1-2023

From market making to matchmaking: Does bank regulation harm market liquidity?

Gideon SAAR
Cornell University

Jian SUN
Singapore Management University, jiansun@smu.edu.sg

Ron YANG
Stanford University

Haoxiang ZHU
NBER

From Market Making to Matchmaking: Does Bank Regulation Harm Market Liquidity?

Gideon Saar, Jian Sun, Ron Yang, and Haoxiang Zhu*

This version: September 2020

Abstract

Post-crisis bank regulations raised market-making costs for bank-affiliated dealers. We show that this can, somewhat surprisingly, improve overall investor welfare and reduce average transaction costs despite the increased cost of immediacy. Bank dealers in OTC markets optimize between two parallel trading mechanisms: market making and matchmaking. Bank regulations that increase market-making costs change the market structure by intensifying competitive pressure from non-bank dealers and incentivizing bank dealers to shift their business toward matchmaking. Thus, post-crisis bank regulations have the (unintended) benefit of replacing costly bank balance sheets with a more efficient form of financial intermediation.

Keywords: bank regulation, market making, matchmaking, financial crisis, corporate bonds, liquidity, over-the-counter markets, broker-dealers, Basel 2.5, Basel III, Volcker Rule, post-crisis regulation, market microstructure

JEL Classification: G01, G12, G21, G24, G28

*Gideon Saar is from the Johnson Graduate School of Management, Cornell University (gs25@cornell.edu). Jian Sun is from MIT Sloan School of Management (jiansun@mit.edu). Ron Yang is from Harvard Business School (rnyang@g.harvard.edu). Haoxiang Zhu is from MIT Sloan School of Management and the NBER (zhuh@mit.edu). We thank Philip Bond, Briana Chang, Christopher Hrdlicka, Stacey Jacobsen, Pete Kyle, Jonah Platt, Dimitri Vayanos, Kumar Venkataraman, Yajun Wang, Yao Zeng, Zhuo Zhong, and Hao Zhou as well as seminar/conference participants at the American Finance Association 2020 meeting, the Chinese University of Hong Kong, the Chinese University of Hong Kong Shenzhen, the City University of Hong Kong, Cornell University, the Financial Industry Regulatory Authority, Manchester Business School, MIT, the PBC School of Finance, University at Buffalo, the University of British Columbia, the University of Illinois, the University of Washington, Warwick Business School, Wharton, the 4th Sydney Market Microstructure Meeting, and the mini-conference on Regulatory Reform at the University of Wisconsin-Madison for helpful comments.

1 Introduction

The aftermath of the financial crisis saw several regulatory initiatives to curtail banks' risk-taking desire, hindering proprietary trading and increasing the cost of market making. In part, these initiatives reflected a widespread belief in banking regulatory circles that the pre-crisis price of immediacy did not adequately incorporate the costs required to ensure that market makers are supported by sufficient capital and do not become a source of illiquidity contagion (see, for example, BIS Committee on the Global Financial System (2014, 2016)). While the regulations were meant to improve market-maker resilience, the post-crisis changes in the Basel framework (Basel 2.5 and Basel III) and the Volcker Rule have reduced banks' willingness to accommodate corporate bond trades on their balance sheets and in general have made their market-making operations more costly. Some market observers have portrayed these regulatory initiatives in the light of a trade-off between market resilience (less severe contagion) in times of stress and liquidity during normal times. Reduced market liquidity during normal times, as the argument goes, may be a necessary compromise for enhanced market resilience during stressful periods.

Such a perspective, however, overlooks the market microstructure aspect of liquidity. In particular, the over-the-counter corporate bond market features two parallel trading mechanisms corresponding to the dual capacity of broker-dealers. In the first mechanism, market making, a bank intermediary functions as a dealer that provides immediacy to customers by taking bonds onto his balance sheet. In the second mechanism, matchmaking, a bank intermediary functions as a broker that searches for couterparties for her customers.[1] While regulations aimed at boosting market resilience increased the cost of taking a bond onto a bank's balance sheet, they did not increase the costs associated with the process of matching customers. In fact, technological innovations over the past two decades have reduced the costs of matching, making this trading mechanism an attractive option. Our focus in this paper is therefore not on the trade-off between resilience in times of stress and liquidity during normal times. Rather, we investigate whether the regulatory push to increase resilience could make investors better off during normal times by changing the market structure they experience when trading over-the-counter securities such as corporate bonds.

[1] The matchmaking mechanism therefore encompasses both "agency trades" and "riskless principal trades."

A growing body of empirical literature investigates the impact of post-crisis regulation on liquidity. The US corporate bond market is the most commonly studied in this context given its large size and dealer-centric nature. On balance, this literature finds improvement or at least no deterioration in the average transaction costs of corporate bond trades (Mizrach (2015), Adrian, Fleming, Shachar, and Vogt (2017), Anderson and Stulz (2017), Bessembinder, Jacobsen, Maxwell, and Venkataraman (2018), and Trebbi and Xiao (2019)). Still, some papers document an increase in the cost of immediacy or transacting in times of stress (Bao, O'Hara, and Zhou (2018), Choi and Huh (2017), and Dick-Nielsen and Rossi (2018)). These studies also find a reduction in the amount of capital that bank dealers commit to market making and a shift in their activity from market making towards matchmaking (Bao, O'Hara, and Zhou (2018), Bessembinder, Jacobsen, Maxwell, and Venkataraman (2018), and Choi and Huh (2017)). In contrast, non-bank dealers appear to increase capital commitments and principal trading.

