model of the interbank market to study this trade-off, and estimate it using novel data on interbank exposures. We show that the network is inefficient, in that a social planner can decrease risk without harming banks' surplus. We show existing measures of banks' systemic importance are misleading, as they don't take into account the fact that banks select large exposures where these exposures are safe. We show existing regulation—capital requirements and caps on banks' exposures—are inefficient, as they don't account for how banks form the network of exposures, and propose more efficient alternatives.

Contents

List of Tables

List of Figures

Chapter 1

Structural Approaches to Studying Over-the-Counter Markets.

In this chapter I review the literature on over-the-counter financial markets, with a particular focus on work that structurally estimates models of trading. I review the theoretical literature studying search and network frictions in these markets. I then review structural approaches, which take theories based on search and network frictions to data to build quantitative models. I consider the merits of these approaches and their usefulness for applied research.

1.1 Introduction

Over-the-counter (OTC) markets are decentralised: there is no central exchange on which trading takes place and trades tend to be organised bilaterally between firms. Trading in these markets is therefore relatively opaque. Many key financial assets trade over the counter, including corporate and government bonds, derivatives, and mortgages. These markets play a critical role in the efficient functioning of the financial system. As a result, there is significant interest amongst academics and policymakers in how these markets work, how efficient they are, and their resilience to shocks.

In this chapter I review the literature on OTC financial markets. The aim of the chapter is to review structural approaches, which combine theoretical models with microdata to produce quantitative models of trading. As a result, when reviewing the theoretical literature I focus on the types of models that have subsequently been taken to data, and when reviewing the empirical literature I focus on those papers that are most closely related to theory. I seek to draw out the strengths and weaknesses of different approaches and assess their usefulness for informing applied research into OTC markets.

1.2 Theoretical Work

OTC markets are not well described by Walrasian equilibrium. In particular, many key OTC financial markets exhibit significant price dispersion (Li and Schürhoff, 2019), transaction costs (Edwards et al., 2007), network relationships (Hendershott et al., 2020), and a role for intermediation (Friewald and Nagler, 2019). Theoretical approaches to studying these markets are based around frictions that can replicate some of these features. These include market power, asymmetric information, frictions in funding markets, frictions that prevent traders accessing the market at certain times, search frictions, and network frictions.[1] In this section, I focus on approaches based on network and search frictions, as these are the approaches that usually form the basis of structural approaches to OTC markets.

[1] See Vayanos and Wang (2013) Vayanos et al. (2012) for a unified framework that encapsulates different frictions studied in OTC markets.

1.2.1 Search Frictions

Search models are based on the observation that when a firm wants to trade, meeting a counterparty and executing a trade is not instantaneous.[2] In these models an opportunity to trade takes time to arrive, and when it does it is with a random counterparty drawn from the distribution of traders. This search friction underlies illiquidity and intermediation in OTC markets, and can rationalise various patterns in prices and trading in the data.

A key paper in this literature is Duffie et al. (2005). In their model two sets of traders—dealers and customers—search for and trade a single indivisible asset with each other. Customers receive shocks to their values of holding an asset that create variation in values across customers and hence gains from trade. Each customer meets a counterparty at a rate governed by a Poisson process, and these agents then engage in Nash bargaining over whether and at what price to trade a single unit of the asset with each other. Dealers trade in a frictionless market where they can immediately offload inventory. They show that this model can explain a number of empirical trading patterns.

Subsequent papers have extended this approach in several directions. Lagos and Rocheteau (2007, 2009), Gârleanu (2009) and Üslü (2019) allow traders to trade and hold an unconstrained amount of the asset, such that they can study the role of inventory in driving trading patterns and explore the intensive margin of trade size as well as the extensive margin of trading frequency. Hugonnier et al. (2020) and Neklyudov (2019) introduce a decentralised inter-dealer market, whilst Neklyudov (2019) additionally introduces heterogeneous dealer meeting rates. Liu (2020) endogenises the frequency with which dealers contact customers. Lagos and Rocheteau (2007) endogenise the set of active dealers in the market to study entry and exit in OTC markets. Vayanos and Weill (2008) and Vayanos and Wang (2007) study settings with two assets with identical cashflows and show liquidity can endogenously concentrate in one asset. Farboodi et al. (2018) study how heterogeneous bargaining powers can determine who becomes an intermediary, whilst Farboodi et al. (2021) endogenise traders' choice of market structure. Weill (2007) studies how OTC markets behave out of steady state.

The search approach is a convenient methodological approach and is well suited to studying a number of phenomena in financial markets. It has yielded several important insights into OTC markets, including understanding why certain traders become intermediaries in

[2]The literature on search in OTC markets builds on vast earlier literatures, most notably in labor (see Pissarides, 2000) and in monetary economics (see Kiyotaki and Wright, 1993).

OTC markets (Farboodi et al., 2018), why liquidity differs between markets for assets with identical cashflows (Vayanos and Weill, 2008), and the role that marketmakers play in supporting market functioning by absorbing sales of assets by other traders (Weill, 2007).

1.2.2 Network Frictions

Network models of OTC markets are motivated by the fact that financial firms persistently trade with some counterparties and not with others. This suggests that in some sense relationships are formed in OTC markets, and these relationships shape trading. The network literature as a whole studies the formation of this network of trading relationships and its implications for trading. Within this, papers typically focus on understanding how and why a network with a particular structure is formed, or take the network as given and study its implications for trading, market functioning and welfare.[3] Much of this literature is focused on networks between banks given their importance in the financial system and the role the interbank network played in the 2008/09 financial crisis.

A small set of papers seeks to understand why banking networks tend to exhibit a core-periphery structure, consisting of a small set of highly connected banks and a periphery of banks with few connections, and what this means for the efficiency of the financial system.[4] These papers set out environments where banks face costs and benefits of forming links to each other, and study the network structure that emerges. In Castiglionesi and Navarro (2020), banks forming links with other banks face a trade-off: forming a link improves diversification but exposes banks to the risk of counterparties going bankrupt. They show that both the equilibrium network and efficient network exhibit a core-periphery structure. In Farboodi (2021), banks with a risky investment opportunity form the core of a network of interbank lending, with other banks on the periphery lending to the core banks. Both these papers find that the equilibrium network structure is inefficient.[5]

A larger literature takes the network structure as given and studies the implications of different types of structure for market outcomes. One key strand of literature studies how network structure impacts welfare and systemic risk. A foundational contribution to this literature was made by Allen and Gale (2000), who show that a complete network

[3]See Allen and Babus (2009) for an earlier survey of network models in finance and Glasserman and Young (2016) for a survey of the literature on contagion in financial networks.

[4]See Langfield et al. (2014) for a description of the interbank market structure in the UK.

[5]Earlier work by Babus (2016) also studies the formation of interbank networks. She shows that banks form risk-sharing networks, and that more connected networks exhibit less risk of contagion.

between banks is more robust than an incomplete network.[6] More recently, Acemoglu et al. (2015) study how the network of connections between banks mitigates or amplifies systemic risk. They show that a denser network is stability-enhancing when shocks are small, but amplifies larger shocks. Elliott et al. (2014) show that integration—how exposed banks are to each other—and diversification—the extent to which these exposures are spread out across multiple banks—have non-monotonic effects on the likelihood of cascades of default. Gofman (2017) studies the efficiency of trading in a network, and shows that decentralised markets are unlikely to be efficient where networks are incomplete and intermediation is required.[7]

Taken together, these papers show that simple models with network frictions can explain a key feature of many OTC markets: a core-periphery structure whereby a key set of firms emerge as intermediaries, connecting a larger set of fringe firms. This structure need not be the efficient structure—other alternative networks of relationships could be welfare improving, and this superior network could be denser or sparser or more or less concentrated. And given the existing network structure, the outcomes of trading in these markets need not be efficient.

1.3 Structural Approaches

A large literature studies OTC markets empirically, documenting features of these markets and using these features to discriminate between different economic theories.[8] A particularly relevant set of papers studies networks in financial markets, including Iyer and Peydro (2011), Langfield et al. (2014), Brunnermeier et al. (2013), Furfine (2003) and Hüser et al. (2021). Empirical studies have also addressed the role of policy in shaping financial markets, for example Kotidis and Van Horen (2018) and Adrian et al. (2017), and changes in trading technologies through time (Biais and Green, 2019).

The bulk of this literature is reduced form, not focusing on a specific theory or market friction proposed to explain trading. In this section I review the growing literature that combines theoretical models of trading with microdata on trading and contractual relationships between financial institutions to produce quantitative frameworks.

Structural approaches to OTC markets are typically based either on search frictions or

[6]Notable subsequent papers include Freixas et al. (2000), Leitner (2005) and Dasgupta (2004).

[7]Other papers to study the impact of network frictions in OTC markets include Colliard et al. (2021), Cabrales et al. (2017), Babus and Kondor (2018) and Chang and Zhang (2021).

[8]See for example Di Maggio et al. (2017), Li and Schürhoff (2019) and Hollifield et al. (2017).

on network frictions.[9] The search-based literature uses moments about trading and prices to infer parameters governing traders' flow utilities from holding the assets, the process by which these utilities are shocked, the technology by which traders contact each other, and the bargaining powers of different agents. The network-based literature typically begins with data on the network of relationships between firms, supplemented with data on prices and quantities traded along on this network. It then either takes this network as given and seeks to recover parameters that govern the game played on this network, or treats the network as endogenous and seeks to recover parameters governing the construction of the network as well as the game on the network.

1.3.1 Search Frictions

A relatively small set of papers structurally estimates models of search in financial markets. Typically, these papers take models that build on Duffie et al. (2005) and estimate them using microdata on trading in a given financial instrument. Feldhütter (2012) estimates a version of the model in Duffie et al. (2005) adapted to studying corporate bonds, whilst Gavazza (2016) estimates a search model of a real asset market. Brancaccio et al. (2020) present a model where traders learn about asset valuations by trading in a market characterised by search frictions, and estimate it using transaction data on the secondary market for municipal bonds in the US. Liu (2020) uses data on corporate bond trading in the US to estimate a model of endogenous dealer search.

In each of these papers traders trade a single indivisible unit of an asset, and meet each other at a frequency that is exogenous. Two papers relax these assumptions.[10] Brancaccio and Kang (2021) estimate a structural model of trading in the market for municipal bonds based on the model presented in Üslü (2019). In the model inventories and trading quantities are unconstrained, and dealers choose how hard to search for customers. They use this model to show that bond underwriters add complexity to bond characteristics when it is issued, in order to subsequently profit from trading the bond in the secondary market. Pinter and Uslu (2022) also estimate a structural model based on Üslü (2019), where quantities and holdings are unrestricted but search intensity is fixed. They use this model to quantify differences in

[9] An exception to this is Allen and Wittwer (2021). This paper studies a model where the frictions are (a) only a subset of traders are willing and able to trade each period and (b) there is a network structure in that investors are connected with a single 'home' dealer who has a degree of market power.

[10] In Chapter 2 I present work that also relaxes these assumptions, that was undertaken contemporaneously to these two papers.

frictions between corporate and government bond markets in the UK.

Structural models of search in OTC markets have yielded a number of valuable results. These include (a) quantification of the roles of frictions in shaping market outcomes and (b) counterfactual analyses relating to policies and shocks. For example, Gavazza (2016) shows that trading frictions reduce prices in the market for business aircraft by 20%, and that whilst the presence of dealers reduces frictions they also extract rents, such that welfare would be higher in a market without dealers. These types of quantitative results are only possible with a structural approach.

1.3.2 Network Frictions

A complementary literature estimates models of financial markets with network frictions. Unlike the literature estimating search models—which generally shares a set of common features found in Duffie et al. (2005)—these papers adopt a diverse set of modelling and estimating approaches. The aspect they do share is the existence of a network friction, which prevents certain sets of traders from trading with each other.

Mirroring the theoretical literature, structural models of the financial network can be divided into those that treat the network as fixed and those that treat it as endogenous. Eisfeldt et al. (2020) present a model of the market for credit default swaps. Traders on a fixed network trade credit default swaps to adjust their exposure to underlying default risk. They show that the market is vulnerable to shocks to key traders. Denbee et al. (2021) estimate a network model of the UK interbank market. Banks take the network as given, and choose the quantity of reserves they hold and their supply and demand of liquidity to other banks. They show that the extent to which the network amplifies or dampens shocks to banks depends on market conditions and network structure.

Very few papers endogenise the formation of the network. Craig and Ma (2021) presents a model of endogenous network formation in the German interbank market. The model rationalises the observed lending relationships between banks with unobserved monitoring costs, and the structural estimation recovers the distribution of these costs. They use their estimated model to show how network relationships can propagate shocks to one intermediary to many borrowing banks. Blasques et al. (2018) estimate a dynamic model of the Dutch overnight interbank lending market. In the model banks endogenously form a network of contacts between each other and lend and borrow on this network. Credit risk and the need to monitor peers leads to a core-periphery structure emerging.

The structural network literature has shed light on the role of the network in determining market outcomes, and the forces that drive network formation. An example of the former is the result in Eisfeldt et al. (2020) that market outcomes are highly vulnerable to shocks to key players. An example of the latter is Craig and Ma (2021), who show that a model of monitoring costs can quantitatively explain how the network of interbank lending is formed and how it changes in a stress.

1.4 Discussion

Structural models of search and networks can yield significant insight into OTC markets. They can test theoretical models and show how well they fit the data in reality. They can take conditional results from the theoretical literature—for example, that a denser network is more desirable when shocks are small, but less desirable when shocks are large (Acemoglu et al., 2015)—and tell us which condition is empirically relevant. They can quantify market inefficiencies, and be used to study welfare and optimal policy. In this section I discuss the strengths and weaknesses of different structural approaches to OTC markets, and the questions and settings they are best suited to addressing.

The key strength of the search paradigm is its tractability: in a dynamic setting with heterogeneous agents it offers clean results. Inevitably, this tractability involves some counterfactual predictions, with these models typically unable to replicate the persistent trading relationships seen in markets. Additionally, its assumption of atomistic traders—critical for its tractability—makes it unsuited to studying issues around market power.

More fundamentally, search frictions are a relatively abstract concept. Which costs do search costs capture? Is it the cost of a trader picking up the phone to call a counterparty, or the cost of hiring a trader to do so? Is it some cost of inspecting the asset or trader, or of executing a trade? For many questions these issues do not need to be addressed—the mere presence of a friction, together with the convenient framework that the search paradigm provides, is sufficient. In work that seeks to make quantitative predictions in applied areas, the need to understand what underlies search frictions is perhaps more pressing.

Taking search models to data presents additional challenges, which limit the questions this approach can answer. Take, for example, papers that seek to quantify the extent of search frictions. The most sophisticated existing model that can be solved in analytical form is Üslü (2019). This paper has a continuous shock distribution from which traders' valuations are drawn, zero transaction costs, and fixed meeting intensities across traders. A consequence of these three assumptions is that when two traders meet, they almost always trade. To rationalise the infrequent trading we typically observe in OTC markets (Coen and Coen, 2021), it must be the case that traders meet extremely infrequently and thus search frictions are high. But an alternative model with non-zero transaction costs could rationalise the same data with very low search frictions: traders could meet very frequently, but trade infrequently due to transaction costs. This observation does not invalidate models with search frictions and no transaction costs, but implies caution should be exercised when interpreting these kinds of findings.

The counterfactual questions that can be answered using search models depend on what can reasonably be held constant in a counterfactual. For example, take a counterfactual that studies what happens in an OTC market when a set of traders steps back from the market and stops trading—a question we consider in Chapter 2. In the long run the market is likely to respond, with new traders entering and other traders adjusting their business models. Estimated parameters are unlikely to be informative about what happens in this scenario. However, in the shorter run these types of adjustment may not be feasible. Thus the results of the counterfactual are much more convincing when interpreted as short run results rather than long run.

Constraints on data availability mean that there are many unanswered questions in this literature, and many avenues for future research. The most commonly used datasets in this field only give the identities of a small subset of traders—firms who traditionally act as dealers—with all other traders grouped together under a single anonymous identifier.[11] As a result most papers are unable to study *who* becomes an intermediary, and instead are forced to exogenously designate dealers as intermediaries. Additionally, theoretical papers point to the importance of a trader's inventory in determining their trading behaviour (Üslü, 2019), but these data have historically unavailable and so this is not studied empirically. Newly available datasets on all firms' trading activity[12] and trader inventories[13] mean these topics can now be studied empirically.

A key challenge for modelling financial networks is how to appropriately model the en-

[11]Existing studies typically rely on TRACE data on US corporate bond transactions (for example Choi and Huh (2018)) or US municipal bond transactions (for example Brancaccio et al. (2020)). In TRACE data, any customer is simply marked as 'C', meaning customers can neither be identified nor tracked through the data. The same is true of the municipal bond data.

[12]Chapter 2 describes such a dataset for the Sterling corporate bond market.

[13]In further work I am studying the impact of dealers' inventories of bonds on their behaviour, using a unique dataset on dealers' holdings of corporate and government bonds.

dogeneity of the network. Models with endogenous network models are difficult to solve, meaning much of the literature treats the network as fixed. In reality, financial institutions form links for a reason. Thus when conditions change in a market, there is incentive for the network to rearrange. Some sense of the dynamics of network formation is likely to be important here. A network exists which is in some sense 'sticky'—if this is not true, then the network has no real significance in a market. This stickiness shapes some shorter-run features of markets, for example trading on this network. But in the longer run, the network is endogenous—people choose their relationships. A discussion of the dynamics of network formation, with network links endogenous but in some sense enduring, would be worthwhile.

The endogeneity of the network poses further problems when estimating these models and using the resulting quantitative models to study counterfactuals. Firstly, in estimation, the endogeneity of the network leads to concern about endogeneity in estimation. Suppose a researcher seeks to understand the impact of the interbank network on banks' risk—a question we consider in Chapter 3. Regressing a measure of a bank's risk on a measure of its network exposures would ignore the fact that banks likely choose their exposures to minimise their risk, and thus would produce inconsistent estimates of the network's impact. Secondly, in counterfactual scenarios traders have incentives to change the network. Keeping the network fixed limits a margin of adjustment that traders are likely to adjust in counterfactuals.

A key avenue for future empirical work is to use the data to understand exactly in which senses the network is an endogenous object. At what time horizon is the network fixed? In response to which events are financial institutions willing to create new relationships? Newly available datasets containing information on contractual relationships of different maturity, and how these evolve through time, are well suited to answering these questions.[14]

Some of the shortcomings of structural models of search and networks are common to both approaches. For example, the difficulty in understanding exactly what underlies search costs is analogous to the difficulty in understanding what drives network formation—both involve taking the existence of the key friction as in some sense given. And the assumption that traders do not adjust their search behaviour in counterfactual scenarios is similar to assuming that the network does not reorganise. These are common issues in structural modelling and in both cases require careful interpretation of any counterfactual analysis.

Additionally, almost all existing structural papers of financial markets—both search models and network models—focus on a single asset market, most commonly bonds or interbank

[14] These datasets are described in Coen and Coen (2019) and Hüser et al. (2021).

lending.[15] In reality, the same sets of traders trade multiple types of assets with each other, both within a broader asset class—the same firms tend to trade many different corporate bonds with each other—and across asset classes—firms that trade corporate bonds may well also interact in credit and derivatives markets, amongst others. These cross-asset interactions may well involve cross-subsidisation; for example dealers may offer better prices in corporate bond trading to firms with which they have valuable derivatives business. Further study of these types of issues, by combining the microdata available on different asset markets, is a worthwhile avenue for research.

The random search approach is perhaps best applied in spot trading markets for simple financial instruments like bonds, as in Brancaccio et al. (2020), Feldhütter (2012), and Liu (2020). This is because search models are well suited to capturing the dynamics inherent in bond market trading, and practically speaking the search approach is difficult to apply to more complex financial transactions than simply buying or selling a bond.[16]

By contrast, the network approach is better placed to study markets like interbank markets (Coen and Coen, 2019; Denbee et al., 2021) or credit default swaps (Eisfeldt et al., 2020). Network relationships seem particularly prevalent in these markets (Coen and Coen, 2019), perhaps because the instruments being traded themselves include a time dimension. Further, in some of these settings dynamics are perhaps less important—for example Denbee et al. (2021) study overnight lending in the UK interbank market, where contracts are effectively reset each morning.

1.4.1 This dissertation

In Chapter 2, I present a paper (joint with Patrick Coen) that structurally estimates a search model of trading in an OTC market. We adopt a search approach because of the setting of our paper: the secondary market for Sterling corporate bonds. Dynamics are critical in these markets: bonds trade rarely, and a given firm's traders are clustered together in time. Search models are well suited to capturing trading dynamics, and so are particularly convenient in these settings. The particular features of the search model we use are informed by patterns

[15] Pinter and Uslu (2022) is a notable exception.

[16] In this simple setting, the only state variables for a trader are their fixed type (e.g. a search intensity), their current valuation of the asset, and their holding of the asset. Where a contract between firms has a time dimension—for example in a lending relationship or a swap—a trader's state will include all outstanding contracts in which they're engaged and the time they mature. This, combined with the key assumption that shocks and trades arrive according to Poisson processes, would render these models wholly intractable.

we observe in the data. In particular, facts we establish around firm heterogeneity and the nature and timing of different firms' trading lead us to adopt a model with endogenous search intensity, heterogeneous search costs, and endogenous intermediation.

In Chapter 3, I present a paper (joint with Patrick Coen) that structurally estimates a network model of the interbank market. The setting is the network of direct exposures between UK banks. There is significant persistence in patterns of exposures in this market through time, indicative of a persistent network of relationships between firms. Despite this persistence, there is both heterogeneity and variation through time in exposures between banks. As a result, we adopt a network approach to modelling this market, where the formation of the network is endogenous.

Bibliography

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

Adrian, T., Fleming, M., Shachar, O., and Vogt, E. (2017). Market liquidity after the financial crisis. *Annual Review of Financial Economics*, 9:43–83.

Allen, F. and Babus, A. (2009). Networks in finance. *The network challenge: strategy, profit, and risk in an interlinked world*, 367.

Allen, F. and Gale, D. (2000). Financial contagion. *Journal of Political Economy*, 108(1):1–33.

Allen, J. and Wittwer, M. (2021). Centralizing over-the-counter markets. Working Paper, Stanford University.

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

Babus, A. and Kondor, P. (2018). Trading and information diffusion in over-the-counter markets. *Econometrica*, 86(5):1727–1769.

Biais, B. and Green, R. (2019). The microstructure of the bond market in the 20th century. *Review of Economic Dynamics*, 33:250–271.

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.

Brancaccio, G. and Kang, K. (2021). 'Intermediaries and product selection in the municpal bond market'. Unpublished Working Paper.

Brancaccio, G., Li, D., and Schürhoff, N. (2020). 'Learning by trading: The case of the US market for municipal bonds'. Working Paper, New York University Stern School of Business.

Brunnermeier, M. K., Clerc, L., Omari, Y. E., Gabrieli, S., Kern, S., Memmel, C., Peltonen, T., Podlich, N., Scheicher, M., and Vuillemey, G. (2013). Assessing contagion risks from the cds market. *ESRB: Occasional Paper Series*, (2013/04).

Cabrales, A., Gottardi, P., and Vega-Redondo, F. (2017). Risk sharing and contagion in networks. *The Review of Financial Studies*, 30(9):3086–3127.

Castiglionesi, F. and Navarro, N. (2020). (in) efficient interbank networks. *Journal of Money, Credit and Banking*, 52(2-3):365–407.

Chang, B. and Zhang, S. (2021). Endogenous market making and network formation. Working Paper, London School of Economics.

Choi, J. and Huh, Y. (2018). 'Customer liquidity provision: Implications for corporate bond transaction costs'. Working paper, University of Illinois Urbana-Champaign.

Coen, P. and Coen, J. (2019). 'A structural model of interbank network formation and contagion'. Working paper, London School of Economics.

Coen, P. and Coen, J. (2021). 'A structural model of liquidity in over-the-counter markets'. Working paper, London School of Economics.

Colliard, J.-E., Foucault, T., and Hoffmann, P. (2021). Inventory management, dealers' connections, and prices in over-the-counter markets. *The Journal of Finance*, 76(5):2199–2247.

Craig, B. R. and Ma, Y. (2021). 'Intermediation in the interbank lending market'. Working Paper, Columbia Business School.

Dasgupta, A. (2004). Financial contagion through capital connections: A model of the origin and spread of bank panics. *Journal of the European Economic Association*, 2(6):1049–1084.

Denbee, E., Julliard, C., Li, Y., and Yuan, K. (2021). Network risk and key players: A structural analysis of interbank liquidity. *Journal of Financial Economics*, 141(3):831–859.

Di Maggio, M., Kermani, A., and Song, Z. (2017). The value of trading relations in turbulent times. *Journal of Financial Economics*, 124(2):266–284.

Duffie, D., Gârleanu, N., and Pedersen, L. H. (2005). Over-the-counter markets. *Econometrica*, 73(6):1815–1847.

Edwards, A. K., Harris, L. E., and Piwowar, M. S. (2007). Corporate bond market transaction costs and transparency. *The Journal of Finance*, 62(3):1421–1451.

Eisfeldt, A. L., Herskovic, B., Rajan, S., and Siriwardane, E. (2020). OTC intermediaries. Working paper, UCLA.

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

Farboodi, M. (2021). 'Intermediation and voluntary exposure to counterparty risk'. Working Paper, Massachusetts Institute of Technology.

Farboodi, M., Jarosch, G., and Menzio, G. (2018). Intermediation as rent extraction. Technical report, National Bureau of Economic Research.

Farboodi, M., Jarosch, G., and Shimer, R. (2021). 'The emergence of market structure'. Working Paper, Massachusetts Institute of Technology.

Feldhütter, P. (2012). The same bond at different prices: identifying search frictions and selling pressures. *The Review of Financial Studies*, 25(4):1155–1206.

Freixas, X., Parigi, B. M., and Rochet, J.-C. (2000). Systemic risk, interbank relations, and liquidity provision by the central bank. *Journal of Money, Credit and Banking*, pages 611–638.

Friewald, N. and Nagler, F. (2019). Over-the-counter market frictions and yield spread changes. *The Journal of Finance*, 74(6):3217–3257.

Furfine, C. H. (2003). Interbank exposures: Quantifying the risk of contagion. *Journal of Money, Credit and Banking*, pages 111–128.

Gârleanu, N. (2009). Portfolio choice and pricing in illiquid markets. *Journal of Economic Theory*, 144(2):532–564.

Gavazza, A. (2016). An empirical equilibrium model of a decentralized asset market. *Econometrica*, 84(5):1755–1798.

Glasserman, P. and Young, H. P. (2016). Contagion in financial networks. *Journal of Economic Literature*, 54(3):779–831.

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

Hendershott, T., Li, D., Livdan, D., and Schürhoff, N. (2020). Relationship trading in over-the-counter markets. *The Journal of Finance*, 75(2):683–734.

