The Industrial Organization of Financial Markets

Robert Clark[1], **Jean-François Houde**[2], **and Jakub Kastl**[3]

[1]**Queen's University**
[2]**University of Wisconsin-Madison and NBER**
[3]**Princeton University, CEPR, and NBER**

ABSTRACT

This chapter discusses recent developments in the literature involving applications of industrial organization methods to finance. We structure our discussion around a simple model of a financial intermediary that concentrates its attention either on (i) the retail market and hence engages in a traditional maturity transformation business by accepting funds that can be used to invest in risky projects (loans), or (ii) the investment business, financing its operations on the "wholesale" market and making markets or investing in higher return riskier projects. Our discussion is centered around the analysis of market structure and competition in each of these markets, focusing in turn on (i) primary and secondary markets for government and corporate debt, (ii) interbank loans, (iii) markets for retail funding, and (iv) credit markets, including mortgages.

Keywords: Financial Markets, Imperfect Competition, Information Frictions, Market Power, Wholesale Markets, Retail Funding, Retail Credit, Regulation

1 INTRODUCTION

Financial markets are at the core of every market economy. Their purpose is to facilitate the transfer of funds from savers (agents with excess funds) to borrowers (agents in need of funds). Efficient financial markets help to ensure the transfer of savings to the highest return investments, increasing productivity and growth. However, it has long been recognized that, like in any other market, important frictions exist that might hinder the efficient functioning of financial markets. Asymmetric information, externalities, systemic risk, and market power are just some of the key frictions that can lead to market failure. Many of these are at the center of current policy debate regarding regulation or other government intervention.

The transfer of funds from savers to borrowers could of course be achieved by having savers buy securities directly from firms. In practice, firms in the financial industry supply the service of transferring funds from savers to borrowers. This process of *financial intermediation* is performed by a variety of players operating at different levels of the supply chain of funds.

Figure 1 illustrates the supply chain of financial markets. In the *upstream* retail funding market, savers select a fund manager (depository institution, mutual fund, pension plan, etc.) with whom to place their savings. These savings represent the *inputs* used in the production of financial assets downstream: loan origination in *retail credit markets*, and the issuance of bonds and stocks in primary *capital markets*. Vertically integrated financial institutions, such as retail and investment banks, operate at both ends of the supply chain. Other financial intermediaries specialize in funding or asset creation. For instance, fund managers such as mutual fund companies use funds to purchase securities created by loan *originators* operating in the downstream markets. Similarly, loan origination companies operate upstream and specialize in the origination of loans to consumers, while dealers in primary markets specialize in acquiring

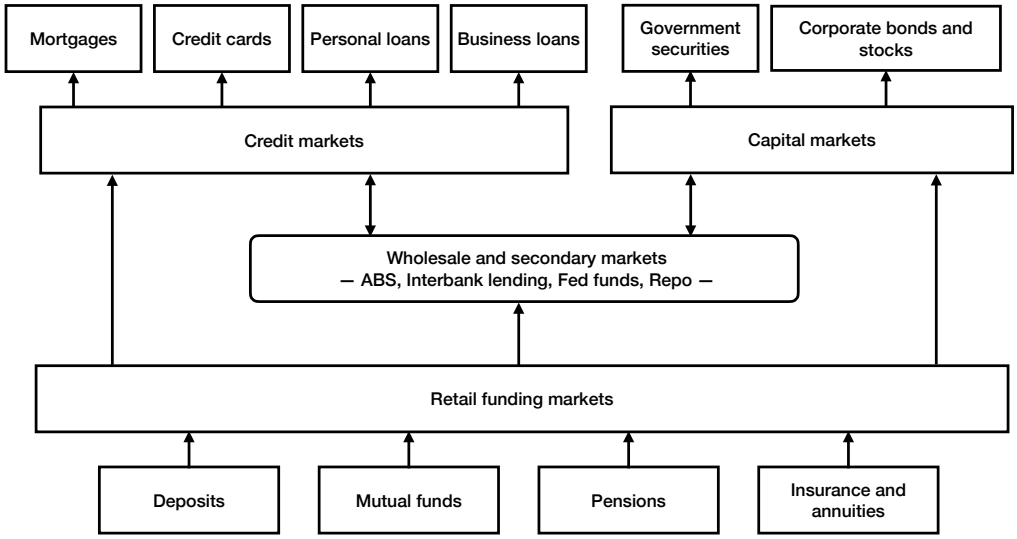


Figure 1. Supply Chain in Financial Markets

and managing inventories of corporate and governmental bonds. These firms finance their operations by buying and selling securities in the secondary market. We refer to the exchange of funds between financial intermediaries as *wholesale* transactions. Examples include the money (or interbank lending) market in which firms buy and sell short-term governmental securities to manage their reserves of liquid assets, as well as markets for long-term asset-backed securities (ABS) such as those for mortgages, car loans, and credit-card debt. Figure 2 provides a characterization of different types of players organized by their position in the supply chain of funds. Pushed by technology innovation, regulation changes, and macroeconomic conditions, the structure of financial markets has experienced important changes over the last 40 years.

First, the extent of vertical intergration in the market has been impacted by technological and regulatory changes. The creation and growth of markets for ABS (including mortgages) since the mid-1980s has increased interactions between funding companies like mutual funds and insurance companies and traditional lenders. Since the 2008 financial crisis, tighter regulation on banks' capital structure have further changed the flow of funds in the system. In response to larger capital requirements, banks have decreased their reliance on interbank lending, and are now relying more on retail deposits to fund asset investments, leading to an increase in the degree of vertical integration among traditional banks. During the same period, the share of loans originated by non-depository institutions (or *shadow-banks*) has increased substantially, contributing to an increased separation between retail funding and lending markets. In 2020, 70% of conforming mortgages sold on the Mortgage-Backed Securities (MBS) market were serviced by shadow-banks, and Quicken Loans has become the largest mortgage lender in the US. Similar trends are observed in other lending markets with access to securitization. In addition to securitization, improved risk screening and more automated loan approval have each contributed to the increased importance of shadow banks. Their rise following the 2008 financial crisis changed how personal loans are funded; since non-depository institutions rely exclusively on wholesale funding and equity to fund their activities.[1]

Second, the financial industry experienced a period of important deregulation in the 1990s. Together with the technological changes mentioned in the previous paragraph and the financial crisis, this caused a significant increase in concentration triggered by a series of mergers, and the

[1] See Jiang et al. (2020) for analysis of the balance sheet of shadow-banks post-crisis.

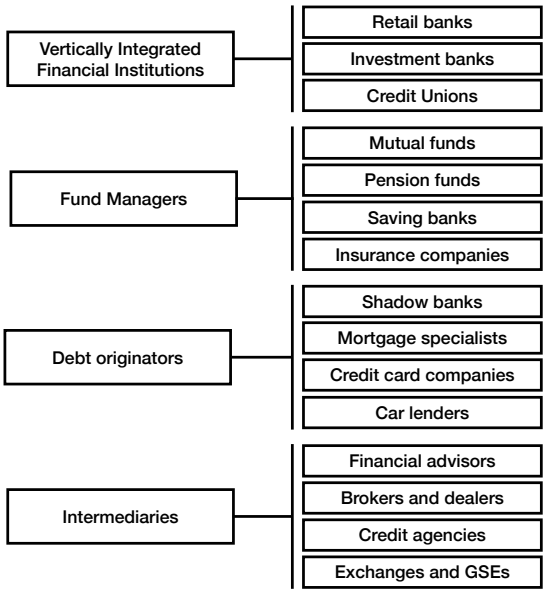


Figure 2. Vertical Integration in Financial Markets

integration of commercial and investment banks. Figure 3 illustrates this point. Between 1991 and 2008, the asset market share of the top 10 banks has doubled to around 60%, and the number of banks went from around 11,000 prior to Riegle-Neal to less than 5,000 banks today. Rising concentration poses questions about efficiency and stability in the financial services industry for policymakers.[2] Moreover, this rise in concentration is associated with a large increase in market power across all sectors of the financial industry. Figure 3b illustrates the rise in markups since the early 1990s documented by De Loecker et al. (2020) (Online Appendix 12). Importantly, the authors find that markups in the financial industry experienced the largest increase across all 2-digit SIC sectors.

Finally, with interest rates close to zero, central banks have had to conduct unconventional monetary policy including increased intervention in primary and secondary markets for government debt. Since these interventions are intermediated by a small number of players (dealers), concerns that market power may limit the flow of money to consumers and firms have increased over time.

In this chapter we address the impact of some of these changes, focusing in particular on the role of imperfect competition and information frictions in determining the flow of funds from savers to borrowers. We are especially interested in the interaction between firms with market power, operating in retail and wholesale markets. As such we concentrate our attention on sub-markets in which concentration is of first-order importance. This includes retail lending and funding markets, as well as in primary and secondary markets for governmental bonds and interbanks lending. Our analysis of wholesale markets largely takes as given the supply of funds in primary markets for corporate bonds and stocks, and abstracts from any competition between firms facilitating the creation of these securities. We mostly leave the discussion of insurance for the chapter in the present Handbook on selection markets (Einav et al. (2021)). See also Koijen and Yogo (2015) and Koijen and Yogo (2016) for recent contributions on imperfect competition

[2]For instance, the Federal Reserve's 2018 Jackson Hole Symposium was entitled "Changing Market Structures and Implications for Monetary Policy".

in life insurance markets.

Over the past two decades, as financial-market data have become more widely available, industrial organization economists have begun to apply the tools and methods used to analyse countless other markets to study the functioning of financial markets. Our objective in this chapter is to review these contributions, focusing especially on the methodologies employed.[3] We also discuss how key institutional features of financial markets, such as regulation and technology innovations, impacted market outcomes.

The rest of the chapter is organized as follows. In Section 2 we present a theoretical model of financial intermediation that introduces the main themes that we cover in the rest of the chapter. Section 3 examines market structure and the measurement of market power in primary and secondary markets for government bonds, which we refer to as wholesale funds. Section 4 studies markets for retail funding (deposits, mutual funds, pension plans, etc.), focusing on the sources of market power. In Section 5 we highlight recent contributions to measuring information frictions and market power in retail credit markets. Section 6 discusses the regulation of financial markets, financial stability and the role of intermediaries. Finally, Section 7 concludes.

2 A MODEL OF FINANCIAL INTERMEDIATION

To illustrate how market structure affects the flow of funds in the industry, we start with the Monti-Klein model of financial intermediation (Klein (1971), Monti (1972)).[4] According to this model, financial intermediaries compete by choosing their optimal portfolio of assets and liabilities, subject to liquidity constraints imposed (in part) by monetary authorities. We add to this classic framework key elements of the industrial organization of the industry: asymmetric information, imperfect competition, and double marginalization.

In a first step we use the model to describe the decisions of a financial intermediary that finances its operations by establishing short positions in equity, retail deposits, and debt, and possibly long positions in regulated loans (mortgages), unregulated loans, safe investments and risky investments. This vertically integrated retail bank engages in maturity transformation and, for regulatory reasons, finances its operation mainly through deposits, and issues mortgages, with surplus funds being supplied back to the wholesale funding market (as described in Gertler et al. (2016)). The model captures the fact that banks have market power at both ends of the supply chain. Banks compete by playing a Cournot game, and interact in three markets: (i) the retail deposit market, (ii) the retail credit market, and (iii) the money market. The balance sheet of bank j includes four elements: equity (E_j), deposits (D_j), money-market position (Q_j), and loans (L_j).

Equity is used by banks to finance asset creation, and it evolves dynamically over time as a function of past profit realization. As mentioned in the introduction to this chapter, we mostly abstract from the determinants of equity value.[5] In this section, since we are focusing on static decisions, we take E_j as a pre-determined state variable, and analyze the portfolio choice between loans and deposits.[6]

On the liability side of the balance sheet, banks bundle deposits with differentiated financial services of quality levels δ_j for bank $j = 1,\ldots,n$. We use a Logit-demand system to model

[3] For other reviews that cover some of these topics see Vives (2016), Degryse et al. (2009) and the Handbook of Financial Intermediation and Banking (Thakor and Boot (2008).)

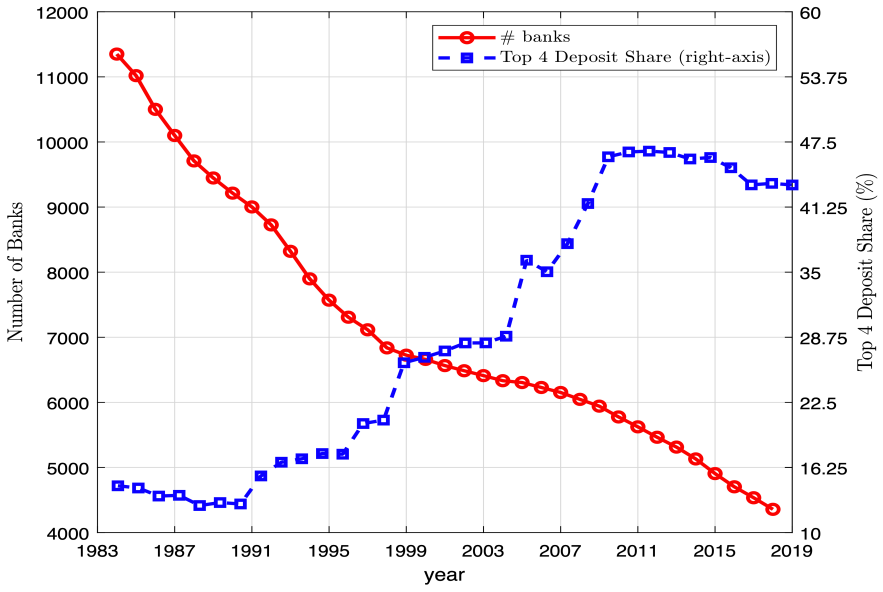
[4] Our description of the Monti-Klein model follows the exposition in Freixas and Rochet (2008).

[5] See Koijen and Yogo (2019) for a model of asset-pricing that borrows ideas from the discrete-choice literature.

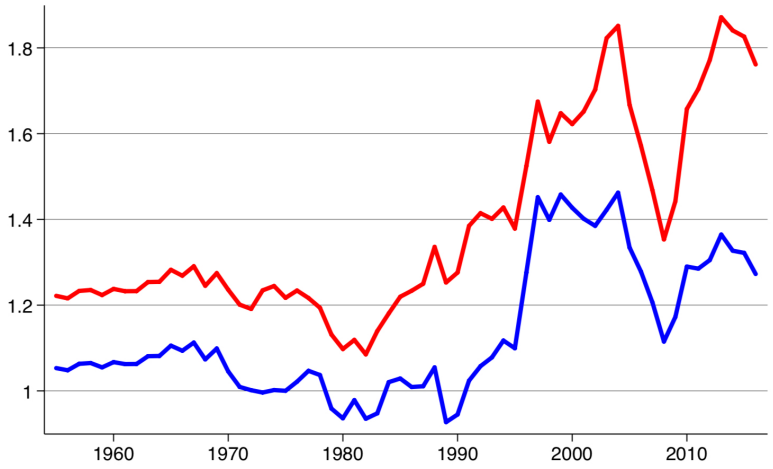
[6] See Corbae and D'Erasmo (2021) for a dynamic model of portfolio choice with imperfect competition and equity formation.

Figure 3. Evolution of concentration and markups in the banking and finance

(a) U.S. Bank Numbers and Concentration



(b) Markups in the finance and insurance sectors



Notes: Figure 3a is taken from Corbae and Levine (2020). Number of Banks refers to the number of bank holding companies. Top 10 Asset share refers to the share of total assets in the hands of the top 10 banks in the asset distribution. Figure 3b is taken from De Loecker et al. (2020) (Figure 12.1 of their online appendix). The Red curve corresponds to the unweighted average markup across firms in the finance and insurance sectors, and the Blue curve the size-weighted average markup.

demand for deposits:

$$D_j(\mathbf{r}_d) = \bar{D} \cdot \frac{\exp(\delta_j + \beta_p r_{j,d})}{1 + \sum_{j'} \exp(\delta_{j'} + \beta_p r_{j',d})} \Rightarrow R_{j,d}(\mathbf{D}) = \frac{1}{\beta_p} \ln\left(\frac{D_j}{\bar{D} - D_j - D_{-j}}\right) - \delta_j/\beta_p, \quad (1)$$

where $\mathbf{r_d}$ is a vector of deposit interest rates, $D_{-j} = \sum_{j' \neq j} D_{j'}$, $\bar{D}$ is total potential deposit size, and $R_{j,d}(\mathbf{D})$ is the inverse-demand function for bank j. Although we refer to D_j as deposits, one can also interpret the model as one of demand for savings, pensions, or annuities.

On the asset side of the balance sheet, firms invest in risky loans that are illiquid (L_j), and trade short-term securities issued by the central bank (Q_j).

The lending market is described by a continuum of potential borrowers demanding loans of unit size. The willingness-to-pay (WTP) of the marginal borrower is given by $R_l(L)$, and we use r_l to denote the realized lending rate. The revenue function is also determined by a repayment probability $0 < M(L) < 1$. The repayment probability is increasing in the total supply of loans $L = \sum_{j=1}^{N} L_j$, consistent with the presence of moral hazard and/or adverse selection. For instance if the borrower's WTP is positively correlated with default probability, an increase in the number of borrowers implies that the marginal borrower is less risky than the average. This can lead to a backward-bending expected return function:

$$\tilde{R}_l'(L) = \underbrace{R_l'(L)M(L)}_{<0} + \underbrace{R_l(L)M'(L)}_{>0}.$$

Finally, banks manage their reserves of liquid assets by investing in government securities. We use the term *money market* to refer to transactions in which banks purchase securities from the government (primary market), and borrow and lend from each other on the fed fund market (secondary market). The net position on bank j in the money market is given:

$$Q_j = E_j + D_j - L_j. \quad (2)$$

Similarly we use K to denote the overall quantity of securities supplied by the central bank and other market participants, including dealers and fund managers, such that: $K = \sum_j Q_j$. Importantly, we refer to the sale/purchase of securities as *wholesale funding* transactions since they involve two financial intermediaries.

The money market rate is given by r_f, and is determined in equilibrium. We abstract from transaction costs, and so r_f denotes both the lending and borrowing money-market rate (i.e. zero spread). We assume that banks are price takers in this market. However, depending on the price setting mechanism (i.e. auctions vs over-the-counter), one or both sides of the transaction may have market power. We investigate this in Section 3 when we analyse the wholesale market for liquid funds.

One reason banks keep government securities on their balance sheet is to guard against liquidity shortages. We introduce liquidity risk in the model by incorporating unexpected deposit withdrawals (or influxes of cash) $\tilde{x}$, realized after the bank commits to originating loans and raising deposit (as in Prisman et al. (1986)). This liquidity shock affects the ability of banks to meet the reserve requirement imposed by the regulator: $Q - \tilde{x} \geq \rho$. If $\tilde{x} > Q - \rho$, banks incur a penalty $r_p > r_f$ to borrow from other financial institutions (or the central bank). For simplicity we assume that banks are price takers in this "emergency borrowing" market, and r_p is determined outside the model. Banks differ in the amount of liquidity risk they face, measured here by differences in the density of $\tilde{x}$, $\phi_j(\tilde{x})$. This is a reduced-form way of measuring factors such as geographic diversification or the risk of bank runs, as well as volatility in equity value. As we will see below, this leads to cost differences across banks.

The expected profit function of bank j is given by:

$$\begin{aligned}\Pi_j(R_j,L_j) &= \tilde{R}_l(L)L_j + r_f Q_j - R_{j,d}(D)D_j - r_p E\left[\max\{0,\tilde{x}+\rho-Q_j\}\right] - C_j(D_j,L_j)\\ &= r_f E_j + \left[\tilde{R}_l(L)-r_f\right]L_j + \left[r_f - R_{j,d}(D)\right]D_j\\ &\quad -r_p\int_{E_j+D_j-L_j-\rho}(\tilde{x}+\rho-E_j-D_j+L_j)\phi_j(\tilde{x})d\tilde{x} - C(D_j,L_j).\end{aligned}$$

The non-financial cost of raising deposits and originating loans is given by:

$$C_j(D_j,L_j) = F_j + c_{j,l}L_j + c_{j,d}D_j. \tag{3}$$

The cost function exhibits economies of scale due to the presence of a fixed cost of lending (F_j). The importance of economies of scale in loan origination and monitoring is central to the theory of banking, as discussed in Diamond (1984)'s model of financial intermediation. The fixed cost can be avoided by setting $L_j = 0$ and investing solely in risk-free securities. Depending on the realization of (c_j,F_j), for some banks, it may be optimal to forgo the investment in loans, and instead invest their deposits directly in securities. Therefore, we can view the choice as essentially between becoming a vertically integrated retail bank or a fund manager. The presence of fixed costs also brings about the possibility of bank failure, which creates a relationship between competition and bank fragility. We discuss the relationship between competition and the fragility of the finance industry later in this chapter.

Abstracting from the possibility of exit, banks solve the following constrained optimization problem:

$$\begin{aligned}\max_{D_j,L_j} \quad & \Pi_j(R_j,L_j) \\ \text{s.t.} \quad & L_j \geq 0.\end{aligned} \tag{4}$$

Conditional on the money market rate, an equilibrium of this game is described by a set of loan size and deposit amounts such that the following two first-order conditions are satisfied:

$$\begin{aligned}R_l(L^*) &= \frac{1}{M(L^*)}\left[c_{j,l} + r_f + r_p\left(1-\Phi_j(E_j+D_j^*-L_j^*-\rho)\right) - \tilde{R}_l'(L^*)L_j^*\right],\\ R_{j,d}(D^*) &= c_{j,d} + r_f + r_p\left(1-\Phi_j(E_j+D_j^*-L_j^*-\rho)\right) - R_{j,d}'(D^*)D_j^*.\end{aligned}$$

These two conditions describe the interior solution of a vertically integrated bank. A corner exists for banks investing solely in the money market:

$$R_{j,d}(D^*) = c_{j,d} + r_f + r_p\left(1-\Phi_j(E_j+D_j^*-\rho)\right) - R_{j,d}'(D^*)D_j^*.$$

This decision depends on banks' access to deposits, as well as on the magnitude of liquidity risk. In particular, smaller banks that do not have access to cheap deposits (low δ_j) and are not geographically diversified, optimally choose to set $L_j = 0$ and earn a return r_f on their deposit.

For larger banks, these two first-order conditions highlight the effect of liquidity risk and market power on deposit and lending rates. The risk of liquidity shortage increases the opportunity cost of lending, and encourages banks to offer more generous deposit rates.[7] In contrast, market power and differentiation in the deposit market create a *markdown* on deposit rates for banks with high service quality. Both create a wedge between the cost of funds and the return on savings, and reduce the equilibrium lending rate. Similarly, the presence of asymmetric

[7] This cost is increasing in the reserve requirement. In practice it gives a cost advantage to financial intermediaries such as *Shadow Banks* that are not regulated by the FDIC.

information in the lending market affects the cost of borrowing by decreasing the average return on loans ($\uparrow r_l$), and decreasing the markup on loans ($\downarrow r_l$).

In equilibrium the position of each bank j on the money market is given by

$$Q_j(r_f) = E_j + D_j^*(r_f) - L_j^*(r_f), \tag{5}$$

where $D_j^*(r_f)$ and $L_j^*(r_f)$ are the Cournot-Nash quantities by bank j for a given money market rate r_f. Q_j can be used to derive a valuation function that determines the strategy of banks when buying or selling short-term securities and trading in the money market: $V_j(q) = Q_j^{-1}(q)$. For instance, in auctions for governmental securities (the primary market), $V_j(q)$ determines the bidding schedule. In secondary markets, heterogeneity in banks' valuations determines the money market rate through a search and bargaining process. In practice, money market transactions are intermediated by dealers that specialize in acquiring securities from the primary market and trade with banks and other financial institutions. This creates a wedge between lending and borrowing costs, and opens up the possibility of market power. For simplicity, we assume that this market operates without frictions which leads to the following market-clearing condition:

$$K = \sum_j Q_j(r_f^*). \tag{6}$$

The model thus far has assumed that financial intermediaries are vertically integrated in retail lending and funding markets. In practice, most lending markets exhibit various degrees of separation between lending and retail funding. This process is facilitated by the ability to securitize loans, so that originators can sell diversified pools of loans to investors in the secondary market. Examples of this include MBS market and ABS markets, such as those for credit-card debt and car loans. In order to limit adverse selection problems and ensure liquid secondary markets, the process of securitization involves additional players, most notably insurance companies and credit-rating agencies.[8] The role of insurance companies, such as Freddie Mac and Fannie Mae for the mortgage market, is to guarantee payment to investors and lenders in the event of default.

To incorporate loan originators into the modeling framework, we assume a sequential timing of moves. In the first stage, banks and fund managers choose the quantity of deposits to raise, as well as their portfolio of asset holdings: long-term asset-backed securities $Q_{j,l}$ and a money market position $Q_{j,f}$. Downstream loan originators (i.e. retailers) take the wholesale price for long-term securities w and insurance premium as given, and decide how many loans to originate. In order to finance loans, these retailers must borrow in the money-market to maintain a certain level of liquidity $\lambda_j \times L_j$. This captures the fact that under this "originate-to-distribute" model of banking, firms finance the creation of new loans by selling loans in the secondary market. The parameter λ measures the speed of the securitization process, which determines the productivity of originators.[9]

The problem of retail originators is defined as follows:

$$\max_{L_j} \quad (R_l(L) - w - g)\, L_j - r_f \boldsymbol{\lambda} \cdot L_j - (F_l + c_l L_j)\,, \tag{7}$$

The actuarially fair insurance premium is given by: $g = 1 - M(L)$. In the mortgage market, this insurance premium is subsidized in order to increase the supply of loans (e.g. $g < 1 - M(L)$).

[8] Chu and Rysman (2019) study competition in the market for ratings.

[9] See Fuster et al. (2019) for a recent analysis of productivity gains realized by Fintech mortgage lenders such as Quickens Loans.

This leads to the following first-order condition:

$$R_l(L) = c_{j,l} + w + g + r_f \lambda_j - L_j R_l'(L). \tag{8}$$

Relative to the equilibrium condition for vertically integrated banks, this equation highlights the relative cost and benefit of separating retail funding and lending. For a given w, lending specialists face limited liquidity risk, and have symmetric funding costs. Both factors reduce the cost of funds, and increase the supply of loans. On the other hand, imperfect competition upstream leads to a double marginalization problem reflected by a markup in w over the cost of funds by upstream fund managers. In addition, retailers do not directly account for the effect of their supply on the riskiness of loans, which enters the model only through the insurance premium.

Aggregating across lenders, this leads to an inverse-demand function for long-term securities determining the wholesale price w:

$$Q_l = L(w) \equiv \sum_j L_j^*(w) \to W(Q_l) = L^{-1}(Q_l). \tag{9}$$

Fund managers allocate deposits between long-term and short-term securities. This leads to the following profit maximization problem:

$$\max_{Q_{j,l}, D_j} \quad (W(Q_l) - r_f) Q_{j,l} + (r_f - R_{j,d}(D)) D_j + r_p \int_{E_j + D_j - Q_{j,l}} (\tilde{x} - D_j + Q_{j,l}) \phi_j(\tilde{x}) d\tilde{x} - c_d D_j. \tag{10}$$

The first-order condition for long-term securities is given by:

$$Q_{j,l}: \quad W(Q_l) = r_f + r_p(1 - \Phi_j(E_j + D_j - Q_{j,l})) - W'(Q_l) Q_{j,l}. \tag{11}$$

As before, the equilibrium wholesale rate is affected by the importance of liquidity risk and (indirectly through the insurance premium) from asymmetric information in the downstream market. However, unlike with vertical integration, only downstream lenders incur the fixed cost of lending, which means that all banks now keep a positive amount of long-term securities on their balance sheet. This increases competition in the market for long-term securities (relative to vertical integration), and in turns attenuates the double marginalization problem. This indirectly affects deposit rates:

$$D_j: \quad R_{j,d}(D^*) = c_d + r_f + r_p \left(1 - \Phi_j(E_j + D_j^* - Q_{j,l}^*)\right) - R_{j,d}'(D^*) D_j^*.$$

The model highlights the main themes that we will explore in the remainder of the chapter: (i) demand for liquidity in wholesale funding markets, (ii) differentiation and liquidity risk in retail funding markets, and (iii) asymmetric information and market power in credit markets. We focus our analysis specifically on the determinant of prices and output in each of these sub-markets. An important focus of the industrial organization literature has been on role that various frictions (e.g. search/switching costs) play in explaining deviations from the law of one price, as well as on the analysis of alternative price-setting mechanisms determining the split of surplus between firms. We discuss how to account for these features when studying the functioning of financial markets.

Unlike in the model above, most papers in the literature study individual sub-markets in isolation. There are a few exceptions. Corbae and D'Erasmo (2021) develop a dynamic entry and exit game in which banks accumulate deposits and equity in order to fund loans. Aguirregabiria et al. (2019) develop a structural model of bank oligopoly competition for both deposits and loans in multiple local geographic markets allowing for interconnections. See also Wang et al. (2019).

3 WHOLESALE FUNDING MARKETS

In this section we describe the methods used to examine market structure and measure market power in markets for government and corporate debt and central banks' operations. The theoretical literature in finance includes several foundational papers that use models with asymmetrically informed traders to provide a rationale for bid-ask spreads (e.g., Glosten and Milgrom (1985)) and market impact of trades (Kyle (1985, 1989)). There has been a recent and burgeoning (theoretical) literature on market microstructure, including Vives (2011) or Rostek and Weretka (2012) that tries to microfound the size of the price impact further, typically employing strict functional form assumptions (e.g., a linear-quadratic setup) to obtain closed form solutions and for comparative statics. However, there has been little work taking these models to the data. One important exception is the case of primary markets for government debt. These markets are often organized as auctions and economic theory has succeeded in developing quite powerful tools for their analysis. We begin by examining the market structure of these auctions, before moving on to interbank markets.

When studying the determinants of prices in the primary market, our focus will be on analyzing willingness-to-pay for these securities, as opposed to the supply of government debt. Based on the framework above, this willingness-to-pay is derived from the resale value of securities in the money market, the regulatory environment and outside investment opportunities. Our interest is in understanding how demand and supply of liquidity affect the price r_f in the interbank market.

3.1 Market for Government Debt

Markets for government debt are, in most countries, organized around a group of agents, called primary dealers, who, among other benefits, have the exclusive right to (i) participate in the issuance of new debt, (ii) route bids on behalf of other players (i.e. who are not primary dealers), and (iii) access various special facilities provided by the central bank.[10] Being a primary dealer also involves certain costs, including (i) stricter regulatory oversight, (ii) a duty to buy a certain fraction of the annual issuance of government debt, and (iii) to continuously "make the markets," (i.e., to stand ready to buy or sell). There are still many open questions regarding the quantification of both the cost and benefit sides of this market.

The number of primary dealers is typically fairly low,[11] which begs the question as to whether the primary dealers can earn large rents from having exclusive participation in government debt auctions. To evaluate the degree of market power, a natural direction to pursue is to build on Bresnahan (1989) and use detailed knowledge of the auction rules to map data on bids from these auctions into underlying values that would rationalize them. We begin by reviewing this method as it will prove useful when discussing further applications.