While the increase in the cost of immediacy is consistent with the regulation-induced higher cost of taking bonds onto banks' balance sheets, the decline in average transaction costs could suggest that a shift from market making to matchmaking is beneficial to customers. Yet, transaction costs are an incomplete measure of overall customer welfare because of two less measurable but potentially important costs. First, the average time it takes to execute a transaction in the post-regulation era is likely longer. These execution delays may be costly to investors. Second, realized transaction costs capture only trades that were executed. If customers forgo transacting in response to the higher costs of immediacy (and their unwillingness to wait longer for execution), their welfare loss cannot be ascertained by analyzing executed trades. To assess overall customer welfare, which is difficult to measure empirically, one needs a model in which the trade-off between the cost of delay (matchmaking) and the cost of immediacy (market making) is considered explicitly, and customers have the option to forgo trading altogether. This is the model we set out to investigate.

Our model features infinitesimal customers (buyers and sellers) who arrive at a constant rate and wish to trade bonds in an over-the-counter market. There are two representative broker-dealers standing for two groups of dealers: bank-affiliated dealers and non-bank-affiliated dealers. Our use of a representative bank dealer and a representative non-bank dealer is meant to acknowledge the market power dealers possess in the corporate bond market relative to customers while at the same time enabling us to explicitly consider the changing nature of competition between the two broker-dealer groups as a result of bank regulations. The bank dealer offers his customers two mechanisms for trading bonds. The first, market making, enables customers to trade immediately at a spread that reflects the bank dealer's cost of market making. In the second, matchmaking, the bank dealer helps customers search for trading counterparties and earns a fee to facilitate such trades.

The bank dealer optimally sets the market-making spread and the matchmaking fee by taking into account his balance sheet cost as well as the search cost and matching rate in the matchmaking mechanism. The non-bank dealer offers market-making services to customers at a spread that he sets to reflect his cost of market making and the competitive environment.[2] The balance sheet costs of the bank dealer and the non-bank dealer are generally different. Customers are price takers and heterogeneous with regard to patience (or the value they attach to immediacy), which we model with private values. They optimize over their trading choices: trade immediately with the bank dealer, trade immediately with the non-bank dealer, use the matchmaking service of the bank dealer and incur delay costs, or forgo trading altogether.

Our main analysis focuses on how customer welfare and market outcomes change when regulation increases the bank dealer's balance sheet cost. If the cost of market making of the bank dealer is lower than that of the non-bank dealer, there are two types of equilibria, depending on whether the bank dealer's spread is constrained by competition from the non-bank dealer. In the unconstrained equilibrium, any increase in the balance sheet cost of the bank dealer is fully passed on to customers, harming customer welfare. This result is reminiscent of warnings made by some market observers that raising the costs of banks would hurt investors in the corporate bond market.

The aforementioned equilibrium is, however, far from providing the complete picture. The more prevalent equilibrium in our model—the constrained bank dealer equilibrium—has a completely different flavor. In this equilibrium, the bank dealer's ability to pass the increase in balance sheet costs on to the customers is constrained by the competitive pressure in market making from the non-bank dealer. A higher balance sheet cost therefore incentivizes the bank dealer to shift more of

[2] Non-bank dealers are typically smaller and are not subject to bank regulations that increase balance sheet costs (Bao, O'Hara, and Zhou (2018), Bessembinder, Jacobsen, Maxwell, and Venkataraman (2018)).

his business to the matchmaking mechanism. As a result, overall customer welfare, which takes into account not just transaction costs but also waiting costs and the welfare of customers who choose not to trade, unambiguously rises. The welfare improvement in the constrained bank equilibrium is driven by the lower matchmaking fee that attracts new customers to trade and benefits all those who use the matchmaking service.

If the balance sheet cost of the bank dealer rises above that of the non-bank dealer, however, customers who demand immediacy switch to the non-bank dealer. The bank dealer, in turn, focuses on maximizing profits in the matchmaking mechanism alone. In this region, increases in regulatory costs reduce the potential for competition between the two representative dealers and therefore result in lower overall customer welfare. Hence, the relationship between the balance sheet costs of the bank and non-bank dealers, which determines the extent to which they influence each other's pricing, is very important to the manner in which overall customer welfare responds to changes in bank regulatory costs in each equilibrium region. The industrial organization aspect, manifested as competitive pressure from the non-bank dealer, and the market microstructure perspective, which introduces possible substitution between the two trading mechanisms, join to deliver the richness of our implications. They are two of the key drivers behind the result that the increase in bank dealer's cost can make customers better off.

The third key driver that is crucial for our results is that bank dealers possess market power relative to customers. This market power impacts both how customer welfare changes in response to increased regulatory costs and the division of the surplus created by technological innovations that have reduced search costs in the matchmaking mechanism over the last two decades. We find that lower search costs unambiguously increase overall customer welfare only in the constrained bank dealer equilibrium. When the bank dealer has unconstrained monopoly power over the provision of market making, the surplus generated by lower search costs can predominantly benefit the bank dealer.

Given the central role played by the market power of bank dealers in both the corporate bond market and our model, we investigate the robustness of our conclusions by analyzing a variant of the model with multiple bank dealers and using the extent of differentiation among the bank dealers to vary the amount of market power they possess relative to customers. We find that increasing regulatory costs can improve overall customer welfare as long as bank dealers have even a small amount of market power, and the parameter region over which this improvement occurs increases in the bank dealers' market power. In other words, only if bank dealers were perfectly competitive would an increase in regulatory costs always lower overall customer welfare, but the evidence in bond markets resoundingly rejects the notion that bank dealers are perfectly competitive.