Hollifield, B., Neklyudov, A., and Spatt, C. (2017). Bid-ask spreads, trading networks, and the pricing of securitizations. *The Review of Financial Studies*, 30(9):3048–3085.

Hugonnier, J., Lester, B., and Weill, P.-O. (2020). Frictional intermediation in over-the-counter markets. *The Review of Economic Studies*, 87(3):1432–1469.

Hüser, A.-C., Lepore, C., and Veraart, L. A. M. (2021). How do secured funding markets behave under stress? Evidence from the gilt repo market.

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.

Kiyotaki, N. and Wright, R. (1993). A search-theoretic approach to monetary economics. *The American Economic Review*, pages 63–77.

Kotidis, A. and Van Horen, N. (2018). Repo market functioning: The role of capital regulation. Bank of England Working Paper 746.

Lagos, R. and Rocheteau, G. (2007). Search in asset markets: Market structure, liquidity, and welfare. *American Economic Review*, 97(2):198–202.

Lagos, R. and Rocheteau, G. (2009). Liquidity in asset markets with search frictions. *Econometrica*, 77(2):403–426.

Langfield, S., Liu, Z., and Ota, T. (2014). Mapping the UK interbank system. *Journal of Banking & Finance*, 45:288–303.

Leitner, Y. (2005). Financial networks: Contagion, commitment, and private sector bailouts. *The Journal of Finance*, 60(6):2925–2953.

Li, D. and Schürhoff, N. (2019). Dealer networks. *The Journal of Finance*, 74(1):91–144.

Liu, S. (2020). 'Dealers' search intensity in US corporate bond markets'. Working paper, Tsinghua University.

Neklyudov, A. (2019). Bid-ask spreads and the over-the-counter interdealer markets: Core and peripheral dealers. *Review of Economic Dynamics*, 33:57–84.

Pinter, G. and Uslu, S. (2022). Comparing search and intermediation frictions across markets. Working paper, Bank of England.

Pissarides, C. A. (2000). *Equilibrium unemployment theory*. MIT press.

Üslü, S. (2019). Pricing and liquidity in decentralized asset markets. *Econometrica*, 87(6):2079–2140.

Vayanos, D. and Wang, J. (2013). 'Market liquidity - theory and empirical evidence'. In *Handbook of the Economics of Finance*, volume 2, pages 1289–1361. Elsevier.

Vayanos, D., Wang, J., et al. (2012). Theories of liquidity. *Foundations and Trends in Finance*, 6(4):221–317.

Vayanos, D. and Wang, T. (2007). Search and endogenous concentration of liquidity in asset markets. *Journal of Economic Theory*, 136(1):66–104.

Vayanos, D. and Weill, P.-O. (2008). A search-based theory of the on-the-run phenomenon. *The Journal of Finance*, 63(3):1361–1398.

Weill, P.-O. (2007). Leaning against the wind. *The Review of Economic Studies*, 74(4):1329–1354.

Chapter 2

A Structural Model of Liquidity in Over-the-Counter Markets.

with Patrick Coen (Toulouse School of Economics).[1]

We study how firm heterogeneity determines liquidity in over-the-counter markets. Using a rich dataset on trading in the secondary market for sterling corporate bonds, we build and estimate a flexible model of search and trading in which firms have heterogeneous search costs. We show that the 8% most active traders supply as much liquidity as the remaining 92%. Liquidity is thus vulnerable to shocks to these firms: if the 4% most active traders stop trading, liquidity falls by over 60%. Bank capital regulation reduces the willingness of these active traders to hold assets and thus reduces liquidity. However, trader search, holdings and intermediation respond endogenously to reduce the welfare costs of regulation by 30%. These costs are greater in a stress, when these margins of adjustment are constrained. The introduction of trading platforms, which homogenise the ability of traders to trade frequently, improves aggregate welfare but harms the most active traders who currently profit from supplying liquidity.

[1]The expressed are those of the authors, and not necessarily those of the Bank of England or its committees. We are particularly grateful to our supervisor, Alessandro Gavazza. We are also grateful for comments by Giulia Brancaccio, Eddie Gerba, Péter Kondor, Will Matcham, Ben Moll, Martin Oehmke, Martin Pesendorfer, Mark Schankerman, Pasquale Schiraldi, Nick Vause and Shengxing Zhang, as well as seminar participants at the Bank of Canada, Bank of England, Brisol, CEMFI, Duke Fuqua, EIEF, the Fed Board, HEC Paris, Imperial, LSE, Tilburg, UNC Kenan-Flagler and the Financial Markets and Corporate Governance Conference 2022. The authors acknowledge the financial support of the Economic and Social Research Council, the Paul Woolley Centre at LSE, the French National Research Agency and the TSE-Partnership foundation.

2.1 Introduction

Key financial assets trade in over-the-counter (OTC) markets, where firms seeking to trade must search for trading counterparties. Market liquidity is the extent to which these firms can trade quickly and at low cost, and is a key part of the efficient functioning of markets. It is an equilibrium outcome of firms' trading behaviour, with firms supplying liquidity by enabling their counterparties to trade. Trading activity varies significantly across firms: most trade very rarely, but some trade frequently.

We study the implications of this firm heterogeneity for market liquidity. What is the distribution of liquidity supply across frequent and infrequent traders? This leads us to study how this heterogeneity interacts with financial stress, policy and technology. How resilient is liquidity to shocks to these frequent traders? What is the impact of bank capital regulation, where banks are amongst the most frequent traders in these markets? What is the impact of introducing trading platforms, which homogenise the ability of traders to trade frequently?

We build our response to these questions around a rich dataset on trading in the secondary market for sterling corporate bonds. We begin by establishing some key empirical facts regarding heterogeneity in how frequently traders trade, who supplies liquidity, and how traders vary their trading behaviour to manage their balance sheets. Based on these facts, we develop a flexible model of search and trading in an OTC market. In the model, traders with *ex ante* heterogeneity in search costs and *ex post* heterogeneity in asset holdings and valuations search for and trade an asset with each other. Traders' search intensity, asset holdings, and trading quantities and prices are endogenously determined. Intermediation is endogenous, in that any trader can supply liquidity. We estimate the model using the identifying information in our data on corporate bonds, and show it fits these data well.

Our model enables us to quantitatively study market liquidity, who supplies it, and how it changes in counterfactual scenarios, whilst estimating the model ensures our results are grounded in the data. This enables us to make the following contributions. We show that 8% of traders supply as much liquidity as the remaining 92%. This asymmetry in the supply of liquidity shapes the impacts of stress, policy, and technology. We show that liquidity is highly vulnerable to shocks to these key liquidity suppliers: if the 4% most active traders stop trading, liquidity falls by over 60%. We show that bank capital regulation negatively affects liquidity by reducing the incentives of these key suppliers to hold assets, but search, holdings, and intermediation adjust to reduce the welfare costs of regulation by 30%. These adjustments are constrained during stress, which causes the negative effects of capital regulation to increase. We show that whilst trading platforms increase aggregate welfare, they harm the most active traders who see their returns from supplying liquidity decrease. We offer this as an explanation for why platforms have not supplanted bilateral trading as the dominant trading mechanism in OTC markets.

Additionally, we are the first to set out a search model of trading in an OTC market with endogenous search and unconstrained holdings. We show that traders adjust their search intensity to manage their balance sheets, searching harder when they are far away from their target holdings. We show this can explain key trading patterns in OTC markets.

The starting point for our analysis is rich data on the transactions, holdings, and identities of traders of sterling corporate bonds. The data are novel, relative to datasets commonly used in the literature, in two key ways: (1) They include the identities of *all traders*,[2] which means we can study how the characteristics of all traders drive their trading decisions; and (2) They include data on the *holdings* of firms as well as their transactions, enabling us to study the role of firms' inventories in driving trading patterns.

We present a number of empirical facts which inform our work. We show that the distribution of trading frequencies across traders is heavily skewed. On average, the 6% most frequent traders in a bond are responsible for 50% of the trades. We then show that the most frequent traders act as intermediaries, channelling assets between traders and sellers and earning a spread by doing so. Intermediation is not limited to firms traditionally seen as dealers in these markets: over a quarter of trades do not involve a dealer, suggesting that customers may be supplying liquidity to each other and to dealers.

Finally we show that firms actively vary their trading frequency to manage their balance sheets. In particular, traders offset trades by buying and then selling (or selling and then buying) the same bond within a short interval. They do this far more frequently than would be the case if trade orders arrived exogenously through time. We then compare dealers' tendency to offset their purchases of bonds to variation in capital regulation across bonds and through time, and show that offsetting is more frequent when capital regulation is tighter. This is consistent with capital regulation leading dealers to manage their balance sheets more tightly.

Our data and empirical findings guide our modelling decisions in the following ways. The heterogeneity in trading frequency across traders leads us to allow for *heterogeneity*

[2] Comparable datasets in the US, for example, only include the identities of firms that traditionally act as dealers in these markets.

in traders' search costs. The fact that a trader's trading frequency appears to respond to their state and the characteristics of the asset being traded leads us to consider *endogenous search intensity*. Given that trading occurs both between dealers and between customers, and the fact that our data contain the trading behaviour of all types of agents, we treat *intermediation as endogenous*.

We develop and estimate a model of trading in a decentralised financial market, building on Üslü (2019) by endogenising search intensity. In the model, a continuum of forward-looking traders trade an asset with each other in a market characterised by search frictions. Traders' holdings of the asset are unrestricted. Traders face random shocks to the utility they derive from holding the asset, which creates heterogeneity in liquidity needs and thus gains from trade. Traders choose how frequently to meet a counterparty subject to a convex cost of searching. These search costs vary across traders. A trader who meets a counterparty draws this counterparty randomly from the trading population. Intermediation in this model is endogenous, in the sense that rather than designating traders as intermediaries, we study how *ex ante* heterogeneity in search cost and *ex post* heterogeneity in liquidity needs across traders determine intermediation.

We estimate the model by matching theoretical moments to those we observe in the data. The novel aspects of our data, notably the presence of identifiers for all market participants, allows us to exploit the identifying information contained in a broad range of moments across and within traders. The search cost distribution is identified by the trading frequency distribution, whilst utility parameters and shock frequency are identified by the distribution of prices, trading quantities, and asset holdings.

We study how traders search in equilibrium. Traders increase their search intensity when their trading needs are high. This creates a convex pattern of trading frequency in traders' asset holdings, with traders searching harder when they are far from a target asset holding. In equilibrium there is wide heterogeneity in trading costs, with some traders finding it much more costly than others to meet counterparties. Traders with lower search costs are more willing to trade to extreme asset positions, as they are able to return to target holdings faster. This means that traders with low search costs naturally emerge as intermediaries.

We then run counterfactual analyses about how firm composition, financial stress, regulation, and technology interact with firm heterogeneity to determine liquidity. In the first counterfactual, we quantify traders' contributions to market liquidity. To do this, we 'withdraw' sets of traders from markets, which entails them selling their asset and stopping searching for counterparties. By studying how market liquidity changes when traders are withdrawn, we can quantify how much liquidity they supply.

The 8% most frequent traders supply as much liquidity as all other traders combined. As a result, market functioning is highly vulnerable to shocks that cause these traders to withdraw from markets. If only the 4% most frequent traders withdraw, liquidity falls by over 60% and price volatility more than doubles. This result is driven by the skewed distribution of estimated search costs, which in turn is driven by the skewed distribution of trading we observe in the data.

Second, we examine the potential effects of capital regulation on banks. Many of the most active traders in OTC markets are banks. Thus whilst there is broad agreement that tighter capital regulation enhances the stability of the banking system, there are concerns that, by reducing banks' willingness to hold assets, it also reduces market liquidity. To study this potential unintended consequence of capital regulation, we simulate an increase in the cost of holding assets for the most frequent traders only, whom we term dealers.

Liquidity declines, as dealers are less willing to take large positions in a bond. Dealers shrink their asset holdings and adjust their search intensity to more tightly manage their inventories. Unregulated traders pick up the slack by increasing their asset holdings, causing their welfare to increase. These endogenous responses mean that the welfare cost of regulation is 30% lower than it would be if traders did not adjust their behaviour. However, these impacts increase in periods when non-dealers are subject to stress, as they are less able to offset dealers' reduced willingness to hold the asset.

These findings rationalise a number of trends in markets in recent years, reported both in this paper and in the literature. Firstly, dealers have reduced their corporate bond inventories (Dick-Nielsen and Rossi, 2019). Secondly, dealers increasingly tightly manage their holdings, organising trades that offset incoming trade orders either instantly or soon after the trade (Schultz, 2017). Third, liquidity during stress events appears to have deteriorated relative to before the financial crisis (Bao, O'Hara and Zhou, 2018; Dick-Nielsen and Rossi, 2019). Fourth, traders who traditionally do not operate as intermediaries can make money by supplying liquidity (Choi and Huh, 2018; BlackRock, 2015; Li, 2021). Each of these can be rationalised as the product of regulation increasing dealers' costs of holding inventory, and the endogenous responses of traders.

In our third counterfactual analysis we study the impacts of technologies that decrease search costs, such as electronic many-to-many trading platforms. The bulk of corporate bond trades are organised bilaterally, between traders who communicate and bargain by phone. The same is true of the markets for derivatives and structured products. In recent years trading platforms have emerged as alternative trading mechanisms in these markets, where trades take place electronically and offered trades are posted to all participating traders rather than communicated bilaterally. These types of trading mechanisms offer clear efficiency benefits, though their progress has been relatively slow (The Economist, 2020).

We study the effects of these platforms by undertaking two counterfactuals: (a) reducing and homogenising search costs across traders; and (b) studying the outcome in a frictionless, Walrasian case. In each simulation, we find that platforms improve aggregate welfare, but reduce the welfare of traders with the lowest search costs. We argue that this helps to explain the relatively slow adoption of these technologies: the most frequent traders—whose participation is required for a platform to be viable—would lose out under platform-based trading.

Below we discuss our contribution to the literature. In Section 2.2 we describe our data and describe the institutional setting. In Section 2.3 we set out some key patterns in trading, intermediation and regulation in the sterling corporate bond market. In Section 2.4 we describe our model, and in Section 2.5 we describe how we estimate the model. In Section 2.6 we show the results of our structural estimation and discuss some key implications. In Section 2.7 we undertake counterfactual analyses, before concluding in Section 2.8.

2.1.1 Literature

This paper's contribution is to quantify the heterogeneity in the supply of liquidity across traders, and to study the implications of this for the resilience of liquidity to stress, the impact of bank capital regulation, and the impact of changes in trading technologies. By doing this we contribute to three strands of literature: (a) empirical work documenting the structure of OTC markets; (b) empirical work on the determinants of market liquidity, including the roles of trader stress, regulation and technology; and (c) work modeling search frictions in OTC markets.

An empirical literature uses transaction-level data to document features of OTC markets. Most notably, a series of papers document a robust pattern across OTC markets: there exists a core of traders who trade frequently and a periphery who trade rarely (Di Maggio, Kermani and Song, 2017; Li and Schürhoff, 2019; Hollifield, Neklyudov and Spatt, 2017). Our contribution is that, with our estimated model, we can quantify what this implies for the heterogeneity in firms' *supply of market liquidity*. Liquidity is not something that can be measured directly in the data, and a firm's role in determining liquidity is effectively a counterfactual question about what liquidity would be if that firm did not exist. Quantifying this, as our paper does, is key to understanding how liquidity changes when conditions in markets change.

Our ability to quantify traders' roles in supplying market liquidity enables us to contribute the study of OTC markets under stress. Eisfeldt, Herskovic, Rajan and Siriwardane (2020) study the impact of intermediary exit in the market for credit default swaps (CDS). They develop a network model of the CDS market, and show that the removal of a single intermediary can increase credit spreads by 20%. Our contribution is to quantify the reliance of *liquidity* on a set of key firms.

Our paper contributes to a literature studying banking regulation and market liquidity.[3] A set of empirical papers including Adrian, Fleming, Shachar and Vogt (2017), Bessembinder, Jacobsen, Maxwell and Venkataraman (2018), Bao, O'Hara and Zhou (2018), Schultz (2017), Dick-Nielsen and Rossi (2019) and Choi and Huh (2018) seeks to understand the effect of post-crisis regulatory changes on market liquidity by comparing measures of liquidity before and the after the financial crisis. A related set of papers studies violations of no-arbitrage conditions—including covered-interest parity (Du, Tepper and Verdelhan, 2018) and the relationship between bond yields and credit default swap rates (Duffie, 2010*b*)—and relates these to post-crisis banking regulation. Finally, two theoretical papers study the impact of capital regulation on liquidity. Cimon and Garriott (2019) and Saar, Sun, Yang and Zhu (2020) show that capital regulation incentivises dealers to intermediate in a way that minimises the inventory they hold. In particular, both show that capital regulation increases the extent to which dealers operate as 'matchmakers' between buyers and sellers, and decreases their incentives to take assets onto their balance sheets.

To the best of our knowledge, we are the first to study banking regulation and market liquidity in a structural context. This enables us to quantitatively identify the mechanisms by which regulation affects liquidity, and thus explain a number of recent trends in markets. It also enables us to study the impact of capital regulation in counterfactual scenarios, most notably in times of stress.

We also contribute to a literature studying how different trading mechanisms impact financial market outcomes (Allen and Wittwer, 2021; Hendershott and Madhavan, 2015;

[3] See Vayanos and Wang (2013) for a survey.

Plante, 2018). We show that distributional effects mean platform-based trading—even if it improves efficiency—may not be implemented.

A large theoretical literature studies search frictions in financial markets. Examples include Afonso and Lagos (2015), Brancaccio, Li and Schürhoff (2020), Duffie, Gârleanu and Pedersen (2005, 2007), Farboodi (2021), Farboodi, Jarosch and Shimer (2021),Gavazza (2016), Gârleanu (2009), Gromb and Vayanos (2018), Hugonnier, Lester and Weill (2020), Lagos and Rocheteau (2007), Lagos and Rocheteau (2009), Liu (2020), Neklyudov (2019), Pinter and Uslu (2022), Sambalaibat (2018) and Vayanos and Weill (2008).[4] Our model builds most closely on Üslü (2019), who sets out a theoretical model of endogenous intermediation between traders with different fixed trading speeds. We show empirically that traders appear to condition their trading frequency on their state and the state of the market. We thus endogenise firms' trading speeds, which gives us a more flexible model with which to run counterfactuals and enables us to explain key trading patterns.[5] We show how to identify and estimate this class of model, and in so doing are able to provide quantitative results. We demonstrate how firms adjust their search intensity to manage their trading portfolios, and how this changes in counterfactual scenarios.[6]

2.2 Data and Institutional Setting

2.2.1 Data

Our primary dataset is a database on corporate bond transactions maintained by the Financial Conduct Authority (FCA).[7] This contains trade-level data on secondary-market trades

[4]An alternative literature takes a network-based approach to studying OTC markets, relaxing the assumption that traders search randomly and studying the network of relationships between traders. Examples include Babus and Kondor (2018), Chang and Zhang (2021), Coen and Coen (2019) and Craig and Ma (2021).

[5]Liu (2020) also studies endogenous search, but in a setting with binary asset holdings, where traders are pre-assigned as dealers—who do not receive liquidity shocks—or customers. The fact that we leave holdings unconstrained means we can study how traders condition their search intensity on their asset holdings and how this impacts the distribution of their asset holdings, the fact that ours is a model of exogenous intermediation means we can study how search costs determine who intermediates, and the fact that we allow all traders to be hit by shocks enables us to study how traders use search to respond to shocks. Farboodi, Jarosch and Shimer (2021) study homogeneous traders' ex ante investment in a search technology, whereas we allow a trader's choice of search intensity to depend on both their type and their state.

[6]In independent, contemporaneous work Brancaccio and Kang (2021) study endogenous search with unconstrained holdings in the context of the US municipal bond market. They estimate a model of exogenous intermediation to study how search frictions interact with the features of bonds that municipalities issue.

[7]Other studies using these data include Czech and Roberts-Sklar (2019), Czech, Huang, Lou and Wang (2021) and Mallaburn, Roberts-Sklar and Silvestri (2019).

in bonds where at least one of the firms is an FCA-regulated entity. In practice this means that almost all financial firms with a legal entity in the UK (including subsidiaries of foreign banks) appear in the dataset, including both banks and non-banks. Our sample covers trading in sterling corporate bonds from January 2012 to December 2017. For each trade, we see who is buying and selling the bond, the price, the quantity traded, the instrument traded, and the time of the trade.

Relative to other datasets typically used in the literature, the advantage of this dataset is that it includes the identity of *all traders*. Traders in OTC markets are typically referred to as belonging to one of two groups: *dealers*—who traditionally supply liquidity in these markets—and *customers* consisting of all other traders. In our data, we take dealers to be firms that are permitted to trade in the primary market with national banks as well as inter-dealer brokers, who exist to facilitate trades between dealers. We refer to all other traders as customers. Dealers tend to be banks. Existing studies based on US data typically observe the identity of dealers, but not customers.[8] In our data we observe both dealers and customers.

This feature of the data is crucial for our paper. It means we can characterise the trading activity of all traders, and use this to study which traders take on different roles in the market. In particular, it allows us to study intermediation as an endogenous outcome, rather than assuming that dealers are the intermediaries. Our estimated impacts of stress, bank capital regulation and trading platforms on liquidity will depend critically on parameters in our model that are identified by the distribution of trading frequencies across traders. Without knowing the identities of all traders, we could not characterise this distribution.

We match these three datasets with information on bond characteristics and primary issuance from Thomson Reuters' Eikon database.

[8]Existing studies typically rely on TRACE data on US corporate bond transactions (for example Choi and Huh, 2018; Trebbi and Xiao, 2017; Kargar, Lester, Lindsay, Liu, Weill and Zuniga, 2021) or US municipal bond transactions (for example Brancaccio, Li and Schürhoff, 2020; Hugonnier, Lester and Weill, 2020). TRACE data comes in two forms: an academic version which includes anonymised dealer identifiers, and a regulatory version which reveals the dealer identities. In both cases, any customer is simply marked as 'C', meaning customers can neither be identified nor tracked through the data. The municipal bond dataset includes dealer identities, but again provides no information on the customer.

2.2.2 The Secondary Market for Sterling Corporate Bonds

The sterling corporate bond market is a key source of financing for both UK and non-UK firms. It has increased in importance since the financial crisis, with virtually all net financing raised by UK private non-financial firms from 2009 to 2016 coming in the form of bonds rather than bank lending (Bank of England, 2016). It is largely an over-the-counter market, with traders determining the terms of trade bilaterally, typically by phone (Anderson, Webber, Noss, Beale and Crowley-Reidy, 2015; Czech and Roberts-Sklar, 2019). Traders in the market consist of dealers, asset managers, insurance companies, hedge funds, and non-dealer banks. Dealers are a counterparty in around three-quarters of trades in our dataset. 80% of trades are carried out on a principal basis—where a firm who holds a bond sells the bond directly to a counterparty—rather than on an agency basis—where a firm acts as a middleman between two other trading counterparties, and at no point takes the bond onto its own balance sheet. Of these agency trades, the bulk are a trading counterparty buying for a non-trading client—for example a wealth manager buying bonds on behalf of their clients. 92% of trades between trading firms are carried out on a principal basis.

Table 2.1 gives key summary stats from our dataset. The average trading price is 108% of a bond's face value with significant variation around this figure. The median trade size is £100,000 with a tail of larger 'block trades' that skews the distribution right. Across the market there are around 1,000 firms trading each month. There is significant heterogeneity in trading activity both across instruments and across traders.

The market is illiquid. Trading tends to be rare, with the median bond trading once a month. There is large variation in price across bonds, and dispersion in price for the same bond: the R-squared from a regression of trading price on instrument fixed effects is 72%, with the remaining 28% reflecting within-bond price dispersion. Where a trader buys and sells a bond in the same week—a measure of the spread the trader earns—the median difference between the purchase price and the sale price is 0.13% of the bond's face value.

In the rest of this paper, our analysis is performed at the bond level. As such, we only include bonds where we have a meaningful number of trades with which to characterise trading. In what follows, we restrict the dataset to bonds that have been traded at least 10 times over our sample period of 6 years. This removes under 1% of the trades in our dataset. Further details on how we prepare the data are given in Section A1.

Table 2.1: Summary Statistics

	Mean	Std. Dev.	Median	25^{th} pctile	75^{th} pctile
Aggregate					
Price (% par)	108	13	106	101	114
Trade size (£000)	475	989	100	14	405
Monthly volume (£bn)	22	4	22	19	25
Monthly traders	975	68	980	929	1,021
Instrument-level					
Issuance (£mn)	135	403	10	3	200
Trades per month	10	26	1	0	7
Number of traders	35	66	4	2	31
Trader-level					
Monthly volume (£000)	10,456	102,915	30	2	458
Instruments traded	64	210	7	2	27
Trades per instrument traded	5	34	2	1	3

Note: This table gives some key descriptive statistics from our data. Aggregate statistics are computed across all instruments and all traders. Instrument-level statistics show how issuance and trading vary across instruments. Trader-level statistics show how trading activity varies across traders. Trades per instrument shows the distribution of the ratio between a trader's total trades and the number of instruments in which they trade.

2.2.3 Banking Regulation

Banks are subject to two types of capital regulation: a leverage requirement and a risk-weighted capital requirement. The leverage requirement states that a bank's equity must exceed a given fraction of its total assets. The risk-weighted capital requirement requires that a bank's equity exceeds a given fraction of its total risk-weighted assets.[9] Risk weights vary across bonds according to their creditworthiness, and capital requirements have more than doubled since the 2008 global financial crisis.[10]

2.3 Empirical Facts

We exploit the richness of our data to uncover a number of empirical facts that motivate our questions and guide our modelling. A key theme in these facts will be heterogeneity in the roles of different traders, both in terms of how frequently they trade and the roles they play.

[9] Where the measures of equity differ across the two requirements (BCBS, 2017).
[10] See Figure A1 for details.

This heterogeneity will shape our modelling assumptions, and in our results it will play a major role in determining the effects on markets of shocks to traders, capital regulation, and the introduction of trading platforms.

Fact 1: 6% of traders trade as frequently as the remaining 94% combined.

Figure 2.1 shows that in each bond a small subset of traders are responsible for the majority of trading. To show this, for each bond we compute the average trading frequency of each trader, compute the distribution of this across traders, and plot the average of this across bonds. The distribution is heavily positively skewed: a small set of firms trades very frequently, and a long tail of firms trades infrequently.

This fact leads us to treat traders' search costs as heterogeneous in our model, and in estimation to model these search costs as a Gamma distribution, which is well suited to capturing positively skewed distributions. In our results the shape of this distribution will play a major role in determining the effects of shocks to traders, capital regulation, and the introduction of trading platforms.

Fact 2: Fast traders intermediate.