In this section we focus mostly on papers studying demand for short-term bonds issued by central banks. Similar methods have been used to study other bond markets. Garrett et al. (2020) study auctions of municipal bonds that are typically used to fund local infrastructure, such as schools, bridges etc. They point out that the various tax incentives, together with imperfect competition play an important role in determining municipalities' borrowing costs. They estimate a model that includes equilibrium bidding and endogenous participation, both as functions of the tax incentives. They find that reductions in the tax advantage of the sort that have been frequently discussed by policy makers in recent years would likely lead to increases in bidder markups and lower competition, and hence ultimately might result in an increase in municipal borrowing costs of more than the tax savings to the government.

[10] See Arnone and Iden (2003) for international comparison.

[11] For example, there are 24 PDs in the U.S., 10 in Canada and 18 in the UK.

3.1.1 Using Auction Data to Study Financial Markets

In any auction mechanism, whether for a single item or for multiple items, we can relate the unobservable willingness-to-pay and the observed bids as:

$$BID = WTP - SHADING. \tag{12}$$

The object of interest is the willingness-to-pay (WTP) and there is an abundant literature describing empirical methods for recovering it from the observed bids (Hendricks and Porter (2007), Athey and Haile (2007)). While in most standard analyses of auction markets (e.g., timber auctions, eBay etc) bidders draw their WTP from an exogenous distribution, in financial markets it can often be further microfounded. For example, as we will discuss in one of the applications below, the WTP for a loan being offered in an auction by a central bank should be determined by the next best alternative, i.e., securing a similar loan elsewhere. These links often allow us to provide a particularly useful interpretation of the estimates of the WTP.

The "SHADING" term in (12) might be non-negative (i.e., in first price auctions), zero (in Vickrey auctions) or even negative (as in a 3^{rd} price auction). The actual auction rules and the characteristics of the environment (the information structure etc) contribute to its particular form. The leading examples for auctions in financial markets are a discriminatory auction and a uniform price auction. We discuss these two formats in more detail and then move to the applications.

The underlying model is based on the share auction model due to Wilson (1979). Kastl (2012) extends this model to settings in which bidders' strategies are restricted to a class of step-functions with a given number of steps, which is the empirically relevant case and hence we will restrict our attention to such strategies here as well. We begin with the basic symmetric model with private information and private values and later introduce asymmetries. The papers discussed below in detail spell out all required formal assumptions. For our purposes here, it is enough to keep in mind a standard model with: (conditionally) independently and identically distributed private signals, supply uncertainty and a well-behaved marginal valuation function $v_i(q,S_i)$ where q is the share of the supply obtained. Note that this implicitly also imposes what is in auction terminology called "private values," since $v_i(\cdot)$ is unaffected by the realization of $S_{j\neq i}$.

Let $V_i(q,S_i)$ denote the gross utility: $V_i(q,S_i) = \int_0^q v_i(u,S_i)\,du$. The expected utility of a bidder i of type s_i employing a strategy $y_i(\cdot|s_i)$ in a discriminatory auction can be written as:

$$EU(s_i) = \sum_{k=1}^{K_i} \left[\Pr\left(b_{ik} > P^c\left(Q,\mathbf{S},\mathbf{y}(\cdot|S)\right) > b_{ik+1}|s_i\right) V(q_{ik},s_i) - \Pr\left(b_{ik} > P^c\left(Q,\mathbf{S},\mathbf{y}(\cdot|S)\right)|s_i\right) b_{ik}\left(q_{ik}-q_{ik-1}\right)\right] \tag{13}$$

$$+ \sum_{k=1}^{K_i} \Pr\left(b_{ik} = P^c\left(Q,\mathbf{S},\mathbf{y}(\cdot|S)\right)|s_i\right) E_{Q,s_{-i}|s_i}\left[V\left(Q_i^c\left(Q,\mathbf{S},\mathbf{y}(\cdot|S)\right),s_i\right) - b_{ik}\left(Q_i^c\left(Q,\mathbf{S},\mathbf{y}(\cdot|S)\right) - q_{ik-1}\right)|b_{ik} = P^c\left(Q,\mathbf{S},\mathbf{y}(\cdot|S)\right)\right],$$

where we let $q_{i0} = b_{iK_i+1} = 0$. The random variable $Q_i^c\left(Q,\mathbf{S},\mathbf{y}(\cdot|S)\right)$ is the (market clearing) quantity bidder i obtains if the state (bidders' private information and the supply quantity) is $(Q,\mathbf{S})$ and bidders submit bids specified in the vector $\mathbf{y}(\cdot|S) = [y_1(\cdot|S_1),...,y_N(\cdot|S_N)]$. Since bidders potentially have private information and since the supply is random, the market clearing price is a random variable, denoted by P^c. This random variable maps the realization of the random state of the world (private information of bidders and the random quantity) into prices.

Turning to a uniform price auction and using the notation from above for the (random) market clearing quantity and price, the expected utility of a bidder i who is employing a strategy

$y_i(\cdot|s_i)$ given that other bidders are using $\left\{y_j(\cdot|\cdot)\right\}_{j\neq i}$ can now be written as:

$$\begin{aligned} EU_i(s_i) &= E_{Q,S_{-i}|S_i=s_i} u(s_i, S_{-i}) \\ &= E_{Q,S_{-i}|S_i=s_i}\left[\int_0^{Q_i^c(Q,\mathbf{S},\mathbf{y}(\cdot|S))} v_i(u, s_i, S_{-i})\, du - P^c(Q,\mathbf{S},\mathbf{y}(\cdot|S))\, Q_i^c(Q,\mathbf{S},\mathbf{y}(\cdot|S))\right]. \end{aligned}$$

A Bayesian Nash Equilibrium in this setting is a collection of functions such that (almost) every type s_i of bidder i is choosing her bid function so as to maximize her expected utility: $y_i(\cdot|s_i) \in \arg\max EU_i(s_i)$ for a.e. s_i and all bidders i. With restrictions of strategies to at most K steps, Kastl (2012) labels such Bayesian Nash Equilibria as $K-$step Equilibria. The system of necessary conditions implicitly characterizing such a BNE is the link between the observables and unobservables that we seek to establish.

For a Discriminatory Auction, the key equation is:[12]

$$\Pr(b_{ik} > P^c > b_{ik+1}|s_i)\left[v(q_{ik}, s_i) - b_{ik}\right] = \Pr(b_{ik+1} \geq P^c|s_i)\left(b_{ik} - b_{ik+1}\right), \tag{14}$$

while for a Uniform Price Auction, it is:

$$\Pr(b_{ik} > P^c > b_{ik+1}|s_i)\left[v(q_{ik}, s_i) - E\left(P^c|b_{ik} > P^c > b_{ik+1}, s_i\right)\right] = q_{ik}\frac{\partial E\left(P^c I\left[b_{ik} \geq P^c \geq b_{ik+1}\right]|s_i\right)}{\partial q_{ik}}. \tag{15}$$

Note that the first condition shows that the trade-off in the multi-unit environment is essentially the same as in the first price sealed bid auction as characterized in Guerre et al. (2000). The expected surplus on the marginal (infinitesimal) unit is traded off against the probability of winning it. The trade-off in a uniform price auction, on the other hand, resembles the decision of an oligopolist facing an uncertain demand. The trade-off is between the (expected) surplus on the marginal unit sold versus the loss of surplus on inframarginal units due to (expected) price impact.

Equations (14) and (15) define the mapping between observables (bids) and the WTP $v(q, s_i)$. Since (14) and (15) are necessary conditions for an optimal choice of q_k in each respective auction format, the estimation approach described in the next subsection can follow along the lines of Guerre et al. (2000). For each observed bid b_k, we estimate the marginal willingness to pay $\hat{v}_k$ that rationalizes that bid. With the estimates of marginal willigness to pay in hand, we can start evaluating rents (both interim and ex post) that market participants enjoyed and efficiency of the allocation. This is one of the key ingredients to the puzzle motivating our analysis: what is the value of being a primary dealer.

3.1.2 Estimation of willingness-to-pay

We now discuss an estimation approach that is based on leveraging the relationship between observables and objects of interest, i.e., preferences, valuations or willingness-to-pay, derived from the equilibrium of an economic model. In a strategic environment, each participant's behavior depends on the behavior of her rivals. Typically, we assume that players play a (Bayesian) Nash Equilibrium, which essentially restricts their beliefs about rivals' play to be consistent with equilibrium strategies. Inspecting (14) and (15), the two random variables, whose distributions need to be estimated, are the market clearing price, P^c, and the market clearing quantity, Q^c. These distributions can be estimated by employing a bootstrap-like procedure: we resample with replacement from the observed data different samples of bids (that are consistent

[12] For more details refer to Kastl (2012).

with the information structure) and evaluate the market outcome. This idea originated in Hortaçsu (2002b).

Let the object we aim to estimate be the distribution of P^c, defined as $H(X) = \Pr(P^c \leq X)$. The market clearing condition implies that:

$$H(X) \equiv \Pr(X \geq P^c | s_i) = E_{\{Q, S_{j \neq i}\}} I\left(Q - \sum_i y(X|s_i) \geq 0\right), \tag{16}$$

where $I(\cdot)$ is the indicator function and $E_{\{.\}}$ is an expectation over the random supply and other bidders' private information. We are essentially interested in the fraction of states of the world in which the supply exceeds the demand and thus the market clearing price must be lower than X.

The estimation approach used in the literature begins by defining an indicator of excess supply at price X (given bid functions $\{y_j(X|s_j)\}_{j \neq i}$ and i's own bid $y_i(X|s_i)$) as follows:

$$\Phi\left(\{y_j(X|s_j)\}_{j \neq i}; X\right) = I\left(Q - \sum_{j \neq i} y_j(X|s_j) \geq y_i(X|s_i)\right). \tag{17}$$

The dependence on i's own bid, $y_i(X|s_i)$, is made explicit, since it follows from (15) that in a uniform price auction one needs to estimate how the expected market clearing price would change if one's own bid were to change. The distribution in (17) can then be estimated by constructing a V-statistic as follows:

$$\xi\left(\hat{F}; X, h_T\right) = \frac{1}{(NT)^{(N-1)}} \sum_{\alpha_1=(1,1)}^{(T,N)} \ldots \sum_{\alpha_{N-1}=(1,1)}^{(T,N)} \Phi\left(y_{\alpha_1}, \ldots, y_{\alpha_{N-1}}, X\right), \tag{18}$$

where $\alpha_i \in \{(1,1),(1,2),\ldots(1,N),\ldots,(T,N)\}$ is the index of the bid in the subsample (where an index (t,n) corresponds to the bidder n^{th} bid in auction t) and $\hat{F}$ is the empirical distribution of bids, i.e., the empirical probability distribution over points in 2K-dimensional space.

Since the number of all subsamples might be too high to be evaluated, one can construct a simulator of $\xi\left(\hat{F}; X, h_T\right)$ by drawing only M subsamples rather than all $(NT)^{(N-1)}$. It is easy to show that such a simulator is consistent as $T \to \infty$ (and under appropriate conditions on the rate at which the number of simulations, M, increases). Hortaçsu and Kastl (2012) and Cassola et al. (2013) analyze the asymptotic properties of these estimators. One important result is that this estimator is also consistent as $N \to \infty$ provided an additional technical condition holds. This result is useful in settings where pooling data across actions might be problematic due to suspected unobserved heterogeneity. The estimator can be easily modified to allow for (ex ante) asymmetries or for presence of covariates by introducing weighting of observations into the resampling procedure, similarly to a kernel-based nonparametric regression.

3.1.3 Evaluating the Performance of the Auction Mechanism

The models and estimation methods described in the previous subsection were first used to compare auction formats. Hortaçsu (2002a) employs Wilson's (1979) model, which assumes continuously differentiable bidding strategies, and evaluates Turkish Treasury auctions. He argues that the discriminatory auction format seems to perform well, since revenue from it exceeds the upper bound on revenue from a hypothetical uniform price auction. Kastl (2011) writes down the model sketched above and argues that when bidders submit step functions, especially those with few steps, the strategies might be quite far from those implied by Wilson's (1979) model and hence the upper bound proposed in Hortaçsu (2002a) might not be valid. He

uses data from the Czech Republic and proposes to evaluate the performance of the auction mechanism by focusing instead on the two sources of inefficiency of the mechanism that contribute to lower revenue for the auctioneer: (i) the inefficiency of allocation that arises from treasury bills not being allocated to bidders who value them the most and (ii) the information rents that accrue to bidders due to their private information. Since these rents cannot be fully eliminated in any incentive compatible mechanism, if the estimated losses are not "too large," the mechanism can be viewed as performing well. For the case of the Czech auctions he finds that these rents amount to about 8 basis points.

Kang and Puller (2008) analyze Korean treasury auctions, which have used both discriminatory and uniform price auction formats. They argue that the auctions seem, in general, fairly competitive, but that the discriminatory auction format leads to a lower allocational inefficiency. These papers, together with Hortaçsu et al. (2018), which considers the case of US treasury auctions and is discussed in more detail below, all suggest that the auction format might not be a first order consideration in treasury auctions. Of course there are still interesting open questions that would address the role of the auction format with endogenous participation and with different market structure than the current one centered around primary dealers.

3.1.4 Quantifying Market Power

Most markets for government debt that are organized around primary dealers also involve two other classes of participants. First, there are large indirect bidders (sometimes labeled "customers" - large buyers like BlackRock or large pension funds etc), whose bids are first submitted to a primary dealer, who is in turn required to route them separately into the auction under the indirect bidder's unique identifier. Second, some auctions allow direct participation even by bidders who are not primary dealers. This is particularly relevant for the case of the US where a large portion of the debt is typically purchased by foreign central banks.

Hortaçsu et al. (2018) study the US market, which is organized as a uniform price auction with three classes of bidders. Their data set covers all auctions between July 2009 and October 2013. They begin by documenting that primary dealers consistently bid higher yields (i.e., offer lower prices) than both direct and indirect bidders. As the model discussed above suggests, this could be because primary dealers have lower marginal values (and thus require a higher yield) or it could be because they are larger and "shade" their bids more as illustrated by equation (15). Using the machinery developed in the previous section, augmented to allow for the ex-ante asymmetry between the different bidder classes, they estimate marginal values that rationalize the observed bids and thus quantify these two channels separately. Table 1 reports the analysis of shading factors.

It follows that primary dealers enjoy some market power, especially relative to direct and indirect bidders. Their expected impact on the market clearing price makes them optimally shade their bids more: depending on maturity the extra shading could amount to several basis points. Using the estimated valuations, Hortaçsu et al. (2018) further report that primary dealers enjoy significantly higher rents than direct and indirect bidders as these could average as much as 22 basis points on a 10-year bond. They also show that these rents are increasing in the number of indirect bidders' bids a primary dealer gets to route in an auction. This suggests that the ability to observe the bids of their "customers" might be an important benefit of being a primary dealer. As we will discuss below, Hortaçsu and Kastl (2012) utilize timestamps on primary dealers' bids and the fact that they tend to update their bids prior to the auction as information about customers' bids to quantify this benefit further.[13]

There are two important takeaways from their analysis: (i) short term maturity auctions

[13]The US data do not currently have these exact timestamps. As soon as these become available, there is certainly a worthwhile paper to be written.

Table 1. Analysis of Bid Shading

	Bills				Notes			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dependent Variable	Shade 1		Shade 2		Shade 1		Shade 2	
Direct	-1.904***	-1.434***	-0.862***	-0.771***	-4.696***	-2.203***	-0.0954***	-0.0480***
	(0.0851)	(0.0969)	(0.0727)	(0.0884)	(0.305)	(0.255)	(0.0103)	(0.0105)
Indirect	-3.511***	-2.996***	-1.125***	-1.025***	-13.36***	-10.15***	-0.122***	-0.0608***
	(0.105)	(0.117)	(0.0813)	(0.0978)	(0.684)	(0.469)	(0.0116)	(0.0129)
%Q Total		3.085***		0.600*		30.73***		0.584***
		(0.353)		(0.330)		(4.399)		(0.108)
Constant	0.841***	0.265***	1.174***	1.062***	-1.888***	-5.394***	0.125***	0.0579***
	(0.0543)	(0.0819)	(0.0441)	(0.0756)	(0.478)	(0.353)	(0.00883)	(0.0122)
Observations	41,264	41,264	41,264	41,264	13,692	13,692	13,692	13,692
R-squared	0.095	0.097	0.015	0.015	0.158	0.162	0.062	0.069
Number of auctions	822	822	822	822	153	153	153	153

Shade 1 is defined as $B(\theta_i) = \frac{\sum_{k=1}^{K_i} q_k[v_i(q_k,\theta_i) - b_k]}{\sum_{k=1}^{K_i} q_k}$ and Shade 2 defined as $S(\theta_i) = \frac{\sum_{k=1}^{K_i} q_k[v_i(q_k,\theta_i) - \mathbb{E}(P^c | b_k > P^c > b_{k+1}, \theta_i)]}{\sum_{k=1}^{K_i} q_k}$.

Bid shading is reported in basis points.

Auction fixed effects are controlled for in every specification.

Robust standard errors, clustered by auctions, are reported in the parentheses.

*** p<0.01, ** p<0.05, * p<0.1

are very competitive, with small surpluses accruing to the bidders, which suggests that the role of private information and market power is not very important, and even for long maturities the rents are not too high and (ii) the customers' order flow is an important source of rents for primary dealers.

3.1.5 Quantification of Front-running and Testing for Private Values

Hortaçsu and Kastl (2012) utilize detailed data from Canadian treasury auctions to quantify the value of observing customer bids. Unlike in the US, Canada uses the discriminatory auction format. The link between the observed bids and the unobserved willingness-to-pay in the basic model is given by equation (13). Hortaçsu and Kastl (2012) modify this model of bidding to allow dealers to observe customer bids. They assume that private information is independent both across dealers and across customers. More importantly, they assume that values are still private. This assumption may be easier to maintain for certain securities (such as shorter-term securities, which are essentially cash substitutes) than others. Fortunately, the feature of the data that allows them to quantify the value of the order flow also allows them to test this important assumption itself. As will become clear below, even with private values, observing customers' bids is valuable for the dealers as it allows them to update their beliefs about the competitiveness of the auction, or, somewhat more precisely, about the distribution of the market clearing price. The key feature of the data is illustrated in Figure 4. As the deadline for the auction approaches, primary dealers often submit bids under their own identifiers (the solid most outward line in the figure). These bids typically aggregate both trades on their own accounts and for smaller customers, for whom the central bank does not require separate identification. Occasionally, with some time still remaining before the deadline, an additional order arrives from a customer, who needs to be separately identified for the central bank (the solid line closest to the origin in the figure). The dealer thus obtains information about the exact shape of the bid of one of the auction participants and can thus update her beliefs about where the market might clear. Such updated beliefs will lead to a new optimal bid and hence, if time permits, this dealer submits an updated bid (the dashed line in the figure). This setup allows for two analyses: (i) if there is a step at the same quantity in both the original and the updated bids and if we were to estimate the marginal value that rationalizes that bid (given the available information), these two estimates should not be statistically distinguishable under the null hypothesis of private values, and (ii) if we can calculate the expected profits (given the available information) corresponding to both the original and the updated bids, the difference can be attributed to the information contained in the observed customer order flow (assuming that the two bids were submitted close to each other).

In their Figure 1 Hortaçsu and Kastl (2012) report that many customers do indeed submit their bids very close to the auction deadline. Such "last minute" bidding behavior by customers can be rationalized as a strategic response by customers who do not want dealers to utilize the information in their bids. There may be reasons for customers to voluntarily share their information with dealers as well. For example, Bloomberg Business published an article on 4/4/2013 in which a representative of BlackRock, the world's largest asset manager, described why BlackRock chooses to participate indirectly by submitting bids through primary dealers: *"While we can go direct, most of the time we don't. We feel that the dealers provide us with a lot of services. Our philosophy at this point is, to the extent we can share some of that information with trusted partners who won't misuse that information, we prefer to reward the primary dealers that provide us all that value."*[14] Whether or not large bidders such as BlackRock should be allowed to participate indirectly and thus share their bids with a particular primary dealer, especially in uniform price auctions, where all winners benefit from a lower market clearing price, is an important open question.

[14] `http://www.bloomberg.com/news/articles/2013-04-04/bond-traders-club-loses-cachet-in-`

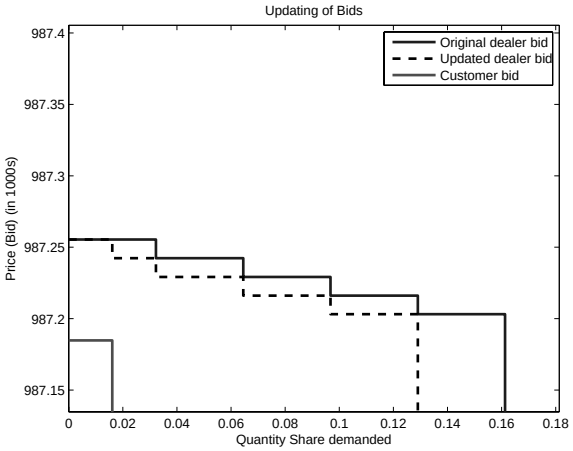


Figure 4. Bid Updating in Canadian Treasury Bill Auctions

Another important issue is the quantification of the many complementary services that primary dealers provide to their customers. Ignoring these channels would lead to an inconsistent estimate of rents dealers and customers accrue just from the Treasury auctions. For example, primary dealers might step in when customers need to off-load some illiquid asset; they might have valuable information about demand for such assets and thus may act as intermediaries. In other markets primary dealers may engage in trade in order to gather information about the fundamentals of assets, for which, for a variety of reasons including illiquidity, prices may not aggregate information properly (see Brancaccio et al. (2019) or Allen and Wittwer (2020) for more details).

To formally address the impact of information updating, let us slightly modify the notation. Let $\mathscr{C}$, $\mathscr{P}$ denote the index sets containing indexes of customers and primary dealers, respectively, and let the type of a dealer be (S,Z), where the random variable S summarizes the signal as before and Z contains the order flow, i.e., all customer bids revealed to this dealer. The distribution of the market clearing price from the perspective of primary dealer j, who observes the bids submitted by customers contained in an index set $\mathscr{C}_j$, is given by:

$$\Pr\left(p \geq P^c | s_j, z_j\right) = E_{\left\{S_{k\in\mathscr{C}\setminus\mathscr{C}_j}, S_{n\in\mathscr{P}\setminus j}, Z_{n\in\mathscr{P}\setminus j} | z_j\right\}} I\left(RS\left(p, Q, \vec{S}, \vec{Z}\right) \geq y^P\left(p|s_j, z_j\right) + \sum_{m\in\mathscr{C}_j} y^C\left(p|s_m\right)\right), \tag{19}$$

where $RS\left(p, Q, \vec{S}, \vec{Z}\right) = Q - \sum_{k\in\mathscr{C}\setminus\mathscr{C}_j} y^C\left(p|S_k\right) - \sum_{n\in\mathscr{P}\setminus j} y^P\left(p|S_n, Z_n\right)$, i.e., the residual supply at price p given supply realization Q and realization of private information $\left(\vec{S}, \vec{Z}\right)$. This expression simply says that the dealer "learns about competition" – the primary dealer's expectations about the distribution of the market clearing price are altered once she observes a customer's bid. If $\mathscr{C}_j$ is an empty set, then the dealer needs to integrate over the types of all rivals, whereas if she observes some of them, she simply conditions on them.

Finally, the distribution of P^c from the perspective of a customer is very similar to an uninformed primary dealer, but with the additional twist that the indirect bidder recognizes that her bid will be observed by a primary dealer, d, and can condition on the information that she provides to this dealer. The distribution of the market clearing price from the perspective of an indirect bidder j, who submits her bid through a primary dealer d is given by:

$$\Pr\left(p \geq P^c | s_j\right) = E_{\left\{S_{k\in\mathscr{C}\setminus j}, S_{n\in\mathscr{P}}, Z_{n\in\mathscr{P}} | s_j\right\}} I\left(Q - \sum_{k\in\mathscr{C}\setminus j} y^C\left(p|S_k\right) - \sum_{n\in\mathscr{P}} y^P\left(p|S_n, Z_n\right) \geq y^C\left(p|s_j\right)\right), \tag{20}$$

where $y^C(p|s_j) \in Z_d$ and $d \in \mathscr{P}$.

Note that the probability distributions from the perspective of an uninformed primary dealer, or an informed primary dealer (given her observation of customer order flow), or an indirect bidder can then be estimated using equations (19) and (20) using the resampling technique described above. With the estimates of the probability distributions of market clearing price in hand, we can use equation (14) to estimate the willingness-to-pay (or, equivalently, the shading factor) at every observed bid.

Private versus Interdependent Values in Treasury Bill Auctions

As mentioned, the observed updating of bids allows us to investigate whether primary dealers are just learning about competition in the upcoming auction (and hence not updating their valuation estimates) or whether they may also be learning about fundamentals, and hence updating their value estimates after observing customers' bids. Equation (13) shows that beliefs about where the market will clear are key for determining the optimal bid. A primary dealer forms these beliefs by integrating over all uncertainty: available supply, the signals of rival primary dealers, the signals of all customers, which primary dealer a customer might route her bid through, etc. By observing a customer's order, part of this uncertainty gets resolved: a primary dealer can therefore update her belief about the distribution of the market clearing price by evaluating (19). Such updating can be incorporated into the resampling technique by fixing the observed customer bid instead of resampling that customer's bid. We can therefore establish a formal hypothesis test about whether values are indeed private. In particular, let $\left\{v_k^{BI}(q_k,\theta)\right\}_{k=1}^{K}$ denote the vector of the estimated willingness-to-pay that rationalizes the observed vector of bids before the customer's order arrives (hence, "BI"), $\left\{b_k^{BI}(q_k,\theta)\right\}_{k=1}^{K}$, and $\left\{v_k^{AI}(q_k,\theta)\right\}_{k=1}^{K^{BI}}$ denote the vector of the estimated willingness-to-pay that rationalizes the observed vector of bids after the customer's bid arrives (hence, "AI"), $\left\{b_k^{AI}(q_k,\theta)\right\}_{k=1}^{K^{AI}}$. If we were to observe a bid for the same quantity being part of both the bid before the customer's information arrives and the bid after, we can simply formulate a statistical test at that quantity. Let $T_j(q) = v^{BI}(q,\theta) - v^{AI}(q,\theta)$ be the difference between the rationalizing willingness-to-pay for a given quantity, q, before and after information arrives (taking into account the updating about the distribution of the market clearing price during the estimation). Testing the null hypothesis (of no learning about fundamentals) then involves testing that $T_j(q) = 0, \forall j, q$.

Hortaçsu and Kastl (2012) report that the null hypothesis of no learning about fundamentals from customers' bids cannot be rejected based on several alternative tests, including those that take into account the multiple hypothesis issue. While this still does not preclude interdependency of values between primary dealers themselves, given that many of the customers are large players (such as BlackRock), this evidence is at least suggestive that modeling the information structure in Treasury auctions as private values is reasonable.

Quantifying the Order Flow

Now we are ready to evaluate the expected profits corresponding to the initial bid and, after appropriately modifying the beliefs about the market clearing price, also those corresponding to the updated bid. If we assume that nothing else is changing between the original bid and the updated bid other than the dealer observing the customers' orders, we can quantify the value of order flow by comparing the two.

Let $\Pi^I(s_i,z_i)$ denote the expected profit of a dealer d, when using the bidding strategy $y^I(p,s_i,z_i)$, i.e., after incorporating the information from customers' orders which is contained in the realization of the random variable Z_i. Similarly, let $\Pi^U(s_i,)$ denote the expected profit corresponding to the bidding strategy $y^U(p,s_i)$, i.e., before customers' orders arrive. The value

of information in terms of this notation is as follows:

$$VI = \int_0^\infty \Pi^I(s_i, z_i)\, dH\left(P^c, y^I(s_i, z_i)\right) - \int_0^\infty \Pi^U(s_i)\, dH\left(P^c, y^U(s_i,)\right), \tag{21}$$

where $H\left(P^c, y^x(\cdot)\right)$ is the distribution of the market clearing price, P^c, given other bidders using equilibrium strategies and dealer d following the strategy $y^x(\cdot)$, where the superscript $x \in \{I, U\}$ indexes the "informed" and the "uninformed" dealer. Recall that the distribution of the market clearing price is different for uninformed and informed dealers: the former case needs to integrate out the uncertainty with respect to all customers, whereas the latter takes one customer's realized bid as given and only integrates out the remaining uncertainty.

Hortaçsu and Kastl (2012) report that about one third of PDs' profits can be attributed to the customers' order flow information. Overall, however, the auctions seem fairly competitive and as a result, the profits do not appear to be excessive. Hortaçsu et al. (2018) reached a similar conclusion for the US Treasury auctions, even though the setting is not as clean since the data are missing precise timestamps.

3.1.6 Mapping Bids into a Demand System for Government Securities

Allen et al. (2020) analyze another peculiar feature of treasury bill auctions - the fact that treasury bills of multiple maturities are being sold in simultaneous auctions. This feature is surprisingly common as most countries, including the US, Canada, the UK and many others, run their primary issuance in this way. In the case of Canada, there are auctions for 3-month, 6-month and 12-month Treasury bills that are run at the same time. However, the bidding language does not allow the bidders to condition their bids for one maturity on the allocation of others. It is straightforward to see that marginal values for treasury of one maturity depend on the allocation of other maturities, since there are budget constraints and the structure of dealers' portfolios also typically dictates a certain most preferred maturity structure. Due to the lack of bid conditioning, the bids for maturity m will thus depend on the expectations of how much of the other maturities a dealer might (expect to) win. Since through the bid inversion technique discussed above we can recover the willingness-to-pay rationalizing each bid, for any triplet of auctions we can recover a corresponding triplet of willingnesses-to-pays and hence we can study how these depend on each other.

The big advantage relative to standard demand estimation settings is that, since the bids in one auction typically consist of several price-quantity pairs, applying the bid inversion technique results essentially in observations of several points on the demand curve at any point in time. This substantially simplifies the problem relative to the applications based on the tools developed in Berry et al. (1995), where one needs to link together price-quantity pairs that are results of equilibrium intersections of supply and demand curves, i.e., solutions to a system of simultaneous equations.

To be more precise, let a simultaneous auction be indexed by τ and let bidder i of type $s^g_{m,i,\tau}$ in bidder group $g \in \{d = \text{dealer}, c = \text{customer}\}$ have the following willingness to pay for amount q_m in auction m conditional on winning q_{-m} of the other two maturities and keeping a share $(1-\kappa_m)$ on its own balance sheet

$$v_m(q_m, q_{-m}, s^g_{m,i,\tau}) = \alpha + (1-\kappa_m)s^g_{m,i,\tau} + \lambda_m q_m + \delta_m \cdot q_{-m}, \tag{22}$$

where δ_m measures the interdependencies across maturities. For example, if $\delta_{3M,6M} < 0$, bidders are willing to pay less for any amount of the 3-month maturity the more they purchase of the 6-month bills, and so the bills are substitutes. When $\delta_{3M,6M} > 0$ they are complementary, and independent if $\delta_{3M,6M} = 0$.