Our model demonstrates the complex and interesting economic interactions that evolve as bank regulatory costs increase, pointing to subtleties not fully recognized in the extant literature. Our contribution lies in characterizing how changes in bank regulatory costs, both in absolute terms and relative to the market-making costs of non-bank dealers, affect the nature of equilibrium in the market, with each equilibrium region giving rise to a set of implications for customer welfare and market outcomes. These complex interactions also highlight why empirical work thus far has struggled to articulate the overall impact of increased bank regulatory costs on customer welfare in the corporate bond market.

Can our theory contribute to evaluating how the post-crisis regulatory initiatives impacted customer welfare in the corporate bond market? We address this question by relating the wealth of empirical implications generated by the model to findings reported in empirical papers. As we increase bank regulatory costs in the model, our predictions for observable market outcomes change as we move from one equilibrium region to another. We believe that the patterns uncovered by the empirical literature suggest that the suite of post-crisis bank regulations has moved the corporate bond market into and across the constrained bank dealer equilibrium region in which an increase in bank regulatory costs improves overall customer welfare.[3] While bank dealers are clearly worse off in the post-crisis regulatory environment, our model vividly demonstrates that regulations that negatively impact dealer profitability and capital commitment need not translate into lower overall customer welfare.

The regulation of banks' proprietary trading remains a work in progress. In 2019, for example,

[3]It is important to stress that the empirical patterns documented in these papers reflect the overall effects of post-crisis regulations. Hence, our results do not imply that a particular rule (or any specific feature of a particular rule) is beneficial for customer welfare.

the five agencies responsible for administering the Volcker Rule approved changes in the rule. While most of the changes appear to be focused on simplifying reporting as opposed to significantly altering the operations of bank trading desks, some have the potential to affect market-making incentives. We believe that understanding how bank regulations that change the cost of market making impact customer welfare and the functioning of the corporate bond market is of paramount importance. We hope that our insights shed clarifying light on this important question.

2 Background and Literature

2.1 Post-Crisis Bank Regulations and the Corporate Bond Market

The Basel regulatory framework and the Dodd-Frank Act are the cornerstones of post-crisis bank regulations that heavily impacted market-making activities in the United States, including in the corporate bond market. Interviews with market participants suggest that the revision of the Basel II market-risk framework ("Basel 2.5"), finalized in June 2012 in the United States, increased inventory costs for corporate bonds by boosting regulatory capital charges through the incremental risk capital (IRC) charge and the trading book's stressed VaR requirement (BIS Committee on the Global Financial System (2014)). Furthermore, the Basel III framework, finalized in July 2013 in the United States, not only raised the risk-based capital requirements on banks but also raised the non-risk-based capital requirement through the supplementary leverage ratio (SLR). For example, global systemically important banks (G-SIBs) are required to maintain an SLR of 5% or higher at the bank-holding-company level and an SLR of 6% or higher at the depository-subsidiary level (see Davis Polk (2014) for more details). Greenwood, Hanson, Stein, and Sunderam (2017) and Duffie (2018) find that the leverage ratio requirement is the most tightly binding constraint for most U.S. G-SIBs, according to data derived from the Federal Reserve's stress tests in 2017. Results from the 2019 Dodd-Frank Stress test show that the leverage ratio remains the most binding constraint for the largest U.S. banks.[4] In addition to capital requirements, another pillar of Basel III is higher liquidity standards, including the liquidity coverage ratio (LCR) and the net stable funding ratio

[4]See https://www.federalreserve.gov/publications/files/2019-dfast-results-20190621.pdf.

(NSFR), which require that banks should hold enough high-quality liquid assets and use sufficiently stable financing sources to guard against adverse conditions in the funding markets.

Another major regulatory measure impacting the corporate bond market is the Volcker Rule. While the Dodd-Frank Act was signed into law in July 2010, the implementation of the Volcker Rule was delayed until April 2014, with full compliance required by July 2015. The Volcker Rule aims at discouraging banks with access to FDIC insurance or the Federal Reserve's discount window from engaging in proprietary trading of risky securities. While it provides an exception for market making, the distinction between market making and proprietary trading is inherently blurry. The rule, therefore, mandates the reporting of various measures (e.g., inventory turnover, the standard deviation of daily trading profits, the customer-facing trade ratio) as proxies for the underlying motive behind the trades. The Volcker Rule's restrictions and associated compliance work have increased the market-making costs that bank-affiliated dealers incur.[5]

By strengthening banks' resilience to various risks (e.g., market, counterparty credit, and funding), the suite of post-crisis bank regulations now in place changes the economics of trading activities undertaken by banks, including market making. Some market participants therefore warned that these rules would increase transaction costs in the corporate bond market, and these claims motivated subsequent empirical research examining post-crisis changes in corporate bond liquidity.

2.2 Post-Crisis Changes in Corporate Bond Liquidity: Empirical Evidence

Most empirical papers that examine corporate bond market liquidity following the financial crisis find no deterioration, indeed even improvement, in liquidity subsequent to the regulatory interventions. For example, Mizrach (2015) and Adrian, Fleming, Shachar, and Vogt (2017) find that measures of execution costs dropped after the crisis to below pre-crisis levels. Adrian et al. also find that

[5] The regulatory costs we model in this paper probably best represent the explicit costs imposed by the Basel framework. In fact, an informal survey of market makers in the corporate bond market conducted by the Bank for International Settlements found that "...revisions to the Basel II market risk framework (Basel 2.5) were seen to have had the largest impact on regulatory charges" (BIS Committee on the Global Financial System (2016)). Still, most empirical findings concerning the impact of post-crisis regulations on the corporate bond market reflect the combined effects brought about by all these rules and therefore we model a single regulatory cost of market making rather than investigating specific features of these rules.