Figure 2.2 shows that the traders that trade the most intermediate in these markets. Figure 2.2a shows how the ratio of a firm's net trading volume to its gross trading volume in a bond depends on its trading frequency. A firm's net-to-gross trading ratio is decreasing in the frequency with which it trades a bond. This implies that firms who trade a bond frequently trade in a balanced fashion—buying and selling the bond—whilst those who trade less frequently tend to trade in one direction only. Figure 2.2b shows how the price at which a trader traders a bond depends on how often they trade it. The purchase price is decreasing in a trader's trading frequency, whilst the sale price is increasing. As a consequence, infrequent traders typically pay a spread, whilst frequent traders earn a spread. This is consistent with evidence of a centrality premium documented by Di Maggio, Kermani and Song (2017) in the US corporate bond market, where more central dealers earn higher spreads. Intuitively, Figure 2.2b shows a similar relationship between prices and trading frequency holds across all traders rather than just dealers.

These patterns are consistent with fast traders intermediating: trading to facilitate other traders' trading needs and earning a spread as a result. In our model we will show that fast traders emerge as intermediaries.

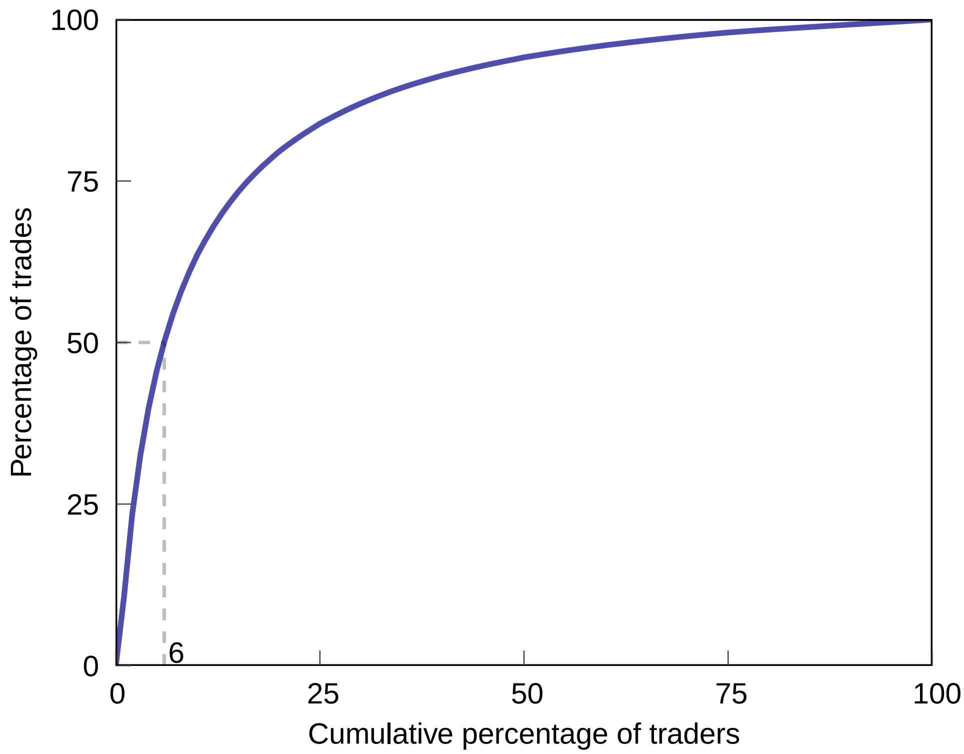


Figure 2.1: Trade frequency across traders

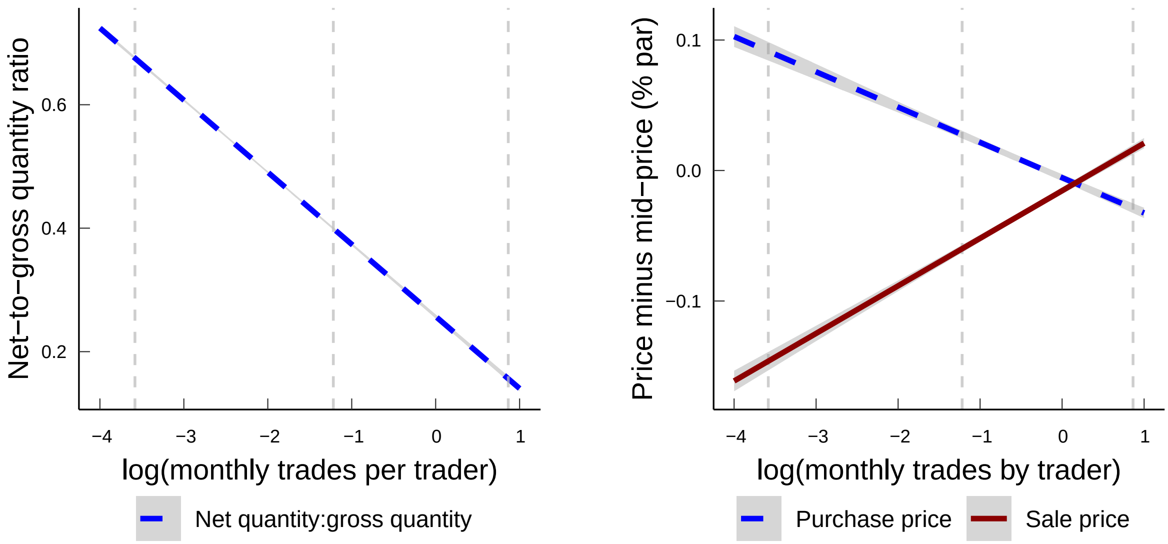
Note: This chart shows the distribution of trades across traders. For each instrument, we calculate the percentage of total trades carried out by the x% most frequent traders in that instrument. We then take the cumulative sum of this across all traders in the instrument. The chart shows the average of this across instruments, weighted by the number of firms trading in each instrument. The dashed lines show that on average the top 6% of traders are responsible for half of the trades in a bond.

Fact 3: Dealers and customers both supply and demand liquidity.

This fact shows that liquidity supply is not restricted to traders traditionally thought of as dealers, whilst liquidity demand is not restricted to traders thought of as customers. Table 2.2 summarises the percentage of trades between dealers and customers. 23% of trades do not involve a dealer as counterparty, whilst 14% of trades are between two dealers. On the basis that all trades are the result of liquidity demand and liquidity supply, it follows that in some trades customers supply liquidity and in others dealers demand liquidity.

This is consistent with the institutional features of the firms that trade in bond markets. The entities typically identified as dealers are generally parts of universal banking groups.[11] These firms do seek to supply liquidity and make money by doing so, but have a number of other motives for trading that they share with other types of financial firm. For example,

[11] See Duffie (2010*a*) for a list of major dealer banks.



(a) Net-to-gross trading ratio

(b) Trading prices

Figure 2.2: Intermediation and trading speed

Note: These figures show that the most frequent traders tend to intermediate. For the left figure, we take the absolute value of a firm's net trading in an instrument and divide it by their gross trading in the instrument. We then regress this figure—which lies between 0 and 1—on the log of the firm's average trades per month in that instrument. For the right figure, we take each trade by a trader, and subtract the mean trading price for that instrument in that month. We then regress this variable on the log of the average trades per month in that instrument by the firm buying the asset (blue line) and selling the asset (red line), and plot the fitted values in the figure. Firms on the left-hand side of the figure on average pay a spread, whilst those on the right-hand side earn a spread. The dashed vertical lines show the 25^{th}, 50^{th} and 75^{th} percentile of the variable on the x-axis, and the grey shaded areas are 95% confidence intervals.

these banks are big players in derivative markets. As their derivative positions change in value banks are subject to margin calls, which mean they must post cash collateral to their counterparties (Heller and Vause, 2012). Such shocks give banks reason to demand, rather than supply, liquidity in bond markets. Likewise, firms typically treated as customers can and do supply liquidity. As described in Choi and Huh (2018) and BlackRock (2015), firms such as asset managers increasingly seek to make money by supplying liquidity in these markets. This is consistent with evidence in other markets of non-dealers supplying liquidity (Biais, Declerck and Moinas, 2016; Franzoni, Plazzi and Çötelioğlu, 2019).

Much of the literature treats dealers and customers as distinct entities, with the former supplying liquidity and the latter demanding liquidity.[12] This is formalised in models by

[12]Theoretical papers such as Üslü (2019) and Farboodi, Jarosch and Shimer (2021) study endogenous intermediation in theory, but papers that take search models to data typically treat intermediation as exogenous, partly due to a lack of data on the trades of non-dealers (Hugonnier, Lester and Weill, 2020;

Table 2.2: Trading by type (%)

Buyer\ Seller	Customer	Dealer
Customer	23	30
Dealer	33	14

Note: This table summarises the trading frequencies between dealers and customers.

(a) assuming customers receive shocks to their asset valuations that give them a motive to trade, whilst dealers do not (Grossman and Miller, 1988; Duffie, Gârleanu and Pedersen, 2005); and (b) assuming customers do not trade with each other. Intermediation is thus exogenous in these models: dealers supply liquidity and customers demand it. Our data paint a more nuanced picture of the sterling corporate bond market. As a result, we will treat intermediation as endogenous: we will not pre-specify who intermediates, but will instead study which traders intermediate in equilibrium. In particular, all traders will face shocks that lead them to demand liquidity and will be able to supply liquidity. And rather than separating traders out into dealers and customers, we will instead capture the differences between traders with heterogeneous model parameters.

Fact 4: Traders vary their trading frequency to manage their balance sheets.

Table 2.3 shows that traders vary their trading frequency to manage their balance sheets. In particular, we show that traders trade the same bond on the same day much more frequently than would be implied by a model where trades arrive exogenously. To do this, we first compute the average empirical probability that, conditional on trading a given bond on a given day, a trader (a) trades that bond more than once that day ('paired trade'); (b) both buys and sells that bond that day ('offsetting trade'); and (c) trades the bond more than once that day where the trades are either all purchases or all sales ('amplifying trades'). We then show how often this would occur if trades arrived exogenously according to a Poisson process, where the arrival rate is equal to the trader's average trading frequency in that bond, and a trade is equally likely to be a buy or a sell. The probability of paired trades is much higher in the data than according to the model of exogenous trading. This is mostly driven by traders offsetting their trades.

Most models of search assume traders contact each other exogenously. In particular, a trader's meeting rate does not depend on the gains to trade.[13] In models with unconstrained

Brancaccio, Li and Schürhoff, 2020). Our novel data on customer identities allows us to consider endogenous intermediation empirically.

[13]A notable exception is Liu (2020), who endogenises dealer search intensity in a model with binary asset

Table 2.3: Endogenous trading frequency

	Probability: Theory (%)	Probability: Data (%)
Paired trades	6.9	44
Offsetting trades	3.5	31
Amplifying trades	3.4	13

Note: This tables shows that traders vary their trading frequency to manage their balance sheets. Paired trades are when a trader trades the same bond more than once on the same day. Offsetting trades are when a trader both buys and sells the same bond on the same day. Amplifying trades are when a trader trades more than once in the same bond on the same day, and all these trades are in the same direction. The theoretical probabilities are computed assuming that trades arrive at Poisson rate λ, where λ is the average trading rate in the data, and buys and sells are equally likely. All probabilities are computed conditional on the trader trading that bond at least once on that day, and are computed separately for each bond before averaging across bonds.

asset holdings such as Lagos and Rocheteau (2009) and Üslü (2019) every meeting results in a trade, and so trading rates are constant through time. Table 2.3 suggests traders vary their trading rates to manage their inventory. To rationalise this, in our model we will treat search intensity as an endogenous variable that traders choose and can condition on both their type and their state.[14]

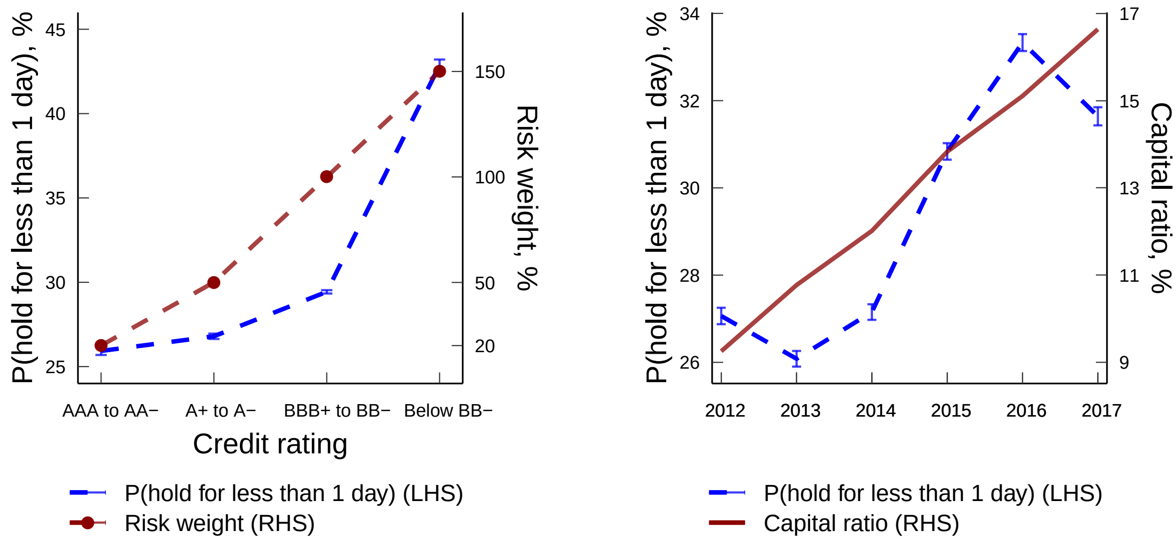
Fact 5: Dealer trading frequency varies with capital regulation.

Figure 2.3 shows that dealers manage their bond inventories more tightly when capital requirements are high. To show this, for each purchase of a bond by a dealer, we compute the probability it offsets this purchase by selling the same bond on the same day. We then compare how this probability varies through time and across bonds, after controlling for the average frequency with which a bond trades. Figure 2.3a shows how the average offsetting probability varies across bonds with different credit ratings (blue line), and how these credit ratings map into different risk weights in bank capital regulation (red line). Figure 2.3b shows how the average offsetting probability varies through time (blue line) as well as UK banks' average capital ratios through time (red line). Traders were significantly more likely to offset bond purchases at the end of the sample—when capital regulation was higher—than at the start of the sample, as well as in bonds with the highest capital requirement.

Capital regulation, to the extent that it increases traders' costs of holding inventory, increases the incentive for traders to minimise their asset holdings. Figure 2.3 is consistent

holdings. Farboodi, Jarosch and Shimer (2021) allow traders to make an ex ante investment to choose their meeting rate, but this meeting rate is then fixed through time.

[14] We note that this is not the only approach that could rationalise Table 2.3. Instead, one could introduce a fixed cost of executing a trade, which would mean that not all trader meetings result in a trade.



(a) Holding period by rating

(b) Holding period through time

Figure 2.3: Patterns in asset holding periods

Note: In this figure we show that the period of time for which a trader holds a bond is negatively correlated with capital regulation, both in the cross-section and inter-temporally. For each purchase by a dealer, we compute the length of time before they next sell that bond, or the bond matures. We create a dummy variable equal to 1 if this is less than 1 day, average this across all purchases and all dealers, and show how this varies by year (blue line in right panel) and by the risk weight the bond receives in banking regulation (blue line in left panel), after controlling for the bond's average trading frequency. The vertical lines show a 95% confidence interval around the point estimate. The red line in the left panel shows the average risk weight by bond credit rating according to the standardised approach to risk weighting, as described in BCBS (2006). The red line in the right panel shows UK banks' average capital ratios through time, taken from Bank of England (2021).

with capital regulation increasing dealers' costs of holding assets, and dealers adjusting their trading behaviour to minimise these costs. This motivates our counterfactual in Section 2.7, where we model the impact of capital regulation on banks as an increase in dealers' costs of holding an asset. The fact that dealers appear to respond to these higher costs is a fact that we will seek to rationalise with our model.

2.4 A Model of Trading in an Over-the-Counter Market

In this section we set out a model of search and trading in a decentralised asset market. Informed by the empirical facts above, our model has the following features: (1) Trading frequencies are endogenously determined and heterogeneous; (2) Intermediation is endogenous, in that any trader in any trade can supply or demand liquidity; and (3) Traders' asset holdings are unlimited and endogenous. Üslü (2019) presents a model of endogenous intermediation with unlimited holdings and exogenous meeting rates between traders. We take this model and endogenise trader search intensity.

In our model, a set of traders trade a single asset. Traders face random shocks to their marginal utility of holding the asset that lead them to want to trade. Search frictions mean it takes time for a trader to find a counterparty with whom to trade. Traders can choose the rate at which they meet counterparties, but increasing this rate is costly. The costs of searching vary across traders, and at each point in time traders choose how hard to search to balance the costs of searching against the prospective gains from trade. Once two traders meet, they bargain over the trading price and quantity. An equilibrium is a steady state of this system, where search, the terms of trade, and asset holdings are the results of traders' optimising decisions.

2.4.1 The Setting

Agents and assets Time is continuous and the horizon infinite. There is a continuum of infinitely-lived traders with measure 1, who discount the future at rate $r > 0$. These traders derive utility from holding a numéraire good with marginal utility equal to 1, and can hold an unconstrained amount of a durable asset with fixed supply a. A trader whose asset holding is h and net consumption of the numéraire good is c receives utility flow $u(\beta, h) + c$, where

$$u(\beta, h) = \beta h - \kappa \frac{h^2}{2} \tag{2.1}$$

is the utility flow from holding the asset. A trader's lifetime utility is the present discounted value of expected utility flows, net of payments for transactions.

A trader draws a new valuation parameter β from the distribution $G(\beta)$ at Poisson arrival rate η. These random shocks to β create a motive to trade. These shocks can be interpreted as shocks to a firm's funding position. For example, a trader who receives a shock that lowers their valuation could be an asset manager who faces investor outflows, and thus needs to sell the asset to raise funds. An increase in valuation would then represent an asset manager facing investor inflows.

This utility function over holdings can be interpreted as a reduced form of a more fundamental setting where traders exhibit constant absolute risk aversion, and utility is derived from consumption rather than asset holdings (Duffie, Gârleanu and Pedersen, 2007; Üslü, 2019). In this setting, risk-averse traders seeking to maximise their utility choose how much wealth to allocate between a risky asset and a risk-free asset, and how much to consume. They have a stochastic income stream which is correlated with the returns to the risky asset, and this correlation changes stochastically, changing the desirability of holding the risky asset. These shocks underly the shocks to β in the reduced-form utility in (2.1). The parameter κ in equation (2.1) is a linear function of the trader's coefficient of absolute risk aversion. In the rest of this paper we will refer to κ as risk aversion. We work with the reduced-form utility, rather than the more fundamental setting, as this captures the main dynamics at play, and when we turn to estimation the parameters we can identify are those of the reduced-form utility.

Endogenous search, matching and trade Traders search for and meet counterparties with whom to trade. Traders choose the rate at which they meet counterparties, but meeting counterparties frequently is costly. Let γ denote a trader's search intensity. A trader searching with intensity γ incurs cost $s(z, \gamma)$, where parameter $z \sim F(z)$ capture a trader's cost of searching, which varies across traders and is fixed through time. The search cost function $s(z, \gamma)$ is a twice continuously differentiable, increasing and convex function of γ. Optimal search γ will be a function of the search parameter z, as well as the trader's valuation β and their current holding h.

The benefit to searching is that it results in meetings with other traders, and potential gains from trade. Meetings between traders searching with intensity γ and γ' are governed by a matching function $m(\gamma, \gamma')$, which is symmetric and linearly increasing in both arguments. This captures the fact that traders can both contact and be contacted by other traders, and that a trader can increase its meeting rate by increasing its search intensity γ.

Together, the search cost function and matching function capture the trade-off a trader faces in choosing how hard to search. A trader benefits by increasing the frequency at which it can trade, as this enables it to better optimise its portfolio and respond to shocks. But doing so, for example by hiring more people to the firm's trading desk or paying them to work overtime, is costly. Firms will choose search intensity to balance these costs and benefits. The benefits to trading are state-dependent—for example a trader may be eager to trade after receiving a shock to their asset valuation, whilst after having traded they may have little desire to trade again. As a result, traders will condition their search intensity on their state—their asset holdings h and their valuation β—as well as their search parameter z.

Once two traders have been matched they engage in bilateral Nash bargaining over the quantity traded q and the unit price p. All traders have bargaining strength of one half. When a trader of type (z, β, h) meets a trader of type (z', β', h'), the signed trading quantity is $q((z, \beta, h), (z', \beta', h'))$ and the unit price is $p((z, \beta, h), (z', \beta', h'))$.

2.4.2 Solving the Model

In this section we define and characterise a stationary equilibrium. Doing so requires that we analyse the three-way feedback between traders' search decisions, the distribution of the asset across traders, and the terms at which they trade. A trader's search decision depends on the distribution of the asset across traders and the terms of trade as these determine the gains to searching. Distributions depend on traders' search decisions and the terms of trade as these determine the flows of the asset between traders. Finally, the terms of trade depend on distributions and search as they determine future trading opportunities. To proceed we write down expressions for traders' value functions, their trading decisions, their search decisions, and the distributions of assets across traders. An equilibrium is then a mutually consistent, stationary relationship between these endogenous variables.

Trading and Value Function

We begin by setting out the equations that govern the quantity and price at which traders trade, before writing down their value function. Let $V(z, \beta, h)$ be the value function of a trader with search parameter z, valuation β and holding h. We assume $V(z, \beta, h)$ is differentiable, increasing and concave. In equilibrium it will inherit these properties from the flow utility function $u(\beta, h)$. When a trader of type $\Delta \equiv (z, \beta, h)$ meets a trader of type $\Delta' \equiv (z', \beta', h')$, the total surplus from trading quantity q is given by the sum of the two traders' changes in value after the trade:

$$V(z, \beta, h - q) - V(z, \beta, h) + V(z', \beta', h' + q) - V(z', \beta', h'). \tag{2.2}$$

Nash bargaining implies that the quantity they trade $q(\Delta, \Delta')$ and the unit price $p(\Delta, \Delta')$ jointly solve:

$$\begin{aligned} \underset{p,q}{\text{maximise}} \quad & (V(z,\beta,h-q)-V(z,\beta,h)+pq)(V(z',\beta',h'+q)-V(z',\beta',h')-pq) \\ \text{subject to} \quad & V(z,\beta,h-q)-V(z,\beta,h)+pq \geq 0, \\ & V(z',\beta',h'+q)-V(z',\beta',h')-pq \geq 0. \end{aligned} \tag{2.3}$$

The terms of trade maximise the product of the traders' surpluses, subject to each trader weakly preferring the agreed trade to not trading at all.[15]

The quantity that solves this problem must by definition maximise total trading surplus. The asset is thus sold by the trader who values the asset less to the trader who values it more, and the quantity sold is that which equalises the post-trade marginal value of the two traders:

$$V_3(z,\beta,h-q(\Delta,\Delta')) = V_3(z',\beta',h'+q(\Delta,\Delta')),$$

where $V_3(z,\beta,h)$ is the derivative of the value function with respect to holdings h. Intuitively, this means trading quantity is larger when the pre-trade marginal values of traders are more spread out, and when the curvature of the value function is lower (as then the trading quantity that equalises traders' marginal values must be greater). When we estimate the model in Section 2.5, these facts will mean the parameters governing the slope and curvature of the value function (value β and risk aversion κ) are identified by trading quantities.

The trading surplus when two traders meet is then simply the surplus (equation (2.2)) evaluated at the optimal trading quantity:

$$S\big((z,\beta,h),(z',\beta',h')\big) = V(z,\beta,h-q(\Delta,\Delta'))-V(z,\beta,h)+V(z',\beta',h'+q(\Delta,\Delta'))-V(z',\beta',h').$$

Evaluating the first-order-condition of equation (2.3) with respect to price and plugging in the optimal quantity $q(\Delta,\Delta')$ gives the following expression for the optimal unit price:

$$p(\Delta,\Delta') = \frac{1}{2}\Bigg(\underbrace{\frac{V(z',\beta',h'+q(\Delta,\Delta'))-V(z',\beta',h')}{q(\Delta,\Delta')}}_{\approx\text{ slope of }\Delta'\text{ value in }h'} + \underbrace{\frac{V(z,\beta,h)-V(z,\beta,h-q(\Delta,\Delta'))}{q(\Delta,\Delta')}}_{\approx\text{ slope of }\Delta\text{ value in }h}\Bigg).$$

The price thus depends on the slopes of the traders' value functions. As the optimal trading

[15] Given the traders' bargaining strengths are equal, they do not appear in the optimisation.

quantity approaches zero, the price approaches the average of the two traders' marginal values. The variance of trading prices will therefore be greater when traders' marginal values are more spread out, and when the curvature of the value function is greater (as this increases the variation in the slope of the value function). These facts will mean that trading prices, as well as trading quantities, will help to identify the value shocks β and risk aversion κ in Section 2.5.

We can now write down an expression for traders' value functions, taking as given their trading decisions, their search decisions and the distribution of the asset across traders. The Hamilton-Jacobi-Bellman equation that governs traders' optimal behaviour can be written as:

$$r \underbrace{V(z,\beta,h)}_{\text{Value}} = \underbrace{u(\beta,h) - s(z,\gamma(z,\beta,h))}_{\substack{\text{Flow utility \&} \\ \text{search costs}}} + \underbrace{\eta \int (V(z,\beta',h) - V(z,\beta,h)) G(d\beta')}_{\text{Switch type}} + \\ \frac{1}{2} \iiint \underbrace{m(\gamma(z,\beta,h),\gamma(z',\beta',h'))}_{\substack{\text{Meeting} \\ \text{probability}}} \underbrace{S((z,\beta,h),(z',\beta',h'))}_{\text{Surplus}} \Phi(dz',d\beta',dh'), \quad (2.4)$$

where $\Phi(z,\beta,h)$ is the cdf of traders of type (z,β,h), $\gamma(z,\beta,h)$ is the optimal search intensity for trader type (z,β,h), and $s(z,\beta,h) \equiv s(z,\gamma(z,\beta,h))$ are the corresponding search costs.

The value function has the usual asset pricing interpretation. A trader of type (z,β,h) gets flow utility $u(\beta,h)$ from holding the asset and incurs search costs $s(z,\gamma(z,\beta,h))$. At intensity η the trader switches type, drawing a new value β' from $G(\beta')$. At intensity $m(\gamma(z,\beta,h),\gamma(z',\beta',h'))\Phi(dz',d\beta',dh')$ the trader meets a counterparty with type (z',β',h'), in which case they trade and extract half the total surplus.

Traders' values thus depend on the search and trading decisions of both themselves and their potential future counterparties, and the distributions of the asset across traders. A trader's search determines the search cost it incurs as well as the expected gains to trade it enjoys. The distribution of the asset across traders, as well as these traders' search decisions, then determine how frequently the trader meets a counterparty, and which type of counterparty they meet. The Nash bargaining protocol then determines the gains to trade when they do meet a counterparty.