Since bidders cannot condition their bids for any given maturity on quantities of other maturities won, these quantities of other maturities won, $q^*_{-m,i}$, are unknown at the time the bid is submitted and thus need to be integrated out:

$$\tilde{v}_m(q_m, s^g_{m,i,\tau}) = \mathbb{E}[v_m(q_m, q^*_{-m,i}, s^g_{m,i,\tau})| \text{ win } q_m].$$

Hence, following the bid inversion technique corresponding to equation (14), after the first stage of the estimation procedure we recover $\tilde{v}_m(q_m, s^g_{m,i,\tau})$. In addition, the resampling approach allows us to estimate the joint distribution of market clearing prices (and quantities). With this we can estimate the conditional expectation $\mathbb{E}[q^*_{-m,i}| \text{ win } q_m]$. Finally, given the linearity assumption in (22) we can estimate the parameters of interest, δ_m, by a linear regression with bidder-auction-time fixed effects that control for $\alpha + (1 - \kappa_m)s^g_{m,i,\tau}$.

Allen et al. (2020) find non-negligible interdependencies in primary dealers' demands for government securities. There is quite a bit of heterogeneity among dealers, but roughly there are two main "types." For one group of dealers, treasuries of different maturities are complementary to each other. For these dealers, the money market is roughly their primary business. They have lots of clients and run large money desks. For the second group, the treasuries of different maturities are substitutes to each other. These dealers have comparatively smaller money desks and the focus of their business models is on other financial markets.

Aggregating these individual demands into a market demand results in significant dependencies, although not as clear cut as at the individual dealer level. Allen et al. (2020) estimate that the marginal valuation for a 3-month T-bill increases when going from an allocation with no other T-bills of other maturities at all to an allocation corresponding to the average observed allocations of 6 and 12-month bills (about 200 million each) by about 0.5 basis points. While this effect may seem small in absolute size this "cross" effect is actually relatively large compared to the "own" effect: the marginal value for the 3-month bill drops by about 1 basis point when going from none to an allocation of 400 million.

3.2 Secondary market and demand for liquidity

3.2.1 Auction mechanisms

Most large financial institutions rely on very short-term loans to finance their operations. These loans can originate either from central banks, such as discount windows or various refinancing operations, or from other financial institutions. They can be either secured by collateral or unsecured. In a secured loan, the so called repurchase agreement (or "repos"), the borrower essentially transfers ownership rights to a security used as collateral and simultaneously enters in a repurchase contract that guarantees the return of these rights after the loan matures. Alternatively, the loans can be unsecured. Of course the price for each loan type needs to reflect the risk assumed by the lender when issuing a loan to a particular borrower. A panel data set on interest rates required from each bank for different types of collateral, together with the actual transactions, is thus an invaluable source of information on riskiness of individual financial institutions, the structure of their balance sheets and ultimately the risk faced by the whole financial system.

One problem is that such data are typically not readily available. There have been two approaches to address this. The first utilizes data from a central clearing house such as the Fedwire. The downside is that such data only include quantities transferred between two counterparties, but not the agreed upon maturity or interest rate. Several papers have proposed to invert the transactions for "similar" amounts between the same two counterparties over a very short time horizon (overnight or at most few days) using an algorithm originally proposed

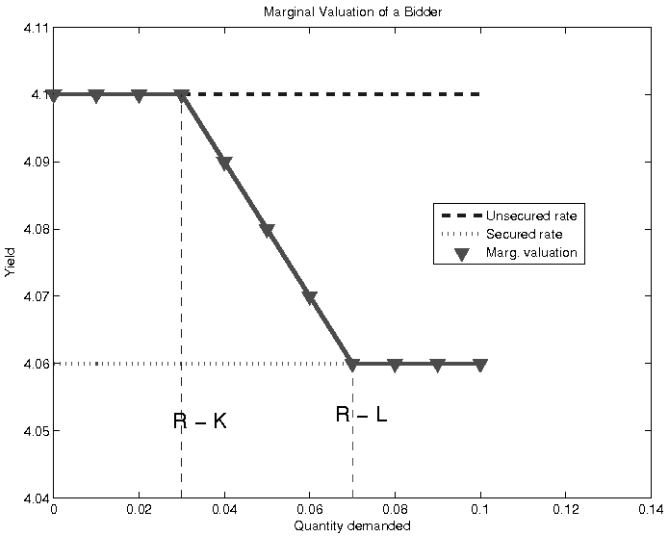


Figure 5. Marginal Willingness-to-pay for Repo Loans

by Furfine (1999).[15] The second approach uses bids for repo loans offered by central banks' liquidity auctions. The starting point of such analysis is the observation that the interest rate a bank would be willing to pay for a repo loan originated by the central bank must be equal to the interest rate that could be obtained for a similarly collateralized loan on the private market that is typically run over-the-counter. Figure 5 (which is Figure 4 in Cassola et al. (2013)) provides an example of a marginal valuation curve, with which a bank might arrive in a liquidity auction. It consists of three segments. In this example, a bank would like to obtain 10% of the available supply of funds. It has L units of highly liquid collateral that would be accepted as collateral in the private market virtually at a risk-free rate (e.g., German treasury bills) - this is the right-most segment. Then it has $K - L$ units of less liquid collateral, which would be accepted as collateral in the private market at a premium over the risk-free rate. This is captured by the middle segment. Finally, this bank then could have additional securities that might be acceptable to the central bank as collateral with an appropriate haircut, but virtually unusable as collateral in a private transaction. This is captured by the first segment. Given the rules of the auction and the bank's beliefs about the behavior of its rivals, it would transform this marginal valuation curve into an optimal bid. Their bid inversion technique therefore allows us to invert the observed bids into a panel of interest rates that can potentially reveal a lot about the dynamics of the balance sheet of each bank.

Cassola et al. (2013) analyze liquidity auctions of the European Central Bank (ECB). Prior to December 2008 these were weekly discriminatory auctions of 1-week loans that were collateralized by pre-specified securities, through which the ECB managed liquidity in the market. They document that virtually all banks started bidding much more aggressively for loans from the ECB - suggesting that the loans from the central bank became a really attractive option for most of them. However, they use the bid inversion technique to construct a panel of interest rates that individual banks were willing to pay to secure a loan from the ECB and they argue that, while for two thirds of banks the outside option indeed deteriorated, the remaining third simply responded strategically and adjusted their behavior.

Using a similar argument Allen et al. (2021b) use data from Canadian liquidity auctions to

[15] Some questions about the reliability of this algorithm have been raised in Armantier and Copeland (2015) due to both high type I and type II errors. However, Allen et al. (2021b) argue that it may be reasonably precise when applied to the Canadian payment system.

argue that Canadian banks' willingness-to-pay for repo loans did not change much during the crisis, other than a brief episode around the collapse of Lehman brothers, and hence the crisis did not seem to have impacted the Canadian financial system in any important way.

3.2.2 Over-the-counter markets

Debt securities purchased in the primary market can be sold on secondary markets. A key friction in these decentralized, over-the-counter (OTC) markets is that the security holder must search for, and match with, a buyer. Duffie et al. (2005) develop a model of search and bargaining in decentralized markets with symmetric information, building on Diamond (1982). Agents are able to trade only if they are matched together. They do so by searching for opportunities to trade. Once matched, agents bargain over transaction prices, understanding that, should they be unable to come to an agreement, costly delay would ensue as each searched for another match. In equilibrium, agents agree to trade any time there are gains from doing so, and prices depend on search frictions.[16]

Gavazza (2016) combines elements from Duffie et al. (2005) and Rubsinstein and Wolinsky (1987) and extends them to capture key features of real asset markets, such as the heterogeneity of assets due to depreciation. His application is the secondary market for business aircraft. Using his model he compares the outcome with search and matching frictions to a Walrasian benchmark. He finds that, relative to the benchmark, 18.3 percent of assets were misallocated, prices were 19.2 percent lower, and welfare was lower by 23.9 percent. He also finds that frictions can be reduced through the use of intermediaries.[17]

Policy-makers have recently been calling for more "transparency" in secondary markets for debt, basically requiring that all trades be made observable in real-time or with a very short lag. Brancaccio et al. (2019) study the decentralized trading in the secondary market for US municipal bonds. They build a model of an environment where the value of the traded objects is stochastic and public information about this value is limited, but there might be private information. In such a setting there may be an incentive to engage in trading for information-gathering reasons. Such information acquisition incentives would endogenously increase liquidity (as measured by the number of executed trades) relative to a world where all trades would be immediately revealed. If it is the objective of the regulators to ensure sufficient liquidity in the market, transparency (defined by making all trading information immediately available) might not necessarily be a good idea. They first document that when the "transparency" regulation was put in practice in 2003 and 2005, the number of trades significantly dropped and the intermediation spread increased.

In order to isolate and quantify the information acquisition incentive for trading, Brancaccio et al. (2019) build a dynamic model of trading. The underlying value of the asset follows a Markov process and all dealers learn about the current value by collecting signals which get generated by trading. A trade with an experienced counterparty (like a dealer) is likely going to be more informative than a trade with a small customer – and hence the history of trades by the counterparty might be an important part of the state space. After estimating the model by maximum likelihood they use it to quantify of the value of information (and of the precision of the signals). For example, they estimate that experimentation increases the precision of the estimates by about 20% relative to a world where these experimentation incentives were shut down and that experimentation explains roughly 15% of the volume of trade in this market. They also simulate the trading behavior under transparency regulation, when all trades, and therefore signals, immediately become public. They find that dealers would execute roughly 4% fewer trades.

[16] See also Hugonnier et al. (2012).

[17] See also Gavazza (2011a,b) for empirical tests of the model.

Allen and Wittwer (2020) study the role of primary dealers in intermediating trades and quantify potential benefits from moving the transactions from OTC financial markets, which rely on bilateral deals between dealers ("intermediaries") and other financial institutions ("investors"), to centralized platforms like the New York Stock Exchange. Such centralized platforms exhibit a much higher degree of transparency. The Canadian context is quite attractive to study this aspect as both markets operate simultaneously: much of the market still relies on OTC transactions, but a non-trivial share of the volume transacts on an online centralized platform. Allen and Wittwer (2020) document two important facts. First, dealers enjoy non-negligible markup in the OTC market even for homogeneous securities such as treasury bills. Second, investors who have access to the centralized platform enjoy better terms in their OTC transactions, suggesting that access to the centralized platform allows them to benefit from a better bargaining position due to the improved outside option. The key trade-off in the model of Allen and Wittwer (2020) is that, on the one hand, accessing the platform is costly (primarily because the investor benefits from the private nature of the OTC transactions), but on the other hand the platform leads to competition among dealers and thus to more competitive quotes. Their results suggest that, overall, investors gain from the platform, but perhaps not as much as one would have expected given that even if platform access is made free, not all investors find it profitable to access it as some of them already enjoy competitive prices in the OTC market.

The above results speak to the importance of information and matching. A series of papers have also looked at the role of information in stock exchanges and at credit agencies. Budish et al. (2015) and Budish et al. (2019) study the design of financial exchanges. Budish et al. (2015) point out that there is a design flaw in financial exchanges, namely that time is considered to be a continuous variable and exchanges process requests to trade in a serial fashion. This setup leads to important arbitrage opportunities across exchanges as new information gets reflected in one exchange, but not necessarily immediately in all other ones. The authors suggest that time should be discrete instead and exchanges should operate frequent batch auctions equivalent to uniform price double auctions conducted at (relatively) high frequency. This fix forces competition to be based on price rather than on speed. Budish et al. (2019) then examine whether the market on its own will fix itself. To do so they develop a model of stock exchange competition, extending the Budish et al. (2015) model to competing exchanges. The model is able to reproduce key institutional features, and is then used to investigate incentives for market design innovation. Their main finding is the existence of a Bertrand trap in design incentives. The private incentives of incumbent stock exchanges to innovate are not aligned with social interests. Since the main source of revenues for the exchanges originates in sales of ancillary services, such as data feed subscriptions or co-location services, rather than from transaction fees, they profit more from the race for speed as this leads their customers to invest in such ancillary services.

4 RETAIL FUNDING MARKETS

In this section we analyse the first link in the vertical chain described in Section 1. Managers raise funds in retail markets by issuing deposits (banks), or claims on future returns (pension plans, mutual funds, life insurance, etc.) Managers then invest these funds in risky projects – issue loans, invest in government and corporate securities – originated by downstream firms in credit and capital markets.

Most of the industrial organization literature has focused on the exercise of market (or monopsony/oligopsony) power by fund managers vis-à-vis savers. In contrast, there has been much less discussion about the exercise of market power towards entities operating in the credit or capital markets. A model of market power in retail funding markets would ideally also include a framework for thinking about competition along the vertical chain. Of course, in some cases

(e.g. banks), entities managing assets and providing credit are vertically integrated. Instead we assume throughout our discussion that fund managers earn some exogenously-set rate of return on the investments they make using the funds they raise.

Market power vis-à-vis savers arises for a variety of reasons, including differentiation, search frictions, and switching costs. A number of papers have been written that examine these different sources of market power in deposit markets, pension/annuities markets and the market for mutual funds. These papers all start with a discrete-choice framework in which savers in each of the different funding markets choose with which financial institution / product to place their savings. To our knowledge, no paper allows savers to consider the portfolio allocation decision across different savings devices / funding markets. In what follows we present a framework for examining competition in these settings, before examining in detail the methods used to study sources of market power.

4.1 Framework

Consider a set of products (savings/chequing accounts, mutual funds, pension plans, etc.) offered by managers (financial institutions) indexed by j, that compete a la Bertrand Nash for savers by charging them a fee p_j to manage their savings. Note that in the case of retail deposit markets, p_j is typically negative, with managers offering savers a positive rate of return on their savings, (r_d in the terminology of the model presented in Section 2). Managers also earn profits by lending out the invested savings, earning a return r_ℓ. We write the profit function for firm j as

$$\begin{aligned} \Pi_j &= (r_\ell + p_j - mc_j) D_j(p,X) \\ &= (r_\ell + p_j - mc_j) \bar{D} s_j(p,X), \end{aligned} \tag{23}$$

where, D_j is demand for savings, $\bar{D}$ is total market size, mc_j is firm j's cost of servicing savings, and s_j is its market share given price and characteristics (X) of all products.

The first order condition yields the following pricing equation

$$p_j = mc_j - r_\ell - s_j(p,X)\Big(\frac{\partial s_j(p,X)}{\partial p_j}\Big)^{-1}, \tag{24}$$

that characterizes price as a markup over marginal cost that depends on the own price elasticity.

It might seem that, from the perspective of investors, the different savings products in each funding market are quite similar in terms of their financial characteristics. For instance, in Hortaçsu and Syverson (2004), a large number of managers compete for investors by offering mutual funds that each track the S&P 500. Similarly, banks offer savings accounts, which are financially homogeneous. This similarity of products would imply that funding markets should be quite competitive, with little or no price dispersion and markups approaching zero. However, this is not what we observe in practice. According to Hortaçsu and Syverson (2004), in the year 2000 the highest-price S&P index fund posted fees that were roughly 30 times greater than the fees associated with the lowest-cost fund. Similarly, Egan et al. (2017) find that the mean returns banks earn on deposits exceed the interest payment by 3.5 percent–6.5 percent.[18] Moreover, a number of studies have shown that changes in concentration affect deposit rates, implying some degree of market power (see for instance Berger and Hannan (1989) and Prager and Hannan (1998)).

To explain these observations, the literature studying retail funding markets has considered three main explanations. The first is information frictions. Savers may not be aware of all products on offer, or at least what are their list prices, and so may need to search for this information. The second is product differentiation. Finally, there may exist important costs for

[18] See also Xiao (2019) who estimates a median markup of 1.37% for US commercial banks.

switching from one financial institution to another, and these may also generate market power. It should be pointed out that market power might also arise were managers to act jointly, either explicitly or tacitly, in a coordinated manner to increase their collective profits. There is little evidence of managers acting in such a manner and so we will only briefly describe work that has investigated this additional possible source of market power.

4.2 Sources of market power

The starting point for our discussion of market power in funding markets is the model of Hortaçsu and Syverson (2004), developed to examine competition for savers by mutual funds managers. We begin here because in this setting there is strong reason to believe that all three main sources of market power (differentiation, search frictions, and switching costs) interact, and Hortaçsu and Syverson consider the possible role played by each one. Their approach is to embed a discrete choice framework within a search model.

A saver choosing mutual fund j gets utility

$$u_j = X_j\beta - p_j + \xi_j, \tag{25}$$

where X_j are attributes of fund j, p_j is its price, and ξ_j represents an unobservable component. Note that the coefficient on price is normalized to -1, such that utilities are expressed in terms of basis points. In this setup, products are vertically differentiated, but a nondegenerate market share distribution still arises because search cost variation across savers creates a type of horizontal differentiation.

In order to learn the indirect utility of a particular fund (other than the first they visit) saver i must pay a cost c_i. These search costs are heterogeneous and drawn from distribution $G(c)$. Search is sequential in the sense that a saver will continue searching across funds as long as

$$\mathrm{c}_i \leq \int_{u^*}^{\bar{u}} (u - u^*) dH(u), \tag{26}$$

where u^* is the highest utility found so far, and $\bar{u}$ is the upper bound of $H(u)$, the distribution of funds' indirect utilities. Hortaçsu and Syverson label the N funds by ascending indirect utility order $u_1 < ... < u_N$ and make the simplifying assumption that the empirical distribution of indirect utilities is observed by savers, such that

$$H(u) = \frac{1}{N}\sum_{j=1}^{N} I[u_j \leq u]. \tag{27}$$

The optimal search rule generates thresholds in the search cost distribution given by

$$c_j = \sum_{k=j}^{N} \rho_k (u_k - u_j). \tag{28}$$

The right hand side is the expected benefit of an additional search for a saver who has already identified fund j: with probability ρ_k sampling yields a higher-utility fund.

Using these cutoffs, Hortaçsu and Syverson write down expressions for market shares. The market share of the lowest utility fund (u_1) is given by:

$$s_1 = \rho_1(1 - G(c_1)) = \rho_1 \left(1 - G\left(\sum_{k=1}^{N} \rho_k (u_k - u_1)\right)\right). \tag{29}$$

Since the expected benefit from continuing to search is large for savers with fund 1, only those who happen to draw this fund first (which occurs with probability ρ_1) and with very high search costs (i.e. those with $c > c_1$) will end up selecting this fund.

More generally the market share of fund j for funds 2 through N is given by:

$$\begin{aligned} s_j &= \rho_j[1+\frac{\rho_1 G(c_1)}{1-\rho_1}+\sum_{k=2}^{j-1}\frac{\rho_k G(c_k)}{(1-\rho_1-...-\rho_{k-1})(1-\rho_1-...-\rho_k)} \\ &\quad -\frac{G(c_j)}{1-\rho_1-...-\rho_{j-1})}]. \end{aligned} \tag{30}$$

These equations map observed market shares to the population fractions with search costs less than the critical values.

Consistent with the discussion above, on the supply side, Hortaçsu and Syverson assume that funds select prices to maximize profits given by equation (23). Profit maximization implies the standard first-order condition for p_j:

$$s_j(p,X)+(p_j-mc_j)\frac{\partial s_j(p,X)}{\partial p_j}=0 \tag{31}$$

with

$$\begin{aligned} \frac{\partial s_j}{\partial p_j} &= -\frac{\rho_1\rho_j^2 g(c_1)}{1-\rho_1} \\ &- \sum_{k=2}^{j-1}\frac{\rho_k\rho_j^2 g(c_k)}{(1-\rho_1-...-\rho_{k-1})(1-\rho_1-...-\rho_k)} \\ &- \frac{p_j(\sum_{k=j+1}^{N}\rho_k)g(c_j)}{(1-\rho_1-...-\rho_{j-1})}. \end{aligned} \tag{32}$$

From the first order condition, with data on market shares and prices and some knowledge of marginal costs, these derivatives can be computed and form a system of equations that can be used to recover the values of $g(c)$ at the critical values $c_1,...,c_{N-1}$. With these in hand, the search cost distribution can be recovered using a normalization for $g(c_N)$ and the fact that the difference between the CDF evaluated at c_{j-1} and c_j can be approximated using the trapezoid method:

$$G(c_{j-1})-G(c_j)=0.5[g(c_{j-1})+g(c_j)](c_{j-1}-c_j). \tag{33}$$

Lastly, from equation (28), the critical values (the c_k's) can be used to calculate u_j. Hortaçsu and Syverson then estimate the attribute loadings (the β's) with the following

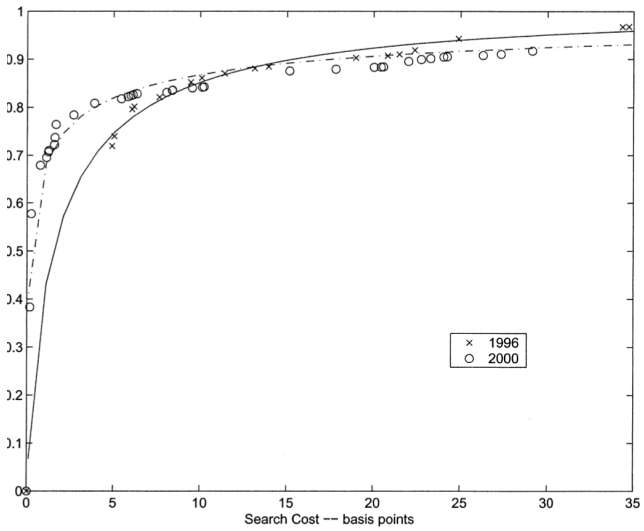
$$u_j+p_j=X_j\beta+\beta_{age}\ln(age_j)+\eta_j, \tag{34}$$

where X_j are observed fund characteristics other than age, such as load, number of family funds, and manager experience, and η_j is a fund-specific error term. Note that η_j includes ξ_j, which may be correlated with fund age, and so Hortaçsu and Syverson estimate equation (34) using and IV approach. Following Berry et al. (1995) they use current-year own-fund attributes and current-year summary measures of funds from other sectors as their instruments.

They estimate the model on S&P 500 index funds using data from 1995 to 2000. Their empirical analysis highlights the importance of differentiation. A simple model in which funds are assumed to be homogeneous (i.e. with search alone) is rejected by the data. Instead a model with both search and differentiation is best able to rationalize the data.

Search costs and differentiation: Together, search frictions and differentiation provide fund managers with the ability to charge higher prices. Differentiation implies that funds can raise their prices and some consumers are willing to pay these higher prices to buy funds with

Figure 6. Estimated search cost distributions, 1996 and 2000



Source: Figure 4 in Hortaçsu and Syverson (2004) (page 435).

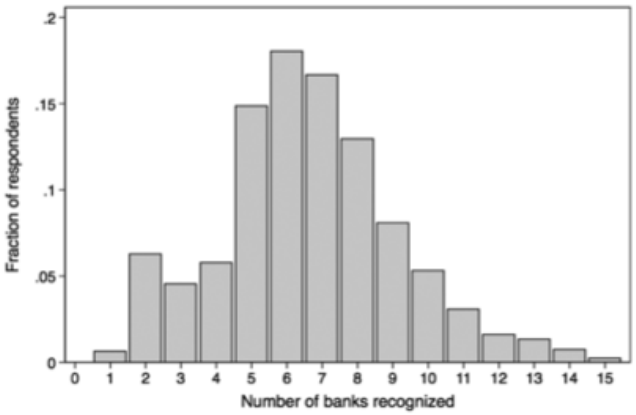
particular characteristics. Similarly, search costs make it difficult for investors to identify lower-priced funds, providing all funds with the opportunity to raise prices. Hortaçsu and Syverson find that investors value the different funds' non-financial characteristics. In particular, they value fund age, the total number of funds in the fund family, and tax exposure. Controlling for differences in these attributes, the observed price dispersion can be rationalized with minimal but heterogeneous search costs. They also find that search costs diverge over time as financially inexperienced investors enter the market towards the end of the 1990s. The high end of the search cost distribution witnessed an increase in costs, while for the remainder of the distribution, search costs declined. This can be seen in Figure 6, which plots the CDFs of the implied search cost distributions for 1996 and 2000.

Switching costs: Hortaçsu and Syverson also consider the possibility that market power and price dispersion in this industry may stem from costs assumed when investors switch assets from one family of mutual funds to another. They note that these costs could either be formal or informal. Formal switching costs could arise for instance from deferred or rear loads, since these imply explicit costs to removing assets from the fund. Informal switching costs, on the other hand, can be equated with the hassle costs of moving funds around. In either case, incumbent fund managers provide higher value and thereby create a form of *lock in* to a particular fund manager, which can lead to both market power and price dispersion.

Since lock-in could also be the result of search frictions, separate identification of search and switching costs is not obvious. Without investor level data, Hortaçsu and Syverson cannot observe any variation in investors' information sets, and so separate identification of search and switching costs is not possible. Nonetheless, in their model Hortaçsu and Syverson account for switching costs by allowing investors' fund decisions to depend on the size of funds' management companies and by letting investors react to the presence of a rear/deferred load. They find a positive effect of rear/deferred load, suggesting that these sorts of funds have larger market shares than would be implied by their other characteristics. Hortaçsu and Syverson attribute this to the presence of switching costs. They also point out that their estimates of search frictions may be picking up some costs stemming from switching.

Hortaçsu and Syverson also provide reduced-form empirical evidence on the importance

Figure 7. Size of Awareness Sets



Source: Figure 1a in Honka et al. (2017) (page 619).

of switching costs. They study the extent to which asset flows into S&P 500 funds respond to the performance of non-S&P funds from the same fund family. They hypothesize that spillover effects should exist, but that they would be damped within load families if there are important switching costs. Their analysis suggests that there are indeed spillover effects, with non-S&P fund performance affecting flows into S&P 500 funds of the same family. However, and most importantly, they find no difference in response for load fund families and no-load fund families, suggesting that switching costs are not sizeable in this context. In other settings switching costs may be more important. We return to this discussion to below.

Awareness and consideration sets: In Hortaçsu and Syverson (2004), some investors do not consider all products when making their discrete choice because search frictions prevent them from investigating all options. An alternative explanation as to why some investors do not consider all products is that they may not be aware of some of them. Marketing activities might help to determine which products investors are aware of, or possibly influence the set they consider more carefully. Honka et al. (2017) consider these possibilities. Unlike Hortaçsu and Syverson (2004), Honka et al. (2017) have access to individual-level survey data from a market research company that allow them to determine which banks consumers are aware of and which they consider when trying to decide where to deposit their funds. According to their data investors are on average aware of just 6.8 banks and there is important heterogeneity in this measure (see Figure 7). The authors develop a model of investor demand for depository institution that contains, not only the purchase stage, but also the awareness and consideration stages. In their setup, investor i is assumed to always be aware of their home bank, but their awareness of any of the other banks j, denoted a_{ij}, is probabilistic and assumed to depend, among other things, on advertising and branch presence. More specifically they model a_{ij} as

$$a_{ij} = \zeta_{0j} + \zeta_{1j} adv_j + f(b_{ij})\zeta_{2j} + D_i\zeta_{3j} + \varepsilon^a_{ij}, \tag{35}$$

where adv_j denotes bank j's advertising intensity, D_i represents demographic variables, and b_{ij} captures local bank presence. The error term ε^a_{ij} is assumed to follow an extreme value Type I distribution, allowing the authors to estimate a binary logit regression for each bank j. The probability that investor i is aware of j is then

$$\Pr(a_{ij} > 0) = \frac{exp(\zeta_{0j} + \zeta_{1j} adv_j + f(b_{ij})\zeta_{2j} + D_i\zeta_{3j})}{1 + exp(\zeta_{0j} + \zeta_{1j} adv_j + f(b_{ij})\zeta_{2j} + D_i\zeta_{3j})}, \tag{36}$$

and the probability that they are aware of a particular set of banks, A_i, is given by

$$P_{iA_i} = \prod_{j=1}^{J} \Pr(a_{ij} > 0)^{\phi_{ij}} (1 - \Pr(a_{ij} > 0))^{\phi_{ij}}. \tag{37}$$

From this the authors can write the unconditional choice probability as

$$P_{ij} = P_{iA_i} \cdot P_{iS_i|A_i} \cdot P_{ij|S_i}, \tag{38}$$

where $P_{iS_i|A_i}$ is the probability that saver i searches set S_i given awareness set A_i, and $P_{ij|S_i}$ is the probability of choosing product j given consideration set S_i.

Their findings suggest that branch presence plays an important role for awareness as does advertising. If advertising increases by 1%, the probability of awareness increases by 0.1%. In their model advertising can also influence choices directly, but their findings suggest that it matters less at this stage. The same 1% increase in advertising leads only to a 0.02% increase in demand at the choice stage of their model. In other words, in their setting, advertising has a five times stronger effect on awareness than it does on choice conditional on awareness.

In related work, Roussanov et al. (2020) study the mutual fund market and extend Hortaçsu and Syverson (2004) to allow funds to influence the likelihood they are considered by investors through costly marketing activities. Similarly, Egan (2019) studies the convertible bond market, where, after controlling for a small number of observable features, there is a clear ranking in terms of which bonds are better from an investor's perspective. Despite this, investors are often found to purchase dominated products. To explain this observation Egan considers the role of broker intermediaries who direct investor search towards products that yield them (the broker) higher expected profit, even if these are dominated for investor's. Egan's data set includes information on broker compensation (kick-backs).

Hastings et al. (2017) also consider the role of marketing activities in the context of the retirement savings plan market in Mexico. In their setting, retirement plans hire agents to market their product and enrol savers. Hastings et al. (2017) model utility as depending on the number of sales agents employed by a plan. They point out that this simple model nests the two roles described above for advertising, to generate awareness/consideration or to act at the choice stage by persuading savers that the product in question is somehow more preferred. This point is relevant for the discussion that follows in which product attributes such as branch and ATM networks are treated as utility shifters, when they might also be making a bank or fund manager more salient, thereby affecting awareness.

4.2.1 Measuring product differentiation

As mentioned, Hortaçsu and Syverson (2004) find that a number of non-financial characteristics play an important role in determining a fund's demand, confirming the relevance of differentiation for market power. In this section we investigate further the contribution of differentiation to market power, focusing especially on heterogeneity in service quality. Hortaçsu and Syverson (2004)'s model is one of vertical differentiation in which all investors share a common utility function, and horizontal differentiation is built in via heterogeneity in search costs. In contrast, much of the literature studying demand and pricing for retail funding products has introduced differentiation by incorporating horizontal taste differences via an iid random utility error term. Following McFadden (1974), Berry (1994), and Berry et al. (1995) the standard assumption is that these utility shocks are distributed i.i.d. Type 1 extreme value, leading to standard logit market shares. According to this setup demand for product j will depend on prices and both observed and unobserved characteristics. Investors choose the product that yields the highest indirect utility, with indirect utility for investor i from choosing product j given by:

$$u_{ij} = X_j\beta - \alpha p_j + \xi_j + \varepsilon_{ij}, \tag{39}$$

for $j = 0, ..., J$, where $j = 0$ is the outside good (i.e. u_{i0} is the utility from not investing in one of the products). Recall from above that in the case of retail deposit markets, price (deposit rates, r_d) will instead typically enter positively, since banks offer savers a positive rate of return on their savings. The term ξ_j captures the unobserved quality of product j, and ε_{ij} is the iid logit shock. The extreme value assumption allows for the predicted market share for product j to be expressed as follows

$$s_j(\mathbf{x}, \mathbf{p} \mid \theta) = \frac{exp(X_j\beta - \alpha p_j + \xi_j)}{\sum_{k=0}^{J} exp(X_k\beta - \alpha p_k + \xi_k)}. \tag{40}$$

This leads to the following estimating equation

$$\ln s_j - \ln s_0 = X_j\beta - \alpha p_j + \xi_j, \tag{41}$$

which can be estimated using data on prices and product characteristics, along with aggregate market share data.