trading volume and issuance are at record levels, although trade size and turnover have declined. Anderson and Stulz (2017) look at a variety of price-based measures of liquidity, and find that they are marginally better following the onset of regulation (2013-2014) relative to before the crisis. Bessembinder, Jacobsen, Maxwell, and Venkataraman (2018) also document that average customer trade execution costs have not increased after these regulations were imposed. Trebbi and Xiao (2019) look for a structural break in various liquidity measures but find no evidence of deteriorating liquidity during the period that corresponds to the Dodd-Frank and Basel III regulatory interventions.[6]

Several papers find worsening in a particular dimension of trade execution: the cost of immediacy. Bao, O'Hara, and Zhou (2018) use downgrades of bonds to junk status as stress events to examine the cost of immediacy, and find that it increased following the implementation of the Volcker Rule. Similarly, Dick-Nielsen and Rossi (2018) use exclusions from the Barclays Capital investment-grade bond index as events that create demand for immediacy, and document an increase in the cost of immediacy after the financial crisis. Choi and Huh (2017) show that trading costs for market-making trades increased substantially in the post-regulation period relative to pre-crisis levels, and this increase is driven by bank dealers.

Our model is able to reconcile these two seemingly conflicting lines of empirical evidence, namely that average transaction costs in corporate bonds have declined while the cost of immediacy has gone up. The driving force behind these results in our theory is the bank dealers' endogenous shift from market making to matchmaking. The empirical papers we have cited indeed provide ample evidence that matchmaking has increased following the crisis and the implementation of post-crisis regulations (e.g., Bao, O'Hara, and Zhou (2018), Choi and Huh (2017), Schultz (2017)).[7] As a result of the shift to matchmaking, the execution of large trades now tends to require more time (BIS Committee on the Global Financial System (2014, 2016)).

[6]Two papers find more nuanced effects. Allahrakha, Cetina, Munyan, and Watugala (2019) find higher markups for a subset of the trades when looking at Volcker Rule exemptions (e.g., trades in newly issued bonds for which a bank dealer is part of the bond's underwriting group) to infer cost differentials. Chernenko and Sunderam (2018) develop an indirect measure of aggregate corporate bond market liquidity by relating mutual funds' cash holdings to the volatility of their fund flows. They find that, while the liquidity of investment-grade bonds in the post-crisis period essentially recovered to the pre-crisis level, liquidity for speculative grade bonds has not.

[7]Ederington, Guan, and Yadav (2014), Randall (2015), and Anand, Jotikasthira, and Venkataraman (2020) provide additional evidence about matchmaking and the provision of liquidity by customers in the corporate bond market.

It is important to stress the heterogeneous manner in which bank dealers and non-bank dealers responded to the regulatory changes. Bao, O'Hara, and Zhou (2018) find that dealers affected by the Volcker Rule increase their matchmaking activity while committing less capital to market making. Goldstein and Hotchkiss (2020) investigate the trades that dealers seek to offset in comparison with those that they hold in inventory overnight. At the same time, competition in market making from smaller non-bank dealers appears to intensify: they increase capital commitments and the amount of principal trading (see also Bessembinder, Jacobsen, Maxwell, and Venkataraman (2018)). Non-bank-dealer matchmaking, on the other hand, has decreased after the implementation of the Volcker Rule. We return to the findings of the empirical literature to motivate our modeling approach and guide our discussion of how bank regulations following the financial crisis affected customer welfare.

2.3 Market Making versus Matchmaking: Theory

Several recent theoretical papers recognize the importance of the dual mechanisms for trading bonds, namely market making and matchmaking, although each of these papers adopts a distinct approach to studying the two mechanisms. An, Song, and Zhang (2017) study intermediation chains by modeling the interaction between one seller, a finite number of dealers, and an infinite number of buyers. Their model shows how an intermediary rat race gives rise to an inefficient amount of principal trading. Our paper does not feature an inter-dealer market or intermediation chains, but rather focuses on what happens to customer welfare and the market environment if the cost of market making increases for bank dealers.

An and Zheng (2017) look at how the dual capacity of broker-dealers (principal and agency trading) gives rise to a conflict of interest, which results in dealers' holding too much inventory as a tool for extracting rent from customers. Unlike in our framework, customers in their model do not optimize and matchmaking is effortless and costless, leading An and Zheng to focus on inventory as a strategic variable. In contrast, we model a two-way market with balanced customer order flows and abstract away from inventory management—an approach that is orthogonal to that of An and Zheng (2017). Furthermore, we investigate the endogenous evolution of trading mechanisms by (i) having customers optimally choose whether and how to trade and (ii) having dealers optimize the pricing of their services.

Li and Li (2017) model a trade-off between inventory costs (in market making) and verification costs (in matchmaking). Moral hazard in matchmaking arises in their model when a dealer gains by providing worse executions for a customer. Because dealers have better information than customers, transparency influences the prevalence of market making over matchmaking.[8] Transparency plays no role in our model because we assume homogeneous common-value information. Our emphasis, instead, is on competition from non-bank dealers and the role it plays in determining customer welfare and the extent of matchmaking.[9]

The paper closest to ours in objective is Cimon and Garriott (2019). In their model, market makers compete for quantity (Cournot) in separate buyer and seller markets and issue equity and debt to fund their operations. Market making is modeled as a more efficient form of trading than matchmaking, and therefore increased agency trading implies a higher price impact of trades. As a result, regulations that increase the cost of market making must hurt liquidity. In contrast, we show that investment in technology can make matchmaking a less expensive form of intermediation than market making, and the market power of dealers provides a role for bank regulations in enhancing competition and improving customer welfare.