Trader Search

We now characterise traders' optimal search decisions, given their trading decisions and the distribution of the asset. To do so, we take the first-order condition of the value function in equation (2.4) and apply the envelope theorem, yielding the equation governing the optimal search of type $\Delta \equiv (z, \beta, h)$:

$$\underbrace{s_2(z, \gamma(\Delta))}_{\substack{\text{Marginal cost} \\ \text{of search}}} = \frac{1}{2} \int \underbrace{\frac{\partial m(\gamma(\Delta), \gamma(\Delta'))}{\partial \gamma(\Delta)}}_{\substack{\text{Increase} \\ \text{in meetings}}} \underbrace{S(\Delta, \Delta')}_{\substack{\text{Surplus from} \\ \text{meeting}}} \Phi(d\Delta'). \tag{2.5}$$

The left-hand side is the marginal cost of searching, which is increasing in the trader's search intensity. The right-hand side is the marginal benefit to searching. Increasing search intensity increases the rate at which the trader meets a counterparty and enjoys the expected gains from trade. The expected gains from trade depend on the trading surplus when the trader meets a given counterparty, and the likelihood of meeting each counterparty type. Traders will thus search harder if the expected gains from trade are higher, or if their search costs are lower. When we estimate the model in Section 2.5, this negative relationship between search costs and meeting rates will mean that the distribution of search costs $F(z)$ is identified by the distribution of average trading rates across traders.

Type Distributions

We now close the model by providing expressions for the equilibrium distributions of assets across trader types. For the system to be in steady state, the net flows from trading and value shocks must be equal to zero across all trader types (z, β, h). The measure of traders with search costs z, valuation $\beta \leq \beta^*$ and holding $h \leq h^*$ must satisfy:

$$\begin{aligned}
&\iint_{\underline{\beta}}^{\beta^*} \int_{h^*}^{\infty} m(\gamma(z, \beta, h), \gamma(\Delta'))\phi(z, \beta, h)\phi(\Delta')1(q(z, \beta, h, \Delta') \geq h - h^*) dh d\beta d\Delta' - \\
&\qquad \iint_{\underline{\beta}}^{\beta^*} \int_{-\infty}^{h^*} m(\gamma(z, \beta, h), \gamma(\Delta'))\phi(z, \beta, h)\phi(\Delta')1(q(z, \beta, h, \Delta') < h - h^*) dh d\beta d\Delta' \\
&\qquad\qquad = \eta(1 - G(\beta^*)) \int_{\underline{\beta}}^{\beta^*} \int_{-\infty}^{h^*} \phi(z, h, \beta) dh d\beta - \eta G(\beta^*) \int_{\beta^*}^{\bar{\beta}} \int_{-\infty}^{h^*} \phi(z, h, \beta) dh d\beta
\end{aligned} \tag{2.6}$$

for all (z, β^*, h^*). The left-hand side represents net inflows due to trade. This consists of the flow of traders with type $(z, \beta \leq \beta^*, h > h^*)$ who meet a counterparty and sell enough of the asset such that their holdings fall below h^*, minus the flow of traders with type $(z, \beta \leq \beta^*, h \leq h^*)$ who meet a counterparty and buy enough that their holdings exceed h^*. The right-hand side consists of net outflows due to value shocks. The outflows consist of the stock of traders with type $(z, \beta \leq \beta^*, h \leq h^*)$, who with intensity η receive a new value shock, and with probability $1 - G(\beta^*)$ draw a new value above β^*. The inflows are traders with type $(z, \beta > \beta^*, h \leq h^*)$ who with intensity $\eta G(\beta^*)$ receive a shock that takes their value below β^*.

Finally, the distribution of types in equilibrium must be consistent with the ex ante distribution of types:

$$\iint \Phi(z, \beta, h) d\beta dh = F(z) \quad \forall z. \tag{2.7}$$

Equilibrium

We now define an equilibrium based on the system of equations derived above. An equilibrium is a set of value functions, search intensities, terms of trade and distributions of agents and assets that are mutually consistent in steady state. Let the space of trader types (z, β, h) be $\mathcal{T} = R^+ \times R \times R$. A steady state equilibrium is:

- value function $V : \mathcal{T} \to R$;
- pricing function $p : \mathcal{T}^2 \to R^+$;
- trading quantity function $q : \mathcal{T}^2 \to R$;
- density function $\phi : \mathcal{T} \to R^+$; and
- search intensity functions $\gamma : \mathcal{T} \to R^+$

that solve the following equilibrium conditions:

- value functions (equation (2.4));
- distribution functions (equations (2.6) and (2.7));
- optimal search (equation (2.5));
- optimal prices and quantity (equation (2.3)); and
- market clearing:

$$\iiint h\Phi(dz, d\beta, dh) = a. \tag{2.8}$$

An equilibrium is thus a set of equations that solves the three-way feedback between traders' search decisions, the distribution of the asset across traders, and the terms at which they trade. We solve for this equilibrium numerically in the sections that follow.

2.4.3 Properties of the Equilibrium

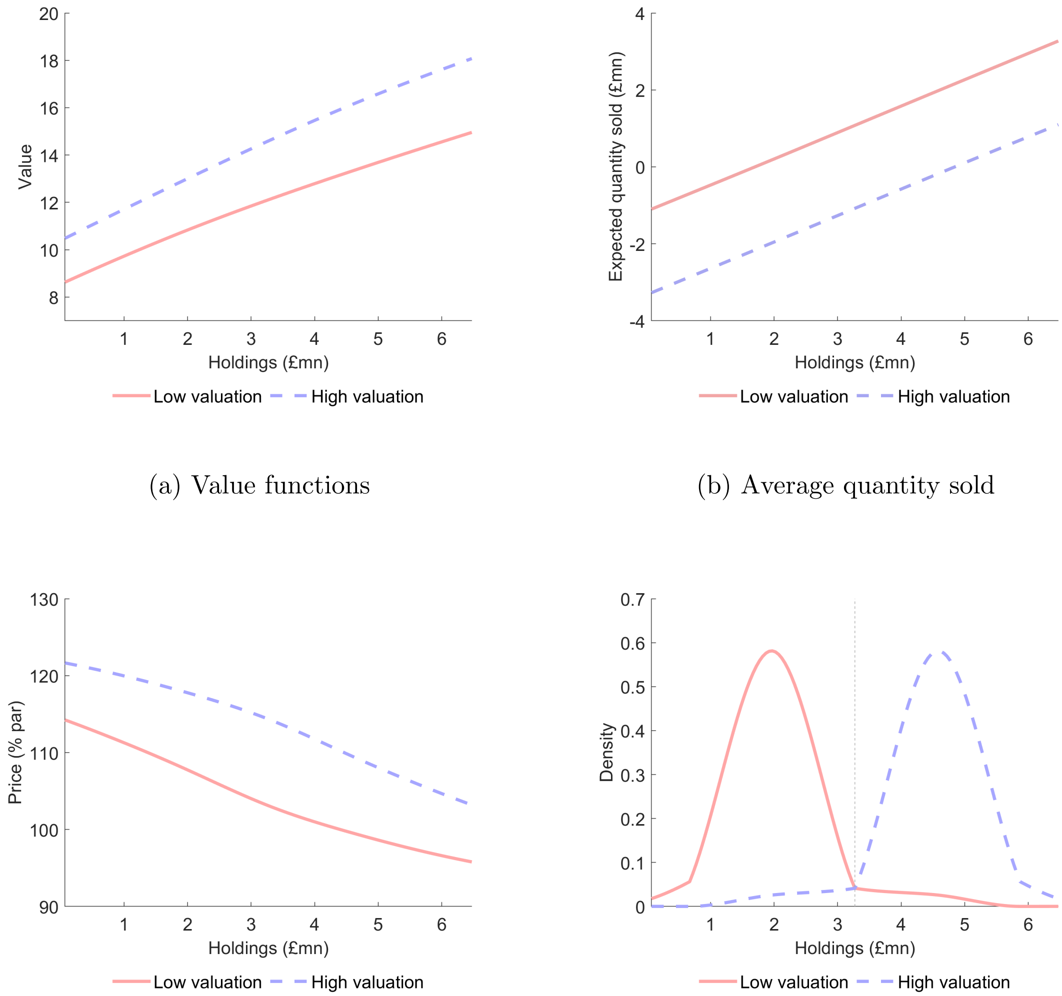
Figure 2.4 displays some properties of the equilibrium, showing the value functions, prices, trading quantity and distributions of traders with different valuations β as a function of holdings, for the parameter vector we estimate in Section 2.5. The value function is increasing and concave in holdings, inheriting these properties from the quadratic flow utility (Figure 2.4a).[16] A trader with a low valuation β values the asset less, and so at any level of holding is more likely to be a net seller than a trader with a high valuation (Figure 2.4b), and trades the asset at a lower price (Figure 2.4c). As a result of these differences in trading patterns, the low-value trader holds less of the asset (Figure 2.4d). With quadratic utility and a symmetric distribution of β, the equilibrium distributions are symmetric around the per-capita supply of the asset a, with the distributions of holdings for traders with low valuation the mirror image of those with high valuation.

Search and Trading

Figure 2.5a shows how traders vary their search and trading behaviour according to their valuation and asset holdings. For a given valuation β a trader's search intensity is a convex function of its asset holdings. Search intensity takes its minimum at a quantity which we call the trader's *target holding*, defined as the level of asset holding $h^*(z,\beta)$ at which the returns to search are minimised.

Trader search thus takes an intuitive form: traders search least when the gains from trade are low, and search harder as the gains from trade increase. Traders' values are concave in their holdings, which means that in equilibrium the gains to trade are higher when a trader has very low or high holdings. Search intensity is thus U-shaped in holdings. This mechanism in our model can rationalise the empirical fact in Table 2.3, which shows that traders offset their trades more frequently than if their search was exogenous. When search is chosen optimally, traders increase search intensity whenever a trade takes them away from their

[16] This concavity is necessary for the model to have a solution, as it ensures traders will always trade and hold a finite amount of the asset.



(a) Value functions

(b) Average quantity sold

(c) Average trading price

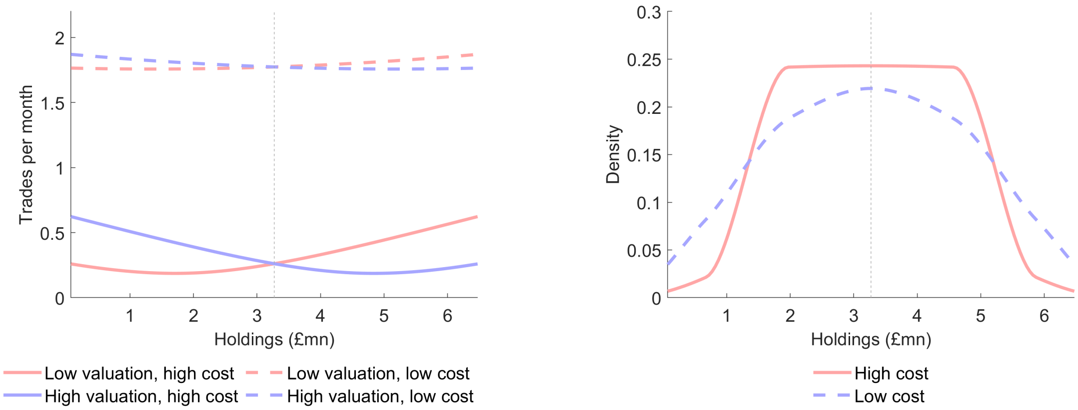
(d) Distributions

Figure 2.4: Terms of trade and distributions by trader type

Note: These figures summarise key features of the model's equilibrium, for the parameter values estimated in Section 2.5. Each panel plots one of the model's outcome variables against a trader's holdings for different levels of valuation β, for a given search parameter z. Panel (a) shows the value function $V(z, \beta, h)$. Panel (b) shows the average quantity sold per trade. Panel (c) shows the average trading price. Panel (d) shows the distribution of holdings. The grey vertical line in panel (d) shows the per-capita supply of the asset a.

target holdings, or a shock changes this target.

Traders with low search costs search harder, and trade more frequently as a result (Figure 2.5a). Their search intensity is less sensitive to their valuation β. Put differently, the difference in search intensity between low- and high-cost searchers is greatest when traders are at their target holdings, and have no reason to demand liquidity. Thus low-cost traders



(a) Trading frequency by shock & holding

(b) Holding distribution by type

Figure 2.5: Search and distributions by trader type

Note: These figures summarise some key quantities in our estimated model. Panel (a) shows trading frequency by asset holding h and valuation β, for a trader with a given search cost. Panel (b) shows the distribution of agents across different holdings for low-search-cost traders and high-search-cost traders. The grey vertical lines show per-capita supply of the asset.

do not leverage their cost advantage to respond faster to shocks, but instead to supply liquidity.

Figure 2.5b shows how a trader's search cost influences its asset holdings. Traders who are able to search at lower cost allow their holdings distribution to spread out more.[17] They are willing to trade to more extreme holdings because their superior search technology means they will be able to offset this trade relatively quickly.

Liquidity

Liquidity in this context is the ease with which traders can change their holdings when they wish to. Traders who wish to change their asset holdings demand liquidity, and those who facilitate this supply liquidity. Traders demanding liquidity face the following types of friction: (a) the time they must wait to trade, (b) the cost they pay to search, (c) the extent to which the quantity they trade is rationed relative to what they would trade in a market without frictions, and (d) the difference in the price at which they trade relative to that in a market without frictions. Aggregate welfare is determined by how the assets are distributed

[17] Üslü (2019) finds a similar result in a setting with exogenous trading frequency.

across traders with different values and the search costs traders incur, with trading profits netting out across traders.

In what follows, we will track a summary measure of liquidity which we call *market depth*. We define this to be the amount a trader could hypothetically sell per unit time without moving the price by more than a given amount D. This measure is in the spirit of definitions of market depth used in the literature on market microstructure (Foucault, Pagano and Röell, 2013). To construct this measure, we first note that the maximum price a trader of type $\Delta \equiv (z, \beta, h)$ would be willing to pay to buy amount q of a bond is:

$$\frac{V(z, \beta, h+q) - V(z, \beta, h)}{q}.$$

As q goes to 0, this becomes the marginal value of holdings in h, denoted $V_3(z, \beta, h)$.

We can then define the maximum amount $q^\pi(\Delta, D)$ that could be sold to this trader at a price no more than D below $V_3(z, \beta, h)$:

$$\frac{V(z, \beta, h + q^\pi(\Delta, D)) - V(z, \beta, h)}{q^\pi(\Delta, D)} = V_3(z, \beta, h) - D,$$

where $V_3(z, \beta, h)$ is the derivative of the trader's value with respect to their holdings.

Intuitively, $V_3(z, \beta, h)$ is the 'current' price at which the trader would be willing to trade a trivial amount ϵ before the trade, and the term on the left-hand side is the price at which they are willing to trade amount $q^\pi(\Delta, D)$. The difference between the two represents a cost for the trader selling the asset, as it captures the decline in the price they receive for the bond that is caused by their own sale. Our measure $\Pi(D)$ is then the trade-weighted average of this across traders multiplied by 2Γ, which is the average frequency with which a trader gets to trade:

$$\Pi(D) = 2\Gamma \int \frac{\gamma(\Delta)\phi(\Delta)}{\Gamma} q^\pi(\Delta, D) d\Delta. \tag{2.9}$$

This measure captures both the extensive margin of liquidity—how often traders get the opportunity to trade—and the intensive margin—how much they are able to trade upon meeting a counterparty without incurring large costs. A larger value of $\Pi(D)$ indicates a higher level of market depth, and hence a more liquid market.

2.5 Estimation

We first set out the parametric assumptions we make to take the model to the data. We then describe our estimation procedure, before summarising the moments we use. Finally we describe the key variation that identifies each of our model's parameters.

2.5.1 Parametric Assumptions

We use the following linear matching function:

$$m(\gamma, \gamma') = 2\gamma \frac{\gamma'}{\Gamma}, \tag{2.10}$$

where

$$\Gamma = \iiint \gamma(z, \beta, h) \Phi(dz, d\beta, dh).$$

This technology captures the fact that a trader can both contact and be contacted by another trader. Conditional on contact the likelihood of meeting a given type of counterparty is proportional to their search intensity.

Another common linear matching function is $m(\gamma, \gamma') = \gamma + \gamma'$ (Shimer and Smith, 2001), which assumes that counterparties are drawn uniformly from the distribution of traders. This matching function could not fit our data. In particular (1) Trading frequency is heavily right-skewed as shown in Figure 2.1; and (2) Many traders have a monthly trading frequency per instrument that is close to 0. As a result, this matching function would imply negative search intensities for the left-tail of the trading frequency distribution.

We take the search cost function to be:

$$s(z, \gamma) = (\gamma - z)^2, \tag{2.11}$$

where $z \sim F(z)$. A trader can costlessly meet traders at rate z, but must pay a constant marginal cost to meet traders more frequently. Intuitively, z can be thought of as a flow of contacts a trader makes as part of its everyday business. For example, if the trader is an investment bank it meets its clients regularly as part of its broader investment banking activity. However, to increase its meeting rate above this minimum level is costly, as it must contact other traders or hire more staff to make the contacts. For simplicity, we take the marginal cost of these contacts to be linear. In what follows, we will refer to z as search efficiency, and refer to traders with high z as having high search efficiency or low search costs.

We assume the search parameter z follows a Gamma distribution, with shape parameter k_z and scale parameter θ_z. We choose this distribution as it can match the skewness of the trading rate distribution (Figure 2.1), has strictly positive support, and is relatively parsimonious. We assume the shock distribution $G(\beta)$ is uniform, with mean μ_β and variance σ_β^2. We choose a uniform distribution as it is simple and parsimonious.

2.5.2 Estimation Procedure

We estimate the model using the data described in section 2.2.1, assuming these data are generated by the model in steady state. The unit of time is 1 month and the monthly discount rate is 0.5%. We fix the asset supply a to be equal to the mean total amount outstanding in a bond, normalised by the number of traders trading in that bond.

We estimate the parameter vector $\psi = \{\mu_\beta, \sigma_\beta, \kappa, \eta, k_z, \theta_z\}$, where $\{\mu_\beta, \sigma_\beta\}$ are the parameters of the β distribution, κ is risk aversion, η is shock frequency and $\{k_z, \theta_z\}$ are parameters of the distribution of the search parameter z. We use a minimum-distance estimator that matches theoretical moments implied by the model to their empirical counterparts. In practice this takes the form of a nested loop: for any given ψ we numerically solve the model. We then calculate a vector of theoretical moments $m(\psi)$ and compare this to its empirical counterpart m_0. We choose the parameter vector that minimises the sum of squared percentage differences between the theoretical and empirical moments:

$$\hat{\psi} = \arg\min_{\psi}(m(\psi) - m_0)'\Omega^{-1}(m(\psi) - m_0)$$

where $\Omega = m_0 m_0'$.

To find the model's equilibrium we solve the model's equations at a finite set of points in the type space $\mathcal{T}$. The distributions of valuations β and search cost parameters z are discrete versions of the continuous distributions in section 2.5.1. We use interpolation splines to ensure continuity of functions in holdings h. We then use a nonlinear solver to numerically solve the equilibrium equations. For full details of the estimation procedure, see Appendix

A2.

The model is defined at the *instrument level*. When computing the empirical moment vector we compute moments separately for each bond, before averaging across bonds to give m_0.

2.5.3 Moments

Our data contain the key information we need to identify the model: trading frequencies, trading prices and quantities, and asset holdings. The data contain the identity of all traders, and not just dealers as is commonly the case, which means we can characterise trader heterogeneity across all traders.

We match two sets of moments: *across-trader* moments and *within-trader* moments. Within-trader moments measure the distribution of quantities for a given trader in a given bond—for example how much its inventory varies through time. Across-trader moments measure the distribution of average quantities across traders for a given bond—for example comparing how the average trading frequency of one trader differs from the average trading frequency of other traders.[18] We choose moments covering the means, variances and correlations of traders' trading frequencies, trading prices, trading quantities and holdings to identify our parameters. We give analytical expressions for each of these moments below.

Theoretical Moments

Expectations

1. Average trading frequency:

$$\mathbb{E}(n) = 2\int \gamma(\Delta)\phi(\Delta)\pi(\Delta)d\Delta,$$

   where $\pi(\Delta)$ is the fraction of Δ's meetings that result in a trade:

$$\pi(\Delta) = \int 1(q(\Delta,\Delta') \neq 0)\frac{\gamma(\Delta')\phi(\Delta')}{\Gamma}d\Delta'.$$

2. Average trade size:

[18] Note that the means of a quantity are the same within and across traders.

$$\mathbb{E}(|q|) = \int \frac{\gamma(\Delta)\phi(\Delta)}{\Gamma} \int |q(\Delta, \Delta')| \frac{\gamma(\Delta')\phi(\Delta')}{\Gamma} d\Delta' d\Delta.$$

3. Average price:

$$\mathbb{E}(p) = \int \frac{\gamma(\Delta)\phi(\Delta)}{\Gamma} \int p(\Delta, \Delta') \frac{1(q(\Delta, \Delta') \neq 0)}{\pi(\Delta)} \frac{\gamma(\Delta')\phi(\Delta')}{\Gamma} d\Delta' d\Delta.$$

Across-trader moments

4. Standard deviation of trading frequency across traders:

$$SD^A(n) = \sqrt{\int (n(z) - \mathbb{E}(n))^2 f(z) dz},$$

where $n(z)$ is type z's average trading frequency:

$$n(z) = 2 \iint \gamma(z, \beta, h) \frac{\phi(z, \beta, h)}{f(z)} \pi(z, \beta, h) d\beta dh.$$

Within-trader moments

5. Standard deviation of holdings, within traders:

$$SD^W(h) = \int \sqrt{\mathbb{V}(h|z)} f(z) dz,$$

where $\mathbb{V}(h|z)$ is the variance in holdings for a trader with search parameter z:[19]

$$\mathbb{V}(h|z) = \iint (h - a)^2 \frac{\phi(z, \beta, h)}{f(z)} d\beta dh.$$

6. Standard deviation of trading prices, within traders:

$$SD^W(p) = \int \sqrt{\mathbb{V}(p|z)} f(z) dz,$$

where $\mathbb{V}(p|z)$ is the variance of trading prices for a trader with search parameter z:

[19]Note that the symmetry of the equilibrium shown in Figure 2.4d means that each trader on average holds the asset's per-capita supply, a.

$$\mathbb{V}(p|z) = \iint \frac{\gamma(z,\beta,h)\phi(z,\beta,h)}{\Gamma(z)f(z)} \times \int \big(p(z,\beta,h,\Delta') - \mathbb{E}(p|z)\big)^2 \frac{1(q(z,\beta,h,\Delta') \neq 0)}{\pi(z,\beta,h)} \frac{\gamma(\Delta')\phi(\Delta')}{\Gamma} d\Delta' d\beta dh;$$

$\mathbb{E}(p|z)$ is type z's mean trading price:

$$\mathbb{E}(p|z) = \iint \frac{\gamma(z,\beta,h)\phi(z,\beta,h)}{\Gamma(z)f(z)} \times \int p(z,\beta,h,\Delta') \frac{1(q(z,\beta,h,\Delta') \neq 0)}{\pi(z,\beta,h)} \frac{\gamma(\Delta')\phi(\Delta')}{\Gamma} d\Delta' d\beta dh;$$

and $\Gamma(z)$ is type z's average search intensity:

$$\Gamma(z) = \iint \gamma(z,\beta,h)\phi(z,\beta,h) d\beta dh.$$

7. Correlation between quantity sold and holdings, within-traders:

$$corr^W(h,q) = \int \frac{cov(h,q|z)}{\sqrt{\mathbb{V}_{TW}(h|z)\mathbb{V}(q|z)}} f(z)dz,$$

where expressions for the conditional (trade-weighted) covariance of holdings and quantity sold $cov(h,q|z)$, trade-weighted holdings variance $\mathbb{V}_{TW}(h|z)$ and variance of quantity sold $\mathbb{V}(q|z)$ are given in Appendix A3.

8. Correlation between absolute inventory $inv \equiv |h-a|$ and trading frequency, within traders:

$$corr^W(inv,n) = \int \frac{cov(inv,n|z)}{\sqrt{\mathbb{V}(inv|z)\mathbb{V}(n|z)}} f(z)dz,$$

where expressions for the conditional covariance of absolute inventory and trading frequency $cov(inv,n|z)$, variance of absolute inventory $\mathbb{V}(inv|z)$ and variance of trading frequency $\mathbb{V}(n|z)$ are given in Appendix A3.

Empirical Moments

We compute the empirical counterpart of each of these moments to construct the empirical moment vector m_0. As the model is defined at the instrument level, we compute each of the moments separately for each bond, and the empirical moment vector m_0 contains the averages of these moments across bonds. The size of our sample is then the number of bonds in our data. To compute bootstrap standard errors, we resample from the set of bonds and repeat the estimation for each bootstrap sample.

The data are likely to include the effects of economic forces outside of our model. Firstly, there may be heterogeneity in preferences for a given bond across traders. For example, a trader who was an underwriter on a bond's initial issuance may hold more of it throughout its life. To control for this, we compute our within-trader moments based only on the residual variation after stripping out trader-level variation. This is equivalent to including trader-level fixed effects. In the example of a trader holding more of a bond it underwrote than other traders, our approach strips out this variation when computing the variance of trader holdings. Secondly, news about the fundamentals of a bond will affect its price. To control for this, when computing the within-trader variance in prices we further control for the current credit rating of the instrument, and compute the variation in prices based on the residual variance after controlling for this. This strips out any news about a bond that is captured in its credit rating.

2.5.4 Identification

The moments defined above provide a mapping from the model's parameters to the data. In this section we show that there is sufficient variation to invert this mapping and infer the parameters from the data. The model's nonlinearity means that all but one parameters affect all moments, but for each of the model's parameters there are key moments that pin it down. Table 2.4 summarises this key variation for each parameter.

The distribution of the search parameter z is identified by the distribution of average trading frequencies across traders, as the optimal search equation (2.5) ensures that a trader's average trading frequency is a monotonic function of its search parameter z. As a result the parameters of the distribution of z are pinned down by the moments of the distribution of traders' average trading frequencies.

The mean of the value shock distribution μ_β is identified by the average trading price.