Using this method, a number of papers have been written to examine the importance of differentiation in service quality for market power and other outcomes. Heterogeneity in service quality enters through the X's in the above equations.

To our knowledge, the earliest application of the logit approach to study differentiation in financial markets is Dick (2008), who used it to study the demand for deposit services and the impact of the changes in the structure of the US banking market following the implementation of the Riegle-Neal Act. Investors are assumed to choose from the set of banks operating in their local market (defined to be a Metropolitan Statistical Area). Individual investors are assumed to be endowed with some amount of deposits, and to make a discrete choice of depository institution in which to place their savings. Dick includes a number of bank characteristics such as age and size, the extent to which it is geographically diversified, and the density of its branch network. Data are from the Summary of Deposits, collected by the Federal Deposit Insurance Corporation (FDIC) each year and the Report on Condition and Income (the so called *Call Reports*) from the Federal Reserve Board. The Call Reports provide information that allow her to determine the average deposit rate offered at the bank, along with the account fees (service charges). As mentioned above, for deposit markets, p_j is negative, since managers must offer savers a positive deposit rate to ensure a positive return on their savings.

In addition to the simple logit model described above, Dick uses a nested logit approach in which she considers banks operating in multiple states, and single-state banks as being separate nests:

$$u_{ij} = X_j\beta - \alpha p_j + \xi_j + \gamma_{ig} + (1 - \sigma)\varepsilon_{ij}, \tag{42}$$

where γ_{ig} represents a shock common to all banks in group g, with a distribution function that depends on $\sigma \in [0, 1)$. The estimating equation is then

$$\ln s_j - \ln s_0 = X_j\beta - \alpha p_j + \xi_j + \sigma \ln(\bar{s}_{j|g}). \tag{43}$$

Dick finds that investors respond to both deposit rates, and to account fees. The median service-fee elasticity is between 0.3 and 0.4 and the median deposit-rate elasticity is between 1.8 and 3.0. She also finds that consumers value staffing and geographic density of local branches, as well as the age, size, and geographic diversification of banks. She finds that a 1% increase in branch density in an urban area would lead to a 0.1-0.2% increase in market share.[19]

[19] Aguirregabiria et al. (2016) find similar effects in a model of in which banks compete a la Nash-Cournot by choosing for each market in which they operate a deposit level to maximize profits. Their findings suggest that the volume of deposits of a bank increases by 22% when the number of branches goes from one to two, by 7.9% when the number of branches goes from nine to ten, and by just 2.2% when adding anything about twenty branches. See also Aguirregabiria et al. (2019)

Complementary work by Ishii (2007) and Ho and Ishii (2011) takes more seriously the role of branch networks by geocoding the branch networks of banks (assigning to each a latitude and longitude). Unlike Dick (2008) who includes average bank branch density as a utility shifter, the demand models in Ishii (2007) and Ho and Ishii (2011) include the distance from each consumer's home to the closest and second-closest branches of each bank. More specifically, they incorporate census data that provides them with *block*-level information on income, population, and location to calculate distances from particular census tracts to branches. These two papers also consider a richer specification of utility that incorporates heterogeneity in consumer tastes for bank characteristics, following Berry et al. (1995). Specifically, they allow investors' deposit levels to affect their preferences for deposit interest rates. Their estimates suggest a mean deposit-rate elasticity of 1.19.[20,21]

In addition to branch networks, investors value ATM networks. Travel costs must be incurred when using an ATM and so when choosing a bank, one attribute investors care about is its ATM network. Furthermore, there is an indirect network effect stemming from the ATM network of rival banks, when these are compatible with the network of the chosen bank, since investors may sometimes need to use these *foreign* ATMs to access their deposits.[22] To capture this, the demand models estimated in papers studying the role of ATMs for bank choice, such as Ishii (2007) and Gowrisankaran and Krainer (2011), specify utility functions that incorporate the number or density of ATM networks or the distance from consumer to ATM. Gowrisankaran and Krainer model demand for withdrawals and so also incorporate an additional *surcharge* that consumers must pay in the event they elect to make a withdrawal at a foreign ATM. Gowrisankaran and Krainer find that a consumer who used a particular ATM with probability 50%, would use it with probability 46% if it were located 1km further away. Ishii finds that the effect of a one-standard deviation increase in the interest rate evaluated at the average income in her sample is equivalent to an increase in the number of ATMs of 8.5 standard deviations.[23]

In related work, Adams et al. (2007) develop a model that allows for differentiation between single- and multi-market banks and also between banks and thrifts. Their results suggest that there is only limited substitutability between single- and multi-market banks and between banks and thrifts, suggesting that even if these institutions operate in proximity, differentiation dampens competition.

Entry models

Similar results to those in Adams et al. (2007) were obtained by Cohen and Mazzeo (2007) using an alternative approach for assessing the competitiveness of local markets. The authors base their analysis on the entry model approach pioneered by Bresnahan and Reiss (1991) that makes inferences about the degree of competition based on the observed structure across a number of markets. These models all employ a revealed preference approach whereby the assumption is that if a firm is observed to be operating in a market, it is because it is profitable to do so.

The profits for a firm operating in market m are given by:

$$\begin{aligned} \pi_m(n) &= M_m VP_m(n) - FC_m(n) \\ &= M_m(X_m\beta - \alpha(n)) - (W_m\gamma + \delta(n) + \varepsilon_m), \end{aligned} \tag{44}$$

[20] Wang and Ching (2019) use a similar approach, but incorporate workflow data to take into account the possibility that investors care not just about branches located close their place of residence, but also their place of work.

[21] Xiao (2019) also studies the role of branch networks. He models banks as being differentiated in terms of their transaction convenience and yields. His focus is on the rise of shadow banks who compete on yields, but offer inferior transaction convenience to traditional banks, since they do not operate a branch network.

[22] See Saloner and Shepard (1995) for a discussion of the network effect in the case of ATMs.

[23] These papers assume that savers already have an ATM card, but there is also a literature studying the saver's adoption decision. See for instance Yang and Ching (2014) and Huynh et al. (2017).

where $\alpha(n)$ and $\delta(n)$ are parameters capturing the competitive effect of having more firms in the market, and X_m and W_m are demand and cost side characteristics. The error term ε is assumed to be independent of M_m, X_m, W_m and iid over markets with distribution $N(0,1)$. Then, since firms are active only if they earn nonnegative profits, the probability of observing exactly n firms in equilibrium is given by

$$\begin{aligned} \Pr(\pi_n \geq 0 \text{ and } \pi_{n+1} < 0) &= \Phi[M_m(X_m\beta - \alpha(n)) - (W_m\gamma + \delta(n))] \\ &\quad -\Phi[M_m(X_m\beta - \alpha(n+1)) - (W_m\gamma + \delta(n+1))], \end{aligned} \tag{45}$$

which can be estimated by maximum likelihood.

Mazzeo (2002) extends this method to allow firms to choose not just which market to enter, but also the type of product they enter with. Cohen and Mazzeo (2007) use this model to study entry of depository institutions as either multimarket banks, single market banks, or thrifts. Profits of each type of institution are allowed to depend differently on the number of competitors of each type present in the market. In other words, competition can be less intense between institutions of each type if investors view these products as being differentiated. This provides banks with the incentive to enter the market with a differentiated product, and helps to explain why multimarket banks can coexist with single market banks and thrifts. It also implies more choice for consumers. Counterfactual results suggest that in their setting there were 24% more depository institutions operating in the markets they consider than if only multimarket banks had operated. In other words, differentiation generates more options for investors.

Mazzeo's method is intractible when there are many types of firms. Seim (2006) proposes an approach that also allows for product differentiation, but that nonetheless makes estimation with multiple types feasible by incorporating private information about the profitability of entering a market. Gowrisankaran and Krainer (2011) use a similar approach to model the decision to open an ATM. Adding an ATM allows a bank to reduce its costs (since ATMs are cheaper than tellers) and to collect interchange fees and surcharges from non-depositors. They assume that banks obtain realizations of their cost chocks for all of their branches and then simultaneously decide at which of their branches to install an ATM.

The approach used in Seim and Gowrisankaran and Krainer considers the problem of a firm deciding on where to put a store/ATM. In contrast, we might want to think about how to determine the entire network or chain. This necessitates relaxing the assumption of independence across markets. This is the challenge faced by Aguirregabiria et al. (2016) who study a bank's decision of where to operate its branches, noting that this decision is similar to a portfolio choice between risky assets, where the risky assets are different geographic local markets. When choosing the network banks take into account not only the expected profits, but also the associated geographic risk. More specifically they incorporate into the banks profit function a cost of liquidity shortage that is affected by the deposit risk and return generated by the chosen network configuration.

Estimating models of network formation is complicated due to the dimensionality of the problem. Jia (2006) solves the problem for the two player case by using a solution method that exploits the lattice structure of the two player problem. The two-player restriction make the application of Jia's approach unappealing in the context of funding markets. Ellickson et al (2013) propose an alternative approach that is based on the idea that the profits stemming from the observed location decisions of a firm should be greater than those from any alternative configuration. This is a revealed preference approach in the spirit of moment inequality estimation as in Pakes (2010) and Pakes et al. (2015).

This is the approach taken in Aguirregabiria et al. (2016) to study whether banks take into account financial stability when establishing their branch network. We discuss this paper in further detail in Section 4.3 below. Ishii (2007) uses a similar approach in her study of ATM networks. To model the network formation stage Ishii (2007) adopts the approach developed in

Pakes et al. (2015), developing a two-stage model in which in a first stage simultaneous move game banks choose the size of their ATM networks given expectations about their rivals' choices. Then in a second stage, given this network, they compete on prices for deposits in the manner described above. She finds that demand-stealing effects in the deposit market associated with ATM networks and surcharging seem to be large enough to cause overinvestment in ATMs. This finding is similar to Massoud and Bernhardt (2002) who show from a theory standpoint that banks choose ATM networks that are socially excessive. In related work, Ferrari et al. (2010) study adoption of ATMs in Belgium. Unlike in the US, in Europe banks jointly owned the ATM network and so coordinated their investment in its expansion. The authors build an entry model along the lines of those discussed in this section, but allow for compatibility and coordinated investment. Without the strategic considerations that arise when firms develop their networks independently, the authors find evidence of underinvestment in ATM network coverage, in stark contrast to the findings in Ishii (2007) .

4.2.2 Measuring switching costs

As mentioned above, Hortaçsu and Syverson (2004) assume that the main friction leading to price dispersion and market power is search costs, but they acknowledge that their estimates of search costs may to some extent be capturing a switching cost effect. Given the nature of their data, they cannot separately identify the role of the two frictions, and so they focus on the one they deem more important. In other settings, the role of switching costs is more first order. For instance, for deposit markets Kiser (2002) suggests that switching costs are quite high. Switching costs could arise because the old account must be closed and a new one opened, which could involve significant time costs for instance to set up direct deposit. Switching costs can provide managers with market power. If we assume that savers have existing relationships with a manager and have to pay a one-time utility cost to change providers, then managers may be able to charge higher prices. The extent to which they can do so may depend on the size of the locked-in population versus the size of the population of new unattached customers. These two types of consumers can generate incentives for a manager to employ an *invest and then harvest* pricing strategy, where new consumers are charged lower prices than are locked-in customers.

Kiser uses survey data for the US and shows that average tenure is about ten years. According to the survey, of the consumers with accounts thirty-two percent reported that they had never switched banks. The decision to switch was most often the result of a move or job change, but also depended on prices, the quality of the service being offered, and income. Limited switching has also been noted in pension systems. Luco (2019), Illanes (2017), and Krasnokutskaya et al. (2018) all study the Chilean pension fund administration market where a market-based private system was established in the 1980s and for which detailed administrative data are available from the Chilean Pension Superintendency. Luco (2019) notes that despite that fact that switchers paid less for the same returns than those who did not switch, only 44% of enrollees switched funds between 1988 and 2001. Similarly, Illanes (2017) points out that despite the fact that loads, defined as the ratio between commission rates paid to the plan administrator and the amount of money remitted to the system, ranging from 10% to 20%, only 0.31% of customers change administrators per month.

A number of papers have tried to quantify switching costs in both deposit and pension markets. The utility function would now include a term reflecting the incumbent bank/pension plan:

$$u_{jt} = X_{jt}\beta - p_{jt} + \gamma \mathbb{1}[d_{t-1} \neq j] + \xi_{jt}, \tag{46}$$

where d_{t-1} represents last period's plan choice and γ the cost of switching to fund j today. In other words, last period's purchases have a causal effect on today's utility. This is the approach

taken in Luco (2019). His results show that eliminating all switching costs intensifies competition and decreases equilibrium fees by 58 percent relative to the case in which enrollees face all switching costs. Luco assumes that enrollees are myopic in the sense that they do not anticipate future changes in prices and make predictions regarding future returns based on current returns. He justifies this by pointing out that there was significant volatility in the Chilean pension system during his sample period such that it would have difficult for enrollees to form expectations regarding the future.

In contrast, Illanes (2017) analyzes a later more stable period and uses a dynamic setup in which in every period individuals select a pension fund administrator to maximize the sum of the present discounted value of flow utilities over time and the expected present discounted value of their retirement balance. Ho (2015) also uses a dynamic model, in his case to study the impact of switching costs on market power in the Chinese deposit market.[24] The important rural-urban migration experience in recent years in China meant that there was an explosion of new depositors. While existing customers are locked in, new customers are not, leading to an invest and then harvest pricing motive. Ho develops a dynamic model of consumer demand. Existing customers choose which bank to use and when to switch, while incoming customers simply choose in which bank to invest their deposits.

In the dynamic problem, consumers must decide to stay or switch managers depending on the switching cost, idiosyncratic shocks, and the current values of the product attributes, along with their expectation for the future evolution of these. Suppose that there are N competing managers. The flow utility of being with bank j is given by

$$u_{jt} = \underbrace{X_{jt}\beta - p_{jt} + \xi_{jt}}_{\delta_{jt}} + \varepsilon_{ijt}. \tag{47}$$

Furthermore, let $\varepsilon_{i\cdot t} \equiv (\varepsilon_{i1t}, ..., \varepsilon_{iJt}, \varepsilon_{iot})$, where o is the outside option. Then the value function for saver i currently with manager j is given by

$$\begin{aligned} V_i(\varepsilon_{i\cdot t}, j, \Omega_t) \quad = \quad & \max(\delta_{jt} + \varepsilon_{ijt} + \beta EV_i(\varepsilon_{i\cdot t+1}, j, \Omega_{t+1}|\Omega_t), ..., \delta_{Jt} + \varepsilon_{ijt} \\ & + \beta EV_i(\varepsilon_{i\cdot t+1}, J, \Omega_{t+1}|\Omega_t) - \gamma, \varepsilon_{iot} + \beta EV_i(\varepsilon_{i\cdot t+1}, O, \Omega_{t+1}|\Omega_t) - \gamma). \end{aligned} \tag{48}$$

In words, saver i does not have to pay a switching cost γ if it chooses manager j in period t, but does for any other choice of manager (or the outside option).

To estimate switching costs it can be noted that the fact that savers must incur costs to change managers implies that choices should be forward-looking, such that the problem is akin to the decision to purchase a durable good. Gowrisankaran and Rysman (2012) provide a structural model of dynamic demand for durable goods that can be employed to quantify switching costs. This is the approach used by Ho (2015). Using this methodology and data from four state commercial banks covering the 1994 to 2001 period, he estimates switching cost equal to roughly 0.8% of deposit value. Illanes (2017) proposes an alternative method based on the moment inequalities approach of Pakes et al. (2015). He estimates switching costs on the order of $1200, which is roughly equivalent to the present discounted value differences in commissions paid across firms.

4.2.3 Other sources of market power

As mentioned in the introduction to this section, an additional reason that market power can arise is that firms coordinate their actions. Many retail markets have features that facilitate collusion.

[24] See also Shy (2002) for an approach for estimating switching costs in deposit markets.

In particular, in these markets managers (i) interact repeatedly, (ii) come into contact in multiple markets, and (iii) at least in some funding markets, pricing is transparent.

Mólnar et al. (2002) consider this possibility in the context of the the Italian deposit market. They estimate both the demand side and also the supply side in an effort to examine market conduct. Specifically, they investigate whether multi-market conduct by banks facilitates collusion and allows them to markup prices over marginal cost. On the demand side they estimate a nested-logit model similar to Dick's. On the supply side they allow banks to choose deposit rates and branch networks for every region in each period. Similar to what we specified in equation (23) above the profit function for firm j in market m is given by

$$\Pi_{jm} = (r_{\ell jm} + p_{jm} - mc_j)\bar{D}_m s_{jm}(p,X) + \sum_{k \neq j} \lambda_{jk}(r_{\ell km} + p_{km} - mc_k)\bar{D}_m s_{km}(p,X), \tag{49}$$

where, unlike above, firm j's profits now also depend on the values of λ_{jk}, which capture different models of competition. The first order condition is given by

$$\bar{D}_m s_{jm}(p,X) + \sum_{k \neq j} \lambda_{jk}(r_{\ell km} + p_{km} - mc_k)\bar{D}_m \frac{\partial s_{km}(p,X)}{\partial p_{jm}} = 0. \tag{50}$$

Following Bresnahan (1987) they estimate parameters for a number of different models of bank conduct and then test the fit of each of these using non-nested tests (see Rivers and Vuong (2002)) to select among possible conducts. They consider not only the two extremes, a competitive and a perfectly collusive model of the supply, but also partially collusive models in which the cartel depends on market overlap.

Their findings suggest that the extreme models of conduct, competition and perfect collusion, are rejected, as is differentiated product Bertrand. These conducts are rejected in favor of coalitions involving subsets of banks based on their contact in multiple markets.

4.3 Market power in retail funding markets and financial stability

Competition in funding markets has important implications for financial stability. Egan et al. (2017) examine the retail banking sector and examine the influence of markups for stability. The authors study demand for insured vs uninsured deposits and incorporate banks' financial distress into the standard differentiated-demand model. As in the models described above, investors' discrete choices over banks depend on the offered interest rates and various services. In addition, Egan et al. allow demand for uninsured depositors to also depend on the probability that the bank will default.[25] They use data on 16 of the largest retail banks in the US during the 2002-2013 period and measure financial distress with a bank's credit default swap spread.

The authors specify separate indirect utility functions for uninsured (N) and insured (I) depositors. For uninsured depositors

$$u^N_{ijt} = X_j\beta^N - \alpha^N p_{jt} - \rho_j\gamma + \xi^N_j + \varepsilon^N_{ijt}, \tag{51}$$

where ρ_j represents the probability of default and γ is the lost utility flow experienced by uninsured depositors. For insured depositors indirect utility is given by

$$u^I_{ijt} = X_j\beta^I - \alpha^I p_{jt} + \xi^I_j + \varepsilon^I_{ijt}. \tag{52}$$

[25] Financial-institution stability is also crucial in the context of variable annuities offered by life insurers. As pointed out in Koijen and Yogo (2021), life insurers have stepped in to fill the void left as a result of the decline of the defined pension plan by offering products providing minimum return guarantees over long horizons. The extent to which withdrawals are guaranteed depends on the ability of the issuer to pay claims. Koijen and Yogo (2021) develop a model in which insurers compete for savers by setting fees and rollup rates. The latter are guaranteed returns on investment in the years before the saver begins taking withdrawals. To capture reputation in the retail market and allow it to vary across insurers and over time Koijen and Yogo (2021) include the A.M. Best rating and insurer fixed effects. Investors who are concerned about the stability of one issuer can substitute towards another with a better rating, and if they are worried about the stability of the sector as a whole, they can substitute to *outside* mutual funds.

Banks take into account the preferences of uninsured depositors for safety when setting interest rates. The bank profit function specified above in equation (23) is modified to reflect the fact that profits can come from the two sorts of depositors.

$$\Pi_{jt} = (r_{\ell jt} + p^I_{jt} - mc_j)\bar{D}^I s^I_{jt}(p,X) + (r_{\ell jt} + p^N_{jt})\bar{D}^N s^N_{jt}(p,X), \tag{53}$$

where in this case *mc* represents an additional cost of servicing insured depositors. Egan et al. (2017) assume that banks are financed through deposits, equity, and bond coupons b_j. Returns $r_{\ell j}$ are assumed to be stochastic and distributed $N(\mu_j, \sigma_j)$ and if they are sufficiently low such that in any given year $\Pi_j - b_k < 0$, equity holders can inject funds to repay deposits and the bond coupon or can decide to default. Equity holders will keep the bank afloat so long as next-period franchise value exceeds the size of the shortfall that would have to be financed:

$$\Pi_{jt} - b_j + \frac{1}{1+\iota} E_j > 0, \tag{54}$$

where E_j is the franchise value of bank j. This yields a cutoff strategy for equity holders that depends on returns: if $r_{\ell jt}$ is less than some cutoff $\bar{r}_{\ell j}$, then they will not inject funds into the bank. In equilibrium $\bar{r}_{\ell j}$ corresponds to the risk-neutral probability of default $\rho_{jt} = \Phi\big(\frac{r_{\ell j} - \mu_j}{\sigma_j}\big)$. Egan et al. (2017) specify an optimal cutoff rule that highlights a tradeoff between the amount of funds equity holders must inject and the future value of the bank, which depends on its deposits, its survival probability and its expected returns. Banks choose deposit rates to maximize the expected return to equity holders. Egan et al. (2017) get the following first order conditions

$$\mu_j + \sigma_j \lambda \big(\frac{\bar{r}_{\ell j} - \mu_k}{\sigma_j}\big) - (mc_j - p^I_{jt}) = \frac{1}{(1 - s^I(p^I_{jt}, \mathbf{p^I}_{-jt}))\alpha^I}, \tag{55}$$

for the insured deposit rate p^I, where the RHS is the mark-up and the LHS is the difference between the marginal benefit and marginal cost. For uninsured deposits they get

$$\mu_j + \sigma_j \lambda \big(\frac{\bar{r}_{\ell j} - \mu_k}{\sigma_j}\big) - (-p^N_{jt}) = \frac{1}{(1 - s^N(p^N_{jt}, \mathbf{p^N}_{-jt}, \rho_{jt}, \rho_{-j,t}))\alpha^N}, \tag{56}$$

Deposit pricing of the two products is different because of the additional marginal cost of insuring deposits, the different price elaticities of the two types of depositors, and the different shares amongst the two populations. In particular, demand for uninsured deposits depends, among other things, on the probability of default. Banks in financial distress must offer higher deposit rates in order to attract uninsured deposits.

Aguirregabiria et al. (2016) study whether banks consider stability when establishing their branch networks. The authors study a bank's decision of where to operate its branches, noting that this decision is similar to a portfolio choice between risky assets, where the risky assets are different geographic local markets. When choosing the network banks take into account not only the expected profits, but also the associated geographic risk. More specifically they incorporate into the banks profit function a cost of liquidity shortage that is affected by the deposit risk and return generated by the chosen network configuration.

As mentioned above, Aguirregabiria et al. (2016) adopt a revealed preference approach in the spirit of Ellickson et al. (2013) and based on Pakes (2010) and Pakes et al. (2015). They set up a static model of branch choice in which bank i decides on its branch network $\mathbf{n}_{it}$ in every period t to maximize expected value $E(V_{it}|\mathbf{X}_t)$, where $\mathbf{X}_t$ is a vector of variables containing all information relevant to banks available in period t, and specified as

$$E(V_{it}|\mathbf{X}_t) = -AC_{it}(\mathbf{n}_{it}, \mathbf{n}_{it-1}) + \frac{1}{1-b}[VP_{it}(\mathbf{n}_{it}) - FC_{it}(\mathbf{n}_{it}) - \lambda_{it} \Pr(D_{it} \leq L_i - E_i|\mathbf{X}_t)], \tag{57}$$

where b is the time discount factor (fixed at 0.95), $VP_{it}(\mathbf{n}_{it})$ is the bank's variable profits from all local markets where it is active, $FC_{it}(\mathbf{n}_{it})$ is the fixed cost of operating network $\mathbf{n}_{it}$, and $AC_{it}(\mathbf{n}_{it}, \mathbf{n}_{it-1})$ represents the costs of adjusting the network. The latter are allowed to depend on whether branch-network expansion occurs via merger or de novo branching. Finally, $\lambda_{it} \Pr(D_{it} \leq L_i - E_i | \mathbf{X}_t)$ represents the cost of liquidity shortage, with D_{it} capturing the total volume of deposits of bank i and L_i and E_i its volume of loans and equity respectively.

The moment inequalities are based on the idea that the expected value to a bank from its observed choice of network $\mathbf{n}_{it}$ must be at least as great as from any other feasible network. For estimation the authors specify the expected value of a bank as $E(V_{it}|\mathbf{X}_t) = W_{it}(\mathbf{n}_{it})\theta + \varepsilon(\mathbf{n}_{it})$, where $W_{it}(\mathbf{n}_{it})$ is the vector of known functions, θ is the vector of parameters, and $\varepsilon(\mathbf{n}_{it})$ captures factors unobservable to the researcher. Then the moment inequalities take the form

$$E\left(W_{it}(\mathbf{n}_{it})\frac{\theta^0}{\sigma_\varepsilon} + \frac{\varepsilon(\mathbf{n}_{it})}{\sigma_\varepsilon} - W_{it}(\mathbf{n})\frac{\theta^0}{\sigma_\varepsilon} - \frac{\varepsilon(\mathbf{n})}{\sigma_\varepsilon}|\mathbf{X}_t\right) \geq 0, \tag{58}$$

where $\mathbf{n}$ is any other feasible branch network for bank i, θ^0 is the true value of the vector of structural parameters, and σ_ε is the standard deviation of the unobservables.

An important challenge is that the total number of possible branch configurations (and therefore inequalities) is extremely large. Aguirregabiria et al. (2016) therefore consider a subset C_{it} of possible configurations and then follow Chernozukov, Hong, and Tamer (2007) to define the estimator for $\frac{\theta^0}{\sigma_\varepsilon}$ as

$$\hat{\theta} = \arg\min_{\tilde{\theta}} \sum_{h,n \in C_{it}} \left[\max\left\{-\sum_{i=1}^{I_t}\sum_{t=1}^{T} Z_{hit}[(W_{it}(\mathbf{n}_{it}) - W_{it}(\mathbf{n}))\tilde{\theta} + K]; 0\right\}\right]^2, \tag{59}$$

where K is an unidentified constant whose value must be fixed (Aguirregabiria et al. (2016) fix it at $K \in [4,6]$). Aguirregabiria et al. (2016) consider the following configurations to include in the set C_{it} the actual choice n_{it} and a large number of hypothetical branch networks that includes for instance opening (closing) up to five branches in the bank's headquarters-county (HQ). Since the objective of Aguirregabiria et al. (2016) is to study the importance of geographic risk diversification, they also consider configurations in which branches are opened in counties (i) neighboring the HQ county, (ii) with the lowest risk within (HQ) states, (ii) with the highest expected returns with the HQ's state, (iii) with the lowest risk with the HQ's state.

Their estimates imply that banks are concerned about geographic diversification of deposit risk when they select their network, but also take into account economies of density and merging costs.

We return to our discussion of financial stability in Section 6, when we talk about the impact of regulation on financial markets.

5 RETAIL CREDIT MARKETS

An important feature of credit markets is that contract terms reflect the risk and profitability of individual borrowers. Individualized pricing can take the form of detailed rate sheets, and/or bilateral negotiation between borrowers and loan officers. Furthermore, in many markets such as credit cards, lenders offer a fixed menu of loan contracts, but can target promotions and adjust credit limits to ensure that terms reflect observed differences in risk across borrowers.

This richness complicates the analysis, since transaction prices reflect observed and unobserved factors determining the demand and riskiness of each borrower. Moreover, since consumers do not select products and lenders from a common "menu," the distribution of rejected offers is unobserved to the econometrician.

In this section, we start by describing a simple discrete-choice model of demand and supply for loans that accounts for risk-based pricing and market power. We then describe specific applications, focusing in particular on market frictions arising from asymmetric information and search costs.

5.1 Modelling framework

Consider a credit market in which consumers search for loans of unit size. An offer from lender j is given by price p_{ij}. We use a discrete-choice framework to derive the demand function.[26] Depending on the context, the price can either measure the interest rate (r_l), monthly payments or upfront payments. The indirect utility associated with lender j depends on consumer i's ex-ante need for liquidity θ_i, as well as an idiosyncratic valuation capturing factors such as differentiation, past relationships/switching costs, and search costs, which we denote by $x_{ij}\beta + \xi_j + \varepsilon_{ij}$:

$$\text{Surplus}_{ij} = \begin{cases} \theta_i + x_{ij}\beta + \xi_j + \varepsilon_{ij} - p_{ij} & \text{if } j > 0, \\ \varepsilon_{i0} & \text{if } j = 0, \end{cases}$$

where $j = 0$ corresponds to an outside funding source (normalized to zero). The demand function depends on the distribution of willingness-to-pay, conditional on borrower characteristics x_i. Assuming that consumers search for a loan of unit size, the loan-demand function is given by the probability that a consumer of type x_i accepts an offer p_{ij}:

$$L_j(p_{ij}|x_i) = \Pr\left(p_{ij} < \underbrace{\theta_i + x_{ij}\beta + \xi_j + \varepsilon_{ij} - E\left[\max_{j \neq k} \{\varepsilon_{i0}, \theta_i + x_{ik}\beta + \xi_k + \varepsilon_{ik} - p_{ik}\} \,\middle|\, x_i\right]}_{v_{ij}} \,\middle|\, p_{ij}, x_i \right), \quad (60)$$

where v_{ij} corresponds to the willingness-to-pay of consumer i for lender j. The expectation operator accounts for the possibility that consumers might be uncertain about the value or availability of competing offers and must incur search or switching costs. Below, when the identity of the lender is not relevant for the discussion, we refer to v_i as the willingness to pay for the chosen bank, and p_i as the accepted offer. We use x_i to summarize the information that lenders have about each borrower when setting the price. Importantly, this vector can include characteristics that are unobserved to the econometrician, which leads to unobserved heterogeneity across borrowers.

The present value of profits from a loan is a function of credit risk, and the lending cost realization. Credit risk can be caused either by the decision of borrowers to default, or the decision to prepay the loan early. On the cost side, we distinguish between two types of expenses. The upfront cost of origination is denoted by c_{ij}, and includes marketing, labor and funding costs. In the case of securitized loans, c_{ij} also includes the resale value of the loan in the secondary market. The per-period cost w_{ij} corresponds to the cost of "servicing" the loan prior to maturity, which includes the coupon and fees paid to investors and insurers if the loan is securitized.

The present-value of of profits generated from a fixed-rate loan of unit size amortized over T periods is given by:

$$\pi_{ij}\left(p|\tilde{T}, P_1, \ldots, P_{\tilde{T}}\right) = \sum_{\tau=1}^{\tilde{T}} \delta^{\tau}\left(U_{\tau-1} - P_{\tau}\right)(p - w_{ij}) - c_{ij} = S \times (p - w_{ij}) - c_{ij},$$

where P_τ is the principal payment made in period τ, $U_\tau = U_{\tau-1} - P_\tau$ is the unpaid loan balance at the end of τ ($U_0 = 1$). The loan duration ($\tilde{T} \leq T$) and sequence of payments ($P_1, \ldots, P_{\tilde{T}}$)

[26] Similar predictions emerge from a model in which consumers choose both the size of loans and the identity of the lender.

are random variables determined by the borrower's decisions to prepay the loan early (i.e. $P_{\bar{T}} = U_{\bar{T}-1}$), or default (i.e. $P_{\bar{T}} < U_{\bar{T}-1}$). Both actions reduce the lender's revenue:

$$R(p) = S \times p.$$

Because of repayment risk, the realized revenue is a random variable lower or equal to $\bar{S} \times p$, where $\bar{S}$ is the maximum repayment multiplier if all payments are made on time, and the loan is paid in full. We assume that the lender is risk neutral.