Our model shares certain features that are examined in the vast industrial organization (IO) literature, in particular the branch that investigates multiproduct competition. To the best of our knowledge, however, no other paper captures all the important characteristics of bond markets that we wish to model. Mussa and Rosen (1978) characterize the optimal pricing strategy of a monopolist over a range of products that differ in quality. Katz (1984) analyzes competition between various multiproduct firms. He shows that because competition in one product spills over to another, endogenous specialization can arise. Johnson and Myatt (2003) consider the duopoly competition between a multiproduct incumbent and a multiproduct entrant, where both face the same costs. The entrant in their model is assumed to focus on low-quality products, and the incumbent's

[8]Li and Li also provide empirical results pertaining to the share of matchmaking around the financial crisis and how this share relates to transparency and volume.

[9]Chang and Zhang (2020) model endogenous network formation and discuss how network structure may change as a result of OTC reforms. Their framework also does not consider competition from non-bank dealers, which is central to our results.

equilibrium products are shown to be of weakly higher quality than those of the entrant. Nocke and Schutz (2018) show that in a fairly general multiproduct setting, increasing competition leads to expanded product offering because a firm worries less about cannibalizing its other products when facing more intense outside competition.

Our results are distinct from this line of IO literature in at least two ways. First, we allow differentiated costs between the bank dealer and the non-bank dealer. This phenomenon of "same activity, different costs" is salient in financial markets that are regulated based on the type of entities involved. Our results predict that the bank dealer expands into matchmaking even when it still has a cost advantage. This feature matches the empirical fact that bank dealers are changing their business models even when they maintain overall dominance in liquidity provision. Second, we focus on consumer welfare. Under some conditions, consumer welfare increases when production costs rise. That message is not present in the papers cited above.

The increased balance sheet cost in our model resembles an increase in taxes. Weyl and Fabinger (2013) characterize the pass-through of taxes to consumers when firms compete imperfectly in an oligopoly market of a single good, showing that, under some conditions, the pass-through can exceed one. In a more stylized setting of two goods (market making and matchmaking), we show that a higher tax (balance sheet cost) on market making can lead to a net *negative pass-through* to customers, manifested by a lower quantity-weighted average transaction cost and higher customer welfare.

3 The Model

Time is continuous, $t \in [0, \infty)$. The traded asset has an expected fundamental value of v. All customers and dealers are risk-neutral and have the same information about the fundamental value of the asset. The discount rate is $r > 0$.[10]

Customers and dealers. Infinitesimal buyers arrive in the market at the rate μ; that is, the mass of buyers arriving during the time interval $(t, t + dt)$ is μdt. Each buyer wishes to buy one

[10] We use the discount rate r to capture two effects: the rate at which customers and dealers discount future profits and the rate at which trading opportunities decay over time.

unit of the asset, and her private benefit (or "value") for trading immediately is a random variable $x \in [0, \infty)$, with cumulative distribution function G. Heterogeneity in this private value that describes the customer's need for immediacy is the manner in which we model differences across customers in their degree of patience. The buyers' arrival process is time-invariant in the sense that the types of buyers arriving during each small time interval $(t, t+dt)$ are distributed according to G. Likewise, infinitesimal sellers arrive in the market at the same rate μ, and their private benefit for selling the asset immediately is also distributed according to G. A customer's need for immediacy is not observable by others, and the customer exits the market upon trading.

To make the model more general, we impose only modest structure on the distribution of private values: the familiar monotone-hazard-rate assumption. While not entirely innocuous, the assumption of a non-decreasing hazard rate has been used extensively in the mechanism-design literature (see Fudenberg and Tirole (1991), Chapter 7). A non-decreasing hazard rate is equivalent to the log-concavity of the reliability function $1 - G(\cdot)$, and is satisfied by many common distributions, including uniform, normal, exponential, logistic, extreme value, Laplace, and, under some parametric restrictions, power, Weibull, Gamma, Chi-squared, and Beta (see Bagnoli and Bergstrom (2005)). In our context, this assumption simplifies the proofs by guaranteeing a unique equilibrium in some parameter ranges and helping to sign comparative statics when bank regulatory costs increase.[11] For convenience, we state this assumption in terms of the inverse hazard function (or Mills ratio) of G,

$$\zeta(x) = \frac{1 - G(x)}{G'(x)},$$

and specify that $\zeta(x)$ is non-increasing in x. We stress that while customers' desire to trade in the model—motivated by risk-sharing, liquidity needs, and other non-informational reasons—is specified exogenously as is standard in many models, both the quantity of trading and its composition (market making versus matchmaking) arise endogenously.

Trading institutional-sized orders in the OTC market for corporate bonds requires the active

[11] There are studies showing that some counterintuitive results may obtain when this assumption is violated. For example, Bulow and Klemperer (2002) show examples in which a decreasing hazard rate may lead to a higher per-unit auction price as supply increases. Chen and Riordan (2008) show that, under a decreasing hazard rate, prices can be higher in a duopoly market than in a monopoly market.

intermediation of broker-dealers. An important aspect of liquidity provision in this market that we choose to model explicitly is the competition between bank and non-bank dealers. While bank dealers have long dominated the corporate bond market, the potentially important role that non-bank dealers could play if post-crisis regulations were to curtail bank dealer trading was emphasized even before these regulations were implemented (e.g., Duffie (2012)). Therefore, our model features two representative yet distinct strategic intermediaries, called dealers, who help customers trade this asset. One of the dealers is a bank affiliate, subject to bank regulations, whereas the other dealer is unaffiliated with any bank and hence is not subject to bank regulations. In practice, all intermediaries face some types of regulatory constraints, but the specific post-crisis regulations we discuss in Section 2.1 (Basel 2.5, Basel III, and the Volcker Rule) apply only to banks and their broker-dealer affiliates.