Table 2.4: Identification

Parameter	Moment
Search efficiency $z \sim \Gamma(k_z, \theta_z)$	
$\mathbb{E}(z) = k_z\theta_z$	$\overset{[+]}{\mathbb{E}(n)}$
$\mathbb{V}(z) = k_z\theta_z^2$	$\overset{[+]}{SD^A(n)}$
Utility $u(h) = \beta h - 0.5\kappa h^2$; $\beta \sim U(\mu_\beta, \sigma_\beta)$	
μ_β	$\overset{[+]}{\mathbb{E}(p)}$
σ_β	$\overset{[+]}{SD^W(p)}$, $\overset{[+]}{\mathbb{E}(\lvert q\rvert)}$, $\overset{[+]}{SD^W(h)}$
κ	$\overset{[+]}{SD^W(p)}$, $\overset{[-]}{\mathbb{E}(\lvert q\rvert)}$, $\overset{[-]}{SD^W(h)}$
Shock frequency η	
η	$\overset{[+]}{corr^W(\lvert h-a\rvert, n)}$, $\overset{[+]}{corr^W(h, q)}$

Note: This table summarises the key variation that identifies our parameters. n is the trading frequency, p is the trading price, q is the amount sold by a trader and h is a trader's holding. $SD^W(x)$ is the standard deviation of x for a given trader, whilst $SD^A(x)$ takes the average value of x for each trader and computes the standard deviation of this across traders. $corr^W(x, y)$ is the correlation of x and y for a given trader. Shocks β are distributed according to a uniform distribution with mean μ_β and variance σ_β^2. Parameter κ governs traders' risk aversion. Parameter η is the frequency (per month) at which traders draw new values β. Parameter z in search cost function $z = (\gamma - z)^2$ is distributed according to a Gamma distribution with shape parameter k_z and scale parameter θ_z. A [+] above a moment indicates the moment is increasing in the relevant parameter, whilst a [−] indicates it is decreasing.

This is because it affects the mean price, but no other moments in the model. As set out in Section 2.4, when two traders meet, the quantity q they trade is such that their post-trade marginal values for the asset are equalised: $V_3(z, \beta, h - q) = V_3(z', \beta', h' + q)$. Moving the location of the valuation distribution $G(\beta)$ has no effect on this, as it affects all traders equally. As a result, the location of the β distribution does not affect trading surplus, and so does not affect search intensity or the asset distributions. The average trading price, which depends on the average slope of traders' value functions, does depend on the location of β, and as a result pins it down exactly.[20]

The variance of the value shock σ_β^2 and the level of risk aversion κ are identified by the distributions of holdings, trade sizes and price. A decrease in σ_β^2 causes the value functions across different levels of β to move closer together, whilst an increase in risk aversion increases

[20] In practical terms this makes estimation a simpler task, as we can estimate the mean of the shock distribution separately from the other parameters. In a first stage we search over $(\sigma_\beta, \kappa, \eta, k_z, \theta_z)$ to match all moments except the mean price. In a second stage we choose μ_β to perfectly fit the mean price.

the curvature of the value function in holdings for a given β. Both these shifts cause traders to shrink the variance of their asset holdings. Intuitively, both a smaller shock variance and increased risk aversion reduce a trader's incentive to trade to extreme asset holdings. For the same reason, they also both decrease the average trade size. As explained in Section 2.4 trade size is smaller when traders' marginal valuations are less spread out (as is the case when σ_β^2 is small) and when the curvature of the value function is greater (as is the case when κ increases).

An increase in risk aversion κ and a decrease in the variance of shocks σ_β^2 have opposite predictions, however, for the variance of prices. As set out in Section 2.4, the price at which two traders trade is governed by the slopes of their value functions. A reduction in σ_β^2 reduces the variance of these slopes across different valuations β and hence reduces the variance of prices. An increase in κ, however, increases the curvature of utility for a given β, which increases the variation in the slope of the value function across holdings. This increases the variance of prices, and enables us to separate the variance of shocks from traders' risk aversion.

The shock frequency, η, is identified by the correlation between traders' holdings of the asset and their trading price and quantity. An increase in the frequency of shocks moves the value functions of traders with different levels of β closer together. As a trader's value becomes more transient, they place less weight on their current value in their trading decisions and more on their expected future value, which is simply the mean of the distribution of β. In the context of Figure 2.4b and Figure 2.5a, this causes the average trading quantities and trading frequencies of traders with different β to move closer together. This strengthens the positive correlation between their holdings and the amount they sell, and their inventories—defined as the absolute difference between their holdings and the per-capita supply of the asset—and the amount they trade. These moments thus pin down the frequency of shocks.

2.6 Results

Table 2.5 shows the estimated parameters. Below we discuss the interpretation of each of the parameters in turn.

Figure 2.6 shows the estimated distribution of the search parameter z, and how the estimated trading rates fit those observed in the data. The model fits the empirical distribution of trading quite well, replicating the heterogeneous and skewed distribution of trading rates

Table 2.5: Parameter Estimates

Parameter	Estimate
Search efficiency $z \sim \Gamma(k_z, \theta_z)$	
k_z	0.545
	(0.46; 0.632)
θ_z	0.374
	(0.333; 0.429)
Shock frequency η	
η	0.04
	(0.038; 0.042)
Utility $u(h) = \beta h - 0.5\kappa h^2$; $\beta \sim U(\mu_\beta, \sigma_\beta)$	
μ_β	0.031
	(0.029; 0.034)
σ_β	0.015
	(0.014; 0.016)
κ	0.008
	(0.007; 0.008)

Note: Parameter z in search cost function $z = (\gamma - z)^2$ is distributed according to a Gamma distribution with shape parameter k_z and scale parameter θ_z. Parameter η is the frequency (per month) at which traders draw new values β. Shocks β are distributed according to a uniform distribution with mean μ_β and variance σ_β^2. Parameter κ governs traders' risk aversion. 95-percent confidence intervals are shown in parentheses, and are obtained by bootstrapping the data using 100 replications.

we observe in the data (Figure 2.1). The model rationalises this with a heterogeneous distribution of search efficiency (the inverse of search costs). Intuitively, this means some traders have a meeting technology that enables them to meet counterparties frequently at low cost, whereas others do not. Institutionally, these active traders could be large investment banks that have large client networks who they contact regularly. This large client network could come from some more efficient technology—these traders are simply more effective at being middlemen—or could be due to the fact that they often have other business lines which help them create contacts. The less active traders do not enjoy the same technologies or client contacts, and are thus less able to trade frequently.

Shocks are infrequent—a trader faces a shock on average 0.5 times a year. These infrequent shocks are consistent with the infrequent trading we observe in corporate bond markets. Traders trade more frequently—0.4 times a month (Table 2.6)—than they are shocked. In a frictionless market these two frequencies—the trading frequency and the shock frequency—would be the same. This difference can be explained by two factors: (1) When it meets a counterparty, a trader is unable to trade the full amount it would like, and will thus trade

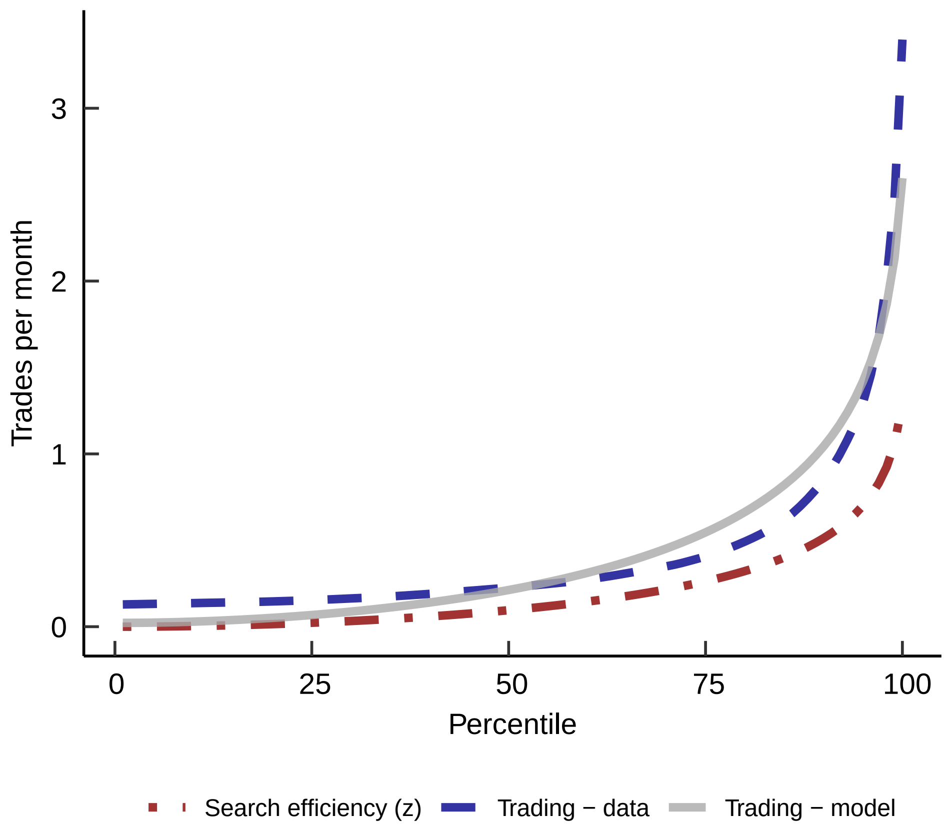


Figure 2.6: Distribution of trading frequencies and search parameter z

Note: This chart shows (a) the estimated distribution of the parameter z in traders' search cost function $s(z,\gamma) = (\gamma - z)^2$, where z follows a Gamma distribution (red dashed line), (b) the distribution of traders' average trading rates in the estimated model (grey line); and (c) the empirical distribution of traders' average trading rates (blue dashed line). To compute the empirical distribution, we first compute the distribution of traders' average trading rates per bond, and then take the weighted average of this across bonds, weighted by the number of traders in each bond.

again; and (2) Traders do not just trade to demand liquidity when shocked, but also to supply liquidity to other traders.

The parameters of the utility function $u(\beta, h) = \beta h - 0.5\kappa h^2$ are reduced-form parameters, and thus their magnitudes are difficult to interpret in terms of fundamental preferences.[21] Here we limit ourselves to a discussion of the relative magnitudes of the utility parameters β and κ, and note that below we will show the model fits the data very well. Both the variance of β and the level of κ govern variation in the slope of the utility function: κ creates variation

[21] As explained in Section 2.4 $u(\beta, h)$ is a reduced-form version of a more fundamental model with CARA preferences over consumption. However, given we cannot identify the parameters of this more fundamental setup, it is hard to map our parameter estimates into this setting in a quantitative sense.

Table 2.6: Model fit

Moment	Data	Model
Mean price, %	1.09	1.09
Std. dev. price within traders, %	0.04	0.04
Mean holdings, £mn	3.27	3.27
Std. dev. holdings within traders, £mn	1.23	1.31
Mean trade size, £mn	0.67	0.55
Mean trading frequency, per month	0.44	0.43
Std. dev. trading frequency across traders, per month	0.55	0.55
Correlation inventory & trading frequency	0.08	0.09
Correlation holdings & quantity sold	0.33	0.30

Note: This table shows empirical moments and simulated moments calculated using the parameter estimates in Table 2.5.

in marginal utility across holdings for a given β, whilst a shock to β changes the marginal utility for a given holding. The variation caused by β is quantitatively larger than that caused by κ. Concretely, a one standard deviation shock to β causes a change in marginal utility that is 50% greater than that caused by a one standard deviation change in holdings, which operates via κ. The concavity of a trader's utility function, which is governed by κ, is thus small relative to the distance between the utility functions of traders with different valuations β. This limited concavity of traders' flow utility is what lies behind the limited concavity of the value function displayed in Figure 2.4a. As a result, transferring the asset between traders with the same valuation β has a limited effect on their welfare.

Table 2.6 summarises the model fit. The model fits price, holdings, trading quantity and trading frequency moments well. Two moments merit further discussion. In the data, variation in price is more limited than variation in asset holdings: the standard deviation of the prices a trader pays for a bond is around 4% of the mean, whilst the standard deviation of their holdings of the bond is 40% of the mean. Fitting these two moments is what drives the risk aversion parameter κ to be quite low in our results, as low risk aversion reduces the variability of prices but increases the variability of holdings.

2.7 Counterfactuals

In this section we run counterfactual simulations to study the determinants of liquidity. We begin by establishing which traders supply liquidity. We show that a small set of traders with low search costs supplies the bulk of liquidity in these markets, and that as a result market outcomes are highly sensitive to the actions of these traders. We then study two counterfactuals that affect these traders' incentives to supply liquidity. The first – bank capital regulation – changes the cost to these liquidity suppliers of holding the asset. The second – the introduction of trading platforms – changes the cost advantage these liquidity suppliers enjoy over the rest of the market.

2.7.1 Resilience of Market Liquidity

In this section we study who supplies liquidity in these markets, and what this means for the resilience of liquidity to shocks. To do this, we 'withdraw' sets of traders from markets, which entails them selling the asset and stopping searching for counterparties.[22] We then solve for the new equilibrium in the market with the remaining traders, and show how liquidity has changed. This offers a convenient way to quantify each trader's contribution to market liquidity, and also to study the impact of shocks to traders that lead them to sell their assets and stop trading.

Figure 2.7 shows the relative contributions to market liquidity of the lowest- and highest-search cost traders in our model. The blue dashed line shows the fall in market depth when we withdraw the low-cost traders. Liquidity decreases sharply as we withdraw the traders with the lowest costs, with the pace of the decline in liquidity decreasing as we remove more and more traders. The red line shows the fall in market depth when we withdraw the high-cost traders. Liquidity barely changes as we withdraw the least efficient traders, but progressively deteriorates as we withdraw more traders.

The top 8% most frequent traders supply as much liquidity as all other traders combined (Figure 2.7). This asymmetry in the provision of liquidity is driven by the data. Our model maps traders' average trading frequencies into search efficiency parameters z. The stark heterogeneity observed in traders' average trading frequencies (Figure 2.1) thus drives the highly skewed distribution of z we obtain in Figure 2.7. When we counterfactually remove low- and high-cost traders from the market, this skewed distribution in z results in very different effects on market liquidity.

This result has direct implications for the resilience of market liquidity to shocks to traders. As a result of their disproportionate contribution to market liquidity, the market is

[22]This is equivalent to assuming they still meet counterparties, but refuse to trade with them.

Figure 2.7: Trader contributions to market liquidity

Note: In this figure we show that low-search-cost traders are disproportionately important for market liquidity. We summarise each trader's contribution to market liquidity by withdrawing sets of traders and computing the effect on liquidity. The x-axis shows the percentage of traders we withdraw. The y-axis shows the resulting fall in market depth relative to its level in the estimated equilibrium, where depth $\Pi(D)$ is defined in equation (2.9), and D is set to 1%. The red line shows the fall in liquidity when we sequentially remove traders starting from those with the highest search costs. The blue dashed line shows the fall in liquidity when we start from those with the lowest search costs. The grey dashed lines show the points where removing the 8% of traders with the lowest costs has the same effect on liquidity as removing the 92% with the highest costs.

vulnerable to shocks to the lowest-cost traders that lead them to withdraw from the market. We demonstrate in Table 2.7 by showing how welfare and liquidity would change if the 4% of traders with the lowest search costs withdrew from the market. The impacts of this small change in the trader population are extreme. The remaining traders, with their inferior search technologies, are able to meet each other less frequently. Market depth falls by 62%, whilst price volatility increases by 140%. The remaining traders are forced to absorb the exiting traders' sales, which, due to the concavity of their value functions, lowers the price. In response to the poorer liquidity conditions traders shrink the variance of their holdings. In total the welfare of traders falls by 5%.

Table 2.7: Effects of 4% of traders withdrawing

	% change in counterfactual
Market depth	-62
Holdings variance	-9
Mean trading frequency	-17
Mean price	-23
Price variance	140
Aggregate utility	-5

Note: This table shows the impact of the 4% of traders with the lowest search costs selling their asset holdings and stopping trading. Holdings and price variance are within-trader variances: we take the variance for a given trader and average across traders. Market depth, which we define in equation (2.9) is a measure of how much traders are able to sell without reducing the trading price by more than 1%.

This stark dependence of liquidity on a few key traders can help explain patterns in trading and liquidity during the Covid-19 crisis in March 2020. In the early parts of March, dealers were unwilling to expand their inventories (Kargar, Lester, Lindsay, Liu, Weill and Zuniga, 2021). During this period, market conditions deteriorated rapidly, with prices falling and the cost of transacting increasing. From 18$^{\text{th}}$ March, the Federal Reserve took large scale policy action to boost liquidity in markets, at which point dealers began to allow their inventories to increase and markets stabilised (Kargar, Lester, Lindsay, Liu, Weill and Zuniga, 2021). This is consistent with the findings with Table 2.7, which show that bond markets are highly reliant on the appetite for trading of a small proportion of traders.

2.7.2 Regulating Dealers

Tighter capital regulation increases the fraction of a bank's balance sheet that must be funded by equity, rather than debt. If a bank applies this evenly across all its business, this increases the fraction of any position—short or long—that must be funded by equity.[23] There is relatively broad agreement that capital regulation enhances financial stability by increasing bank resilience, but there are concerns that it also harms liquidity in financial markets (Duffie, 2018). In this section we study this potential cost of capital regulation.

There are three mechanisms by which tighter capital regulation can reduce a bank's incentive to hold assets. The first is via violations of the Modigliani-Miller theorem (Modigliani

[23] Whilst capital regulation applies at the bank level, banks' internal capital markets are typically organised such that increases in capital requirements take effect across the bank's business lines (Bajaj, Binmore, Dasgupta and Vo, 2018).

and Miller, 1958). A bank's cost of equity is typically higher than its cost of debt. In the absence of any frictions, increasing a bank's equity funding will not increase its overall cost of capital, as its costs of equity and of debt will fall. In a world with frictions, its cost of capital will increase.[24] The second is via a violation of the debt overhang problem (Myers, 1977) discussed by Andersen, Duffie and Song (2019) and Duffie (2018), by which capital regulation reduces a bank's willingness to take a new, relatively safe asset onto its balance sheet. Finally, in the short run banks are unable to issue new equity. As a result, a binding capital requirement in the short run sets an upper limit on banks' asset holdings.

Given dealer-affiliated banks have typically acted as intermediaries in fixed income markets and are amongst the most frequent traders in the sterling corporate bond market, there are concerns that bank capital regulation could harm market functioning. In the context of our model, bank capital regulation affects the incentives of the low-search cost traders that supply liquidity to hold the asset. To study this, we simulate an increase in the cost of holding a long or short position in an asset for the traders with the lowest search costs. Their flow utility function becomes:

$$u(\beta, h) = \beta h - \kappa \frac{h^2}{2} - \tau |h|, \tag{2.12}$$

where $\tau \geq 0$ represents the increase in costs due to regulation.

Table 2.8 summarises how market quantities change when we set $\tau = 1\%$ for the 15% of traders with the lowest search costs. These figures are such that if trader behaviour did not adjust at all, dealers' utility would fall by 40%. Given that it is difficult to map a level of τ into a level of capital regulation, the figures that follow are best interpreted in relative terms, rather than as the absolute impact of a given level of capital regulation.

Tighter regulation of dealers leads them to significantly decrease their holdings of the asset, as they find holding inventory more costly. Market clearing implies unregulated traders increase their holdings. Tighter regulation, by reducing the marginal utility of holding the asset for dealers, reduces the average trading price.

[24] Potential violations of the Modigliani-Miller assumptions include the tax advantage of debt (Kashyap, Stein and Hanson, 2010), asymmetric information leading to a 'pecking order' theory of financing (Myers and Majluf, 1984), agency problems of bank management (Diamond and Rajan, 2001) and the possibility that banks' short-term debt enjoys a 'money-like' convenience yield (Stein, 2012; Kashyap, Stein and Hanson, 2010). Note that in the case of the tax shield on debt, a full welfare analysis would need to take into account the increase in tax revenues resulting from higher capital requirements (Admati, DeMarzo, Hellwig and Pfleiderer, 2013).

Table 2.8: Capital counterfactual

	% change in counterfactual
Dealers	
Expected holdings	-33
Spread received	17
Expected utility	-37
Other traders	
Expected holdings	6
Spread paid	17
Expected utility	1
Aggregate	
Price	-19
Market depth	-10
Aggregate welfare	-5

Note: This table summarises the changes in key variables in the capital counteractual, relative to our baseline equilibrium. Capital regulation is modelled as setting $\tau = 1\%$ for the 15% of traders with the lowest search costs, where their flow utility from holding the asset is given by $u(\beta, h) = \beta h - 0.5\kappa h^2 - \tau|h|$. Market depth, which we define in equation (2.9) is a measure of how much traders are able to sell without reducing the trading price by more than 1%.

Dealers adjust their trading frequency in two ways, as shown in Figure 2.8. Firstly, their target holding—where search is at its minimum—shifts to the left, reflecting their reduced desire to hold inventory. Secondly, the slope of the trading frequency curve becomes steeper when they are away from their target, meaning regulation leads them to search harder when away from target inventory. This is the mechanism driving the empirical finding in Figure 2.3 that dealers' tendency to offset trades is related to capital regulation, and stems from the fact that regulation has increased the costs of being away from target holding.

Tighter regulation increases the spreads earned by dealers (Table 2.8). This reflects the balance of two forces. On the one hand, regulation means dealers are less inclined to trade away from target holding, as the costs of taking large positions are greater. All else equal, this would increase the spread required to persuade them to do so. On the other hand, conditional on being away from target holding, dealers are more anxious to trade back to target. This would decrease the spread they pay, as in these cases they are demanding, rather than supplying, liquidity. The former effect dominates the latter, meaning dealers' realised spreads increase.

Tighter capital regulation reduces market depth, our summary measure of liquidity, by 10%. This is driven by dealers' decreased willingness to take large positions in the asset

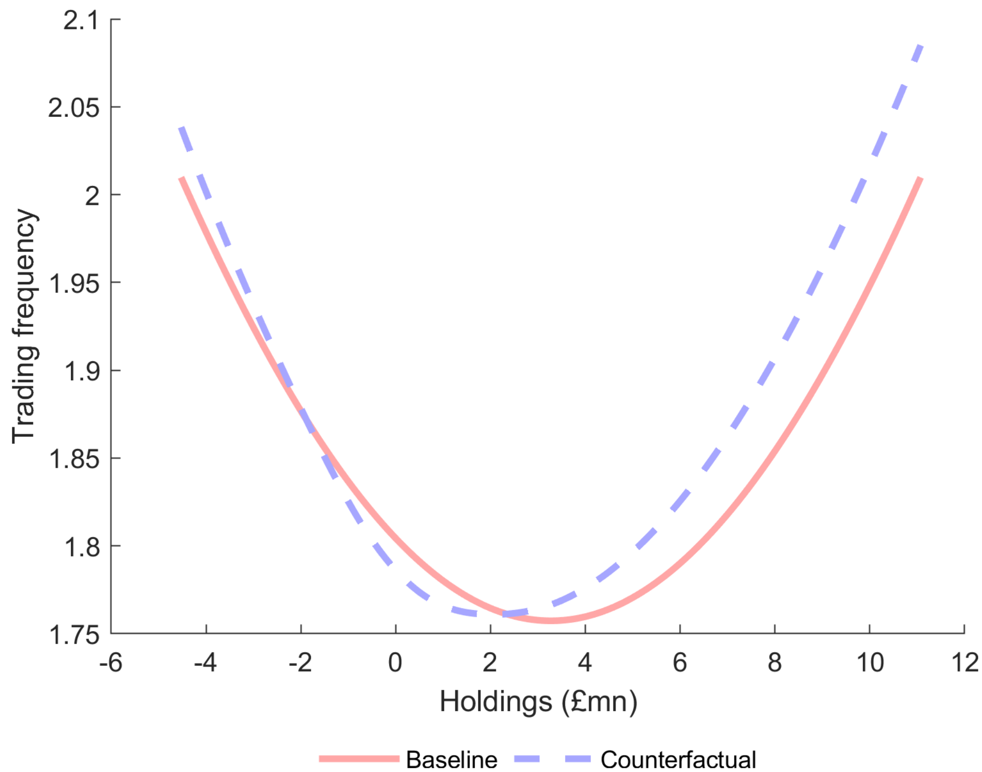


Figure 2.8: Dealer capital regulation and search

Note: This figure shows how capital regulation affects dealer search. We show that capital regulation shifts dealers' target holdings downwards and causes them to search harder when they are away from their target holdings. Capital regulation is modelled as setting $\tau = 1\%$ for the 15% of traders with the lowest search costs, where their flow utility from holding the asset is given by $u(\beta, h) = \beta h - 0.5\kappa h^2 - \tau|h|$.

now inventory costs have increased. Aggregate welfare in this market falls by 5%. If traders were unable to adjust their trading or holdings decision, the welfare cost would be 7%. In equilibrium, the costs of capital regulation are mitigated by the effects of endogenous search—with dealers optimally adjusting their search behaviour to minimise the harm caused by tighter regulation—and the endogenous reallocation of asset holdings from dealers to non-dealers.

The magnitudes of these effects are driven by the distributions of the search efficiency parameter z and the level of risk aversion κ. The fact that search efficiencies are highly skewed, and that capital regulation is applied to dealers, mean that capital regulation affects traders who play a key role in supplying liquidity. If the distribution of search efficiency were more homogeneous, the effect on liquidity would be more limited. Risk aversion κ governs the curvature of the value function. The greater is κ, the less willing non-dealers are to increase their holdings of the asset in response to capital regulation. The fact that κ is relatively low means that the transfer of holdings from dealers to non-dealers is relatively large, which limits the welfare costs of capital regulation.

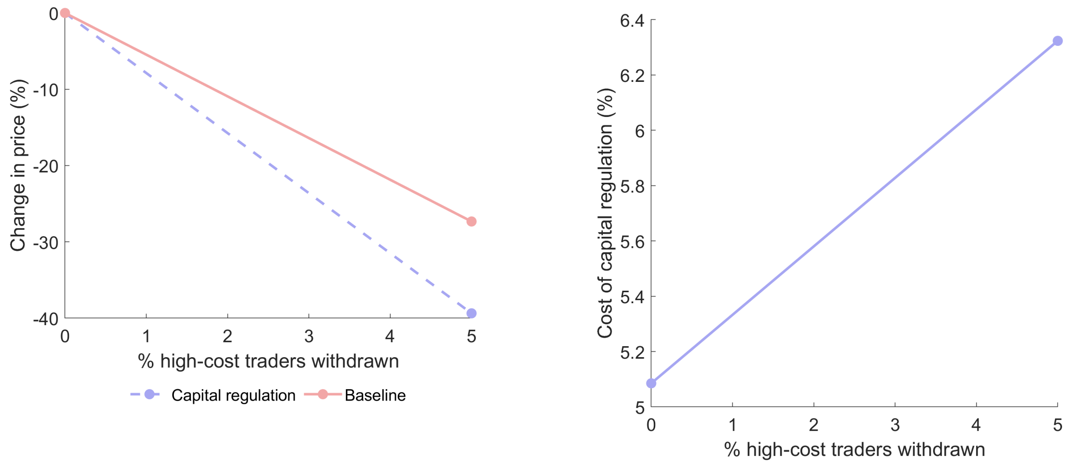
Liquidity, Regulation and Trading in a Sell-off

The negative effects of illiquidity can be particularly large during stress, amplifying the effects of financial shocks (Brunnermeier and Pedersen, 2009; Gromb and Vayanos, 2002). There is some evidence that liquidity during stress events—for example when a bond is downgraded or is removed from a market index—deteriorated after the 2008 financial crisis, as dealers were subject to more stringent regulation (Bao, O'Hara and Zhou, 2018; Dick-Nielsen and Rossi, 2019). In each of these cases, the stress event is a shock that causes traders who are not dealers—such as traders who track market indices—to sell the bond. An important advantage of a structural model is that we are able to counterfactually simulate how the effects of capital regulation vary during these types of sell-offs.