An asymmetric information problem exists in the market whenever consumers have more information about S than does the lender.[27] We distinguish between two sources of asymmetric information: moral hazard and adverse selection. To distinguish between the two, following Einav et al. (2012), we use a reduced-form model to describe the repayment risk facing lenders:

$$S_i = \min\{\bar{S}, \exp(\alpha p_i + x_i\gamma + \eta_i)\}. \tag{61}$$

For simplicity we assume that ν_i and η_i are normally distributed, and use ρ to denote the correlation coefficient between ν_i and η_i conditional on x_i. As before, we use $M(p|x_i)$ to denote the expected repayment rate conditional on accepted offer p:

$$M(p|x_i) = E\left[\min\{\bar{S}, \exp(\alpha p_i + x_i\gamma + \eta_i)\} | \nu_i > p, x_i\right]. \tag{62}$$

In some applications S is binary, and so $M(p|x_i)$ corresponds to the repayment probability: $M(p|x_i) = \Pr(\eta_i > -\alpha p - x_i\gamma | \nu_i > p, x_i)$. Moral hazard and adverse selection imply that a price increase leads to a decrease in the repayment rate (Stiglitz and Weiss (1981)): $M'(p|x_i) < 0$.

Moral hazard is present if $\alpha < 0$. This relationship can arise because higher prices reduce the incentive for borrowers to exert effort that would increase expected returns. Alternatively, moral hazard can arise because higher interest rates reduce the probability that the borrower has sufficient (exogenous) liquidity to repay the loan. Since effort is typically unobserved, the literature does not distinguish between these two cases.

In contrast, the chosen lender is adversely selected if the distribution of willingness-to-pay ν_i is negatively correlated with η (i.e. $\rho < 0$). For instance if consumers with high willingness-to-pay ($\uparrow \nu_i$) are more likely to face an adverse liquidity shock ex-post ($\downarrow \eta_i$), consumers who accept high price offers are more likely to default: $E(S|\nu_i > p_i, x_i) < E(S|x_i)$.

Importantly, the market is adversely selected because contract terms cannot fully condition on borrower unobserved risk characteristics. This can be caused by an asymmetric information problem, or constraints on the pricing decision of lenders such as a ban on price discrimination, or price ceiling regulations leading to a pooling of risk types. For instance, the market is adversely selected if observable characteristics, x_i, affecting both borrowing and repayment decisions are not priced into the contract.

Lender j's profit maximization problem when facing a consumer of type x_i is

$$\max_{p<\bar{p}} L_j(p|x_i) \cdot [M(p|x) \cdot (p - w_{ij}) - c_{ij}] = \tilde{L}_j(p|x_i) \cdot (p - w_{ij}) - L_j(p|x_i)c_{ij},$$

where $\bar{p}$ is a price ceiling (e.g. usury laws). In the interior, the optimal price is described by the

[27] We focus on the asymmetric information problem that exists between borrowers and lenders. See Raisingh et al. (2020) for an analysis of the importance of asymmetric information in the secondary market for mortgages.

following first-order condition:

$$p_{ij} = w_{ij} + c_{ij}\frac{L'_j(p_{ij}|x_i)}{\widetilde{L}'_j(p_{ij}|x_i)} - \frac{\widetilde{L}_j(p_{ij}|x_i)}{\widetilde{L}'_j(p_{ij}|x_i)} = \text{Risk-adjusted Cost} + \text{Markup}, \tag{63}$$

where,
$$L'_j(p_{ij}|x_i) = \frac{\partial L_j(p_{ij})}{\partial p_{ij}}$$

$$\widetilde{L}'_j(p_{ij}|x_i) = \underbrace{L'_j(p_{ij}|x_i)M(p_{ij}|v_{ij} > p_{ij}, x_i)}_{<0} + \underbrace{L_j(p_{ij}|x_i)\frac{\partial M(p_{ij}|v_{ij} > p_{ij}, x_i)}{\partial p_{ij}}}_{\leq 0}.$$

This pricing equation highlights the roles of market power and asymmetric information. As competition between lenders increases, demand for lender j becomes more elastic, and prices reflect the risked-adjusted lending cost (i.e. markup $\to 0$). With moral hazard or adverse selection, the repayment probability decreases with the price, which makes the slope of risked-adjusted demand steeper: $L'_j(p_{ij})/\widetilde{L}'_k(p_{ij}) < 1$. This contributes to an increase in the marginal cost of lending, and leads to higher equilibrium prices.

Imperfect competition can attenuate the importance of adverse selection (and vice-versa) by incentivizing lenders with pricing power to reduce markups compared to an environment without credit risks. To see this, note that an increase in the correlation between willingness-to-pay and S decreases the expected repayment probability of the average borrower ($\partial M/\partial|\rho| < 0$), and makes the repayment elasticity more negative ($\partial^2 M/\partial p\partial|\rho| < 0$). This reduces the loan markups chosen by lenders with market power. Crawford et al. (2018) quantify the importance of this mechanism in the context of business lines of credit. See also Mahoney and Weyl (2017) for a theoretical analysis of the relationship between market power and adverse selection.

5.2 Asymmetric information and default risk

We start with the question of identifying and quantifying moral hazard and adverse selection using loan performance data. The literature on insurance markets has developed a series of tests to identify the presence of asymmetric information. A seminal contribution in this literature is Chiappori and Salanié (2000). They argue that asymmetric information is present if consumers choosing more insurance coverage are more likely to experience an accident, compared to observationally identical consumers facing the same price menu. In the lending context, the analogue of the Chiappori-Salanié "correlation test" is the observation that riskier borrowers are more likely to demand larger loans and accept high rate offers. The challenge for empirical work is that contract terms are typically not randomly assigned across consumers with different risk types, especially in markets with risk-based pricing.

One avenue to get around this identification problem is the use of field experiments. Two prominent examples of this approach are Ausubel (1999) and Karlan and Zinman (2009). Ausubel (1999) analyzes the results of a large scale experiment in which potential consumers were mailed credit offers with randomly selected introductory rates. Consistent with models of asymmetric information, consumers with worse risk attributes are found to be more likely to accept high-rate offers (even conditional on observed characteristics and credit score). This leads to a positive correlation between introductory rates and default probability. Although the magnitude of this correlation is likely too large to be explained by moral hazard, this test cannot distinguish between moral hazard and adverse selection. Karlan and Zinman (2009) conduct a similar field experiment in South Africa designed to separately identify the relative importance of the two sources of asymmetric information. In addition to randomly assigning introductory offers to consumers (as in Ausubel's paper), the experiment also provides more generous ex-post contracts to a randomly selected group of borrowers. As a result, the experiment generates independent

variation in the rate that induces selection (the introductory offer), as well the rate that determines the incentive of borrowers to repay the loan. The first treatment arm identifies the importance of adverse selection on risk. The second is designed to identify the elasticity of repayment with respect to the transaction rate (moral hazard). In their setting, Karlan and Zinman (2009) show that the moral hazard margin is more important than the selection margin for explaining the positive correlation between rate and default.

Adams et al. (2009) and Einav et al. (2012) develop a framework to identify and measure the importance of moral hazard and adverse selection using observational data on subprime car loans. This market is composed of dealers/lenders targeting liquidity constrained consumers who do not have access to normal credit channels. Both papers use administrative data from a large subprime lender. The company offers risk-based pricing contracts to consumers based on credit history and financial characteristics. A contract offer is a combination of interest rate r_i, car price v_i, and minimum downpayment $\bar{d}_i$. The last two variables determine the maximum loan size to which consumers have access. Despite this screening effort, most borrowers end up defaulting, making it an ideal setting to study asymmetric information problems.

The authors abstract from competition between dealers. The loan size is determined by the car price offer (V_i) and the choice of downpayment: $p_i = V_i - d_i$. Since the data exhibit limited variation in interest rates across borrowers (due to usury laws), the authors analyze how loan size can lead to moral hazard and adverse selection. In order to screen high-risk borrowers, the company targets offers by varying minimum downpayment.[28] The minimum downpayment ($\bar{d}_i$) and the car price range are determined by the lender's credit-scoring algorithm.

Two demand shocks, $\nu_i = \{\nu_i^1, \nu_i^2\}$, determine the loan demand function:

$$L(\bar{d}_i|x_i) = \underbrace{E\left[V_i - \max\{\bar{d}_i, \nu_i^2\} | \nu_i^1 > V_i - \bar{d}_i, x_i\right]}_{\text{Loan size}} \times \underbrace{\Pr(\nu_i^1 > V_i - \bar{d}_i|x_i)}_{\text{Accept prob.}}, \tag{64}$$

where ν_i^2 measures the "ideal" downpayment, and ν_i^1 is the willingness-to-pay for borrowering from the dealer.[29] The two shocks are negatively correlated: consumers with small ideal downpayment have high WTP.

The car price varies across consumers based on negotiated discounts and the choice of car model. The latter decision is largely made by the dealers, since high-risk consumers have access to a limited set of models. The presence of negotiated discounts creates a challenge for identification since car price (and therefore loan size) is potentially correlated with unobserved (to the econometrician) differences in risk or willingness-to-pay across consumers, even absent adverse selection. For instance, if high- and low-risk consumers differ in their incentives to compare offers across dealers, this would naturally create a correlation between loan size and risk. The authors abstract from the details of price negotiation/shopping by using a reduced-form pricing equation in which the list price V_i^l is used as an excluded instrument:

$$V_i = x_i b_x + b_v V_i^l + e_i,$$

where e_i is allowed to be freely correlated with the other residuals.

Because of adverse selection, the expected repayment probability depends on the choice of downpayment.[30] For instance, consumers putting more than the minimum are considered "lower

[28] Einav et al. (2013) studies the effect of the company's adoption of electronic risk-scoring technologies on lending practices.

[29] The authors use a slightly different functional form for the acceptance probability. They allow the price elasticity to differ with respect to the car price and minimum downpayment for constrained consumers. The results show that demand is very sensitive to the minimum downpayment, and is inelastic with respect to car price. Because of this, the model assumes that the acceptance probability depends on the maximum loan size, rather than the realized one.

[30] Adams et al. (2009) and Einav et al. (2012) model the repayment behavior of consumers as a continuous variable corresponding the ratio of months with positive payments.

risk" if the correlation between η and ν is positive:

$$E[S_i|\nu_i^1 > (V_i - \bar{d}_i), \nu_i^2 < \bar{d}_i, x_i] < E[S_i|x_i]. \tag{65}$$

This inequality is satisfied if consumers demanding larger loans (lower down-payments) are less likely to default for unobserved reasons, consistent with the hypothesis that borrowers select loan size based on private risk information. In contrast, moral hazard is present if the loan size, p_i, has a direct effect on the default probability. In this context, raising the minimum downpayment increases repayment by screening out consumers with low ν_i^2, and by reducing the loan size of consumers accepting the offer.

Adams et al. (2009) use this framework to develop a reduced-form test for moral hazard and adverse selection by combining plausibly exogenous variation in loan size (due to car prices) and minimum down-payment (due to company policy changes). The test is implemented by estimating the repayment probability using a control-function:

$$M(p_i|x_i) = E\left[\min\{\bar{S}, \exp(\alpha p_i + \lambda \hat{\nu}^2(z_i, p_i) + x_i\gamma)\}|\nu_i^1 > (V_i - \bar{d}_i), x_i\right] \tag{66}$$

where p_i is the observed loan size (determined by the chosen downpayment), and $\hat{\nu}^2(z_i, p_i) = E[\nu_i^2|z_i, p_i]$ is the conditional expectation of the residual from the estimated downpayment choice model define in equation (64).

Without controlling for the proxy variable $\hat{\nu}^2(z_i, p_i)$, a positive correlation between default and loan size ($\alpha < 0$) combines the effect of moral hazard and adverse selection. By conditioning on $\hat{\nu}^2(z_i, p_i)$, the model controls for the private information available to consumers at the time of borrowing, and therefore can isolate the effect of moral hazard. This proxy variable is used to test the adverse selection hypothesis: $\lambda > 0$. A positive correlation between the "ideal" downpayment and repayment is due to selection on risk.

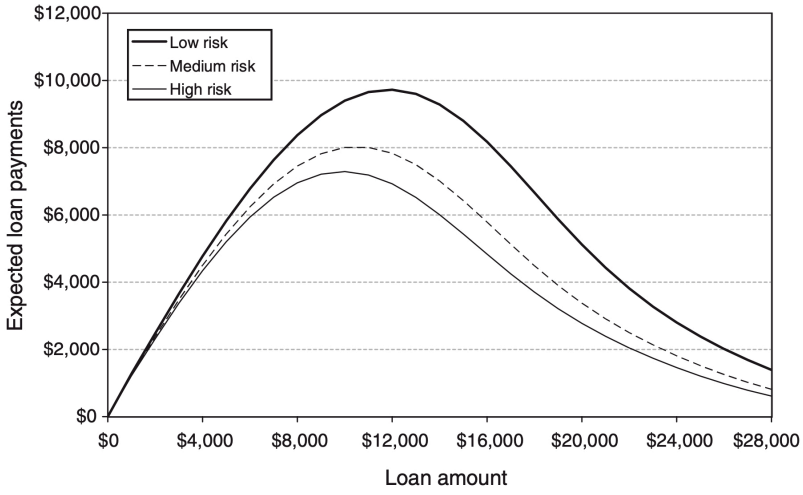
Note that loan size enters both the default probability and the control function. In order to separately identify α and λ, the authors assume that the size of the downpayment does not have a direct effect on default: $z_i = (x_i, \bar{d}_i)$. Since loan size is determined by car price and the minimum downpayment, the data exhibit two (independent) sources of variation in loan size p_i. In order to generate plausibly exogenous variation in car prices, the model controls for a rich set of covariates describing the cost and list price of each car. The remaining variation is mostly driven by changes in markups across cars and time periods (national policy).

The results provide strong evidence that moral hazard and adverse selection contribute to credit risk in this market. A \$1000 increase in loan size leads to a 2.4% increase in the probability of default. Roughly two thirds of this correlation is due to moral hazard. The results also suggest that the company is able to curb the effect of asymmetric information by restricting the loan size that borrowers can obtain, especially for consumers in the lowest risk-score categories whose ideal downpayment is substantially lower than the minimum.

By limiting loan size for risky borrowers, the company is able to lower the size of loans that consumers would have taken absent risk-based pricing, which in turns reduces demand from those consumers (↓ adverse selection), and increases repayment (↓ moral hazard). Figure 8 illustrates the effect of loan-size limits on expected loan repayments. For small loans, the curve follows a 45 degree line. However, as the loan size increases, the marginal effect of loan size on expected revenue decreases and eventually turns negative because of an increase in default. This is especially pronounced for high-risk consumers. The shape of this function implies that the profit function is concave in the quantity of credit allowed which leads to credit rationing in equilibrium, as in the classic theory models of Jaffee and Russell (1976) and Stiglitz and Weiss (1981).

Einav et al. (2012) formalize this relationship by analyzing the lender's profit maximization problem. To do so they estimate jointly the demand function, repayment probability and pricing

Figure 8. Relationship between expected loan repayment and loan size



Source: Figure 6a in Adams et al. (2009) (page 77).

function. The model is estimated by parametrizing the distribution of the random variables, $\{v_i^1, v_i^2, \eta_i, e_i\}$, using a multivariate normal distribution with unrestricted covariance matrix. The correlation coefficients measure the degree of adverse selection in the market, which can arise because of a correlation between willingness-to-pay and default risk (extensive margin), and/or between the ideal loan size and default (intensive margin). The characteristics of consumers and dealers (including the interest rate) enter the model by shifting the mean of each variable.

The results reveal that the correlation between loan demand/WTP and credit risk is relatively small conditional on x_i, implying that adverse selection is not very important. In contrast, the repayment acceptance probabilities are strongly affected by loan size and minimum downpayment (negative), consistent with the presence of moral hazard. Importantly, the risk scores assigned to each consumer are a strong predictor of loan size and default.

The estimated model allows the authors to quantify the revenue and cost components of the dealer's profit function:

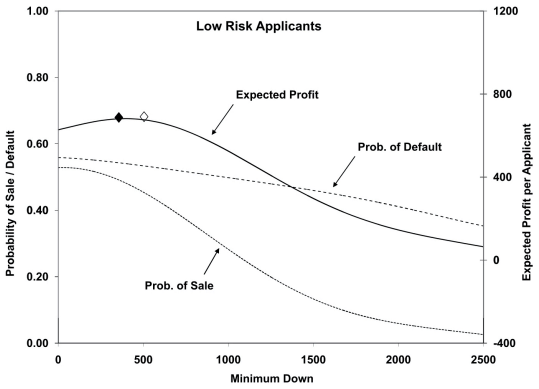
$$E\left[\pi|x_i,\bar{d}_i,V_i\right] = \\ \Pr(v_i^1 > \mathrm{v}_i - \bar{d}_i)\cdot E_{\eta_i}\left[d_i + (1+r_i)p_i M(p_i|x_i) - \text{Car cost}_i - c\middle| v_i^1 > V_i - \bar{d}_i\right].$$

This specification assumes that the recovery value of cars in case of default is zero. In practice, this value can be estimated from data on the observed recovery rate and resale value. The parameter c measures the indirect origination cost, which includes the cost of servicing the loan, and risk-related costs associated with loan securitization. To estimate this financing cost, the authors calculate the value of c that would rationalize the observed minimum down payment schedule, taking as given the other components of the contract (interest rate and price). The revealed preference inequalities yield large indirect financing cost beyond reasonable estimates of the servicing cost. This suggests the presence of importance frictions in the secondary market for subprime car loans, due, for instance, to large aggregate risk premiums.

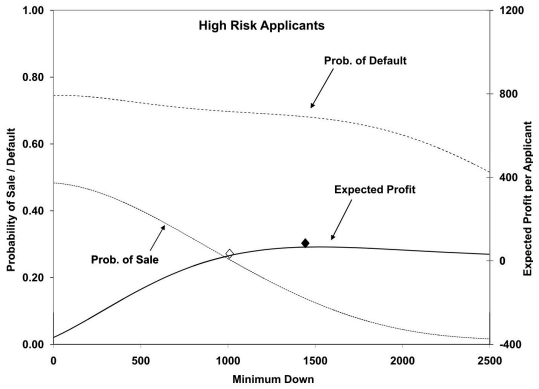
The optimal loan size is the result of a tradeoff between expected revenue and default risk. By increasing the minimum downpayment, lenders lower the probability of a sale, but increase the expected repayment rate because of better risk-screening and a reduction in moral hazard. Figures 9a and 9b illustrate this tradeoff for consumers with low and high risk scores, respectively.

Figure 9. Expected profit per consumer and default probability as function the minimum downpayment

(a) Low risk borrowers



(b) High risk borrowers



Source: Figure 6 in Einav et al. (2012) (page 1423). Dark diamond: optimal contract. Hollow diamond: observed contract.

The observed risk-based pricing schedule is fairly close to the optimal one for both types of consumers. By requiring bigger downpayments from risky consumers, the lender is able to reduce the risk of default (conditional on acceptance), and reduce the overall demand from this group. The difference between the two contracts highlights the importance of risk screening. With a uniform pricing policy, the optimal pooling contract would lower the minimum downpayment for risky borrowers and increase overall default. Average profit would fall by roughly 18%. This result confirms that investment in better credit-scoring technologies is crucial for limiting the cost of asymmetric information.

The empirical analysis in Einav et al. (2012) relies on monopolist lenders, and assumes that consumers face an exogenous outside option. As discussed above, competition between lenders can change the model predictions with regards to the effect of asymmetric information. Crawford et al. (2018) extend the empirical framework to study the interaction between market power and asymmetric information. They use administrative data on business lines of credit from all major Italian banks, including information on the transaction interest rate, credit limits, lender and borrower characteristics, and loan repayment information.

The authors model lender choice using a differentiated product demand system (Berry et al. (1995)). Differentiation arises in this context for instance due to lender-specific transaction costs, or the possibility of offering complementary products/services. Consumer surplus is parametrized as a linear function of lender and borrower characteristics and a T1EV random-utility shock (ε_{ij}):

$$\text{Surplus}_{ij} = \theta_i + x_{ij}\beta_x + \xi_j + \varepsilon_{ij} - \alpha^q p_{ij} = \theta_i + \delta_{ij} - \beta_p p_{ij} + \varepsilon_{ij},$$

where p_{ij} is the interest rate offer from lender j, $\beta_p = 1/\sigma_\varepsilon$ is the common price coefficient, and ξ_j the unobserved (to the econometrician) quality of lender j. As before, the value of the outside option is normalized to zero. The model abstracts from the loan approval margin by assuming that lenders can offer arbitrary high interest rates.

Lenders' expected revenues are determined by the acceptance and repayment decisions of

consumers. This leads to risk-adjusted loan demand function:[31]

$$\tilde{L}(p_{ij}|x_i) = \int \frac{\exp(\theta_i + \delta_{ij} - \beta_p p_{ij})}{1 + \sum_{k \in J_i} \exp(\theta_i + \delta_{ik} - \beta_p p_{ik})} M(p_{ij}|\theta_i, x_i)\phi(\theta_i)d\theta_i,$$

where the choice set of consumer i (J_i) is determined by the branch location of lenders, and $\phi(\cdot)$ is the normal probability density function. The ex-ante and ex-post liquidity shocks are assumed to follow a joint normal distribution, with correlation parameter ρ measuring the importance of adverse selection.

The estimation of the model is complicated by the fact that lenders make individualized offers to potential borrowers, which implies that *rejected* offers are unobserved by the econometrician. As in Einav et al. (2012), differences in transaction prices reflect unobserved factors determining the risk and negotiation abilities of borrowers. Einav et al. (2012) and Crawford et al. (2018) use similar approaches to impute these "missing" prices. The price that each borrower is expected receive is written as the sum of a lender fixed effect measuring the average rate differences across banks, and an idiosyncratic function of observed borrower characteristics. In addition, Crawford et al. (2018) use repeated price observations from the same borrower to estimate a borrower fixed effect capturing time-varying profitability factors. This leads to the following predicted price function for lenders competing in market t:

$$\hat{p}_{ij,t} = \bar{p}_{j,t} + x_{ij}b + \tau_i = \bar{p}_{j,t} + \Delta_{ij}.$$

The predicted price is used directly in the demand system. Implicitly the authors assume that borrowers select a lender based on average transaction prices, but choose to repay the loan based on the realized offers. The fixed-effect parameter τ_i is estimated by OLS in a first stage, and added to the vector of borrower characteristics x_{ij}.

The parameters are estimated by maximizing the joint likelihood of lender and repayment choices $y_i = \{\text{Lender}_i, S_i\}$. The mean ex-post liquidity shock is parametrized as a linear function of lender and borrower characteristics: $E[\eta_i|x_{ij}] = x_{ij}\gamma$. The parameter vector includes the mean utility and repayment parameters (β, γ), the correlation coefficient ρ, and lender-market fixed-effects $\xi_{j,t}$. The likelihood contribution of consumer i choosing lender j is given by:

$$\begin{aligned} l\left(y_i \middle| x_i, \beta, \gamma, \xi, \rho\right) = & \int \frac{\exp(\theta_i + \delta_{ij} + \beta_p \hat{p}_{ij})}{1 + \sum_{k \in J_i} \exp(\theta_i + \delta_{ik} + \beta_p \hat{p}_{ik})} \\ & \times \left[M(p_{ij}|\theta_i, x_i)^{S_i} \left(1 - M(p_{ij}|\theta_i, x_i)\right)^{1-S_i} \right] \phi(\theta_i)d\theta_i. \end{aligned}$$

The likelihood associated with consumers without a line of credit is similarly defined using the probability of choosing the outside option (Lender = 0).

Crawford et al. (2018) address the endogeneity of prices using instruments. The default price coefficient is identified using a control function approach. Transaction prices are instrumented for using average offers in other markets, and the residual of the first-stage regression is included in the repayment probability function to proxy for unobserved risk characteristics. The demand price coefficient is estimated by GMM by inverting the aggregate demand of each lender, as in Berry et al. (1995). The coefficient on the expected transaction price $\hat{p}_{ij}$ is identified using the local deposit share of each lender (a cost shifter) as an instrument.

The authors use the estimated model to simulate the counter-factual distribution of rates and default in an environment with more adverse selection. Doubling the correlation coefficient $\hat{\rho}$ raises the risk-adjusted marginal cost, which leads to higher equilibrium rates and default

[31] The model also incorporates the line of credit using using a Probit model. We abstract from this here to simplify the exposition.

probabilities – 1.87 and 5.85 percentage point increases respectively. As in competitive markets, this reduces the overall quantity of loans originated. However, because banks have market power, they also adjust down their markup to limit the unraveling of the market. In particular, the rate increase is significantly more pronounced for banks that have low average markups in the baseline scenario (a measure of market power). This price difference implies that banks with market power are less affected by the increase in default rate, and experience a smaller decline in loan origination. These results imply that an increase in asymmetric information favors banks that serve more captive consumers, and lead to an increase in market concentration.

Another tool used by lenders to limit the cost of asymmetric information is to design lending platforms in which borrowers can signal private information about their risk types. In practice this can be done by letting consumers choose from a menu of contracts. For instance, mortgage borrowers in the U.S. can choose between higher (or lower) interest rates and lower (or higher) upfront fees. Kawai et al. (2020) analyzes the effect of signalling in the online lending platform Prosper.com. Asymmetric information is particularly severe online, since lenders cannot extract "soft-information" by using loan officers. Prior to 2010, Prosper tried to address this problem by allowing borrowers to reveal the maximum interest rate they were willing to accept. Conditional on this reservation price, the offered rate was determined using an auction mechanism similar to the uniform-price auction analyzed in the Section 3 of this chapter. Using the above notation, the fact that lenders face idiosyncratic lending-cost shocks implies that access to credit is stochastic:

$$M(\bar{r}|x_i)r^* - c_{ij} \geq 0, \tag{67}$$

where r^* is the market clearing interest rate, and $\bar{r}$ is the reservation price chosen by consumers.[32] Let $\Pr(L>0|\bar{r},x_i)$ denote the probability of receiving funding. When this probability is increasing in $\bar{r}$, consumers face a tradeoff between the level of interest rate (since $E[r^*|\bar{r},x_i]$ is increasing in $\bar{r}$) and the probability of receiving funding:

$$\max_{\bar{r}<r^{\max}} \Pr(L>0|\bar{r},x_i)E_{r^*}[u(r^*|x_i,\nu_i)|\bar{r},x_i] + \Pr(L=0|\bar{r},x_i)\nu_i, \tag{68}$$

where $r^{\max}$ is the maximum interest rate determined by usury laws (36% at the time).[33] As before, the market is adversely-selected if their type ν_i is negatively correlated with the repayment rate S_i. However, when there exists an interior solution for the optimal reserve price, consumers are able to signal perfectly their types. When the reserve price is constrained by $r^{\max}$, high- and low-risk borrowers pool at the usury rate, which can lead to an unravelling of the market. Kawai et al. (2020) show that in some cases, pooling leads to a backward-bending credit-supply curve.

The authors simulate the welfare gains of signaling by comparing the equilibrium distribution of rate and credit supply under three information structures: (i) signaling, (ii) perfect pooling, and (iii) symmetric information. For consumers with median and high credit scores, the ability to signal via reserve prices eliminates over 70% of the welfare cost of asymmetric information. This is not the case however for riskier borrowers, who are more likely to be constrained by usury laws, leading to a reduction in the supply of credit.

5.3 Market power and search frictions

A takeaway from the previous section is that consumer-based pricing is an important tool used by lenders to curb the effect of asymmetric information. Combined with the fact that financial transactions can be complicated and difficult to compare, consumers often end up paying very different prices for homogeneous credit products. In this section we discuss methods and data

[32] The paper uses a richer model to determine the supply curve of each potential lender.

[33] The authors parametrize the value function $u(r)$ using a dynamic discrete choice model in which consumers repeatedly make a choice to default or not on their obligations.

used to understand the causes of this price dispersion, focusing especially on the role of search frictions, and their implications for market power.

Evidence: Price dispersion and search

Lenders use different strategies to make targeted offers to consumers: credit-card companies send mail-in or electronic offers to consumers based on credit history and demographic characteristics (Stango and Zinman (2016)), banks set wholesale mortgage-rate sheets that vary across downstream originators and borrowers based on risk and contract attributes (Woodward and Hall (2012), Bhutta et al. (2019)), and loan officers often have considerable authority to set individual offers and respond to local market conditions (Allen et al. (2014b)).

This leads to substantial dispersion in transaction prices, both across borrowers due to differences in risk, and across lenders using different pricing strategies. Some of the observed dispersion can be attributed to differences in borrower characteristics, and so it is important to consider "residual price dispersion," which takes these into account. Residual dispersion has been documented in a wide array of credit markets in different countries. For instance, using data on mortgage transactions and offers in the US, Bhutta et al. (2019) show that the standard-deviation of residual rates is 25 bps, which corresponds to a 90-10 percentile gap of roughly 60bps. Allen et al. (2014b) document somewhat larger magnitudes for Canada (std-dev=55 bps). For credit cards, dispersion is even greater. Stango and Zinman (2016) find that the 90-10 percentile (APR) is on average 1000 bps after conditioning on observed risk factors. Greater dispersion in credit card markets can be explained in part by product differences across cards, namely in terms of reward points and fees. In addition, price dispersion in credit markets has been expanding over time as lenders use more sophisticated pricing models, and are better able to target offers. For instance, interest rates on credit cards adjusted sluggishly and exhibited little dispersion in the 1980s and 1990s (Ausubel (1991) and Knittel and Stango (2003)). Similarly, until the early 1990s, over 80% of mortgage borrowers in Canada paid a common *posted* rate independent of characteristics, compared to less than 20% post-2000 (Allen et al. (2014b)).

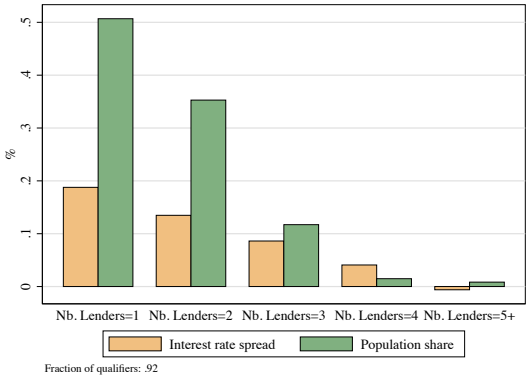
When analyzing the sources of price dispersion, the main challenge is to disentangle differences in prices due to inefficient search effort versus unobserved cost factors. This is particularly relevant in lending markets because of the tendency towards consumer-based pricing (as opposed to posted prices). One approach is to construct indirect proxies for search effort or consumer inattention. Stango and Zinman (2016) for instance document a negative correlation between borrowers' self-reported search effort and APR on credit card debt, while Keys et al. (2016) find that approximately 20% of home owners fail to refinance their mortgage when it is optimal to do so.[34] An important distinction between credit markets and markets for other search goods is the fact that sellers can reject applications based on risk assessment. As a result, consumers who have the most to gain from searching also face higher costs of generating competitive offers. Agarwal et al. (2020) study this mechanism using data on credit inquiries arising from mortgage applications. They find a strong *positive* correlation between the number of loan applications and mortgage rates, especially for consumers in the bottom three quartiles of the FICO score distribution.