Insofar as our focus is on bank regulations that apply to all bank dealers and, by the same token, do not apply to any non-bank dealer, we choose to model each type using a representative dealer that stands for the entire group of dealers. In so doing, we abstract from the inter-dealer market and competition within each group. While other papers (e.g., An, Song, and Zhang (2017)) focus on the inter-dealer market, we believe that the basic economics of intermediation chains did not materially change with the implementation of post-crisis bank regulations.

Furthermore, while we stress competition from non-bank dealers as a factor that is important for understanding the impact of bank regulations, we strongly believe that dealers in the corporate bond market have market power relative to customers. It has long been established that per-share transaction costs in corporate bonds decline in trade size (e.g., Schultz (2001)), even though fixed costs do not appear to be very high. The usual explanation for this empirical regularity is dealer market power: large customers have greater bargaining power and hence obtain better prices or lower fees than small customers can. In the model, the spread charged by a dealer does not depend on the identity of the customer. While this is clearly a simplification, we believe that specifying several types of customers with varying degrees of bargaining power would not materially change the main implications of our model. Hence, we choose to simplify the exposition by having only one type of customer and giving the full market power to the dealer. In other words, each representative dealer optimally sets prices to maximize profits subject to competition from the other representative dealer.

Ascribing market power to the dealer is not an innocuous assumption; in fact, it is a driving force behind our results concerning customer welfare. In Section 6.1, we investigate the role played by market power by examining a variant of our model in which multiple differentiated bank dealers compete in offering market-making and matchmaking services.

Trading protocols: market making and matchmaking. The customers in our model are meant to represent institutional investors who trade large quantities of bonds. These large trades, which account for most of the volume in the corporate bond market, require the active intermediation of OTC broker-dealers either in a principal ("market making") or an agent ("matchmaking") capacity.[12] The market-making mechanism allows customers to trade immediately, while searching for a counterparty using the matchmaking mechanism takes time.

Under the market-making protocol, a dealer who faces a stochastic flow of buy and sell orders immediately fills a customer's order from his own balance sheet by incurring a balance sheet cost. To fill an order immediately, the bank (non-bank) dealer charges customers a per-unit spread of $S_B > 0$ ($S_{NB} > 0$), which is publicly observable.[13] The bank (non-bank) dealer's balance sheet cost is assumed to be c_B (c_{NB}) per unit of the asset regardless of whether he is accommodating a buy or a sell order (that is, the cost is incurred on the gross amount traded). Given this specification of balance sheet costs and the risk-neutrality of dealers in the model, inventory level does not play a role in pricing.[14] As such, there is no loss of generality in assuming equal arrival rates of customers who wish to buy and customers who wish to sell, which simplifies the exposition of the model.

We think about bank regulatory costs as per-share balance sheet costs imposed on all trades accommodated by the bank dealer in his capacity as a market maker. This is a parsimonious way to

[12]In contrast, retail-size orders are often traded on MarketAxess, an all-to-all trading platform that gives access to both dealers and customers. MarketAxess is a centralized system that currently executes almost 20% of the volume in the corporate bond market (investment-grade and high-yield bonds combined), although it is not a big player in executing the large institutional trades that we consider in this paper (O'Hara and Zhou (2020)).

[13]We adopt the usual formulation in market microstructure models of a dealer who faces a stochastic flow of buy and sell orders and quotes a spread (or price concession) for executing these orders immediately (O'Hara (1995)).

[14]See An and Zheng (2017) for a model of the dual capacity of broker-dealers that focuses on the dealer's choice of inventory level.

differentiate market-making activity from trade facilitation via the matchmaking mechanism that is not subject to such regulatory costs. In other words, we view market making as any activity that may require the dealer to hold the position overnight, which is why regulations that increase capital requirements would raise its costs. As such, the bank dealer's balance sheet cost is conceptually comprised of three components:

$$c_B = \mathit{OperatingCosts} - \mathit{ImplicitSubsidy} + \mathit{PostCrisisRegulatoryCosts}.$$

The first component reflects the direct operating costs involved in running the market-making business. We assume that the bank dealer operates optimally to minimize this cost. This component is similar to the cost of the non-bank dealer, c_{NB}, and can be influenced by the same economic forces.

The second component (*ImplicitSubsidy*) has been discussed extensively in regulatory circles in the aftermath of the financial crisis. The Committee on the Global Financial System of the Bank for International Settlements writes in its report on fixed-income market liquidity that, in the pre-crisis era, "Underpriced liquidity services were predicated on expectations of an implicit public sector backstop for major financial institutions" (BIS Committee on the Global Financial System (2016)). This implicit too-big-to-fail subsidy lowers the capital costs of the bank dealer's trading book relative to that of the non-bank dealer and hence could enable the bank dealer to offer cheaper liquidity. Post-crisis bank regulations, represented by the third component of c_B, were aimed at increasing the market-making costs of bank dealers to counteract this implicit subsidy. Fixing the first two components, the imposition of post-crisis regulations unequivocally increases the bank dealer's cost, but it is unclear whether the sum of the three components is lower than, equal to, or greater than the cost of the non-bank dealer. We therefore investigate how post-crisis bank regulations impact customer welfare and market outcomes by examining the effects of increasing c_B over the entire relevant range (below and above c_{NB}).[15]

Under the matchmaking protocol, the bank dealer searches for a counterparty for the customer's

[15] Throughout the paper, when we discuss an increase in c_B we mean an increase in the third component, holding the other two components constant.

order, but the dealer does not use his balance sheet to accommodate the order.[16] The effort to search for such a counterparty is costly, though. In particular, spending I on the search process for each customer results in the customer's being matched with a counterparty at an exponentially distributed time τ with intensity $H \in [0, \infty)$, where I and H are exogenous and commonly known. We examine how changing I affects overall customer welfare and market outcomes to study the second development in the corporate bond market: technological advances that have reduced the costs of matchmaking.