To do this, we subject dealers to capital regulation at the same time as assuming high-cost traders set their asset holdings to zero and withdraw from the market. Figure 2.9 shows the results of this. Figure 2.9a shows the change in prices as a function of high-cost trader withdrawals both in the baseline (red line) and when dealers are subject to capital regulation (blue line). With capital regulation, the fall in price is greater. Figure 2.9b shows the fall in aggregate welfare due to capital regulation as a function of high-cost trader withdrawals. The cost of capital regulation is greater in a stress. Intuitively, the costs of capital regulation are limited by the endogenous responses of customers, who increase their asset holdings by trading with dealers who wish to offload the bond. As these traders stop trading, these margins of adjustment are constrained, and the costs of capital regulation increase.

2.7.3 Trading Technologies

Trading in corporate bond markets is predominantly undertaken via traditional methods: dealers intermediate the market, and trades are organised bilaterally between traders, who often communicate on the phone. In recent years there has been a gradual increase in the use of electronic trading platforms (Anderson, Webber, Noss, Beale and Crowley-Reidy, 2015), which replace bilateral negotiation with a multilateral system where a trade quote is posted to all platform members, any of whom can bid to be counterparty to the trade. These



(a) Prices in a stress

(b) Welfare in a stress

Figure 2.9: Impact of capital regulation in a stress

Note: In this figure we show that the negative effect of capital regulation on welfare and prices is worse in a stress. Stress—along the x-axis—is the percentage of traders that withdraw from the market, selling their asset and stopping trading. Traders are withdrawn in descending order of their search costs, starting from the highest-cost traders. The left-hand chart plots the change in price as a percentage of par against the level of stress, when dealers are subject to capital regulation (blue dashed line) and when they are not (red line). The right-hand chart shows the level of aggregate welfare when dealers are subject to capital regulation vs when they are not, as a function of the level of stress. Capital regulation is modeled as setting $\tau = 1\%$ for the 15% of traders with the lowest search costs, where their flow utility from holding the asset is given by $u(\beta, h) = \beta h - 0.5\kappa h^2 - \tau|h|$.

types of platforms offer potential efficiency savings, particularly for those who do not have large client networks with which to trade. Such trading innovations require the participation of sufficient traders in order to succeed—there needs to be sufficient supply of assets and trading activity in order to make setting up such a platform worthwhile. The successful implementation of these innovations thus depends on the participation of the most efficient traders, as these are the largest and most frequent traders.

The take-up of these new technologies has been relatively slow (The Economist, 2020). In this section we offer an explanation for this: the most efficient traders in corporate bond markets lose out under these new technologies. To show this, we run two counterfactual simulations. First, we *reduce* and *homogenise* search costs in traders. Specifically, we take the highest value of search efficiency z in our (discretised) distribution of search costs, and counterfactually assign all traders this value of z. This captures the fact that platform-based trading has the potential to increase search efficiency, particularly for the most inefficient traders. Secondly, we study the Walrasian equilibrium, where trading is frictionless and prices are no longer bilaterally negotiated. This captures the increase in search efficiency on platforms and also the change in trading mechanism they introduce.

Table 2.9: Counterfactual search technologies

	Baseline		Homogenous	Walrasian
	Low cost	High cost	Aggregate	Aggregate
Trades per month	1.76	0.20	1.75	0.03
Spread, bps	223	-223	0	0
Utility	16.3	13.3	14.5	14.5
Aggregate				
Price variation, bps	378		90	0
Trade size, £mn	0.5		0.5	1.9
Gross volume, £mn	0.07		0.23	0.08
Utility	13.8		14.5	14.5

Note: This table shows the effect of platform-based trading on trading, spreads and welfare. The first two columns show results in our baseline equilibrium. The next column shows results when all traders' search costs are set to those of the lowest-cost trader. The final column shows results in the Walrasian equilibrium. Price variation is the standard deviation of prices within traders. Gross volume is instantaneous trading volume per trader. The spread is the difference between a trader's average selling price and buying price.

Table 2.9 summarises the effect of these changes in search technologies on trading and welfare. The penultimate column shows the effects of reducing and homogenising traders' search costs. Traders who initially have high search costs take advantage of the improvement in their meeting technologies by increasing their trading frequency. These traders benefit from the change in technology. Low-cost traders, however, lose out from it. The reason for this can be seen in the effect on spreads. Initially, low-cost traders on average earn trading revenues by charging a spread—they sell the asset at a high price and buy it at a low price. They are able to do this because of their technological advantage over higher-cost traders. When platforms erode this advantage, their gains from trade decrease. As a result, whilst aggregate welfare increases in the counterfactual, the welfare of the lowest-cost traders decreases. As shown in the final column of Table 2.9, the lowest-cost traders lose out even in the case where trading is frictionless.

This result is driven by the shape of the distribution of traders' search efficiencies z, which in turn is driven by the distribution of trading rates we observe in the data (Figure 2.1). The skewness of the estimated distribution of z means there are traders where the costs of frictionless trading exceed the benefits. All traders derive some benefit from frictionless trading, as they can respond instantly to shocks. The most efficient traders also incur a cost, as they no longer earn spreads. The heterogeneity of traders' search efficiencies means the spreads earned by the most efficient traders—and hence the losses they incur from frictionless trading—are large. Additionally, the fact that these efficient traders can meet counterparties much more frequently than the rate at which they are shocked means the benefits of frictionless markets—being able to respond to shocks instantly—are relatively small. As a result, these traders stand to lose out from platforms. This result is driven by the data in the sense that if we had estimated a distribution of z that was lower and more homogeneous, no traders would lose out from frictionless trading.

The most efficient traders thus have an incentive to resist shifts to trading platforms. This can explain why trading platforms have for a long time struggled to gain a foothold in corporate bond markets (Bessembinder, Spatt and Venkataraman, 2020).[25]

2.8 Conclusion

Liquidity is a key aspect of financial market functioning, but is hard to pin down. It is an equilibrium outcome that cannot be directly observed in the data. It varies with financial market conditions, and is impacted by regulatory and technological change. Liquidity, and how resilient it is in times of stress, is a key area of concern for academics (Duffie, 2018) and policymakers (Powell, 2015).

In this paper we present a quantitative model of liquidity in an OTC market. This enables us to document the reliance of market liquidity on a small set of traders, and to study the implications of this for the resilience of liquidity to trader stress, the impact of bank capital regulation, and the effects of new trading technologies.

[25] For example, in 2001 the online trading platform BondBook closed its operations on the basis that 'the behavioural change among market participants required for the platform to take off' was unlikely to materialise sufficiently quickly (see Finextra, "BondBook shuts down its trading operations', available at https://www.finextra.com/newsarticle/3605/bondbook-shuts-down-trading-operations). Similarly, in 2013 the major asset manager BlackRock shelved plans to develop its own in-house trading platform (see Grind, L. and T. Demos "BlackRock Shelves Platform For Bonds", available at https://www.wsj.com/articles/SB10001424127887323551004578441053526969438).

Bibliography

Admati, A. R., DeMarzo, P. M., Hellwig, M. F. and Pfleiderer, P. C. (2013), 'Fallacies, irrelevant facts, and myths in the discussion of capital regulation: Why bank equity is not socially expensive', *Max Planck Institute for Research on Collective Goods* **23**.

Adrian, T., Fleming, M., Shachar, O. and Vogt, E. (2017), 'Market liquidity after the financial crisis', *Annual Review of Financial Economics* **9**, 43–83.

Afonso, G. and Lagos, R. (2015), 'Trade dynamics in the market for federal funds', *Econometrica* **83**(1), 263–313.

Allen, J. and Wittwer, M. (2021), Centralizing over-the-counter markets, Working Paper, Stanford University.

Andersen, L., Duffie, D. and Song, Y. (2019), 'Funding value adjustments', *The Journal of Finance* **74**(1), 145–192.

Anderson, N., Webber, L., Noss, J., Beale, D. and Crowley-Reidy, L. (2015), 'The resilience of financial market liquidity', Bank of England Financial Stability Paper.

Babus, A. and Kondor, P. (2018), 'Trading and information diffusion in over-the-counter markets', *Econometrica* **86**(5), 1727–1769.

Bajaj, R., Binmore, A., Dasgupta, R. and Vo, Q.-A. (2018), 'Banks' internal capital markets: how do banks allocate capital internally?', *Bank of England Quarterly Bulletin* p. Q2.

Bank of England (2016), 'Financial stability report: November 2016', Technical report, Bank of England.

Bank of England (2021), 'Financial Policy Committee's core indicators', https://www.bankofengland.co.uk/financial-stability. Accessed: 05/08/2021.

Bao, J., O'Hara, M. and Zhou, X. A. (2018), 'The volcker rule and corporate bond market making in times of stress', *Journal of Financial Economics* **130**(1), 95–113.

BCBS (2006), 'International convergence of capital measurement and capital standards', Technical report, Bank for International Settlements.

BCBS (2017), 'Basel III: finalising post-crisis reforms', Technical report, Bank for International Settlements.

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.

Bessembinder, H., Spatt, C. and Venkataraman, K. (2020), 'A survey of the microstructure of fixed-income markets', *Journal of Financial and Quantitative Analysis* **55**(1), 1–45.

Biais, B., Declerck, F. and Moinas, S. (2016), 'Who supplies liquidity, how and when?', Bank for International Settlements Working Paper 563.

BlackRock (2015), 'Addressing market liquidity', Technical report, BlackRock.

Brancaccio, G. and Kang, K. (2021), 'Intermediaries and product selection in the municpal bond market', Unpublished Working Paper.

Brancaccio, G., Li, D. and Schürhoff, N. (2020), 'Learning by trading: The case of the US market for municipal bonds', Working Paper, New York University Stern School of Business.

Brunnermeier, M. K. and Pedersen, L. H. (2009), 'Market liquidity and funding liquidity', *The review of financial studies* **22**(6), 2201–2238.

Cai, Y. and Judd, K. L. (2013), 'Shape-preserving dynamic programming', *Mathematical Methods of Operations Research* **77**(3), 407–421.

Chang, B. and Zhang, S. (2021), Endogenous market making and network formation, Working Paper, London School of Economics.

Choi, J. and Huh, Y. (2018), 'Customer liquidity provision: Implications for corporate bond transaction costs', Working paper, University of Illinois Urbana-Champaign.

Cimon, D. and Garriott, C. (2019), 'Banking regulation and market making', *Journal of Banking & Finance* **109**, 105653.

Coen, P. and Coen, J. (2019), 'A structural model of interbank network formation and contagion', Working paper, London School of Economics.

Craig, B. R. and Ma, Y. (2021), 'Intermediation in the interbank lending market', Working Paper, Columbia Business School.

Czech, R., Huang, S., Lou, D. and Wang, T. (2021), 'Informed trading in government bond markets', *Journal of Financial Economics* .

Czech, R. and Roberts-Sklar, M. (2019), 'Investor behaviour and reaching for yield: Evidence from the sterling corporate bond market', *Financial Markets, Institutions & Instruments* **28**(5), 347–379.

Di Maggio, M., Kermani, A. and Song, Z. (2017), 'The value of trading relations in turbulent times', *Journal of Financial Economics* **124**(2), 266–284.

Diamond, D. W. and Rajan, R. G. (2001), 'Liquidity risk, liquidity creation, and financial fragility: A theory of banking', *Journal of Political Economy* **109**(2), 287–327.

Dick-Nielsen, J. and Rossi, M. (2019), 'The cost of immediacy for corporate bonds', *The Review of Financial Studies* **32**(1), 1–41.

Du, W., Tepper, A. and Verdelhan, A. (2018), 'Deviations from covered interest rate parity', *The Journal of Finance* **73**(3), 915–957.

Duffie, D. (2010*a*), 'The failure mechanics of dealer banks', *Journal of Economic Perspectives* **24**(1), 51–72.

Duffie, D. (2010*b*), 'Presidential address: Asset price dynamics with slow-moving capital', *The Journal of finance* **65**(4), 1237–1267.

Duffie, D. (2018), 'Financial regulatory reform after the crisis: An assessment', *Management Science* **64**(10), 4835–4857.

Duffie, D., Gârleanu, N. and Pedersen, L. H. (2005), 'Over-the-counter markets', *Econometrica* **73**(6), 1815–1847.

Duffie, D., Gârleanu, N. and Pedersen, L. H. (2007), 'Valuation in over-the-counter markets', *The Review of Financial Studies* **20**(6), 1865–1900.

Eisfeldt, A. L., Herskovic, B., Rajan, S. and Siriwardane, E. (2020), OTC intermediaries, Working paper, UCLA.

Farboodi, M. (2021), 'Intermediation and voluntary exposure to counterparty risk', Working Paper, Massachusetts Institute of Technology.

Farboodi, M., Jarosch, G. and Shimer, R. (2021), 'The emergence of market structure', Working Paper, Massachusetts Institute of Technology.

Foucault, T., Pagano, M. and Röell, A. (2013), *Market liquidity: theory, evidence, and policy*, Oxford University Press.

Franzoni, F. A., Plazzi, A. and Çötelioğlu, E. (2019), 'What constrains liquidity provision? Evidence from hedge fund trades', CEPR discussion paper 13645.

Gârleanu, N. (2009), 'Portfolio choice and pricing in illiquid markets', *Journal of Economic Theory* **144**(2), 532–564.

Gavazza, A. (2016), 'An empirical equilibrium model of a decentralized asset market', *Econometrica* **84**(5), 1755–1798.

Goodman, T. (2001), 'Shape preserving interpolation by curves', Working paper, Dundee University.

Gromb, D. and Vayanos, D. (2002), 'Equilibrium and welfare in markets with financially constrained arbitrageurs', *Journal of Financial Economics* **66**(2-3), 361–407.

Gromb, D. and Vayanos, D. (2018), 'The dynamics of financially constrained arbitrage', *The Journal of Finance* **73**(4), 1713–1750.

Grossman, S. J. and Miller, M. H. (1988), 'Liquidity and market structure', *The Journal of Finance* **43**(3), 617–633.

Heller, D. and Vause, N. (2012), Collateral requirements for mandatory central clearing of over-the-counter derivatives, Working paper, Bank for International Settlements.

Hendershott, T. and Madhavan, A. (2015), 'Click or call? auction versus search in the over-the-counter market', *The Journal of Finance* **70**(1), 419–447.

Hollifield, B., Neklyudov, A. and Spatt, C. (2017), 'Bid-ask spreads, trading networks, and the pricing of securitizations', *The Review of Financial Studies* **30**(9), 3048–3085.

Hugonnier, J., Lester, B. and Weill, P.-O. (2020), 'Frictional intermediation in over-the-counter markets', *The Review of Economic Studies* **87**(3), 1432–1469.

Kargar, M., Lester, B., Lindsay, D., Liu, S., Weill, P.-O. and Zuniga, D. (2021), 'Corporate bond liquidity during the Covid-19 crisis', *The Review of Financial Studies* **34**(11), 5352–5401.

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

Lagos, R. and Rocheteau, G. (2007), 'Search in asset markets: Market structure, liquidity, and welfare', *American Economic Review* **97**(2), 198–202.

Lagos, R. and Rocheteau, G. (2009), 'Liquidity in asset markets with search frictions', *Econometrica* **77**(2), 403–426.

Li, D. and Schürhoff, N. (2019), 'Dealer networks', *The Journal of Finance* **74**(1), 91–144.

Li, J. (2021), 'The importance of investor heterogeneity: An examination of the corporate bond market', Working Paper, The University of Chicago.

Liu, S. (2020), 'Dealers' search intensity in US corporate bond markets', Working paper, Tsinghua University.

Mallaburn, D., Roberts-Sklar, M. and Silvestri, L. (2019), 'Resilience of trading networks: evidence from the sterling corporate bond market', Bank of England Working Paper 813.

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

Myers, S. C. (1977), 'Determinants of corporate borrowing', *Journal of financial economics* **5**(2), 147–175.

Myers, S. C. and Majluf, N. S. (1984), 'Corporate financing and investment decisions when firms have information that investors do not have', *Journal of financial economics* **13**(2), 187–221.

Neklyudov, A. (2019), 'Bid-ask spreads and the over-the-counter interdealer markets: Core and peripheral dealers', *Review of Economic Dynamics* **33**, 57–84.

Pinter, G. and Uslu, S. (2022), Comparing search and intermediation frictions across markets, Working paper, Bank of England.

Plante, S. (2018), 'Should corporate bond trading be centralized? Theory and evidence', Working Paper, Wisconsin School of Business.

Powell, J. H. (2015), 'Structure and liquidity in treasury markets', Speech at the Brookings Institution, Washington, DC, August 3, 2015, Technical report, Board of Governors of the Federal Reserve System (US).

Saar, G., Sun, J., Yang, R. and Zhu, H. (2020), 'From market making to matchmaking: Does bank regulation harm market liquidity?', Working Paper, Cornell University.

Sambalaibat, B. (2018), 'A theory of liquidity spillover between bond and CDS markets', Working paper, Princeton University.

Schultz, P. (2017), 'Inventory management by corporate bond dealers', Working paper, University of Notre Dame.

Shimer, R. and Smith, L. (2001), 'Matching, search, and heterogeneity', *The BE Journal of Macroeconomics* **1**(1).

Stein, J. C. (2012), 'Monetary policy as financial stability regulation', *The Quarterly Journal of Economics* **127**(1), 57–95.

The Economist (2020), 'Electronic trading platforms are challenging bond broker-dealers', 2020-03-07.

Trebbi, F. and Xiao, K. (2017), 'Regulation and market liquidity', *Management Science* **65**(5).

Üslü, S. (2019), 'Pricing and liquidity in decentralized asset markets', *Econometrica* **87**(6), 2079–2140.

Vayanos, D. and Wang, J. (2013), 'Market liquidity - theory and empirical evidence', *in* 'Handbook of the Economics of Finance', Vol. 2, Elsevier, pp. 1289–1361.

Vayanos, D. and Weill, P.-O. (2008), 'A search-based theory of the on-the-run phenomenon', *The Journal of Finance* **63**(3), 1361–1398.

A Appendices

A1 Further Details on Data and Empirics

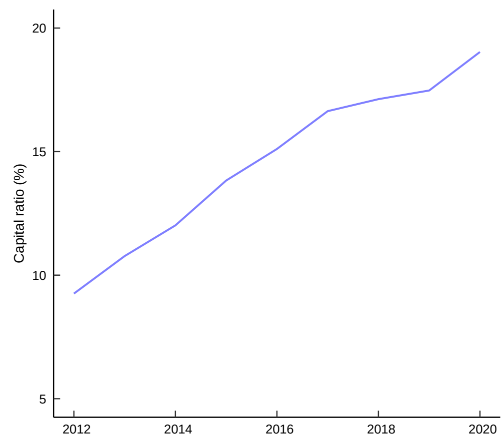
Traders are consolidated to the group level, with the exception of the asset management arms of large banking groups, who we separate from the banking business. We do this as these two business lines report separately in the data, and they tend to operate with separate balance sheets and trading strategies. We winsorize both trading prices and quantities at the 1^{st} and 99^{th} percentiles. We exclude instruments that have been traded fewer than 10 times in total in our dataset, to focus on instruments where we have a meaningful number of observations. Finally, we restrict the sample to trades that take place at least 10 days after a bond is issued, to remove any trading in the primary market.

The data present three practical challenges, common when dealing with transactions data: (a) duplicate reporting where, for example, both the buyer and the seller report the transaction; (b) how to treat agency trades, where a firm trades a bond on behalf of another firm; and (c) how to deal with trades with missing counterparty names and IDs.

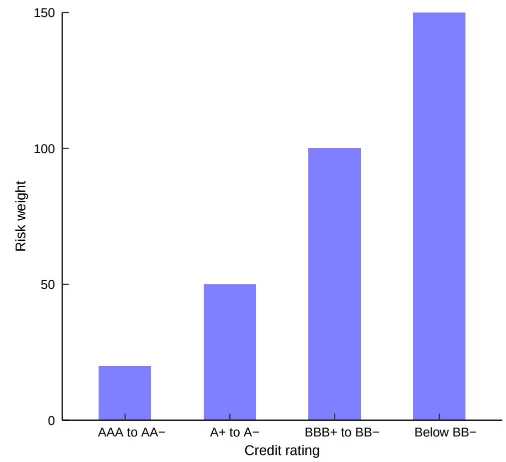
To remove duplicate trade reports, we identify all trades where the instrument, trading quantity and price match across two trade reports, and the firms reporting the trades differ. We then remove one of these duplicate reports.

With agency trades, we aim to distinguish between two types. The first is agency trading where the client firm on whose behalf the bond is being traded is a non-trading firm, for example a client of a wealth manager or a mutual fund within an asset manager's group. For the purposes of the model, we consolidate the trading firm and its clients into a single group represented by its trading entity - the wealth management firm or the asset management firm in the examples above. The second is agency trading between two trading firms. For the purposes of the model, these are trades between two distinct trading firms, and are thus not consolidated in any way. In the empirical results in Section 2.3 we (a) do not include agency trades when computing spreads, as agency trades do not earn a spread; and (b) treat each leg of an agency trade as a distinct trade. For example, if A buys a bond from B for C, we count this as two trades: one between A and B, and one between A and C.

In some cases a counterparty in a trade is identified only by an internal code, and not by name. This will most commonly be the case when the counterparty is not a trading firm but, for example, a client of an wealth manager. When computing firm-level summary



(a) UK banks' capital ratios



(b) Risk weights by credit rating

Figure A1: Bank capital ratios and risk weights.

Note: These figures summarise the risk-weighted capital regulations faced by banks. The left panel shows risk-weighted capital ratios for the UK banking system, taken from Bank of England (2021). Risk weights are for exposures to corporates according to the standardised approach to risk weighting, as described in BCBS (2006).

statistics such as the number of trading firms or the distribution of trading frequencies, we do not include unnamed firms as trading firms. However, we do take into account trades with unnamed firms when computing these statistics for named firms. For example, when computing a firm's holding period for a bond, we include their trades with unnamed counterparties to accurately reflect the changes in their asset portfolios.

Further Details on Institutional Setting

Figure A1 shows intertemporal and cross-sectional variation in capital regulation. Figure A1a shows how UK banks' average risk-weighted capital ratios have increased through time. Figure A1b shows how capital requirements vary according to a bond's credit rating.

A2 Computational Details

In this section we describe in detail how we solve the model.

At each step of the estimation procedure we solve for the unknown value, pricing, trading quantity, density and search intensity functions. The value shocks are assumed to take discrete values $\beta_1, ..., \beta_{n_\beta}$ and the search costs discrete values $z_1, ..., z_{n_z}$. Each of the functions are continuous in holdings h, and take the form of interpolation splines between the values $h_1, ..., h_{n_h}$. We thus need to solve for each of the functions at $n_T = n_z n_\beta n_h$ points.

We use a nonlinear solver to search over the search intensities γ and densities ϕ at each point on our grid. Conditional on these quantities, and the spline functions we fit through the holdings dimension, we can solve for the value functions V, the price function p and the quantity function q directly without resorting to numerical methods. This significantly reduces the dimensionality of the problem. Below we explain how we compute the value, pricing and quantity functions, before setting out equations we solve numerically to solve for the search and density functions.

Given our matching function (equation (2.10)), we can solve for the value function solely in terms of the search intensities. To see this, first plug the matching function into the expression for optimal search (equation (2.5)) and multiply by $\gamma(z, \beta, h)$, yielding:

$$\begin{aligned}\gamma(z,\beta,h)s_2(z,\gamma(z,\beta,h)) &= \\ &\frac{1}{2}\iiint m(\gamma(z,\beta,h),\gamma(z',\beta',h'))S((z,\beta,h),(z',\beta',h'))\Phi(dz',d\beta',dh'). \quad (2.13)\end{aligned}$$

Using this equation we can substitute out the final term of the value function (equation 2.4):

$$\begin{aligned}rV(z,\beta,h) = u(\beta,h) - s(z,\gamma(z,\beta,h)) + \\ \eta\int(V(z,\beta',h) - V(z,\beta,h))G(d\beta') + \gamma(z,\beta,h)s_2(z,\gamma(z,\beta,h)).\end{aligned}$$

All terms involving the value function V enter this equation linearly, the distribution function ϕ does not enter, and the functional form $s(z, \gamma)$ is known. As a result, conditional on knowing the function γ we have a closed form expression for the value function V at each of the points on our grid. Having computed the value function, we then interpolate along the holdings dimension using a cubic spline to give value functions $V(z, \beta, h)$ that are continuous in h. We can then solve exactly for trading quantity and price at each point on our grid using the Nash bargaining solution, which depends only on the value functions of the two traders that meet.

The variables we pass to the nonlinear solver thus consist of a set of search intensities γ and densities ϕ at each point on the grid. We then fit a cubic interpolation spline through the holdings grid to get a function for $\gamma(z, \beta, h)$ that is continuous in h. For the density $\phi(z, \beta, h)$

we fit a cubic Hermite spline through the holdings values, constraining the function to be a valid density function.[26]

The remaining equations that need solving numerically are the market clearing equation (equation (2.8)), the search intensity equation (equation (2.5)) and the distribution equation (equation (2.6)). The market clearing equation is straightforward to compute given the density function. We compute the terms of the search intensity equation based on a version of the discretized version of the density function ϕ at the points of the grid, with the surplus following directly from the value function and trading quantity.

The distribution function (equation 2.6) involves a double integral over the holdings of a trader and their potential counterparties involving the quantity traded, their measures and their search intensities, which is potentially difficult to evaluate. However, a property of the Nash bargaining solution simplifies this significantly. In particular, the post-trade holdings of trader (z, β, h) after meeting another trader (z', β', h') depend only on the sum of the two traders' pre-trade holdings, $h^T \equiv h + h'$. We can thus define $h^1(z, \beta, z', \beta', h^T)$ as the post-trade holdings of trader (z, β, h) after meeting trader $(z', \beta', h^T - h)$. Fitting a cubic spline through the holdings dimension of $h^1()$, for any given h^* we can solve for the level of h^T such that $h^1(z, \beta, z', \beta', h^T) = h^*$. The flows into and out of holdings $h < h^*$ from trades between types (z, β) and (z', β') are then as shown in Figure A2, and to get the trading flows we simply integrate meeting rates over the relevant areas. Given we have splines for the relevant expressions, this is an analytical integral, and is straightforward to compute. This enables us to compute the terms of the distribution function (equation (2.6)).

This process enables us to solve the model at each step of the estimation procedure. We then compute the theoretical moments described in Section 2.5.1 based on the model solution, and choose the parameters to minimise the distance between these moments and their theoretical counterparts.

A3 Moments

In this section we provide further details on some of the moments used in estimation.

- Correlation between quantity sold and holdings, within-traders:

[26] Hermite splines are particularly convenient for modelling densities as they are shape-preserving, meaning the interpolated curve and the points through which we are interpolating have the same local minima. This makes it easy to constrain the functions to be positive whilst still ensuring they are smooth. See Cai and Judd (2013) and Goodman (2001) for further details.