We illustrate the relationship between search effort and rate dispersion using mortgage data from the National Survey of Mortgage Originations (NSMO). The public-use file contains information on roughly 30,000 borrowers between 2013 and 2017, including information on financial characteristics and credit score, ex-post loan performance (pre-payment and default), and a series of self-reported questions measuring the amount of information acquired by borrowers

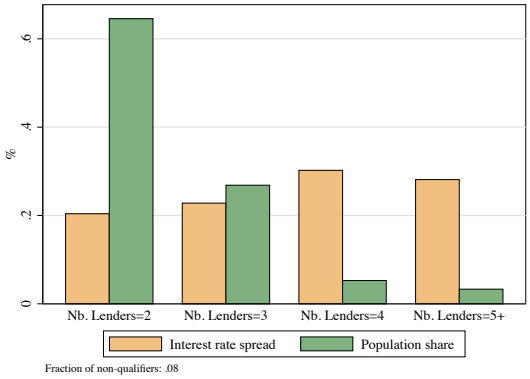
[34] See Ambokar and Samaee (2019) for a structural model of mortgage refinancing that accounts for consumer search cost.

Figure 10. Relationship between the number of lenders considered and mortgage rates

(a) Interest rate spread for qualifying borrowers



(b) Interest rate spread for non-qualifying borrowers



Source: National Survey of Mortgage Originations (NSMO).

during the search process.[35] Figure 10 shows the average interest rate spread for consumers with different levels of search effort, proxied by the number of lenders considered. The first thing to note is that more than half of borrowers consider a single lender or broker when shopping for their mortgage, and about 8% of borrowers report filing loan applications with multiple banks or brokers because of concerns about loan approval.[36] For these consumers the relationship between search and average transaction rate is positive, consistent with the results found in Agarwal et al. (2020). For the group of "qualifying" consumers, there exists a clear negative relationship between search and rate. The magnitude of the correlation is likely biased towards zero because the search decision of consumers in endogenous: informed consumers tend to be more educated and experienced, and therefore also more profitable for lenders.

To address this simultaneity problem, we borrow the identification strategy used in Stango and Zinman (2016), who argue that borrower characteristics that do not enter the pricing policy of lenders (but are correlated with search cost) can be used as instruments for the reported search effort of consumers. For instance, lenders are regularly audited to ensure that offers do not depend on sex and race, but those variables are potentially correlated with information frictions. Table 2 presents the results of this regression. Without instrumenting, consumers who consider a single lender pay on average 4 bps more than those who consider more than 2 lenders. The IV results show that considering multiple lenders leads to a 55 bps reduction in transaction rate, or more than half of the 90-10 percentile gap, in line with the elasticity found in Stango and Zinman (2016). This point estimate represents an upper bound on the gain from search, since the gender and race of borrowers are correlated with omitted characteristics such as household composition, occupation, and location. It does however highlight the fact that consumer awareness is an endogenous outcome to the transaction, and that unobserved borrower heterogeneity plays an important role in determining the decision to search and the outcome of the contract negotiation.

The last three regression coefficients from Table 2 highlight the importance of unobserved

[35]The data are available here: https://www.fhfa.gov/DataTools/Downloads/Pages/National-Survey-of-Mortgage-Originations-Public-Use-File.aspx. This dataset has been used to diagnosed search frictions by Bhutta et al. (2019), Alexandrov and Koulayev (2017) and Ambokar and Samaee (2019).

[36]Importantly, consumers often submit multiple loan applications when requesting a quote from banks or brokers. This is because brokers and mortgage specialists inquire over multiple lenders prior to offering a quote. This helps to reconcile the facts reported in Agarwal et al. (2020) that many consumers apply to dozens of lenders, with self-reported measures obtained from loan origination surveys.

Table 2. Relationship between mortgage rates, search effort and risk characteristics

VARIABLES	(1) OLS	(2) IV
1(Consideration>1)	-0.0377	-0.55
	(0.00593)	(0.11)
FICO at origination	-0.145	-0.13
	(0.00596)	(0.0076)
FICO change: 2019-Origination	-0.0502	-0.043
	(0.00551)	(0.0064)
1(Active mortgage)	-0.125	-0.12
	(0.0123)	(0.014)
Loan age at repayment	-0.0292	-0.030
	(0.00403)	(0.0047)
Observations	28,891	28,891
R-squared	0.242	-0.022
Residual dispersion: p90-p10	0.910	1.14

Robust standard errors in parentheses. Dependent variable: Interest rate spread relative to monthly average from Freddie Mac's survey of mortgage rates. Additional controls: Month-year FEs, loan size, number of borrowers, jump, term FEs, debt-to-income ratio, marital status, property type, loan purpose, number of borrowers, education, agency. Instrument for consideration dummy: sex and race of respondent. Kleibergen-Paap F-statistic: 24.31. Source: National Survey of Mortgage Originations (NSMO).

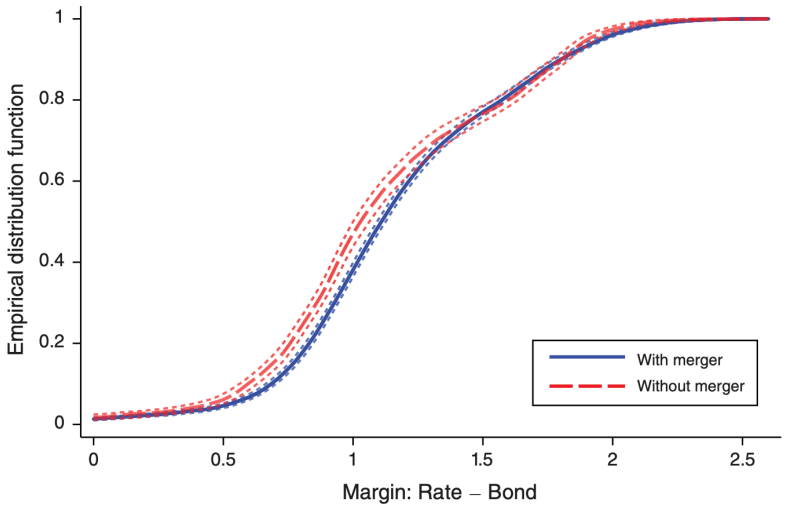
consumer heterogeneity. In addition to borrower characteristics observed by lenders at the time of negotiation, we also control for ex-post measures of risk: (i) the change in the FICO score between 2019 and origination date, and (ii) two measures of early prepayment. Both measures clearly show that consumers who are *ex-post* more profitable pay lower rates on average at origination. This relationship is either due to the presence of asymmetric information (moral hazard or adverse selection), or to signals about default or prepayment risks observed by lenders at origination (and therefore priced in). Because of consumer-based pricing, it is not possible to distinguish between the two interpretations, unlike in the market studied in Adams et al. (2009), which includes aggregate shocks to borrowing cost. Separating unobserved heterogeneity from asymmetric information is an important topic for future research on lending markets.

Given the importance of search frictions, the next question is: to what extent do firms exploit their information advantage to exert market power and price discriminate across consumers? A natural identification strategy to answer this question is to exploit plausibly exogenous variation in market structure and firm strategies that affect the market power of lenders, but are uncorrelated with the unobserved cost of a borrower. For instance, Gurun et al. (2016) exploit variation across cities in the timing of entry of Craigslist to measure the effect of newspaper advertising on mortgage rates. They find that lenders spending large amounts on advertising are more likely to set high reset rates on adjustable-rate mortgages, consistent with the idea that lenders use advertising to obfuscate their product and make high-price offers to uninformed consumers. Similarly, Argyle et al. (2020) exploit discontinuities in the relationship between credit unions' introductory offers and borrowers' FICO scores to show that consumers who receive, for quasi-random reasons, a high initial rate offer on car loans are more likely to search for additional offers, especially in markets featuring more available lenders (proxying for the gain from searching).

Merger retrospectives represent another important source of market structure variation.[37]

[37] See Sapienza (2002) for an early example of this approach studying the effect of mergers on business lines of

Figure 11. Empirical distribution of negotiated mortgage rates with and without the merger



Source: Allen et al. (2014a) Figure 2 (page 3381).

Allen et al. (2014a) study the local impact of a national bank merger in Canada to measure the importance of price discrimination in the mortgage market. When two national banks merge, consumers located in neighborhoods in which both entities are present experience a net reduction in the potential number of quotes. In contrast, consumers living in otherwise identical neighborhoods in which only one of the two banks is present experience the same aggregate effects of the merger, but continue to have access to the same number of potential quotes after the merger (assuming that lending is local).

Allen et al. (2014a) exploit this variation to measure the effect of losing a potential lending option on the distribution of negotiated interest rates. In the Canadian mortgage market, loan officers have considerable freedom to give discounts off the posted rate to consumers in response to market pressures. To the extent that consumers have different information or search costs, competition between lenders benefits only consumers who are willing to search for multiple lenders, either because they obtain multiple quotes or can credibility signal their knowledge of competing offers. The paper tests this hypothesis by estimating the treatment effect of the merger across all percentiles of the distribution of transaction rates, using the "change-in-change" estimator developed by Athey and Imbens (2006). The estimator uses the observed changes in the distribution of negotiated rates in the control group to predict the counter-factual distribution of prices absent the merger in the neighborhoods directly impacted by it. Figure 11 presents the observed and counterfactual distributions of (residual) prices, measured relative to the 5-year bond rate. The horizontal difference between the two curves corresponds to the treatment effect of each percentile. Consistent with theory, the effect of the merger is concentrated at the bottom and middle of the distribution. In other words, high search-cost consumers above the 60th percentile of the (residual) price distribution are unaffected by variation in the number of lenders. This leads to a positive correlation between competition and price dispersion. In this case, the merger decreased the 90-10 percentile gap in residual rates by 9.5 bps (or roughly 10%).

Model: Price competition with search frictions

Allen et al. (2019) develop a model of competition with consumer-based pricing aimed at quantifying the importance of market power due to search frictions in lending markets. The

credit in Italy.

case study is the Canadian market for insured mortgages. In this market, like in many lending markets, consumers receive individual rate offers and can negotiate discounts by generating competition between lenders.[38] The timing of search and price negotiation can be described as follows. First, consumers obtain a "free" quote p^0 from their home-bank – the lender with which they have a prior banking relationship. Second, consumers choose between accepting p^0, or paying a search cost κ_i to obtain additional offers. Finally, if the initial offer is rejected, lenders compete by providing competitive bids, and consumers select the lowest price option.

Abstracting from the outside option of not borrowing, the demand demand curve for the bank making the initial offer is given by:

$$L_j(p^0|x_i) = \Pr\left(p^0 < E\left[\min_{j=1,\dots,n} p_j \middle| p^0, x_i\right] + \kappa_i \equiv v_i\right), \tag{69}$$

where, as before, x_i summarizes the information available to lenders at the time an offer is made. When bringing the model to the data, z_i is used to summarize the variables observed by the econometrician. In this framework, the distribution of willingness-to-pay (v_i) depends on the distribution of search costs and consumers' expectations regarding the gain from search.

Importantly, the gain from search cannot be measured directly from the data. This is because rejected offers are typically not observed by the econometrician, and consumers differ (in unobserved ways) in their consideration set n and/or their lending cost. Given the magnitude of residual price dispersion in the mortgage market, assuming that consumers take random draws from the observed distribution of prices would overstate the gain from search and bias upward estimated search costs and profit margins. The paper incorporates unobserved heterogeneity in the cost of lending and assumes that consumers in a given neighborhood consider the same options. The latter assumption could in principle be relaxed if additional data on consideration sets were available.

The authors impose an equilibrium assumption on how firms compete in the last stage of the game to predict the counterfactual distribution of rejected offers and solve this "missing data" problem. In particular, the model assumes that consumers obtain quotes by conducting an English auction among lenders operating in their municipalities.[39] This is a convenient and flexible modeling choice, meant to approximate the back-and-forth that takes place when consumers negotiate terms with multiple banks simultaneously. In equilibrium, the expected transaction price among searchers is obtained from the cost distribution of the second most efficient lender accounting for the possibility that the initial offer p^0 can be recalled:

$$E\left[\min_{j=1,\dots,n} p_j | p^0, x_i\right] = \left(1 - G_{(2)}(p^0|x_i,n)\right) p^0 + \int_{\underline{c}}^{p^0} c g_{(2)}(c|x_i,n) dc,$$

where $g_{(2)}(\cdot|x_i,n)$ is the conditional distribution of the second lowest cost for a consumer of type x_i among n banks. This corresponds to the Nash equilibrium distribution of offers.

An alternative modelling strategy is to assume that firms compete by making different take-it-or-leave offers (or posted prices) based on the cost of lending to each consumers. Due to search costs, consumers only observe a random set of offers (which can depend endogenously on search effort), and choose the most attractive one. In this case the equilibrium price distribution

[38] Note that the paper focuses on the most popular contract type, and abstracts from some of the choices that consumers make such as LTV, amortization period, and type of contract (fixed vs variable rate). Other papers have focused more on questions of mortgage design. See for instance Campbell and Cocco (2015), Piskorski and Tchistyi (2017), and Guren et al. (2019). See also Gambacorta et al. (2020), who study steering of consumers towards riskier mortgage products.

[39] The assumption that all lenders take part in the auction is made for simplicity and can be relaxed by modelling the number of banks invited to the auction as a function of search costs, as in Allen et al. (2014a) and Salz (2020).

is implicitly defined by a system of differential equations, similar to the bid distribution in first-price sealed-bid auction models. See Galenianos and Gavazza (2020) for an application to the market for credit-cards.

In Canada, roughly 76% of borrowers remain loyal to their home-bank, and "switching" consumers pay on average 10 bps less.[40] In contrast 62% of consumers report considering more than one lender. This suggests that banks with a large consumer base have an *incumbency* advantage over competing lenders. The authors model this by assuming that the home-bank has a cost advantage ω_h:[41]

$$c_{ij} = \begin{cases} \bar{c}_i - \omega_h & \text{if } j = \text{home-bank}, \\ \bar{c}_i - \omega_{ij} & \text{if } j \neq \text{home-bank}, \end{cases}$$

where $\omega_{ij} \sim T1EV(\xi_j, \sigma_\omega)$ is a privately observed match-value shock to lender j's profit, and the location parameter ξ_j is a function of observed lender characteristics. This distributional assumption is particularly convenient here, because it leads to closed-form expressions for the distribution of auction prices and winning probabilities despite the presence of asymmetries between firms (see Brannan and Froeb (2000) for discussion).

The common cost component is observed by all parties, and includes a random-effect term capturing unobserved heterogeneity across borrowers: $\bar{c}_i \sim N(z_i\beta, \sigma_\varepsilon^2)$, where z_i is a vector observed borrower characteristics. The model abstracts from asymmetric information on repayment risk. This assumption is made in part because payments are insured against default, and data on the ex-post performance of each loan are not available. Instead, the common component captures in a reduced-form way the information that lenders have about the expected duration of the loan: $c_{ij} = \bar{c}_i - \omega_{ij} = \text{Origination cost}/M(x_i)$.

The importance of bank heterogeneity determines the gain from search in the market. If $\sigma_\omega = 0$, consumers can limit their search to only two banks in order to generate a competitive offer in which the home bank offers the Bertrand-Nash price: $\bar{c}_i$. If $\sigma_\omega > 0$, the retention probability of the home bank is less than one, and rival lenders earn a positive markup on average. In practice, this heterogeneity reflects differences across lenders and origination time in the pricing model used to evaluate risk, an important source of price dispersion in lending markets documented in Stango and Zinman (2016) and Bhutta et al. (2019).

In the first stage of the game, the home bank chooses an initial offer to screen high search-cost consumers:

$$\max_{p^0 < \bar{p}} \quad L_j\left(p^0|x_i\right)\left(p^0 - c_{ih}\right) + \left(1 - L_j(p^0|x_i)\right) E[\pi^*|p^0, x_i], \tag{70}$$

where $E[\pi^*|p^0, x_i]$ is the expected profit of the home bank at the auction stage. The posted price, $\bar{p}$, acts as a price ceiling in the market. This implies that lenders are constrained in their ability to price risk into the contract. As a result consumers with high realized values of $\bar{c}_i$ face a higher likelihood of being rejected (negative profit), and must incur a search cost to qualify at a least one lender, similar to the model analyzed in Agarwal et al. (2020). The initially (unconstrained) offer is implicitly defined as:

$$\begin{aligned} p^0 &= \bar{c}_i - \omega_h + E[\pi^*|p^0, x_i] - \frac{1}{L_j'(p^0|x_i)}\left[L(p^0|x_i) + (1 - L(p^0|x_i)\frac{\partial E[\pi^*|p^0, x_i]}{\partial p^0}\right] \\ &= \bar{c}_i - \omega_h + \mu(n), \end{aligned} \tag{71}$$

[40] The paper abstracts from the role of mortgage brokers who originate loans mainly from mono-line lenders.

[41] Alternatively, an incumbency advantage can arise because of a switching cost or quality differences between banks. The results suggest that the model with a cost advantage fits the data better.

where $\mu(n)$ is an equilibrium markup function. Intuitively, this markup is a function of the bargaining leverage of consumers in the first stage, which is affected by the level and dispersion of search costs, as well as the expected gain from search. For instance, conditional on qualifying, low-cost borrowers and consumers financing larger loans are more likely to search and, as a result, receive better offers in the first stage.

The distribution of transaction prices corresponds to a mixture of three density functions: the distributions of competitive offers accepted by switching and loyal consumers, denoted by $\psi^s(p|p^0,n,\bar{c}_i,z_i)$ and $\psi^l(p|p^0,n,\bar{c}_i,z_i)$ respectively, and the distribution of initial quotes, $\psi^0(p|n,z_i)$.[42] The mixing probability is derived from the decision of consumers to search and switch lenders. Abstracting from the binding posted price and letting y denote the switching decision, the likelihood contribution is given by:

$$l(y_i,p_i|z_i) = \begin{cases} L(p_i|p_i+\omega_h-\mu(n),z_i)\psi^0(p_i|n,z_i) \\ \quad +\int\left(1-L(p^0|x_i)\right)\psi^L(p_i|p^0,n,x_i)d\phi(\bar{c}_i|z_i)d\bar{c}_i & \text{if } y_i=\text{loyal}, \\ \int\left(1-L(p^0|x_i)\right)\psi^S(p_i|p^0,n,x_i)d\phi(\bar{c}_i|z_i)d\bar{c}_i & \text{if } y_i=\text{switch}, \end{cases}$$

where $x_i=(\bar{c}_i,z_i)$ and $L\left(p^0|x_i\right)$ is the equilibrium probability that the consumer accepts the first quote. The model is estimated by maximum-likelihood using a nested-fixed point algorithm. The likelihood is augmented with auxiliary moments from the probability of considering multiple lenders when shopping for mortgages.

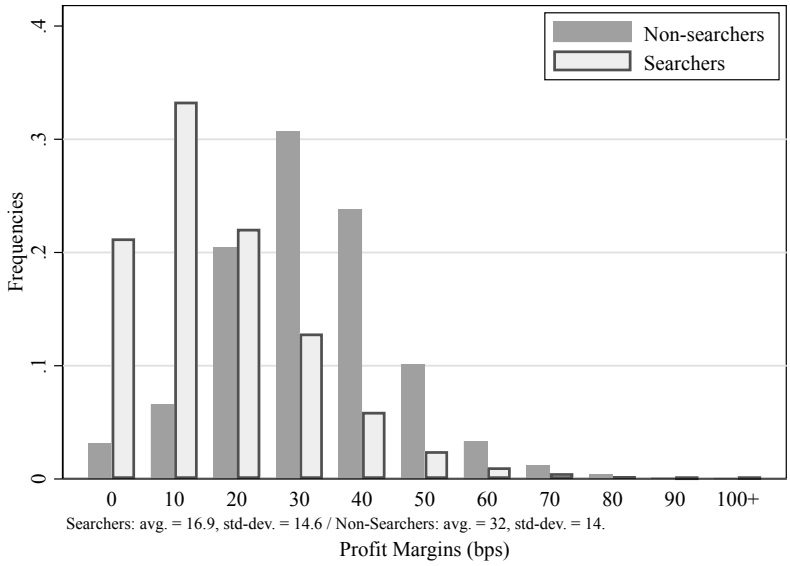
The main challenge for identification is to separate the relative importance of the three sources of consumer heterogeneity: (i) search cost (κ_i), (ii) borrowing cost ($\bar{c}_i$), and (iii) lender heterogeneity (ω_{ij}). A key identifying assumption is that the home bank makes the first offer. Under this assumption, the distribution of prices among switchers reflects the outcome of an auction with unobserved heterogeneity. The distributions of $\bar{c}_i$ and ω_{ij} are separately identified using variation across markets in the number of competitors. In particular, assuming conditional independence, exogenous variation in the number of lenders changes the offer distribution, which identifies the distribution of ω_{ij}). The distribution of prices in markets with one or a small number of lenders identifies the common component (since $p_i=\bar{c}-\omega_h$). In contrast the search cost distribution is identified using variation across markets and consumers in the switching probability. For instance, the probability of switching is increasing in the number of lenders, and low-income and first-time buyers are more likely to switch.

The results suggest that, despite the importance of search costs, the market is fairly competitive for consumers able to obtain multiple quotes. On average, consumers must incur an upfront cost of \$1,200 to search, equivalent to slightly more than a month of mortgage payments. The average profit margin is 22 bps, or 18% of the spread between the transaction rate and the 5-year bond rate (120 bps). Firms have little market power because the bulk of price dispersion is due to unobserved cost differences, common across lenders ($\bar{c}_i$), as opposed to idiosyncratic differences across lenders. However, this masks important difference for consumers. Figure 12 plots the estimated distribution of markups across consumers. The average markup faced by searchers is roughly half of the markup for non-searchers, and the standard deviation of markups ($\approx$ 15 bps) corresponds to one third of the residual dispersion in rates.

The difference in profits from searchers and non-searchers confirms that the ability to make the first offer is quite valuable. In the model, the home bank moves first and is able to price discriminate by offering (up to) two quotes. This, coupled with the cost advantage (ω_h) implies

[42]The distribution prices for switching consumers is given by the distribution of the random variable: $\bar{c}_i - \min\{\omega_h,\omega_{-b}\}$ where $\omega_{-b}=\min_{j\neq b}\omega_j$ is the minimum idiosyncratic cost component among loosing banks. The distribution of competitive prices for loyal consumers is derived from the minimum rival ω's: $p_i=\bar{c}_i+\min_j\omega_{ij}$. The initial price distribution is derived from the density of $\bar{c}_i$: $\psi^0(p|n,z_i)=\phi\left(p_i+\omega_h-\mu(n)|z_i\right)$.

Figure 12. Distribution of markups on mortgage transaction rates for searchers and non-searchers



that lenders with large consumer bases have significantly more market power than small banks. The difference in profit margins between "small" and "large" lenders is equal to 15 bps, and the market share of banks with small networks is an order of magnitude smaller than that of large network banks.

These results highlight the importance of the first-mover advantage in markets with search frictions. In the Canadian mortgage market, this timing advantage is created by the fact that banks are vertically integrated and bundle multiple financial services. In lending markets more generally, firms invest significant resources to be in a position to make the "first-quote" and price discriminate. These investments include advertising, branch presence, and referrals from real estate agents and/or dealers. Understanding how these factors affect the initial match between borrowers and lenders (and therefore market power), is an important avenue for future research.

Extensions: Adverse-selection, price ceilings and repayment risk

The model just described abstracts from the borrower's repayment decision. Uncertainty about loan repayment affects price discrimination and market power by changing the expected lending cost, and potentially creating an averse selection problem. Cuesta and Sepúlveda (2019) and Galenianos and Gavazza (2020) analyze the effect of price ceilings on access to credit and competition. Allen and Li (2020) and Agarwal et al. (2020) study pre-payment and default risk, respectively, in mortgage markets. All three papers abstract from concerns related to moral hazard: $\alpha = 0$.

Cuesta and Sepúlveda (2019) focus on the second-stage of the pricing game described above, taking as given the consumer's outside option. An important feature of the market is the presence of a price regulation that imposes a ceiling on the interest rate that lenders can charge. As with the posted price in the case of mortgages, this implies that consumers are not guaranteed to qualify for a loan conditional on applying. If $\bar{p}$ denotes the price ceiling, this leads to the following loan approval condition:

$$\omega_{ij} > M(\bar{p}|x_i)(\bar{p} - w) - \bar{c}_i = \bar{\omega}_i.$$

In this context, the equilibrium transaction price at the auction stage is given by:

$$p_i = \begin{cases} w + \frac{\bar{c}_i - \omega_{(2)}}{M(\bar{p}|x_i)} & \text{if } \omega_{(2)} > \bar{\omega}_i, \\ \bar{p} & \text{if } \omega_{(2)} < \bar{\omega}_i < \omega_{(1)}, \\ \text{Reject} & \text{otherwise}, \end{cases} \tag{72}$$

where $\omega_{(1)}$ and $\omega_{(2)}$ are the first and second highest match values, respectively. Let $\Psi(p|\bar{p},n,x_i)$ denote the equilibrium distribution of auction prices (conditional on qualifying), and $G_{(1)}(\bar{\omega}_i)$ the probability of not qualifying for a loan conditional on searching. If p_i^0 denotes the value of the outside option (exogenous), the decision of consumers to collect competitive quotes is summarized by:

$$\begin{aligned} L(\bar{p}|x_i) &= \Pr\left((1 - G_{(1)}(\bar{\omega}_i)) \int p d\Psi(p|\bar{p},n,x_i) + G_{(1)}(\bar{\omega}_i) p^0 + \kappa_i < p^0 \right) \\ &= \Pr\left(\kappa_i < (1 - G_{(1)}(\bar{\omega}_i)) \int (p^0 - p) d\Psi(p|\bar{p},x_i) \right). \end{aligned} \tag{73}$$

Note that the demand for loans is a function of the price ceiling (instead of realized offers). This is because the model assumes that consumers commit to accept the lowest price offer after paying the search cost κ_i. This simplifies the auction game, since banks have common beliefs about the repayment risk, which is independent of the cost realizations. In this context, the market is adversely selected if consumers deciding to search for quotes are less profitable than consumers who chose not to apply for a loan. This can be due to a negative correlation between S and the search cost of consumers, and/or a positive correlation between p^0 and S. For instance, in the case of mortgages, low search-cost consumers are more likely to switch lenders and refinance their mortgage before the end of the contract, leading to adverse selection on pre-payment risk. Similarly, private information about job stability or health can generate adverse selection on default risk.

Following Einav et al. (2012), Cuesta and Sepúlveda (2019) parametrize repayment and search costs using bivariate normal distributions:

$$\begin{aligned} \ln\left(\frac{S_i}{\bar{S}}\right) &= \min\{z_i\gamma + \eta_i, 0\}, \\ \ln \kappa_i &= z_i\beta + \nu_i, \end{aligned}$$

where $(\nu_i, \eta_i) \sim N(0,\Sigma)$ and z_i is a vector of observed borrower characteristics. The endogenous selection of consumers into the market leads to the following expression for the expected repayment rate:

$$M(\bar{p}|x_i) = E\left[\min\left\{ \exp(z_i\gamma + \eta_i), \bar{S} \right\} \middle| \kappa_i < (1 - G_{(1)}(\bar{\omega}_i)) \int (p^0 - p) d\Psi(p|\bar{p},x_i) \right]. \tag{74}$$

Note that the qualifying probability and price distribution defined in equation (72) depend on firms' beliefs about the repayment rate of consumers applying for a loan. An equilibrium in this market is therefore defined as a fixed-point $M^*(\bar{p}|x_i)$ (for each consumer type) such that the price distribution is consistent with the selection of consumers in the market.

Cuesta and Sepúlveda (2019) estimate the model using a partial likelihood approach that avoids repeatedly solving for the equilibrium beliefs. In particular, assuming that the common component $\bar{c}_i$ is a deterministic function of observed characteristics (no unobserved heterogeneity), the equilibrium price distribution and qualifying probability can be estimated using reduced-form techniques, $\hat{\Psi}(p|\bar{p},n,x_i)$ and $\hat{G}_{(1)}(\bar{\omega}_i)$, from data on transaction prices and rejections (conditional on z_i). These are used to compute the application probability defined in

equation (73), and estimate the search cost distribution and the value of the outside option using data on loan applications. In a second step, the lending cost parameters are estimated by maximizing the joint likelihood summarizing the pricing and repayment decisions, using data on prices and loan repayment. The main downside of this approach is that the (residual) dispersion in prices is fully explained by heterogeneity in match values across lenders. Relative to Allen et al. (2019), the estimated model therefore predicts more market power, since markups are proportional to the expected difference between the lowest and second lowest cost lender ($\omega_{(1)} - \omega_{(2)}$).

Allen and Li (2020) extend the model developed by Allen et al. (2019) by studying the mortgage renewal decision of consumers. In Canada, most mortgage contracts are amortized over 25 years, but rates are fixed for 5 years and consumers face pre-payment penalties. As a result, very few consumers refinance their mortgage before the 5-year term (except when selling their house), but they are free to search for lower rates at the renewal stage.[43] However, because of search frictions, most consumers renew their mortgages with the same lender, which increases the value of "locking-in" a new consumer at origination, effectively increasing the repayment rate $M(x_i)$. The paper abstracts away from asymmetric information on repayment (i.e. $M(p|x_i) = M(x_i)$).

To illustrate this, consider a version of the model described above in which consumers sign a two period contract that must be renegotiated after the end of the first. The last stage of the game is the same as the static game discussed above, except that the loan balance is smaller (i.e. lower incentives to search). In the first stage, the discounted value of acquiring the consumer is given by:

$$\begin{aligned} V_1(p|x_i) &= M_1(x_i) \times (p - \bar{c}_i + \omega_{ij}) + \delta V_2(x_i) \qquad (75) \\ \text{where,} \qquad & V_2(x_i) = \max_{p_2} \quad M_2(x_i) \times \Big\{ L_2(p_2|x_i)\,(p_2 - \bar{c}_i + \omega_h) + (1 - L_2(p_2|x_i))\,E[\pi^*|p_2, n, x_i] \Big\}, \end{aligned}$$

and where $E[\pi^*|p_2, n, x_i]$ is the expected profit from the auction, and $L_2(p_2|x_i)$ is the probability that the consumer accepts the second-stage offer (see equation (70)). Equation (75) defines the willingness-to-pay of lenders at the auction stage. In particular, conditional on searching, consumers receive an offer corresponding to the value of the second most efficient lender:

$$p_1^* = \bar{c}_i - \omega_{(2)} - \delta \frac{V_2(x_i)}{M_1(x_i)}.$$

This highlights the importance of search costs in determining initial contract prices. Because consumers are unlikely to switch lenders in the last stage of the game, the expected profit margin in $t = 2$ is large. This decreases the auction price in the first stage. Anticipating this, the consumer's gain from search is larger in the first stage, which makes demand more elastic and reduces the importance of price discrimination.