While the dealer conducts the search, having a positive discount rate (r) means that the customer incurs a delay cost when discounting the private benefit of trading. Given the exponential distribution of matching time τ, we define the benefit customers obtain from matchmaking with intensity H as

$$\mathcal{H} \equiv E[e^{-r\tau}] = \int_{\tau=0}^{\infty} He^{-Hu}e^{-ru}du = \frac{H}{r+H}. \tag{1}$$

A higher $\mathcal{H}$ implies a shorter waiting time (with a lower cost of delay) for the searching customers, and hence we refer to $\mathcal{H}$ as the *speed* of matchmaking. When a match is made, the bank dealer receives a publicly observable fee of f from both the buyer and the seller.[17]

To simplify the exposition, only bank dealers operate a matchmaking mechanism in our model, but we have reasons to believe that omitting the matchmaking mechanism for non-bank dealers does not detract from the external validity of our results. First, Bao, O'Hara, and Zhou (2018) and Bessembinder, Jacobsen, Maxwell, and Venkataraman (2018) find that most non-bank dealers are small. Many of these non-bank dealers are regional in focus, and it is reasonable to assume

[16] It is important to acknowledge the difficulty involved in empirically measuring the extent of matchmaking. The empirical papers we cite in Section 2.2 use various forms of the TRACE database. An in-house cross—a dealer buying from a customer and immediately selling to another customer—is reported in TRACE as two transactions. Whether TRACE reports these two transactions as agency or principal depends on the internal accounts of the dealer involved, and can be influenced by a dealer's idiosyncratic preference for reporting the price inclusive of the mark-up/mark-down or as a separate commission. Given that agency and riskless principal have similar balance-sheet implications, we lump them together under our matchmaking mechanism. In other words, matchmaking consists of all dealer-facilitated trading that does not involve taking a trade onto the bank dealer's balance sheet.

[17] As Choi and Huh (2017) note, a dealer searching for a counterparty may need to offer better terms of trade to the counterparty to execute the trade. In other words, when one side is more eager than the other, the two sides of the trade need not be symmetric in the amount they pay for participating in the transaction, and one can view $2f$ as the net compensation earned by a dealer that executes both legs of such an agency (or riskless-principal) cross. Still, any such asymmetries must be temporary because prices adjust so that the order flow is balanced. Hence, we postulate equal arrival rates of buyers and sellers and specify a fee of f from each side.

that most of them do not possess vast networks of customers that would allow them to operate an efficient matchmaking mechanism. Bank dealers, on the other hand, have long dominated the OTC markets and have rich information about a large group of customers. Second, both these papers find that non-bank dealers significantly decreased their matchmaking activity following the implementation of the Volcker Rule, which could suggest that they are being crowded out by the bank dealers. In fact, Bessembinder et al. find that matchmaking constitutes only a very small fraction (12.8%) of the trading facilitated by non-bank dealers after post-crisis regulations were implemented. Therefore, the main dimension along which they inject competition into the corporate bond market following the financial crisis appears to be market making (by boosting both capital commitments and trading). As such, market making is the dimension we choose to model explicitly for non-bank dealers.

Objective Functions. Optimizing customers are an important feature of the model because we seek to contribute insights that are difficult to observe empirically. As we stress in the introduction, the empirical finding of lower average transaction costs can be accompanied by other outcomes that are not easily measured but that negatively impact customer welfare (e.g., delayed execution of trades or the loss of welfare of customers who choose not to trade). Therefore, customers in our model optimize over whether to trade and how to trade. Specifically, customers choose between trading immediately with the bank dealer, trading immediately with the non-bank dealer, searching for a counterparty using the matchmaking service, or not trading. The bank and non-bank dealers' market-making services are identical from the customer's perspective. Therefore, a customer who opts to trade immediately will choose the market-making service that charges the lower spread, which we denote by

$$S = \min(S_B, S_{NB}) \tag{2}$$

Recall that x denotes the private benefit a customer obtains from trading immediately. The customer's profit from using a market-making service is $x - S$. Her expected profit from using the matchmaking mechanism offered by the bank dealer, which takes into account the expected waiting cost, is $(x - f)\mathcal{H}$. Her profit from leaving the market without trading is 0. Therefore, a customer prefers matchmaking to not trading if and only if $x \geq f$.

Let b be the value of the marginal customer who is indifferent between matchmaking and market making. The indifference condition is

$$(b - f)\mathcal{H} = b - S, \tag{3}$$

and because $\mathcal{H} < 1$ we obtain

$$b = \frac{S - f\mathcal{H}}{1 - \mathcal{H}}. \tag{4}$$

The customer's optimization problem therefore results in a very simple behavior: do not trade if $x \in [0, f)$, choose matchmaking if $x \in [f, b]$, or choose market making with the dealer offering the lower spread if $x > b$.[18]

Our main objective in this paper is to analyze the impact of regulations on overall customer welfare. Given the two thresholds f and b, we can write the overall welfare of customers aggregated across the three ranges of x as:

$$\pi_c = \frac{2\mu}{r}\left[\underbrace{\int_{x=0}^{f} 0 \cdot dG(x)}_{\text{no trade}} + \underbrace{\int_{x=f}^{b} (x - f)\mathcal{H}dG(x)}_{\text{matchmaking}} + \underbrace{\int_{x=b}^{\infty} (x - S)dG(x)}_{\text{market making}}\right]. \tag{5}$$