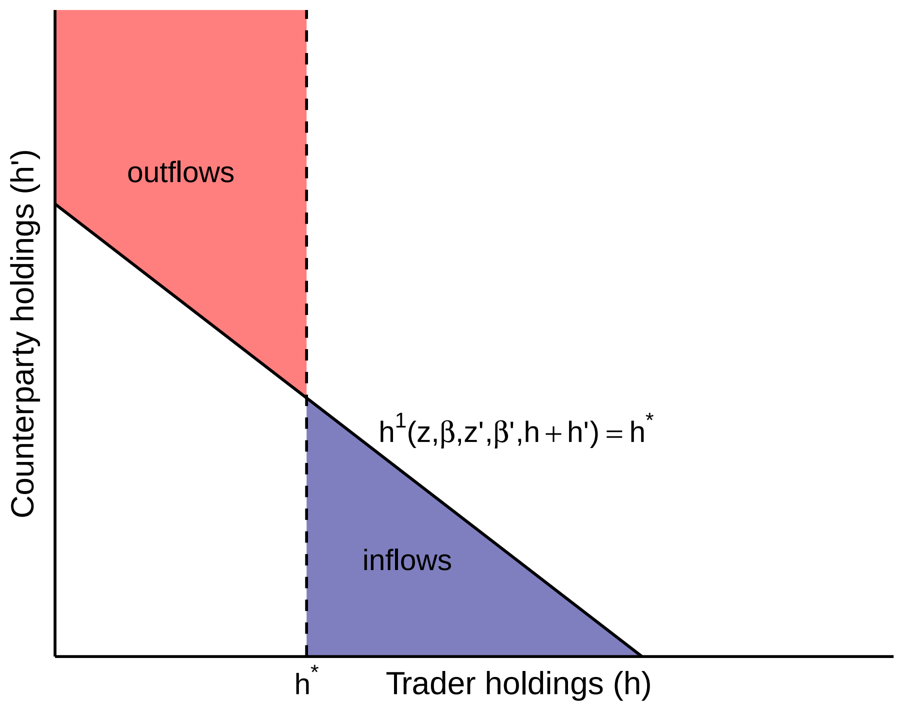


Figure A2: Trading inflows to and outflows from $h \leq h^*$

Note: This figure shows the inflows to and outflows from holdings $h < h^*$. For given (z, β, z', β'), if a trader and counterparty meet with pre-trade holdings in the blue shaded area, the trader's post-trade holdings will be beneath h^*, representing an inflow. If their holdings are in the red shaded area, the trader's post-trade holdings will be above h^*, representing an outflow. The solid line, with gradient -1, denotes meetings that will result in the trader having post-trade holdings of exactly h^*.

$$corr^W(h, q) = \int \frac{cov(h, q|z)}{\sqrt{\mathbb{V}(h|z)\mathbb{V}(q|z)}} f(z)dz,$$

where:

$$cov(h, q|z) = \iint \frac{\gamma(z, \beta, h)\phi(z, \beta, h)}{\Gamma(z)f(z)} \times \int \big(q(z, \beta, h, \Delta') - \mathbb{E}(q|z)\big)\big(h - a\big)\frac{\gamma(\Delta')\phi(\Delta')}{\Gamma} d\Delta' d\beta dh,$$

$$\mathbb{V}_{TW}(h|z) = \iint \frac{\gamma(z, \beta, h)\phi(z, \beta, h)}{\Gamma(z)f(z)} \int \big(h - a\big)^2 \frac{\gamma(\Delta')\phi(\Delta')}{\Gamma} d\Delta' d\beta dh,$$

$$\mathbb{V}(q|z) = \iint \frac{\gamma(z,\beta,h)\phi(z,\beta,h)}{\Gamma(z)f(z)} \int \big(q(z,\beta,h,\Delta') - \mathbb{E}(q|z)\big)^2 \frac{\gamma(\Delta')\phi(\Delta')}{\Gamma} d\Delta' d\beta dh.$$

- Correlation between absolute inventory $inv \equiv |h-a|$ and trading frequency, within traders:

$$corr^W(inv,s) = \int \frac{cov(inv,n|z)}{\sqrt{\mathbb{V}(inv|z)\mathbb{V}(n|z)}} f(z)dz,$$

where:

$$\mathbb{V}(inv|z) = \iint inv^2 \frac{\phi(z,\beta,h)}{f(z)} d\beta dh,$$

$$cov(inv,n|z) = \iint \big(2\gamma(z,\beta,h)\pi(z,\beta,h) - n(z)\big) inv \frac{\phi(z,\beta,h)}{f(z)} d\beta dh,$$

$$\mathbb{V}(n|z) = \iint \big(2\gamma(z,\beta,h)\pi(z,\beta,h) - n(z)\big)^2 \frac{\phi(z,\beta,h)}{f(z)} d\beta dh.$$

Chapter 3

A Structural Model of Interbank Network Formation and Contagion.

with Patrick Coen (Toulouse School of Economics).[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 expressed 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 Áureo de Paula, Péter Kondor, Caterina Lepore, Jack McKeown, Martin Oehmke, Amar Radia, Wolfgang Ridinger, Claudia Robles-Garcia, Caspar Siegert, Rhiannon Sowerbutts, John Sutton, Matthew Willison, Mungo Wilson, Kathy Yuan, Shengxing Zhang and Jean-Pierre Zigrand, as well as seminar participants at the Bank of England, the Federal Reserve Bank of New York, Imperial, the London School of Economics, Princeton, Stanford GSB, the Toulouse School of Economics, Universidad Pompeu Fabra, Queen Mary University of London, and conference participants at the RES Junior Symposium, EARIE and the Econometric Society Winter Meeting. The authors acknowledge the financial support of the Economic and Social Research Council, the Paul Woolley Centre at LSE, the French National Research Agency and the TSE-Partnership foundation.

3.1 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 capital 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 homogeneous capital requirement.

We discuss related literature below. In Section 3.2, we introduce the institutional setting and describe our data. In Section 3.3, we set out our model. In Section 3.4, we describe our approach to estimation. In Section 3.5, we set out our identification strategy. In Section 3.6, we set out our results. In Section 3.7, we undertake counterfactual analyses. In Section 3.8, we conclude.

3.1.1 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.

3.2 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.

3.2.1 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.

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).

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.

3.2.2 Data

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.

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.

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.

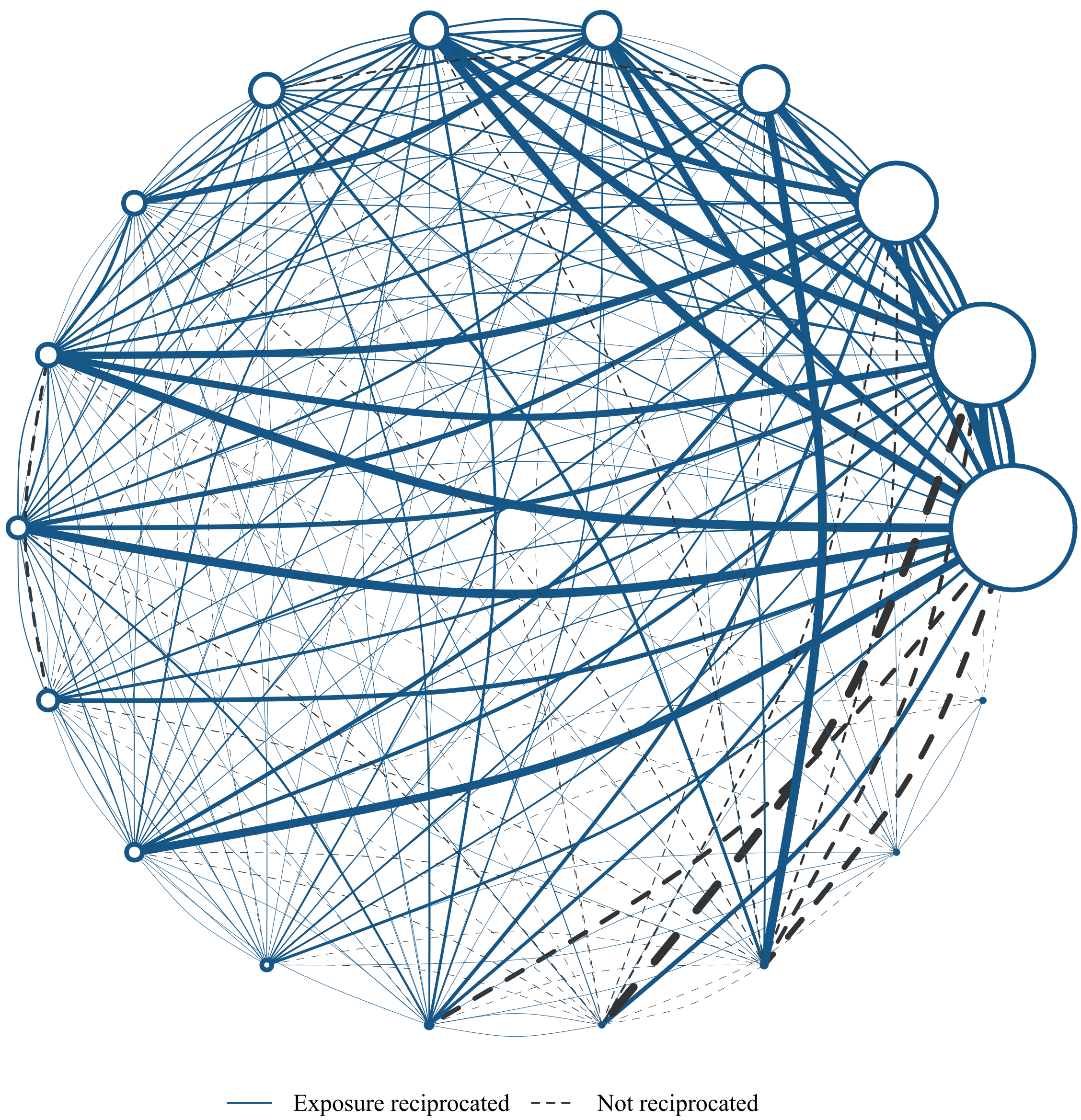
3.2.3 Summary statistics

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 GBP $7,682$m, the largest total exposures to other banks in a given period is GBP $26,367$m.

Figure 3.1: 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.

The mean exposure is GBP 285m and the mean total exposure to other banks in a given period is GBP $4,851$m.

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 3.1 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 3.1. We further demonstrate this in Table 3.1, 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 3.1, 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 product supplied by each bank varies according to currency, maturity and type. For example,

Table 3.1: 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.

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

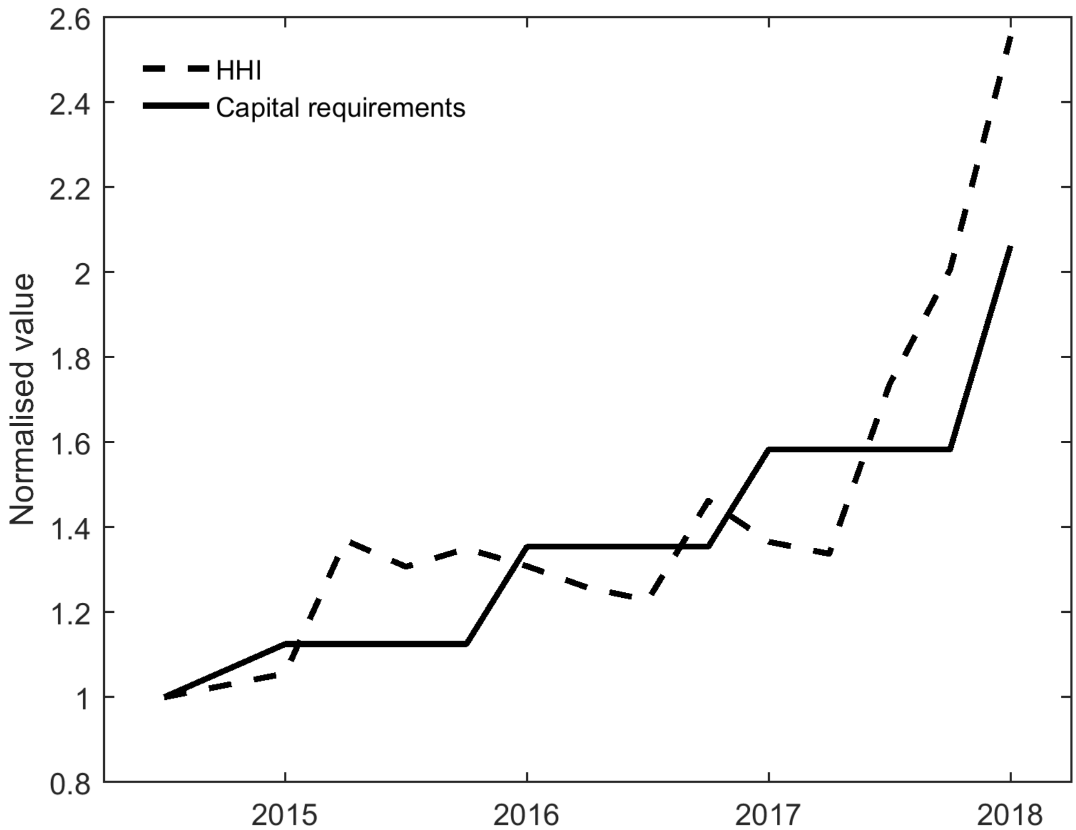
Even though the network is persistent, there is still inter-temporal variation. In particular, concentration in the interbank network has increased over time, in that the Herfindahl-Hirshmann index[10] over exposure supply has increased, as set out in Figure 3.2. In Figure 3.2, we show that the HHI index 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 3.3. Importantly, though, there is cross-sectional variation across banks, and inter-temporal variation in that cross-sectional variation. We show this in Figure 3.3, 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.

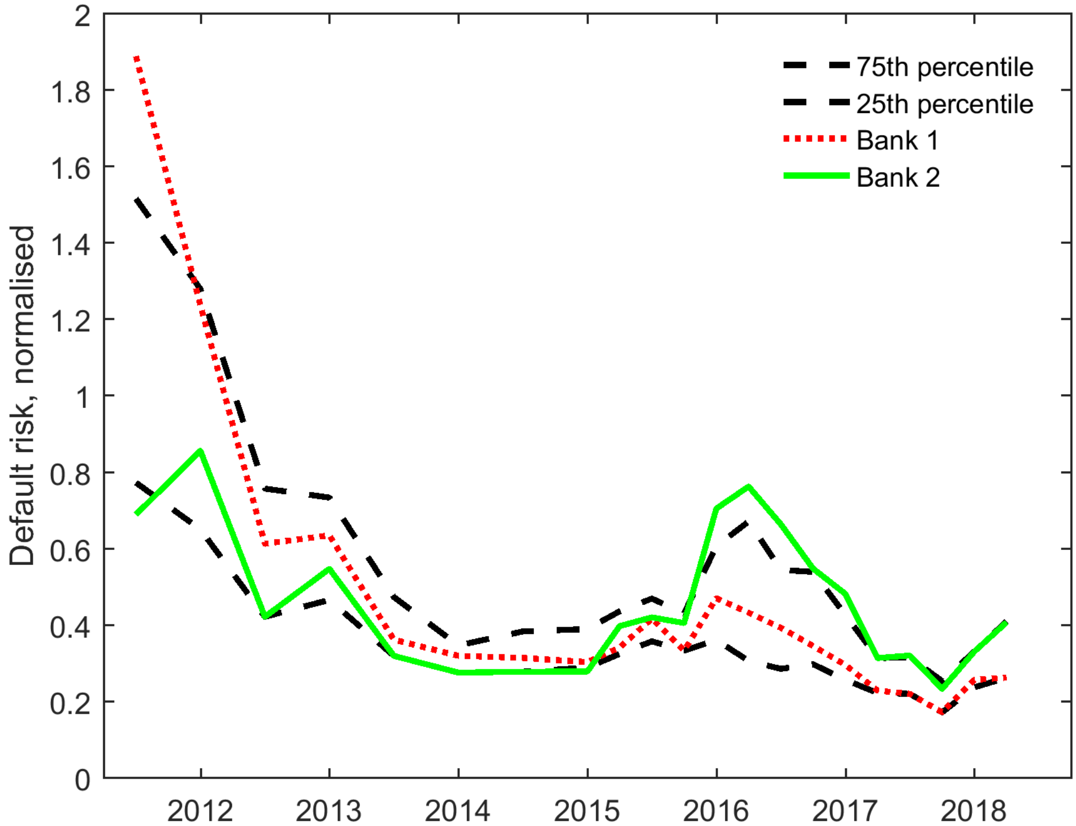
[10] $HHI_t = \frac{1}{N}\sum_j \sum_i s_{ij}^2$, where s_{ij} is the share of bank i in the total supply to bank j: $s_{ij} = \frac{C_{ij}}{\sum_i C_{ij}}$. Larger HHI indicates greater concentration. Because of the group-to-unit measurement issue we describe above, we weight exposures in our calculation 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 3.2: 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 3.3: 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.

3.2.4 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).

3.3 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.

3.3.1 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]

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

[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.

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:

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

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 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.

3.3.3 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:

$$\begin{aligned} \Pi_{jt}^{D} = \sum_{i=0}^{N} (\zeta_{ij} + \delta_{jt} + e_{ijt}^{D}) C_{ijt} \\ -\frac{1}{2}\left(B \sum_{i=0}^{N} C_{ijt}^{2} + 2 \sum_{i=0}^{N} \sum_{k \neq i}^{N} \theta_{ik} C_{ijt} C_{kjt}\right) \\ - \sum_{i=0}^{N} r_{ijt} C_{ijt} \end{aligned}$$

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}.

3.3.4 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

[12]For ease of exposition we have collapsed the risk-weighting (ρ, using the notation from Section 3.2) and the capital required per risk-weighted assets (λ) into a single parameter, λ.

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:

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

such that $C_{ijt} \geq 0$, $E_{it} - \sum_j C_{ijt} \geq 0$ and $puc_{ijt} = \lambda_{ijt}\,(\phi p_{it} + e_{it}^c)$.

For interior solutions the first order condition is as follows:

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

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 '∆ 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 '∆ D-side cost').

3.3.5 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 3.3.1 *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 B2. 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.

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.

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.

3.3.6 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 3.3.2 : 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 3.4.

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 3.3.3 : 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 3.4 shows the frontier and illustrates what conclusions we can draw using this model about different outcomes.

Definition 3.3.4 : 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.*

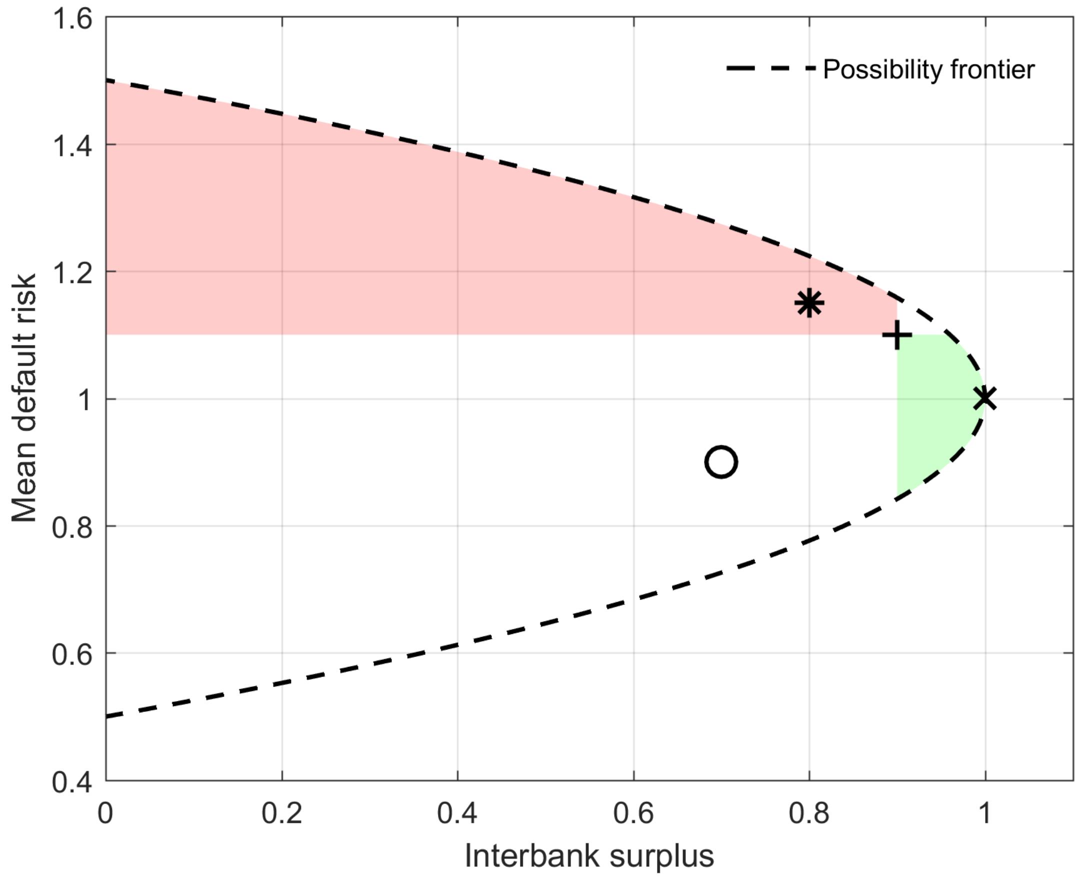
Definition 3.3.5 : TS inefficiency : *The total surplus inefficiency of some allocation $(TS_I, \bar{p})$ is the percentage increase in TS_I that could be obtained without increasing $\bar{p}$. In other words, it is the horizontal distance in percentage terms from the frontier.*

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.

3.4 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.

Figure 3.4: 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.

3.4.1 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 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.

3.4.2 Estimation structure

The parameters we seek to estimate are $\mathbf{\Theta} = (\tilde{\mathbf{\Gamma}}, \tau, \omega\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\zeta}, \tilde{\boldsymbol{\theta}}, \phi)$; 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: $\mathbb{E}[\mathbf{Z}'(\hat{\mathbf{C}}(\mathbf{\Theta}) - \mathbf{C})] = 0$ and (2) contagion: $\mathbb{E}[\mathbf{Z}'(\hat{\mathbf{p}}(\mathbf{\Theta}) - \mathbf{p})] = 0$. We express $\mathbf{p}$ in logs.

3.4.3 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.

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

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 particular the fact that, as described in Section 3.2, 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.

3.5 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.

3.5.1 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 3.2, 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.

3.5.2 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.

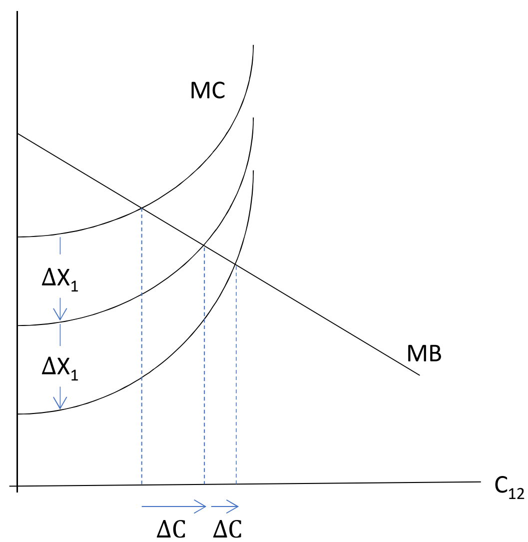
Second, we show that this gives us the pair-specific variation that we need. Third, we set out exactly which variables we use as instruments.

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 3.5. Suppose the fundamentals of banks 1, 2 and 3 improve, worsen and remain unchanged, respectively. In these circumstances, we show in Figure 3.5 that C_{12} changes non-linearly relative to the size of these. In Appendix B2 we show, for a simplified version of our model for which an analytical solution exists, that equilibrium C are a non-linear function of $\mathbf{X}$.

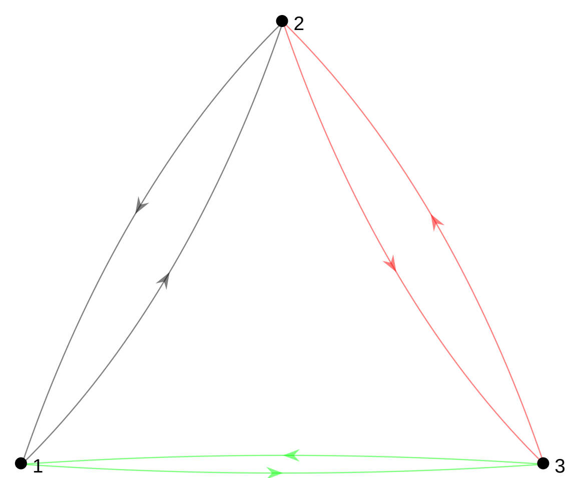
Having shown that exposures are non-linear in fundamentals, it is straightforward, using the same simple example, to show that changes in fundamentals then give us the pair-specific variation that we need for them to be instruments for C_{ijt}. Assume again that the fundamentals of banks 1, 2 and 3 improve, 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 3.3), 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 $[X^2_{it},\ X^2_{jt},\ \tilde{X}^2_{ijt},\ X_{it}/X_{jt},\ X_{it}/\tilde{X}_{ijt}...]$, 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

Figure 3.5: Non-linear bank fundamentals as instruments for C



(a) Non-linear effect of X on C



(b) Pairwise variation

Note to Figure 3.5: 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.

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).

3.5.3 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 3.2. 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 3.2: Key variation

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

3.6 Results

We set out our results in Table 3.3. 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

[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 3.3: 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.

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.

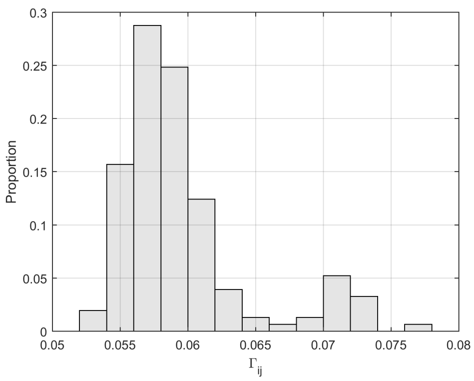
- **Contagion is heterogeneous**: there is substantial pairwise variation in contagion intensity Γ_{ij}: some links are nearly twice as costly as others, in terms of their effect on default risk. We plot the estimated distribution of Γ_{ij} in Figure 3.6.
- **Contagion is time-varying**: there is evidence 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 3.6. 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 3.7. 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 3.6, 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.