The authors rule out the possibility of adverse selection by assuming that the first and second period search costs are independent (conditional on observed characteristics). Assuming instead that the two search costs were positively correlated would reduce the option value of winning the contract at the auction stage. To see this, note that conditional on rejecting the initial quote p^0, and assuming myopic consumers, the continuation value is given by:

$$\begin{aligned} V_2(x_i) = \quad & \max_{p_2} \quad M_2(x_i) \times \Big\{ L_2(p_2|x_i, \kappa_{i1} < \bar{\kappa}_1)\,(p_2 - \bar{c}_i + \omega_h) \\ & \qquad + (1 - L_2(p_2|x_i, \kappa_{i1} < \bar{\kappa}_1))\,E[\pi^*|p_2, n, x_i] \Big\}, \end{aligned}$$

[43] Mortgages in the UK have a similar term structure. See Benetton (2018) for an empirical analysis of the UK mortgage market.

where $\bar{\kappa}_1 = E\left[\min_{j=1,\ldots,n} p_j \middle| p_1, x_i\right] - p_1$ is the threshold defining the search decision of consumers in the first stage. When κ_{i1} and κ_{i2} are positively correlated, the retention probability for searchers is smaller than for consumers who accepted the initial offer p_1 in the first stage. This implies that lenders competing in the first-stage auctions face a higher pre-payment (or non-renewal) probability. This increases the first-period prices (relative to an environment with IID search costs), consistent with the presence of adverse selection.

Finally, Agarwal et al. (2020) use a sequential search model to study the effect of adverse selection in the US mortgage market. Rather than assuming that lenders face different costs, the authors assume that lenders use different risk assessment models when evaluating the repayment risk:

$$E\left[\pi(p|)\right] = M_j p - c, \tag{76}$$

where c is a common financing cost, and M_j is a signal observed by lender j regarding the default probability of the borrower.[44] The model assumes a two-types distribution: the perceived repayment probability is $M_j = M^h$ with probability μ_θ, and $M_j = M^l$ with probability $1-\mu_\theta$. The probability of a "high-type" borrowers (type h) in the population is given by ω_h, and the screening technology is informative if $M_h > M_l$.[45]

The authors assume that loans are profitable only for type h borrowers, and therefore $1-\mu_\theta$ also measures the probability that a loan application is rejected. In the model, rejection occurs because loan prices do not reflect the riskiness of the borrower, contrary to the models used in Allen et al. (2019) and Cuesta and Sepúlveda (2019). The authors justify this assumption by arguing that most lenders in the US mortgage market use posted prices, instead of using targeted offers based on private information. This leads to the following profit maximization problem:

$$\max_p \lambda L_h(p) M_h + (1-\lambda) L_l(p) M_l - c,$$

where $L_h(p)$ is the volume of loans approved by the bank to consumers of type h.

Conditional on entering the mortgage market, consumers search sequentially across lenders, accounting for the probability of being rejected. This leads to the following optimal stopping rule defined by a reservation price $r(\kappa_i, \theta)$ for a consumer with search cost κ_i:

$$\kappa_i = \mu_\theta \int_{\underline{p}}^{r(\kappa_i,\theta)} (r(\kappa_i,\theta) - p)\, d\Psi(p),$$

where $\Psi(p)$ is the equilibrium distribution of prices, and $\theta \in \{H, L\}$ the risk type of consumer i. If $h(\kappa_i)$ denotes the density of search cost, we can use the reservation price to derive an expression for the quantity of loans originated by each bank:

$$L_\theta(p) = \int_{\bar{\kappa}(p,\theta)}^{\infty} \frac{h(\kappa_i)}{\Psi(r(\kappa_i,\theta))} d\kappa_i, \tag{77}$$

where $\bar{\kappa}(p,\theta) = r^{-1}(p,\theta)$ is the inverse reservation price function.

In this framework, asymmetric information between borrowers and lenders induces an adverse-selection problem by changing the reservation price of consumers. In particular, riskier consumers ($\theta = l$) have lower gains from search, and therefore, on average, accept higher prices. This leads to an adverse selection problem, since high-price offers are accepted by consumers with lower expected revenue.

[44] The model abstracts from prepayment risks.

[45] Note that the profit function abstracts from observed differences across borrowers by considering the pricing problem for the "average" borrower and residualizing transaction prices.

In order to generate price dispersion in the market, the authors assume that the cost of originating the loan differs across note rates. In particular, lenders choose from a discrete set of note rates, $p \in \{p_1, \ldots, p_K\}$, and each discrete price is associated with a profit shock $\omega_{j,k}$ distributed according to a $T1EV(0, \sigma_\omega)$ distribution. This implies that the price distribution takes a multinomial logit form:

$$\Psi(p_k) = \frac{\exp(\Pi_k/\sigma_\omega)}{\sum_{k'} \exp(\Pi_k/\sigma_\omega)}, \tag{78}$$

where $\Pi_k = \lambda L_h(p) M_h + (1-\lambda) L_l(p) M_l - c$. This is analogous to an entry game with incomplete information. A Bayes-Nash equilibrium is defined as a fixed point of the above function, where consumers and firms have consistent beliefs about the price distribution.

6 REGULATION OF FINANCIAL MARKETS

6.1 Financial stability and regulation

As discussed in the Introduction to this chapter, borrowers (firms) seeking funds could obtain these directly from savers by selling securities, but often turn instead to financial intermediaries to facilitate this transfer. As shown in Diamond and Dybvig (1983) intermediaries can provide risk sharing among individuals needing to consume at different times. That is, financial intermediaries facilitate the transformation of the illiquid entrepreneurial projects of firms into liquid liabilities via demand deposits. Furthermore, intermediaries provide monitoring services, since these projects are opaque and feature problems of asymmetric information. The ability of banks to create liquidity can lead to instability in the form of bank runs or coordination failures in which depositors panic and attempt to withdraw their funds (Diamond and Dybvig (1983)). An alternative explanation for crises is that they reflect fundamentals (see Gorton (1988)). The global games literature reconciles these two views by noting that weak fundamentals can lead to panic (see for instance Morris and Shin (1998)).

Either way, crises involve the risk of financial institution failure and, if left unchecked, of systemic risk to the financial system. Recent crises have witnessed the failure of large numbers of financial institutions. During the Savings & Loans crisis of the 1980s over 1000 thrifts failed, while the great financial crisis witnessed the failure of over 500 banks.[46] To minimize the risk and disruption to the overall financial system the Federal Deposit Insurance Corporation (FDIC) resolves failed institutions by auctioning them off to healthy banks.

In an effort to avoid costly resolution processes governments throughout the world have enacted regulations designed to avoid crises. One immediate response to bank runs is to freeze deposits. More long-term solutions include deposit insurance and capital requirements. We discuss each in turn.

6.1.1 Deposit insurance

To reduce the risk of bank runs many governments require deposit insurance, whereby intermediaries pay into a fund that will provide insurance to savers on eligible deposits in the event of failure. The existence of insurance is meant to move away from Diamond and Dybvig's panic-equilibrium in which savers run, towards their good equilibrium in which they do not. Typically, not all deposits are insured, since most insurance programs commit to covering only up to a certain threshold. As discussed in more detail in Section 4, Egan et al. (2017) allow for run-prone uninsured depositors in their structural model of the banking sector. As pointed out in Diamond and Dybvig (1983) these depositors may be sensitive to the distress of financial institutions, which can lead to instability of the financial system and runs on banks. Egan et al. (2017)

[46] See for instance James and Wier (1987) for the Savings & Loans crisis and Granja et al. (2017), Vij (2018), and Allen et al. (2021a) for the great financial crisis.

quantify the extent to which uninsured depositors are sensitive to distress and evaluate whether this level of sensitivity is sufficient to generate self-reinforcing runs or multiple equilibria. Their findings suggest that there are indeed multiple equilibria reflecting depositor beliefs that some banks might default. They use their model to examine whether increasing insurance limits, like the FDIC did on two occasions during the great financial crisis, could improve stability. Their counterfactual simulations suggest that doing so would have little impact as the probability of default would not change for most banks. Rents instead accrue to newly insured depositors. They also consider an alternative counterfactual policy in which interest rates on insured deposits are capped. Doing so prevents unstable banks from attracting too many insured deposits and thereby keeps more deposits in more stable institutions.

6.1.2 Capital regulations

There is concern however that the presence of deposit insurance, can lead to a moral hazard problem whereby financial institutions engage in excessive risk taking, knowing that losses will be covered. To address this risk taking behavior banks are typically required to hold a minimum ratio of capital to assets along with enough liquidity to meet funding shocks.[47] Basel III proposed to increase these from 4% to 6%.

A number of papers study the effect of increased capital requirements using the tools of empirical IO. In their study of the US banking system, Egan et al. (2017) also consider a counterfactual scenario in which capital requirements are increased. They find that doing so increases stability in the worst equilibrium, with the probability of default falling dramatically. Benetton (2018) develops a structural model of the UK mortgage market to evaluate the cost of risk-weighted capital requirements and the impact of leverage regulations. On the demand side he uses a discrete-continuous setup in which borrowers choose both the mortgage product and the size of the loan. On the supply side multi-product firms compete on mortgage rates and face capital requirements. His focus is on the effect of these regulations on interest rates, concentration and risk. The UK setting is interesting because policymakers have allowed banks to establish their own internal rating-based models in order to calculate the risk weights on their activities. Since doing so is costly, a two-tier system has developed in which large lenders employ internal rating-based models, while small lenders use the standardized regulatory approach. Using the estimates from his structural model Benetton is able to perform counterfactuals in which either all banks use the standard regulatory approach or all use internal rating-based models. The former is found to increase the equity buffer of larger lenders. Furthermore, their costs rise, and the increase is passed on to borrowers resulting in significant changes in market shares and a reduction in bank and consumer surplus.

Corbae and D'Erasmo (2021) build a dynamic model of the banking industry that extends the static frameworks proposed in Allen and Gale (2004) and Boyd and De Nicoló (2005), in which an exogenous number of banks engage in Cournot competition. Corbae and D'Erasmo endogenize the size distribution of banks by incorporating shocks and dynamic entry and exit decisions. They then solve for the industry equilibrium following Ericson and Pakes (2012) and Gowrisankaran and Holmes (2004). Parameters are calibrated to match long-run industry averages. Using this framework they perform policy counterfactuals in which they evaluate the impact of Basel III reforms to capital requirements, which involved raising them from the Basel II level of 4% to the required minimum risk weighted capital requirement of 6% plus a 2.5% capital conservation buffer. Their findings suggest that doing so leads to a significant reduction in bank exit probabilities, but a more concentrated industry due to less entry. They also find that aggregate bank lending falls and that loan interest rates increase moderately in the long run.[48]

[47] See Cooper and Ross (2002).

[48] See also Begenau (2020) and Begenau and Landvoigt (2020), who examine the role of capital requirements in

6.1.3 Competition and stability

It has long been argued that there exists a tradeoff between competition and stability in the financial sector (see for instance Allen and Gale (2004) and Vives (2011)). Corbae and Levine (2020) explain well this tradeoff, noting that : "although competition boosts efficiency, it reduces banking system stability by squeezing profits, lowering bank valuations, and encouraging bankers to make riskier investments because they have less to lose." On the other hand, Boyd and De Nicoló (2005) point out that the relationship could actually work in the opposite direction with banks in concentrated markets using their greater market power to increase loan rates, to which borrowers react by choosing riskier projects.

Corbae and Levine (2020) examine this question using a dynamic Cournot model with endogenous entry (thereby allowing market structure to vary in response to policy changes) in an effort to determine whether competition lowers stability in the banking sector. Their calibration exercise and empirical results provides evidence that the competition-stability tradeoff exists, and that policymakers can get the benefits of competition without the negative impact on stability by enhancing bank governance and tightening leverage requirements.

Entry regulations

Entry regulations have played an important role in limiting the extent of competition in the financial industry. This has been especially the case in the US banking market, where stringent restrictions on the ability of banks to expand geographically both within and across states led to a banking industry that was and is more fragmented than in other countries. This fragility came at a cost, as a large number of thrifts and small community banks failed in the 1980s. These failures incentivized policy makers to push for the elimination of restrictions on geographic expansion, culminating in 1994 with the passage of the Riegle-Neal Interstate Banking and Branching Efficiency Act. The Act removed all remaining barriers to interstate banking and provided the foundation for the removal of constraints on interstate branching. Over the following twelve years, the US banking sector became much more concentrated. The ten largest banks tripled in size, their fraction of deposits rising from 12% to 36%.[49]

To evaluate the impact of Riegle-Neal, Dick (2008) estimates the change in welfare between 1993 (before) and 1999 (after). Specifically, she quantifies the expected equivalent variation (EV) of the change in welfare over this period:

$$EV = S_{99}(p', x'; \theta_D) - S_{93}(p, x; \theta^D), \tag{79}$$

where $S(p, x; \theta_D) = \ln[\sum_j^J exp(\delta_j(p_j, x_j; \theta_D))]/\alpha$. Dick finds that an investor in the *median* market would experience an annual benefit of between \$8.00 and \$18.00, implying that, despite the significant reorganization in the market, services and prices adjusted such that investors were slightly better off. In related work, Ho and Ishii (2011) estimate welfare gains of \$60 per year from Riegle-Neal, with most of the gains coming from branch location adjustments that shrank the distance between consumers and banks.

While the market became less competitive following Riegle-Neal, it is less clear that it became more stable. Aguirregabiria et al. (2016) use the model described in Section 4 to examine the extent to which banks actually took advantage of Riegle-Neal to reduce their geographic risk. Specifically, the examine whether banks became more geographically diversified following elimination of the entry restrictions. Their findings suggest that, although the Act provided banks with considerable opportunities for diversification, few small and medium banks (i.e. those most at risks of failure) took advantage of these. Using the structural model of branch-network choice

competitive models.

[49] Dick (2006) notes that even as regional concentration increased, local-market (i.e. at the MSA level) concentration has remained fairly stable, with between two and three dominant banks controlling roughly half of market deposits.

described in Section 4 they illustrate that although banks had preference for diversification, this was counterbalanced by economies of scale and density, merger costs and local market power concerns, preventing them from expanding in such a way as to lower their geographic risk.

? also study the impact of Riegle-Neal. They consider the Riegle-Neal to lower the cost of expanding a bank's funding base. They develop an industry equilibrium model in which banks can raise the mean size of their deposits at some cost. The setup is in the spirit of Ericson and Pakes (1995) with banks endogenously moving up a funding base ladder.

Finally, Aguirregabiria et al. (2019) examine Riegle-Neal's impact on access to credit. Doing so requires a model that allows for interconnections across geographic locations and between deposit and loan markets such that local shocks to deposits or loans can affect endogenously the volume of loans and deposits in every local market. The authors develop a structural model of bank oligopoly competition for both deposits and loans in multiple geographic markets. They use the model to perform a counterfactual in which they evaluate the impact of Riegle-Neal on the geographic imbalance of deposits and loans. The authors operationalize this counterfactual by dividing every multi-state bank in into different independent banks, one for each state. They find that the geographic expansion of branch networks allowed for by deregulation had a significant positive effect on the geographic flow of credit, but benefited mostly larger/richer counties.

Competition from shadow banks

In the last fifteen years competition for traditional banks has also started to come from *shadow banks*: non-bank financial intermediaries operating outside the purview of traditional banking regulation. Shadow banks provide credit, but do not rely on deposits for their funding, turning instead to wholesale funding and securtization. They lack access to central bank liquidity or public sector credit guarantees (see FSB.org).

According to Buchak et al. (2018), between 2007 and 2015 shadow banks' share of mortgage originations roughly doubled from 30% to 50%. The authors point out that the growth of online *Fintech* lenders has been particularly dramatic, accounting for roughly 25% of shadow bank origination by 2015. The authors evaluate to what extent technological change rather than regulatory arbitrage can explain the rise of shadow banks. As pointed out in Fuster et al. (2019), Fintech lenders can process mortgage applications significantly faster than can traditional lenders, without leading to a higher incidence of default, and so the use of improved technology is one channel through which, shadow banks could increase their market share. Alternatively, shadow banks may have engaged in *regulatory arbitrage* and filled a void left when traditional bank presence contracted due to the increased legal and regulatory burden faced by traditional banks in the form of increased capital requirements, mortgage servicing rights, mortgage-related lawsuits, and the movement of supervision to Office of Comptroller and Currency following closure of the Office of Thrift Supervision.

Buchak et al. (2018) develop a simple model of intermediary choice that illustrates the migration to shadow banking and highlights the role of technology and regulation. To do so they make use of the HMDA data set. In their model borrowers make discrete choices between three types of banks: traditional, Fintech shadow banks, and non-Fintech shadow banks. This requires them to classify mortgage lenders as either shadow or traditional banks, and then within shadow banks to determine which are Fintech and which are not. According to their classification, a bank is traditional if it accepts deposits and shadow otherwise. It is Fintech if it has a strong online presence and the vast majority of its mortgage application process takes place online without human involvement (from the lender). The choice over lenders is based on the mortgage rate each offers and various attributes, including convenience. On the supply side, lenders set mortgage rates to maximize profits, taking as given the structure of the market. Finally, there is free entry into the market for each of the three types of banks. The model is then calibrated to

the conforming loan market. Their findings suggest that non-Fintech shadow banks offer lower quality service than do traditional banks. Fintech gains market share through higher quality and online convenience. They find little role for funding or entry costs in explaining the rise of shadow banks. The time period coinciding with the implementation of Dodd-Frank and Basel III along with an increase in mortgage lawsuit activity is found to coincide with a period of increased regulatory burden. Overall they find that regulation accounts for approximately 60% of the growth of shadow banks, while technological advantages can be shown to explain roughly 30%.

Buchak et al. (2020) documents that this migration from traditional to shadow banks, occurred mostly for loans that were easily sold, but less so for loans that were more balance-sheet intensive (i.e. jumbo loans), where the market share for traditional banks remained at around 90% over this time period. The authors extend Buchak et al. (2018), developing a more comprehensive model of the lending market that highlights the role of these two features and makes clear that they are important for understanding the consequences of capital-requirement and monetary policies. In their model potential borrowers demand loans from different providers present in their local market. The authors allow borrowers to not only choose a lender, but also the size of their loan. This is important in order to endogenize the choice of a conforming or jumbo loan. On the supply side three types of lenders are present in each local market: traditional, non-Fintech shadow and Fintech shadow. Lenders choose interest rates along with what fraction of mortgage dollars to retain on their balance sheet and how many to finance through GSEs. This split is allowed to depend on the cost of portfolio lending, which, in turn, is allowed to differ across the lender types. Different lender types also face different levels of regulatory burden, which the authors model in a reduced-form way. They then use the model to perform counterfactuals in which they (i) allow for changes to capital requirements, (ii) target secondary market interventions, and (iii) reform conforming loan limits. The main takeaway is that it is important to take into account the shadow bank migration and balance sheet retention margins when considering the impact of these policies. Failure to do so will lead to incorrect conclusions about the link between them and financial stability.

The movement towards shadow banks has important consequences for stability. Jiang et al. (2020) use call report data for shadow banks acquired through access to information requests to point out that the capital structure choices of shadow banks are similar to those of pre deposit insurance banks. There is also a further competition angle since, as pointed out in Jiang (2020), many shadow banks are funded in part through warehouse loans from traditional banks possibly leading to strategic motives on the part of traditional banks.

Similar growth of less regulated players has occurred elsewhere too. Koijen and Yogo (2016, 2015) document that between 2002 and 2012 US life insurance and annuity liabilities ceded to shadow re-insurers increased by $353 billion, from $11 billion to $364 billion.

6.2 Lending price regulations

Price regulations are is very common in retail credit markets, and are used by regulators to alleviate the adverse effects of price dispersion by limiting the ability of lenders to price discriminate, which can reduce access to credit for borrowers with lower wealth. An important example of this is the presence of usury laws. As we saw earlier, the main downside of price ceilings in lending markets is the possibility of market unraveling due to asymmetric information. For instance Einav et al. (2012) note that usrury laws forces subprime lenders to impose limits on loan size in order to screen high-risk borrowers. Kawai et al. (2020) show that this "pooling" of high-risk borrowers can be particular severe in online platform like Prosper.com, which have limited tools to screen risk.

Cuesta and Sepúlveda (2019) measure the tradeoff between market power and adverse selection, and study how interest-rate ceilings affect the equilibrium supply of credit and consumer welfare. Their counterfactual simulations show that, on average, the reduction in price discrimination is more than offset by a large reduction in the supply of credit, measured by the probability of qualifying for a loan. Note that this large supply reduction is mostly driven by heterogeneity in lending costs across borrowers, as opposed to an unravelling effect associated with adverse selection.

Galenianos and Gavazza (2020) use a search framework with posted prices to analyze the effect of credit-card interest rate ceilings. Instead of haggling with banks to negotiate the best terms, consumers are randomly matched with lenders who compete by sending different posted-price offers. This modelling choice is sensible in the context of the credit-card market, in which banks send pre-approved offers to consumers based on a coarse set of risk characteristics. In this environment, price ceilings affect the equilibrium supply of credit by forcing high-cost lenders to exit the market (i.e. fewer unsolicited offers), and by affecting the incentive of consumers to search (which in turns affects the demand elasticity of each lenders). While the paper abstracts from adverse selection, the presence of fixed costs of sending pre-approved offers can in theory lead to a sufficiently large reduction in supply. As in Cuesta and Sepúlveda (2019), this can lead lower surplus for consumers.

Under the second mechanism, the imposition of a price ceiling tends to incentivize more consumers to search, since the regulation eliminates high-rate offers. This force is particularly important for risky borrowers who are more likely to face a truncated offer distribution. The authors find that in equilibrium, the supply reduction is not sufficiently large to off-set the reduction in prices caused by the price ceiling. As a result the imposition of a price ceiling is predicted to increase the surplus of most consumers.

The previous examples study regulations in which lenders are unable to make different offers to consumers based on their WTP and/or cost differences. Other regulations limit the ability of lenders to *raise* prices after origination. This type of policy is specifically designed to curb market power originating from switching cost and repeated interactions between borrowers and lenders. In markets for unsecured loans, such as credit cards or lines of credit, the decision to borrow and repay past loans has dynamic implications for future opportunities, for instance because of switching costs and/or the fact that banks learn about borrowers' risk types over time.

Nelson (2020) analyzes the importance of adverse selection and learning in the context of the U.S. credit card market.[50] Prior to the implementation of the CARD Act in 2009, credit card lenders were free to adjust lending terms as a function of past borrowing and repayment decisions by consumers. The ability to set individual-level prices serves two purposes. On the one hand, due to the presence of switching costs, lenders can price discriminate against consumers that are locked into a lender, for instance by raising fees on late payments. On the other, the ability of lenders to raise prices on risky borrowers after origination alleviates the adverse selection problem created by the fact that borrowers are better informed about default risk when receiving pre-approved credit-card offers. When asymmetric information is severe, the ability to set prices based of private information can prevent the market from unravelling, and, in equilibrium, benefits high-risk consumers.

Nelson studies the relative importance of the discrimination and adverse selection channels by estimating a rich model of demand and supply for credit cards. On the demand side, the model extends the previous framework to a dynamic environment using a dynamic discrete-choice model in which consumers make borrowing/repayment decisions, and choose whether or not

[50] See Xin (2020) and Tian (2021) for recent papers studying the dynamics of lending decisions in environments with incomplete information and learning.

to switch lenders. The market suffers from an adverse-selection problem, since the willingness to pay of consumers for credit-card borrowing is positively correlated with default risk (both observed and unobserved). On the supply side, banks compete by posting introductory offers based on observed risk factors. In contrast, incumbent banks are fully informed about consumers' risk types, and are therefore able to manage their risk by increasing interest rates for consumers with high (unobserved) default risks.

In this environment, adverse selection depends on the importance of competition between lenders at the introductory-rate offer stage. By offering attractive rates, banks "steal" the riskiest borrowers from their rivals. This lemons problem is amplified when banks are unable to adjust prices after observing borrowers' types. The author simulates the effect of a price regulation, similar to the CARD Act, which forces lenders to commit to an interest rate prior to observing borrowers' risk types. In this case, prices reflect solely observed risk factors (e.g. FICO scores), which substantially reduces the amount of dispersion in borrowing terms. The results show that this "pooling" equilibrium induces a partial unravelling of the market segment for sub-prime borrowers. However, the reduction in the ability of banks to price discriminate reduces offers to most borrowers in the medium to high credit-score segments, leading to an expansion in borrowing and consumer welfare. These model predictions echo earlier results obtained by Agarwal et al. (2015), who use a difference-in-differences approach to estimate the impact of the CARD act on transaction rates and lending volume.

6.3 Intermediaries and agency problems

So far, in our discussion of retail lending and funding markets we have assumed that consumers select firms and products based on prices and characteristics. However, in practice, financial transactions are often intermediated by third-party firms responsible for helping consumers to find the best investment or lending options.

In the US mortgage market, roughly 50% of loans are originated via third-party firms, either brokers or correspondent lenders. Mortgage brokers and correspondents have very similar roles in the lending process, with the primary differences being that brokers generally do not fund loans and are working on behalf of the lender or borrower, whereas correspondents fund and close loans in their own name and subsequently sell the loan to a wholesale lender. In both cases, the intermediary works on behalf of borrowers to find the lowest-cost lender and to complete the loan application and closing process. Similarly, more than half of households rely on financial advisors and/or brokers when investing their savings in mutual funds or annuities. These advisors work for one or more fund managers, and are most often compensated by the fund manager in the form of commissions or fixed payments.

The fact that these intermediaries are more informed about the cost and availability of different products creates a potential agency problem. In the lending case, brokers and correspondents are incentivized to find a lender and/or product that yields the largest expected revenue, and not necessarily the largest surplus for consumers. In the case of investments, advisors are incentivized to sell revenue-generating products, and are often accused of misconduct, for instance of not sharing relevant information about expense ratio fees with consumers.

Egan et al. (2019) provide a descriptive analysis of the occurrence of misconduct among financial advising firms in the US. They find that 1 in 13 advisors are guilty of a least one offense, and that financial misconduct is concentrated in a (small) group of firms that have lax hiring and firing standards. Consistent with the hypothesis that financial misconduct is caused by asymmetric information between consumers and advisors, the authors show that misconduct is concentrated among firms primarily targeting small investors and active in counties with a less educated population.

Woodward and Hall (2012) describe the pricing and compensation of brokers in the US.[51] Brokers base their decisions on the rate sheets of different lenders. A rate sheet provides the price that lenders are willing to pay to acquire a loan with fixed characteristics (e.g. FICO, LTV, loan size, etc.), along with a discrete grid of interest rates and lock-in periods (typically between 30 and 90 days). The price, or yield-spread premium (YSP), corresponds to the main compensation for brokers and correspondent lenders, and is increasing in rates and decreasing in the lock-in period. As a result, everything else being equal, brokers have an incentive to offer a high-rate contract to borrowers, subject to the constraint that the contract satisfies the underwriting criteria of wholesale lenders (e.g. debt-to-income ratio). Consumers with limited information about origination costs and broker compensation are likely to search an inefficiently small number of brokers and as a result will fail to negotiate lower closing fees on high-rate mortgages. Woodward and Hall show that consumers in the top percentiles of the distribution of closing fees fail to negotiate discounts off high-rate contracts.

This adverse effect of intermediaries is balanced by pro-competitive effects and lower transaction costs. For instance, a mortgage broker acts as a middle-man between borrowers and lenders. As such, brokers can more efficiently compare the price of multiple lenders at once, which ultimately increases competition in the market and can lead to lower prices. Robles-Garcia (2019) analyses this tradeoff between competition and agency costs.

Her model works as follows. Consumers make a choice between shopping for a mortgage on their own (direct or retail channel), or visiting a broker at a lower cost. When searching on their own, consumers pick the lender and product that maximizes their expected utility. When searching with a broker, consumers are offered a product that maximizes a weighted average of their own utility, and the payoff of brokers (i.e. commission). The weight quantifies the importance of agency frictions. As brokers put more weight on commissions, consumers end up borrowing from lenders with potentially higher rates and/or lower "quality."

A novel feature of Robles-Garcia's model, is the idea that smaller lenders only have access to consumers via brokers, as opposed to vertically integrated banks that can originate loans through their own networks of loan officers. By offering higher commissions to brokers, these lenders can originate more loans. Therefore, the "conflict of interest" that arises because brokers care about commissions also increases the supply of loans in the market. This pro-competition effect attenuates the importance of agency costs in the market.

The relative importance of these two channels depends on how the surplus is split between brokers and wholesale lenders. The author uses a rich data set of broker commission rates that are broker/lender specific, as well as information on the network of lenders dealing with each broker. The model uses this information to infer the relative bargaining power of each party, using a Nash-in-Nash bargaining model (as in Gowrisankaran et al. (2015)).

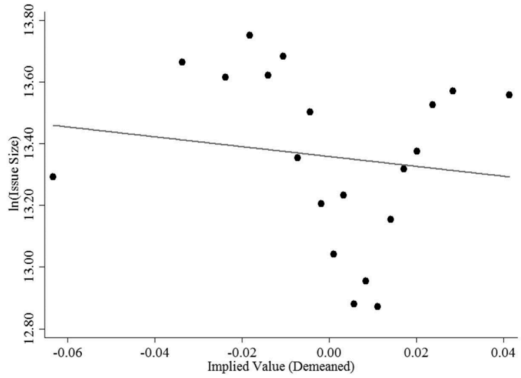
The paper analyzes the effect of price-control regulations imposing ceilings on broker commissions. The goal of these regulations is to reduce transaction costs in the market by attenuating the agency friction. However, accounting for competition between lenders for brokers' business, these proposed regulations can prevent lenders from price discriminating, and limit the ability of small lenders to compete with vertically integrated lenders. The equilibrium effect is therefore ambiguous. In the context of the UK market, Robles-Garcia finds that restrictions on commission rates tend to raise rates for the "Big-6" lenders.

Measuring the importance of financial misconduct using data on the type of products that consumers select is challenging. The fact that two observationally equivalent consumers purchase products with different values can be due to unobserved heterogeneity, and/or unobserved

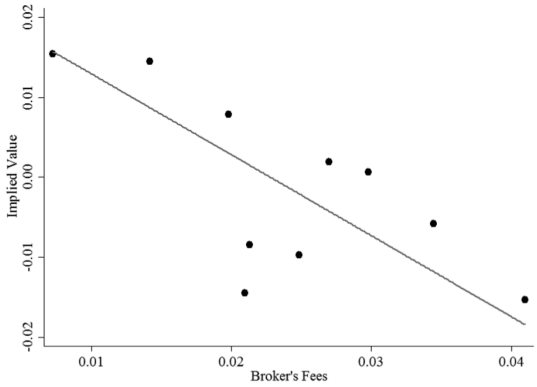
[51] The pricing and compensation of correspondent lenders has a similar structure, but is less transparent from the point of view of the regulator.

Figure 13. Expected return and broker compensation for reversible bonds

(a) Issuance size and bond value



(b) Bond value and broke fees



Source: Egan (2019), Figures 4 and 5.

differences in the availability of financial products. For instance, a consumer dealing with a mortgage broker (as opposed to a vertically integrated lender) could be selecting a high-rate product because he/she did not qualify at the low-rate product available at competing lenders. In the case of financial advisors, limited inventories of bonds can cause consumers to purchase sub-optimal products.

Egan (2019) provides evidence that low-return investment products are purchased at a higher rate. Figure 13a illustrates this point in the context of homogeneous reversible bonds issued by investment banks and sold to retail investors by independent brokers. Assuming that the same products are available to all consumers, this "upward sloping" demand function clearly indicates that many investors are steered towards dominated products. Figure 13a shows that these products are also more profitable for brokers, leading to a conflict of interest between brokers and investors.

As discussed in Section 4, Egan uses a search model à la Hortaçsu and Syverson (2004) to rationalize these two empirical facts. On the modeling-side, the main innovation is to allow brokers to use dominated/superior products as a way to price discriminate across consumers, selling high-fee dominated products to unsophisticated consumers (infinite search cost) and low-fee superior products to sophisticated consumers (positive search cost). In this context, regulatory interventions that would "force" brokers to offer only the best products to consumers can lead to substantial welfare gains.

Bhattacharya et al. (2020) analyze this question by measuring the equilibrium effect of regulations aimed at imposing Fiduciary duties on financial advisors.[52] The paper focuses on the advice given by resident investment advisers (RIAs) and broker-dealers (BDs). Both types of intermediary are compensated via commissions, which potentially creates a conflict of interest. In the United States, RIAs have a fiduciary duty toward clients mandated at the national level, while licensed BDs do not. However, several states have established fiduciary duty for BDs within their borders. Despite this difference in fiduciary duty regulations, RIAs and BDs perform roughly the same function for retail investors and the annuities they offer have similar fees and contract characteristics.

In theory, a federal fiduciary-duty law can impact positively the return that savers earn on

[52] See also Egan et al. (2020) for a related analysis of the impact of a 2016 Department of Labor proposal to impose a fiduciary standard on brokers selling variable annuities.

their investments, raising the cost of offering low-return products to clients ("advice channel"). However, by reducing the profit of financial advisors, these laws also have the potential to limit entry into the market and reduce the overall quality of financial products available in the market ("fixed-cost channel"). The authors provide descriptive evidence in favor of quality improvement by comparing the average returns of consumers living on both sides of state boundaries with different fiduciary duty laws. They find that broker-dealers subject to fiduciary duty sell annuities with risk-adjusted returns that are 25 basis points higher than those who are not subject to the regulation. In contrast, fiduciary duty causes a small and imprecisely estimated reduction in the number of broker-dealers, suggesting that the fixed cost channel is limited.

To quantify the equilibrium effect of a federal fiduciary-duty law, the authors estimate the parameters of a two-stage game in which brokers first choose to enter the market, then compete for retail investors by choosing the level of "advice" quality that they provide. The optimal service quality supplied by advisors depends on an unobserved type capturing the ability of advisors to provide quality advice to consumers, as well as on the regulation, which imposes an additional cost for offering low quality advice. In this context, market structure affects the equilibrium quality of advice by increasing the return on advisors' effort, as well as by reducing the probability of matching with consumers. Depending on the effect of fiduciary duty on the marginal cost of providing quality advice and/or on the fixed cost of operating in the market, these two competitive effects can lead to an improvement or a deterioration of the distribution of quality post-policy reform. The results of this structural exercise show that improving the stringency of fiduciary standards increases the quality of financial advice offered by brokers. This is despite the fact that the policy would lead to the exit of a (small) number of brokers.

7 CONCLUSION

In this chapter, we provided an overview of the literature that exists at the intersection of industrial organization and finance. A series of key factors have to led to a rapid increase in the number of papers that have been written that apply the tools and methods from industrial organization to study financial markets. First, regulatory and technological changes have led to a rise in concentration in a number of key markets and exacerbated market power concerns. Second, the upheaval caused by the financial crisis caused policy makers to implement many important changes in these markets. Finally, the improved availability of financial-market data has made it easier to examine these issues. Together, these have led researchers to study the role that various frictions such as asymmetric information and market power play in explaining deviations from the law of one price, and to consider the appropriateness of alternative price-setting mechanisms.

Our objective has been twofold. First, we provided a detailed description of how empirical industrial organization models and methods can be extended to account for the specificities of financial markets – for instance for the use of empirical auction models to study the demand for government securities, models of demand and supply for differentiated products that account risk and fragility, and models of price competition with endogenous selection and information acquisition costs. Our second objective was to provide a birds-eye view of empirical papers using those techniques, and how the results relate to important policy questions that are central to the regulation of financial markets. Naturally, we have not managed to cover all aspects of financial markets. Our focus has been on the flow of funds from retail funding markets to lending and government bonds. As such, we abstracted away from equity and corporate bonds markets, as well as the supply for annuities and insurance products. Much of the discussion has concentrated on the role of banks, but the tools and methods we describe can and have been applied to study other funding and credit markets too.

Despite these advances, the literature is still in its infancy. The complexity of the institutions

determining the interactions between firms, consumers and regulators still represents an important barrier to entry for researchers. In particular, we identify a number of lines of research that, in our opinion, have not been paid sufficient attention by industrial organization economists.

First, the analysis of primary markets has mostly taken market structure as given, but it would be useful to know more about the effect of market design on participation and concentration. For instance, as mentioned in Section 3, it would be important to understand whether or not large bidders such as BlackRock should be allowed to participate indirectly in primary auctions, thereby revealing their bids to a particular primary dealer.

Second, the models of credit supply that we have discussed take the distribution of lending costs as a primitive. However, most lending markets rely on active secondary markets, which tend to be significantly more concentrated than their retail counterparts. For instance, in the U.S. mortgage markets, the share of the top-4 banks at the securitization stage was 40% in 2017, compared to less than 15% at the origination stage (i.e. retail). Similar levels of upstream concentration are observed in other markets that rely on networks of independent lenders and brokers to originate their loans (see for instance Grunewald et al. (2019)). The extent to which it makes sense to think of the securitization stage as competitive (and hence to take the lending cost as a primitive) is still an open question. As with any market in which firms have market power at both ends of the supply chain, the effect of concentration and the pass-through of cost shocks on retail prices is ambiguous.

Third, an important avenue for future research lies at the intersection of industrial organization and monetary policy. Central banks influence interest rates by intervening in interbank lending markets. Given the two-sided nature of financial markets, this can affect final interest rates either through a "deposit" or "credit" channel. Recent empirical work in finance and macro has provided evidence on the importance of both channels, and highlighted the possible role for imperfect competition and agency frictions (see for instance Drechsler et al. (2017) and Jiménez et al. (2014)). IO models and methods are only starting to be used to analyze the mechanisms explaining these empirical relationships, and hopefully provide a guide for monetary policy authorities.

Fourth, the growth of online platforms for both funding and lending has significantly transformed the industry, perhaps more so than for other retail markets. For instance, while Amazon's share of retail spending is roughly 10%, online lending specialists like Quicken Loans now originate the largest share of mortgages in the US. The entry of online deposit and saving institutions has similarly changed the way banks raise funds from consumers, by limiting the importance of physical networks. While there has been a large number of papers studying the adoption by consumers of online finance services, there exist few studies of the equilibrium effects of these technologies on the supply of credit and on the overall productivity of the industry.

Finally, this discussion highlights the need for more research that examines the interactions between the different markets we have studied in this chapter. As mentioned above, only a small number of papers have considered the linkages between the deposit and credit markets (Corbae and D'Erasmo (2021), Aguirregabiria et al. (2019), Wang et al. (2019)), and there have been none linking either of these markets with wholesale funding markets. Important interconnections exist and so studies that take these into account would be valuable.

Acknowledgements: We would like to thank Victor Aguirregabiria, Jason Allen, Dean Corbae, Ali Hortaçsu, Ralph Koijen, Shaoteng Li, Nicola Pavanini, Eric Richert, and an anonymous reviewer for helpful comments.

REFERENCES

Adams, R., Brevoors, K., and Kiser, E. (2007). Who competes with whom? The case of depository institutions. *Journal of Industrial Economics*, 55(1):141–167.

Adams, W., Einav, L., and Levin, J. (2009). Liquidity constraints and imperfect information in subprime lending. *American Economic Review*, 99(1):49–84.

Agarwal, S., Chomsisengphet, S., Mahoney, N., and Stroebel, J. (2015). Regulating consumer financial products: Evidence from credit cards. *Quarterly Journal of Economics*, pages 111–164.

Agarwal, S., Grigsby, J., Hortaçsu, A., Matvos, G., and Seru, A. (2020). Searching for approval. NBER working paper 27341.

Aguirregabiria, V., Clark, R., and Wang, H. (2016). Diversification of geographic risk in retail bank networks: Evidence from bank expansion after the Riegle-Neal Act. *RAND Journal of Economics*, 47(3):529–572.

Aguirregabiria, V., Clark, R., and Wang, H. (2019). The geographic flow of bank funding and access to credit: Branch networks, local synergies and competition. Working paper, University of Toronto.

Alexandrov, A. and Koulayev, S. (2017). No shopping in the u.s. mortgage market: Direct and strategic effects of providing information. CFPB No. 2017-01.

Allen, F. and Gale, D. (2004). Competition and financial stability. *Journal of Money, Credit and Banking*, 36(3).

Allen, J., Clark, R., Hickman, B., and Richert, E. (2021a). Resolving failed banks: Uncertainty, multiple bidding & auction design. Working paper.

Allen, J., Clark, R., and Houde, J. F. (2014a). The effect of mergers in search markets: Evidence from the Canadian mortgage industry. *American Economic Review*, 104:3365–3396.

Allen, J., Clark, R., and Houde, J. F. (2014b). Price dispersion in mortgage markets. *Journal of Industrial Economics*, 62:377–416.

Allen, J., Clark, R., and Houde, J.-F. (2019). Market power and search frictions in negotiated-price markets. *Journal of Political Economy*, 127(4):1550–1598.

Allen, J., Hortaçsu, A., and Kastl, J. (2021b). Crisis management in Canada: Analyzing default risk and liquidity demand during financial stress. *AEJ: Microeconomics*.

Allen, J., Kastl, J., and Wittwer, M. (2020). Primary dealers and the demand for government debt. working paper.

Allen, J. and Li, S. (2020). Dynamic competition in negotiated price markets. Working paper, Bank of Canada.

Allen, J. and Wittwer, M. (2020). Centralizing over-the-counter markets? working paper.

Ambokar, S. and Samaee, K. (2019). Inaction, search costs and market power in the US mortgage market. Working paper, University of Pennsylvania.

Argyle, B., Nadauld, T., and Palmer, C. (2020). Real effects of search frictions in consumer credit markets. Working paper, MIT Sloan.

Armantier, O. and Copeland, A. (2015). Challenges in identifying interbank loans. *Economic Policy Review*, 21(1):pp.1–17.

Arnone, M. and Iden, G. (2003). Primary dealers in government securities: Policy issues and selected countries' experience. IMF Working paper WP/03/45.

Athey, S. and Haile, P. A. (2007). Chapter 60: Nonparametric approaches to auctions. volume 6 of *Handbook of Econometrics*, pages 3847–3965. Elsevier.

Athey, S. and Imbens, G. (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica*, 74(2):431–497.

Ausubel, L. M. (1991). The failure of competition in the credit card market. *American Economic*

Review, 81(1):50–81.

Ausubel, L. M. (1999). Adverse selection in the credit card market. NBER working paper.

Begenau, J. (2020). Capital requirements, risk choice, and liquidity provision in a business cycle model. *Journal of Financial Economics*, 32.

Begenau, J. and Landvoigt, T. (2020). Financial regulation in a quantitative model of the modern banking system. Working paper.

Benetton, M. (2018). Leverage regulation and market structure: A structural model of the uk mortgage market. Working paper, Haas School of Business.

Berger, A. and Hannan, T. (1989). The price-concentration relationship in banking. *Review of Economics and Statistics*, 71:291–299.

Berry, S. (1994). Estimating discrete-choice models of product differentiation. *RAND Journal of Economics*, 25(2):242–262.

Berry, S., Levinsohn, J., and Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica*, 63(4):pp.841–890.

Bhattacharya, V., Illanes, G., and Padi, M. (2020). Fiduciary duty and the market for financial advice. Working paper, Northwestern University.

Bhutta, N., Fuster, A., and Hizmo, A. (2019). Paying too much? price dispersion in the us mortgage market. Working paper, Federal Reserve Board.

Boyd, J. H. and De Nicoló, G. (2005). The theory of bank risk taking and competition revisited. *Journal of Finance*, LX(3).

Brancaccio, G., Li, D., and Schürhoff, N. (2019). Learning by trading: The case of the US market for municipal bonds. working paper.

Brannan, L. and Froeb, L. M. (2000). Mergers, cartels, set-asides, and bidding preferences in asymmetric oral auctions. *The Review of Economic and Statistics*, 82(2):283–290.

Bresnahan, T. (1987). Competition and collusion in the american automobile industry: The 1955 price war. *Journal of Industrial Economics*, 35:457–482.

Bresnahan, T. and Reiss, P. (1991). Entry and competition in concentrated markets. *Journal of Political Economy*, 99:977–1009.

Bresnahan, T. F. (1989). Empirical studies of industries with market power. In Schmalensee, R. and Willig, R., editors, *Handbook of Industrial Organization*, volume 2. Elsevier.

Buchak, G., Matvos, G., Piskorski, T., and Seru, A. (2018). Fintech, regulatory arbitrage, and the rise of shadow banks. *Journal of Financial Economics*, 130(3):453–483.

Buchak, G., Matvos, G., Piskorski, T., and Seru, A. (2020). Beyond the balance sheet model of banking: Implications for bank regulation and monetary policy. NBER working paper No. 25149.

Budish, E., Cramton, P., and Shim, J. (2015). The high-frequency trading arms race: Frequent batch auctions as a market design response. *Quarterly Journal of Economics*, 130:pp. 1547–1621.

Budish, E., Lee, R., and Shim, J. (2019). Will the market fix the market? A theory of stock exchange competition and innovation. Working Paper.

Campbell, J. and Cocco, J. (2015). A model of mortgage default. *Journal of Finance*, 70.

Cassola, N., Hortaçsu, A., and Kastl, J. (2013). The 2007 subprime market crisis in the EURO area through the lens of ECB repo auctions. *Econometrica*, 81(4):pp. 1309–1345.

Chiappori, P.-A. and Salanié, B. (2000). Testing for asymmetric information in insurance markets. *Journal of Political Economy*, 108(1):56–78.

Chu, S. and Rysman, M. (2019). Competition and strategic incentives in the market for credit ratings: Empirics of the financial crisis of 2007. *American Economic Review*, 109(10):3514–55.

Cohen, A. and Mazzeo, M. (2007). Market structure and competition among retail depository

institutions. *The Review of Economics and Statistics*, (1):60–74.

Cooper, R. and Ross, T. (2002). Bank runs: Deposit insurance and capital requirements. *International Economic Review*, 43:55–72.

Corbae, D. and D'Erasmo, P. (2021). Capital buffers in a quantitative model of banking industry dynamics. Forthcoming, *Econometrica*.

Corbae, D. and Levine, R. (2020). Competition, stability, and efficiency in the banking industry. Working paper.

Crawford, G., Pavani, N., and Schivardi, F. (2018). Asymmetric information and imperfect competition in lending markets. *American Economics Review*, 108(7):1659–1701.

Cuesta, J. I. and Sepúlveda, A. (2019). Price regulation in credit markets: A trade-off between consumer protection and credit access. Working paper, Stanford university.

De Loecker, J., Eeckhout, J., and Unger, G. (2020). The rise of market power and the macroeconomic implications. *Quarterly Journal of Economics*, 135:561–644.

Degryse, H., Kim, M., and Ongena, S. (2009). *Microeconometrics of Banking*. Oxford Univerisity Press.

Diamond, D. W. (1984). Financial intermediation and delegated monitoring. *Review of Economic Studies*, LI:393–414.

Diamond, D. W. and Dybvig, P. (1983). Bank runs, deposit insurance, and liquidity. *Journal of Political Economy*, 91(3):401–419.

Diamond, P. A. (1982). Aggregate demand management in search equilibrium. *Journal of Political Economy*, 90(5):881–894.

Dick, A. A. (2006). Nationwide branching and its impact on market structure, quality, and bank performance. *Journal of Business*, 79:1661–1676.

Dick, A. A. (2008). Demand estimation and consumer welfare in the banking industry. *Journal of Banking and Finance*, 32:1661–1676.

Drechsler, I., Savov, A., and Schnabl, P. (2017). The deposits channel of monetary policy. *Quarterly Journal of Economics*, pages 1819–1876.

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

Egan, M. (2019). Brokers versus retail investors: Conflicting interests and dominated products. *Journal of Finance*, LXXIV(3).

Egan, M., Hortaçsu, A., and Matvos, G. (2017). Deposit competition and financial fragility: Evidence from the US banking sector. *American Economic Review*, 107(1):169–216.

Egan, M., Matvos, G., and Seru, A. (2019). The market for financial adviser misconduct. *Journal of Political Economy*, 127(1).

Egan, M., Shan, G., and Tang, J. (2020). Conflicting interests and the effect of fiduciary duty — evidence from variable annuities. National Bureau of Economic Research working paper 27577.

Einav, L., Finkelstein, A., and Mahoney, N. (2021). The io of selection markets. Forthcoming, Handbook of Industrial Organization.

Einav, L., Jenkins, M., and Levin, J. (2012). Contract pricing in consumer credit markets. *Econometrica*, 80(4):1387–1432.

Einav, L., Jenkins, M., and Levin, J. (2013). The impact of credit scoring on consumer lending. *RAND Journal of Economics*, 44(2):249–274.

Ellickson, P., Houghton, S., and Timmins, C. (2013). Estimating network economies in retail chains: a revealed preference approach. *RAND Journal of Economics*, 44.

Ericson, R. and Pakes, A. (2012). Markov-perfect industry dynamics: A framework for empirical work. *Review of Economic Studies*, 62:53–82.

Ferrari, S., Verboven, F., and Degryse, H. (2010). Investment and usage of new technologies:

Evidence from a shared ATM network. *American Economic Review*, 100(3):1046–1079.
Freixas, X. and Rochet, J. (2008). *Microeconomics of Banking*. MIT Press.
Furfine, C. (1999). The microstructure of the federal funds market. *Financial Markets, Institutions, and Instruments*, 8(5):pp.24–44.
Fuster, A., Plosser, M., Schnabl, P., and Vickery, J. (2019). The role of technology in mortgage lending. *The Review of Financial Studies*, 32(5).
Galenianos, M. and Gavazza, A. (2020). Regulatory interventions in consumer financial markets: The case of credit cards. Working paper, LSE.
Gambacorta, L., Guiso, L., Mistrulli, P., and Tsoy, A. (2020). The cost of steering in financial markets: Evidence from the mortgage market. Working paper.
Garrett, D. G., Ordin, A., Roberts, J., and Serrato, J. C. S. (2020). Tax advantages and imperfect competition in auctions for municipal bonds. working paper.
Gavazza, A. (2011a). Leasing and secondary markets: Theory and evidence from commercial aircraft. *Journal of Political Economy*, 119.
Gavazza, A. (2011b). The role of trading frictions in real asset markets. *American Economic Review*, 101.
Gavazza, A. (2016). An empirical equilibrium model of a decentralized asset market. *Econometrica*, 84.
Gertler, M., Kiyotaki, N., and Prestipino, A. (2016). Chapter 16 - Wholesale banking and bank runs in macroeconomic modeling of financial crises. volume 2 of *Handbook of Macroeconomics*, pages 1345 – 1425. Elsevier.
Glosten, L. R. and Milgrom, P. R. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, 14(1):pp. 71–100.
Gorton, G. (1988). Banking panics and business cycles. *Oxford Economic Papers*, 40.
Gowrisankaran, G. and Holmes, T. (2004). Mergers and the evolution of industry concentration: Results from the dominant-firm model. *RAND Journal of Economics*, 35:561–582.
Gowrisankaran, G. and Krainer, J. (2011). Entry and pricing in a differentiated products industry: Evidence from the atm market. *RAND Journal of Economics*, 42(1):1–22.
Gowrisankaran, G., Nevo, A., and Town, R. (2015). Mergers when prices are negotiated: Evidence from the hospital industry. *American Economic Review*, 175.
Gowrisankaran, G. and Rysman, M. (2012). Dynamics of consumer demand for new durable goods. *Journal of Political Economy*, 120.
Granja, J., Matvos, G., and Seru, A. (2017). Selling failed banks. *The Journal of Finance*, 72(4):1723–1784.
Grunewald, A., Lanning, J., Low, D., and Salz, T. (2019). Auto dealer loan intermediation: Consumer behavior and competitive effects. Working paper, MIT.
Guerre, E., Perrigne, I., and Vuong, Q. (2000). Optimal nonparametric estimation of first-price auctions. *Econometrica*, 68(3):pp. 525–574.
Guren, A., Krishnamurthy, A., and McQuade, T. (2019). Mortgage design in an equilibrium model of the housing market. Working paper.
Gurun, Umit, G., Matvos, G., and Seru, A. (2016). Advertising expensive mortgages. *Journal of Finance*, 16(5):2371–2416.
Hastings, J., Hortaçsu, A., and Syverson, C. (2017). Sales force and competition in financial product markets: The case of Mexico's social security privatization. *Econometrica*, 85(6):1723–1761.
Hendricks, K. and Porter, R. (2007). Chapter 32: An empirical perspective on auctions. volume 3 of *Handbook of Industrial Organization*, pages pp.2073–2143. Elsevier.
Ho, C.-Y. (2015). Switching cost and deposit demand in china. *International Economic Review*, 56(3):723–749.

Ho, K. and Ishii, J. (2011). Location and competition in retail banking. *Industrial Journal of Industrial Organization*, 29(537–546).

Honka, E., Hortaçsu, A., and Vitorino, M. A. (2017). Advertising, consumer awareness, and choice: Evidence from the U.S. banking industry. *RAND Journal of Economics*, 48(3):611–646.

Hortaçsu, A. (2002a). Bidding behavior in divisible good auctions: Theory and empirical evidence from turkish treasury auctions. working paper.

Hortaçsu, A. (2002b). Mechanism choice and strategic bidding in divisible good auctions: An empirical analysis of the turkish treasury auction market. working paper.

Hortaçsu, A. and Kastl, J. (2012). Valuing dealers' informational advantage: A study of Canadian treasury auctions. *Econometrica*, 80(6):pp.2511–2542.

Hortaçsu, A., Kastl, J., and Zhang, A. (2018). Bid shading and bidder surplus in U.S. treasury auction system. *American Economic Review*, 108(1).

Hortaçsu, A. and Syverson, C. (2004). Product differentiation, search costs, and competition in the mutual fund industry: A case study of S&P 500 index funds. *Quarterly Journal of Economics*.

Hugonnier, J., Malamud, S., and Trubowitz, E. (2012). Endogenous completeness of diffusion driven equilibrium markets. *Econometrica*, 80:1249–1270.

Huynh, K., Schmidt-Dengler, P., Smith, G., and Welte, A. (2017). Adoption costs of financial innovation: Evidence from Italian ATM cards. Bank of Canada Staff Working Paper 2017-8.

Illanes, G. (2017). Switching costs in pension plan choice. Working paper, Northwestern University.

Ishii, J. (2007). Compatibility, competition, and investment in network industries: ATM networks in the banking industry. Working Paper.

Jaffee, D. M. and Russell, T. (1976). Imperfect information, uncertainty, and credit rationing. *Quarterly Journal of Economics*, 90(4):651–666.

James, C. and Wier, P. (1987). An analysis of FDIC failed bank auctions. *Journal of Monetary Economics*, 20:141–153.

Jiang, E. (2020). Financing competitors: Shadow banks' funding and mortgage market competition. Working paper.

Jiang, E., Matvos, G., Piskorski, T., and Seru, A. (2020). Banking without deposits: Evidence from shadow bank call reports. Working paper.

Jiménez, G., Ongena, S., Peydró, J.-L., and Saurina, J. (2014). Hazardous times for monetary policy: What do twenty-three million bank loans say about the effects of monetary policy on credit risk-taking. *Econometrica*, 82:463–505.

Kang, B.-S. and Puller, S. (2008). The effect of auction format on efficiency and revenue in divisible goods auctions: A test using korean treasury auctions. *Journal of Industrial Economics*, 56:290–332.

Karlan, D. and Zinman, J. (2009). Observing unobservables: Identifying information asymmetries with a consumer credit field experiment. *Econometrica*, 77(6):1993–2008.

Kastl, J. (2011). Discrete bids and empirical inference in divisible good auctions. *Review of Economic Studies*, 78:pp. 978–1014.

Kastl, J. (2012). On the properties of equilibria in private value divisible good auctions with constrained bidding. *Journal of Mathematical Economics*, 48(6):pp. 339–352.

Kawai, K., Onishi, K., and Uetake, K. (2020). Signaling in online credit markets. working paper, Berkeley University.

Keys, B. J., Pope, D. G., and Pope, J. C. (2016). Failure to refinance. *Journf of Financial Economics*, 122:482–499.

Kiser, E. (2002). Predicting household switching behavior and switching costs at depository

institutions. *Review of Industrial Organization*, 20(4):349–365.

Klein, M. A. (1971). A theory of the banking firm. *Journal of Money, Credit and Banking*, 3(2):pp. 205–218.

Knittel, C. and Stango, V. (2003). Price ceilings as focal points for tacit collusion: Evidence from credit cards. *American Economic Review*, 93(5):1703–1729.

Koijen, R. S. and Yogo, M. (2015). The cost of financial frictions for life insurers. *American Economic Review*, 105(1):445–475.

Koijen, R. S. and Yogo, M. (2016). Shadow insurance. *Econometrica*, 84(3):1265–1287.

Koijen, R. S. and Yogo, M. (2019). A demand system approach to asset pricing. *Journal of Political Economy*, 127(4):pp. 1475–1515.

Koijen, R. S. and Yogo, M. (2021). The fragility of market risk insurance. Forthcoming, *Journal of Finance*.

Krasnokutskaya, E., Li, Y., and Todd, P. E. (2018). Product choice under government regulation: The case of chile's privatized pension system. *International Economic Review*, 59(4):1747–583.

Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica*, 53(6):1315–1335.

Kyle, A. S. (1989). Informed speculation with imperfect competition. *The Review of Economic Studies*, 56(3):pp. 317–355.

Luco, F. (2019). Switching costs and competition in retirement investment. *American economic journal: Microeconomics*, 11(2):26–54.

Mahoney, N. and Weyl, G. E. (2017). Imperfect competition in selection markets. *Review of Economics and Statistics*, 99(4):637–651.

Massoud, N. and Bernhardt, D. (2002). "Rip-Off" ATM surcharges. *RAND Journal of Economics*, 33(1):96–115.

Mazzeo, M. (2002). Product choice and oligopoly market structure. *RAND Journal of Economics*, 33:221–242.

McFadden, D. (1974). *Conditional Logit Analysis of Qualitative Choice Behavior*, pages 1345–1425. Frontiers in Econometrics. Academic Press.

Mólnar, J., Violi, R., and Zhou, X. (2002). Multimarket contact in Italian retail banking: Competition and welfare. *International Journal of Industrial Organization*, 31(5):368–381.

Monti, M. (1972). Deposit, credit and interest rate determination under alternative bank objective functions. Mathematical methods in investment and finance, pages 430–454. Amsterdam: North-Holland.

Morris, S. and Shin, H. (1998). Unique equilibrium in a model of self-fulfilling currency attacks. *American Economic Review*, 88.

Nelson, S. (2020). Private information and price regulation in the US credit card market. Working paper, University of Chicago.

Pakes, A. (2010). Alternative models for moment inequalities. *Econometrica*, 78(6):315–334.

Pakes, A., Porter, J., Ho, K., and Ishii, J. (2015). Moment inequalities and their application. *Econometrica*, 83(1):315–334.

Piskorski, T. and Tchistyi, A. (2017). An equilibrium model of housing and mortgage markets with state-contingent lending contracts. Working paper.

Prager, R. A. and Hannan, T. H. (1998). Do substantial horizontal mergers generate significant price effects? Evidence from the banking industry. *Journal of Industrial Economics*, XLVI:433–452.

Prisman, E., Slovin, M., and Sushka, M. (1986). A general model of the banking firm under conditions of monopoly, uncertainty, and recourse. *Journal of Monetary Economics*, 17:293–304.

Raisingh, D., Houde, J.-F., and Hendricks, K. (2020). Asymmetric information in the wholesale

market for mortgages: The case of ginnie mae loans. working paper, University of Wisconsin-Madison.

Rivers, D. and Vuong, Q. (2002). Model selection tests for nonlinear dynamic models. *Econometrics Journal*, 5:1–39.

Robles-Garcia, C. (2019). Competition and incentives in mortgage markets: The role of brokers. Working paper, Staford GSB.

Rostek, M. and Weretka, M. (2012). Price inference in small markets. *Econometrica*, 80(2):pp.687–711.

Roussanov, N., Ruan, H., and Wei, Y. (2020). Marketing mutual funds. forthcoming *Review of Financial Studies*.

Rubsinstein, A. and Wolinsky, A. (1987). Middlemen. *Quarterly Journal of Economics*, 102(3):581—-593.

Saloner, G. and Shepard, A. (1995). Adoption of technologies with network effects: An empirical examination of the adoption of automated teller machines. *RAND Journal of Economics*, 26(3):479–501.

Salz, T. (2020). Intermediation and competition in search markets: An empirical case study. forthcoming *Journal of Political Economy*.

Sapienza, P. (2002). The effects of banking mergers on loan contracts. *Journal of Finance*, LVII(1).

Seim, K. (2006). An empirical model of firm entry with endogenous product-type choices. *RAND Journal of Economics*, 37(3):619–640.

Shy, O. (2002). A quick-and-easy method for estimating switching costs. *International Journal of Industrial Organization*, 20(1):71–87.

Stango, V. and Zinman, J. (2016). Borrowing high versus borrowing higher: Price dispersion and shopping behavior in the u.s. credit card market. *Review of Financial Studies*, 29(4).

Stiglitz, J. E. and Weiss, A. (1981). Credit rationing in markets with imperfect information. *American Economic Review*, 71(3):393–410.

Thakor, A. and Boot, A., editors (2008). *Handbook of Financial Intermediation and Banking*. Elsevier.

Tian, P. (2021). The role of long-term contracting in business lending. working paper, University of Wisconsin-Madison.

Vij, S. (2018). Acquiring failed banks. Working paper.

Vives, X. (2011). Strategic supply function competition with private information. *Econometrica*, 79(6):pp.1919–1966.

Vives, X. (2016). *Competition and Stability in Banking*. Princeton University Press.

Wang, H. and Ching, A. (2019). Consumer valuation of network convenience: Evidence from the banking industry. Working paper, John Hopkins University.

Wang, Y., Whited, T. M., Wu, Y., and Xiao, K. (2019). Bank market power and monetary policy transmission: Evidence from a structural estimation. Working paper, University of Illinois.

Wilson, R. (1979). Auctions of shares. *The Quarterly Journal of Economics*, 93(4):pp. 675–689.

Woodward, S. and Hall, R. E. (2012). Diagnosing consumer confusion and sub-optimal shopping effort: Theory and mortgage-market evidence. *American Economic Review*, 102(7):3249–3276.

Xiao, K. (2019). Monetary transmission through shadow banks. *Review of Financial Studies*.

Xin, Y. (2020). Asymmetric information, reputation, and welfare in online credit markets. Working paper, CALTECH.

Yang, B. and Ching, A. (2014). Dynamics of consumer adoption of financial innovation: The case of ATM cards. *Management Science*, 60.