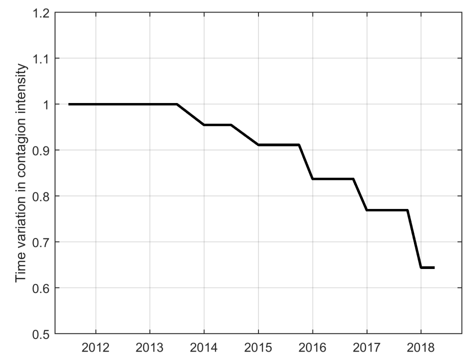
3.6.1 Robustness

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 B3. In both cases, the results are quantitatively and qualitatively similar to our baseline results.

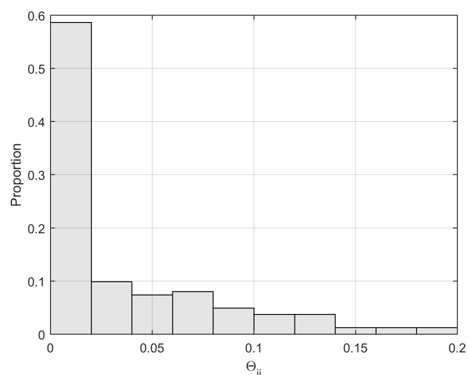
Figure 3.6: Distributions of parameter estimates



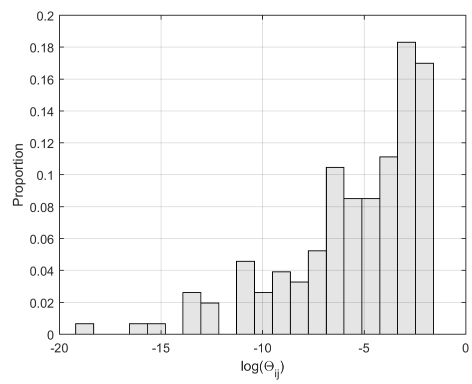
(a) Heterogeneous contagion intensity



(b) Time variation in contagion intensity



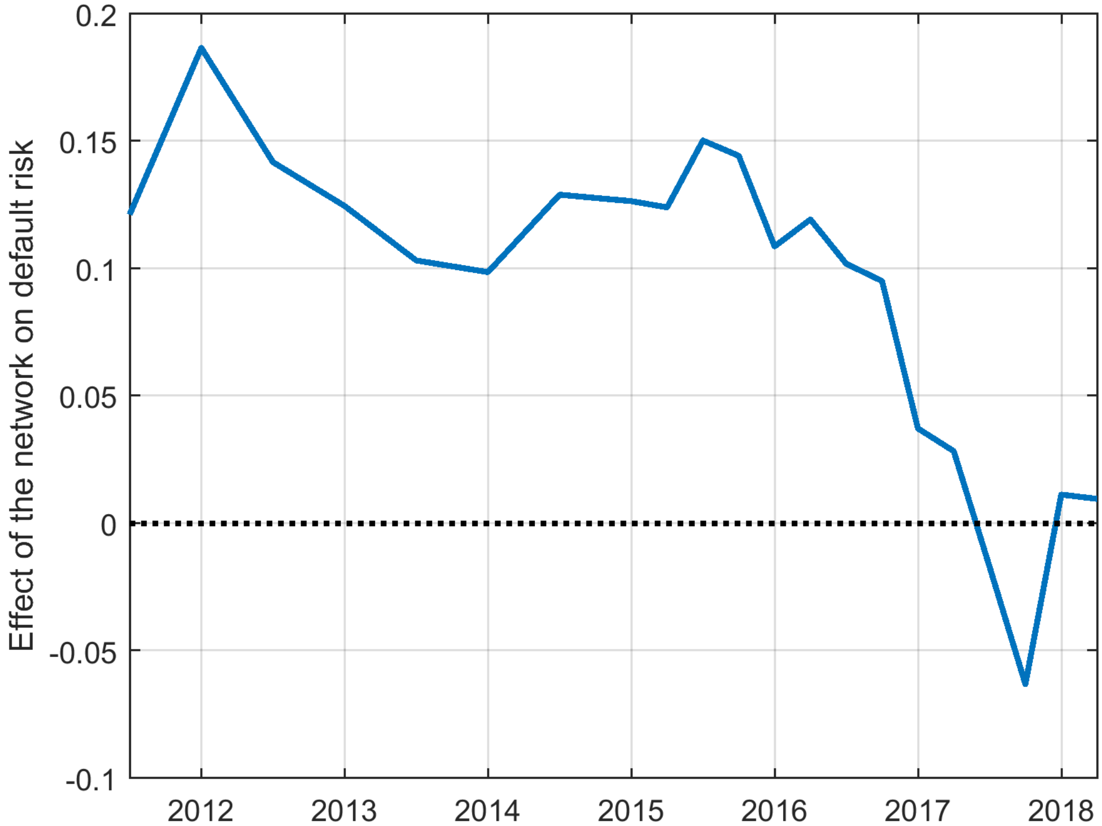
(c) Product differentiation



(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 3.7: 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.

3.6.2 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.

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 B4, 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.

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 3.8, 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 3.8).

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 3.2 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 3.9 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.

3.6.3 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.

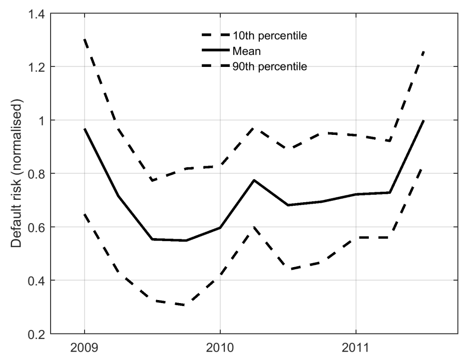
Forward simulation

In Figure 3.9 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.

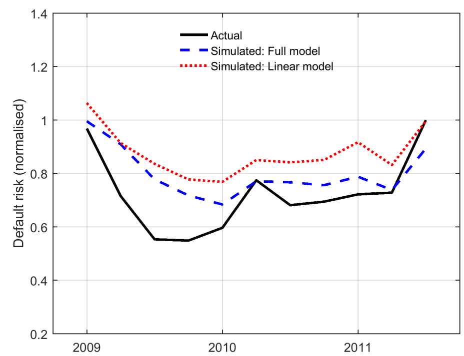
This fact also has implications for predicting the impact of recessions. Suppose, for example, that when modelling the response of default risk p to fundamentals X the endogenous network was ignored, and instead p was simply regressed on X. Because severe recessions are very infrequently observed, a 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 3.9, and show that this bias can be material.

In Figure 3.8 above we show out of sample fit, and show that during periods in which

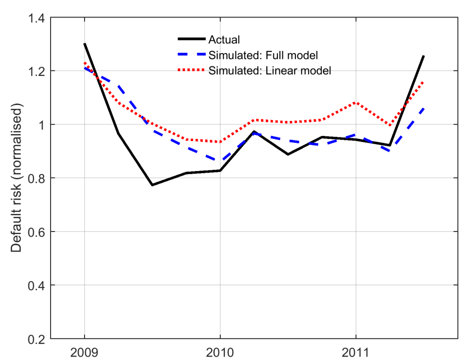
Figure 3.8: Out of sample fit: Bank default risk



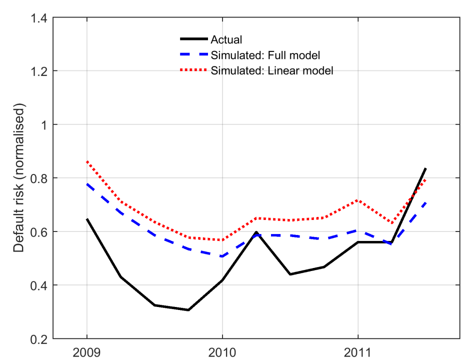
(a) Actual variation in default risk



(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 3.9.

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 3.9, 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.

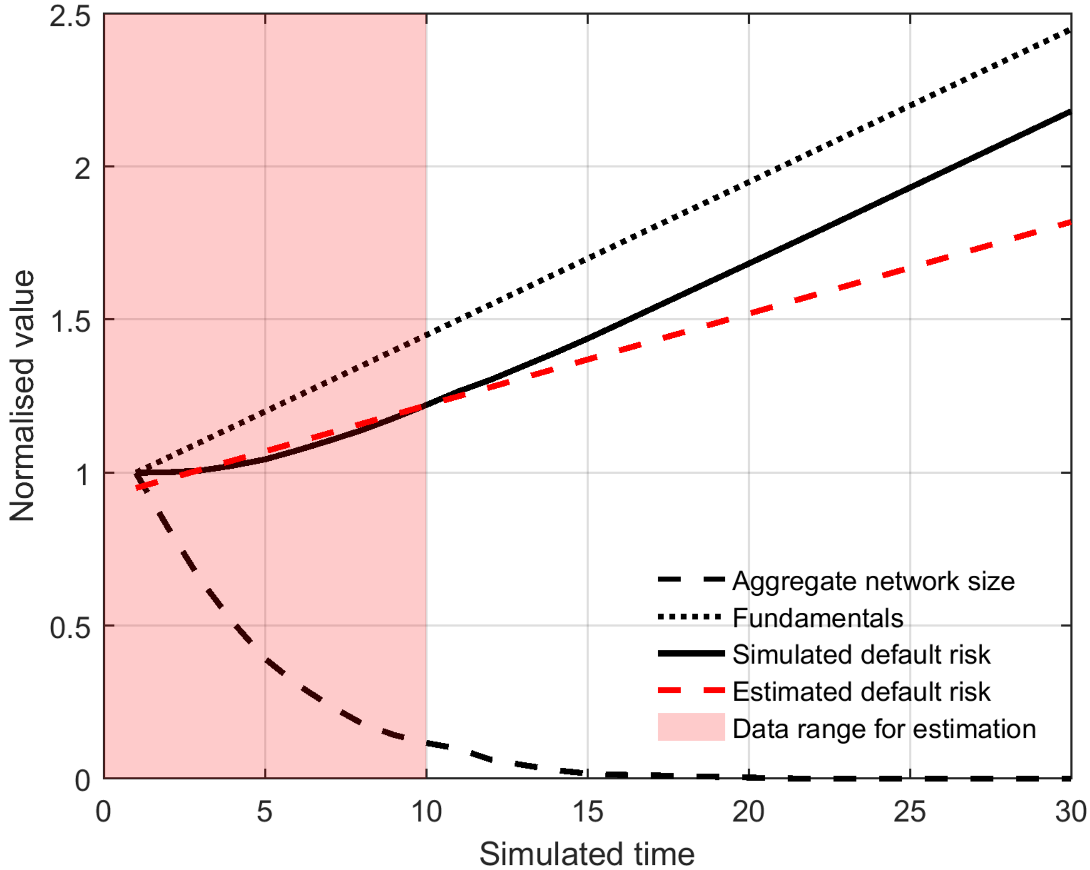
Systemic importance

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 homogeneous, $\Gamma_{ij} = \Gamma$, 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 3.9: 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 3.8.

weighted adjacency matrix $\boldsymbol{\Gamma} \circ \boldsymbol{C}_t$. 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 3.10, 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).

3.7 Counterfactual Analysis

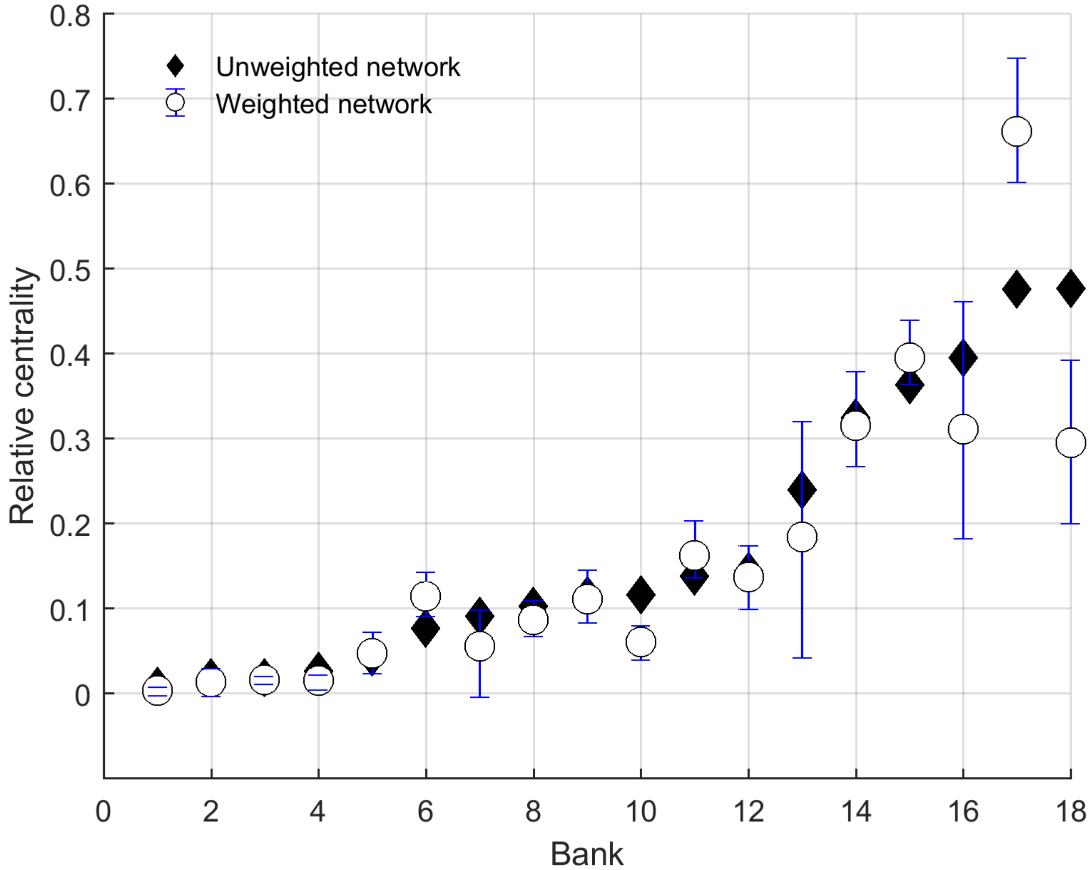
In our counterfactual analyses, we first consider the social planner's solution, and show what that implies for efficiency. We then consider two broad forms of regulation: caps on exposures and capital ratios.

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.

3.7.1 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 3.10: 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 3.11. 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.

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 3.4.

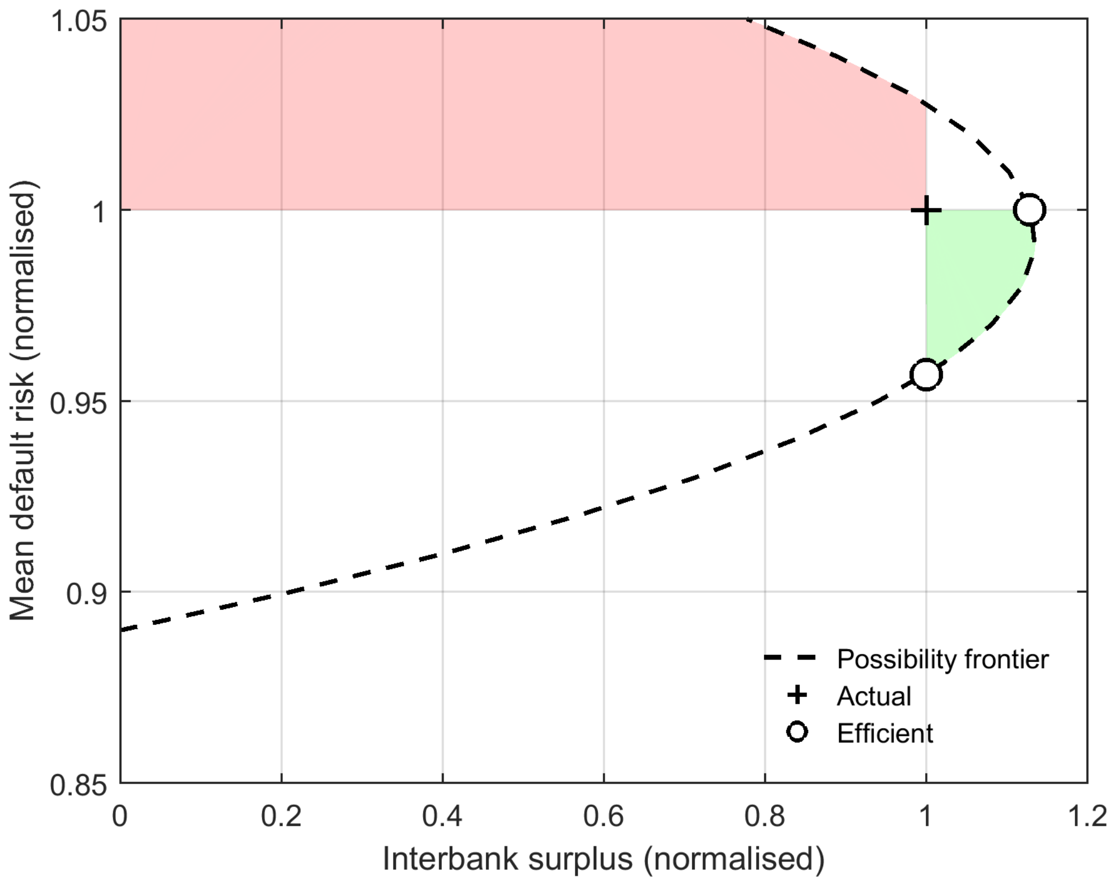
Table 3.4: Comparative statics

	[A] Baseline	[B] $\downarrow mean(\theta_{ij})$	[C] $\uparrow var(\Gamma_{ij})$	[D] $\uparrow mean(\Gamma_{ij})$
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 3.4). As set out in Table 3.4, this increases the distance

Figure 3.11: 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 3.4). 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.

3.7.2 Caps on exposures

As discussed in Section 3.2, 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 3.12 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 3.12, 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.

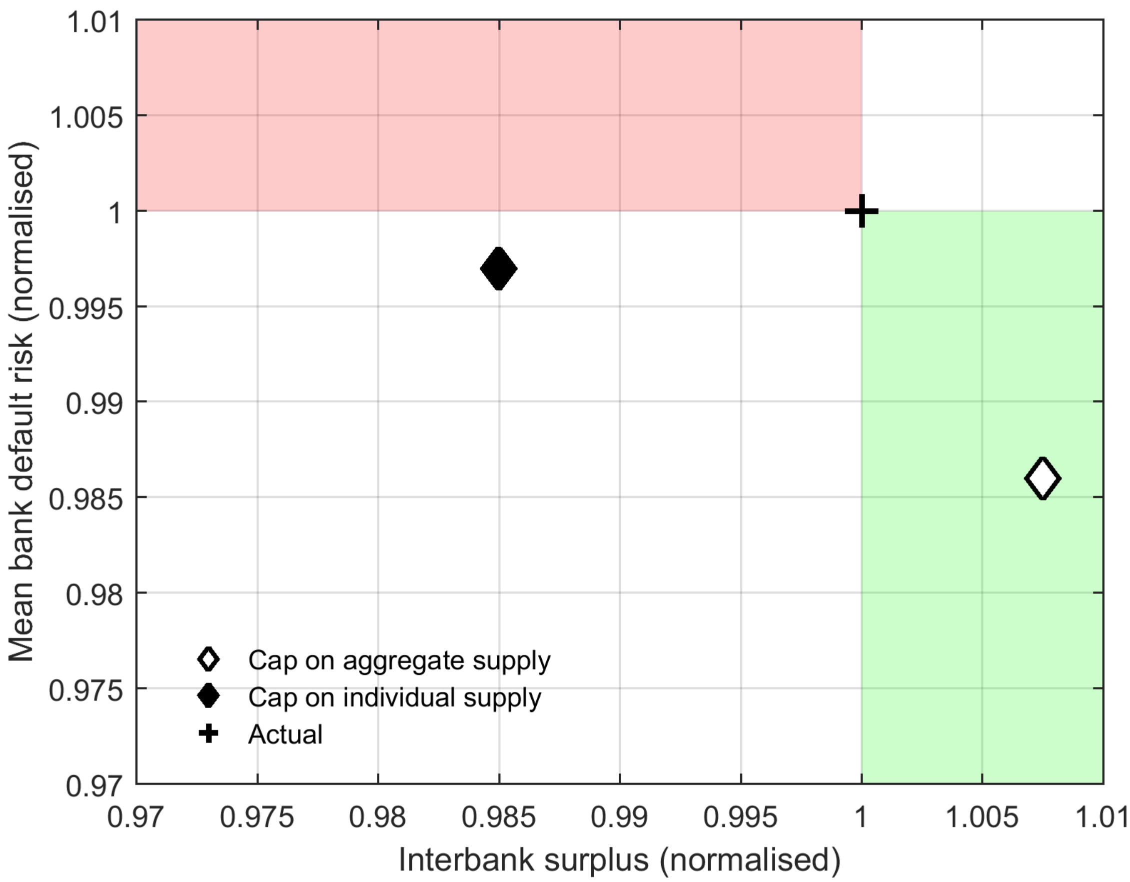
3.7.3 Capital ratios

The second form of regulation we consider is a minimum capital requirement, as applied by regulators since the crisis. As described in Section 3.2, 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 3.13.

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 3.13). 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.

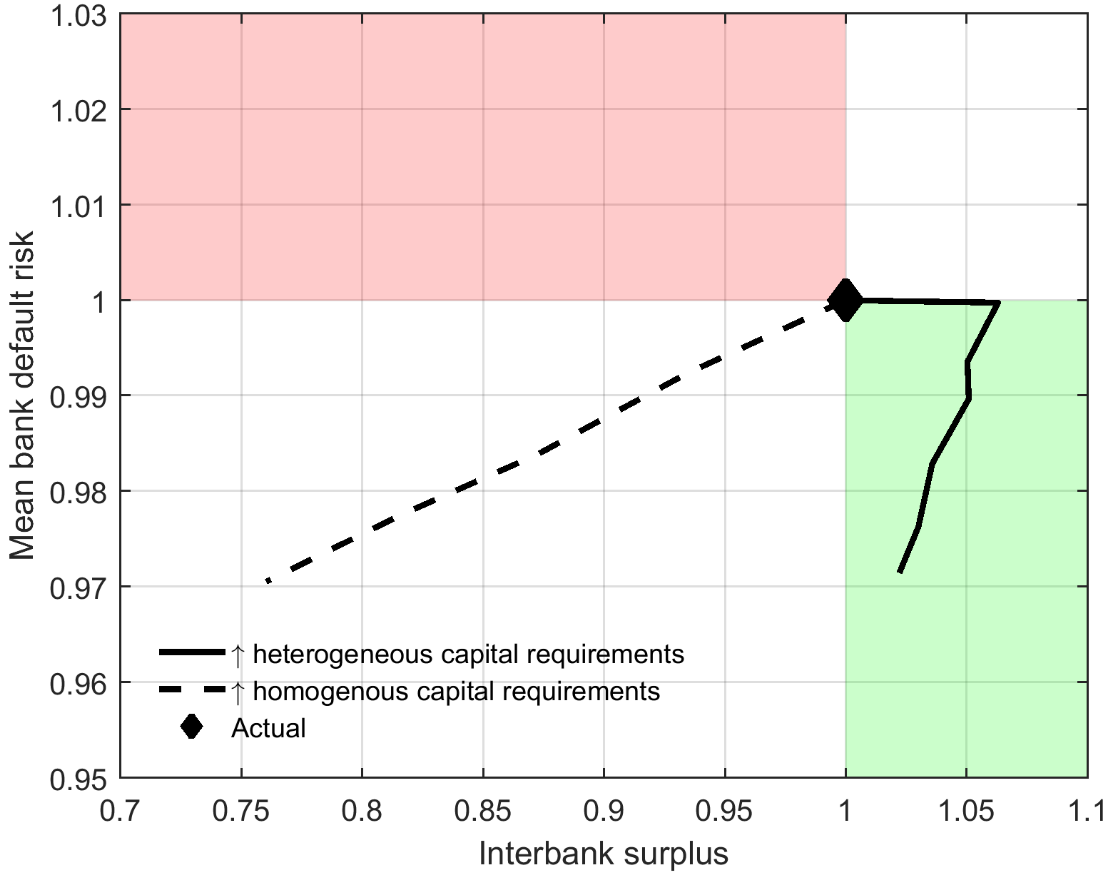
[16] Any spread like this is an improvement over homogeneous capital requirements, this particular spread is one we have chosen arbitrarily as one that produces good results.

Figure 3.12: 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 3.13: 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) homogeneous 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.

3.8 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.

Bibliography

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.

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.*

B Appendices

B1 First stage regression results

Table B1: 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 B2: 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	

B2 Mathematical appendix

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:

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

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.

B3 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.

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 B1. Our estimation procedure is otherwise the same as our baseline specification. In Table B3 we set out our results, which are quantitatively and qualitatively similar to our baseline results.

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.

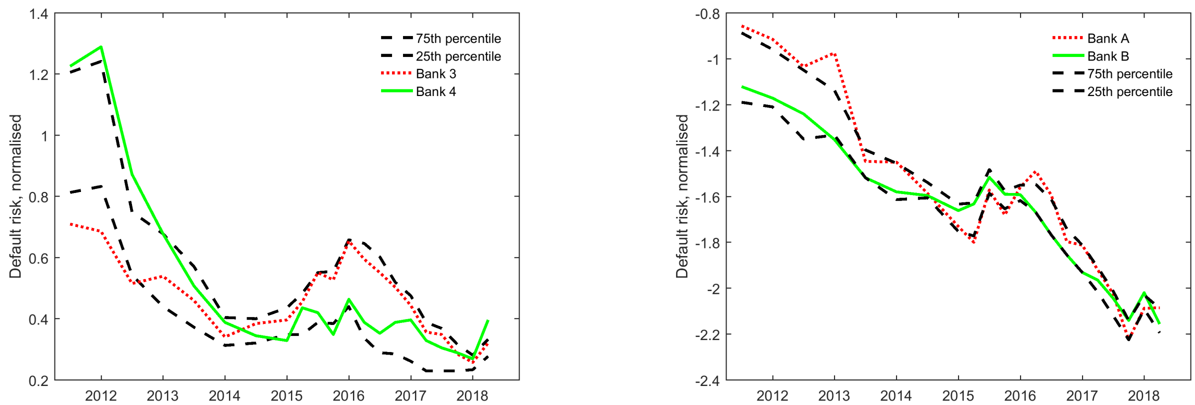
$$\underbrace{\mathbf{p_t}}_{\substack{\text{Default}\\ \text{risk}}} = \underbrace{\boldsymbol{X_t\beta}}_{\substack{\text{Funda-}\\ \text{mentals}}} + \underbrace{(\mathbf{I} - \tau_t \boldsymbol{\Gamma} \circ \mathbf{C_t}) X_{2,t}\beta_2}_{\substack{\text{Risk}\\ \text{premium}}} - \underbrace{\omega\,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

Table B3: 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 B1: 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)).

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 B3, and find that our results are quantitatively and qualitatively similar to our baseline results.

Table B4: 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.

B4 Additional post-estimation tests

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.

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

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 B5: Cost of equity and default risk

	c_{it}^{e}
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.

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 B6: 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.