The bank dealer's profit is comprised of two components: the matchmaking profit, which depends on the fee and the cost of searching, and the market-making profit (if the bank offers the lower spread), such that

$$\pi_B = \frac{2\mu}{r}\left[(\mathcal{H}f - I)\left(G\left(b\right) - G\left(f\right)\right) + (S - c_B)(1 - G(b))\mathbb{I}_{S=S_B}\right], \tag{6}$$

where $\mathbb{I}_{S=S_B}$ is an indicator function that takes the value 1 if $S = S_B$ (equivalently, $S_B \leq S_{NB}$) and 0 otherwise.

[18] It is straightforward to show that any equilibrium in which the bank dealer operates the matchmaking service must satisfy $f < S$ (i.e., the matchmaking fee is lower than the market-making spread) because of the waiting costs associated with the search.

The non-bank dealer's market-making profit can be expressed as:

$$\pi_{NB} = \frac{2\mu}{r} \left[(S - c_{NB})(1 - G(b)) \mathbb{I}_{S=S_{NB}} \right], \tag{7}$$

where $\mathbb{I}_{S=S_{NB}}$ is an indicator function that takes the value 1 if $S_{NB} < S_B$ and 0 otherwise.

Equilibrium Definition. An equilibrium consists of:

1. the bank dealer's choices of market-making spread S_B and matchmaking fee f,
2. the non-bank dealer's choice of market-making spread S_{NB}, and
3. each arriving customer's choice between market making (with one of the dealers), matchmaking, and refraining from trading altogether,

such that dealers and customers maximize expected profits.[19]

A glossary of key model variables is provided in Appendix A for ease of reference.

4 Customer Welfare and Market Outcomes in Equilibrium

Our goal is to investigate what happens to customer welfare in the face of two developments: an increase in bank regulatory costs that affects the bank dealer's market-making business and a decrease in search costs in the matchmaking business. The model we set out to solve takes the parameters of the matchmaking mechanism as exogenous.[20] It is clear that if the search cost is high enough, the bank dealer would never find it optimal to operate the matchmaking mechanism. This, however, is an uninteresting case because we know from empirical research that many customer orders in the corporate bond market have in the past been handled by bank dealers using the matchmaking mechanism, and the market share of this trading mechanism has increased following

[19]If there exist multiple Nash equilibria, then we choose the equilibrium that generates the highest overall customer welfare; if there are multiple equilibria that generate the same customer welfare, then we randomly choose one of them.

[20]In a previous draft of this paper, we solved a version of the model in which the bank dealer optimally chooses the matching speed given a cost structure. The conclusions we obtain from both versions of the model are similar, but using exogenous search cost and matching rate simplifies the model and increases transparency as to the manner in which our assumptions generate the results. The version with endogenous matching speed is available from the authors upon request.

the post-crisis regulatory reform. Therefore, while we have solved the model for all possible levels of search cost, here we focus our exposition on the case in which the bank dealer operates the matchmaking business.[21]

4.1 What Happens when Bank Regulatory Costs Increase?

In this section we focus on the first important development of the past decade in the corporate bond market: the enactment of banking regulations that increase the balance sheet cost of bank dealers. Specifically, we ask what happens to customer welfare and market outcomes when c_B rises.

4.1.1 When $c_B \leq c_{NB}$

The main case we consider is when the balance sheet cost of the bank dealer is lower than that of the non-bank dealer, $c_B \leq c_{NB}$. This is the natural case to consider given the too-big-to-fail subsidy and the dominance of bank dealers in the corporate bond market. In this case, the bank dealer's problem of maximizing expected profit from providing market-making and matchmaking services is

$$\max_{0 \leq f \leq S \leq c_{NB}} \Pi_B (S, f; c_B) \equiv \frac{2\mu}{r} \left[(\mathcal{H}f - I) (G(b) - G(f)) + (S - c_B) (1 - G(b)) \right], \tag{8}$$

where $b = \frac{S - \mathcal{H}f}{1 - \mathcal{H}}$.

Let $\phi(x) \equiv x - \zeta(x) = x - \frac{1 - G(x)}{G'(x)}$.[22] We begin by establishing the existence of an equilibrium and characterizing its structure.

Proposition 1. *When $c_B \leq c_{NB}$ and $I < \mathcal{H} c_B$, the bank dealer operates both market-making and matchmaking services, and the equilibrium is characterized as follows:*

[21]The existence proof and comparative statics for all possible levels of the search costs are available from the authors upon request. We can adopt this focus without loss of generality (or external validity) because we show in the proofs of Proposition 1 and Proposition 4 that, if one starts with an equilibrium in which the bank dealer operates a matchmaking business, an increase in the balance sheet cost or a decrease in the matchmaking cost (the two developments we investigate) will always result in an equilibrium with bank dealer matchmaking. Hence, restricting our analysis in this section to equilibria with bank dealer matchmaking simplifies the exposition without sacrificing the generality of our results.

[22]In the mechanism-design literature, $\phi(x)$ is sometimes called the virtual valuation function.

1. *If $\phi(c_{NB}) \leqslant 0$ and $I \in (0, \mathcal{H}c_B)$, there is a* ***constrained bank dealer*** *equilibrium ($S^\star = c_{NB}$), and $f^\star$ is the minimal solution of*

$$f^\star = \arg\max_f \frac{2\mu}{r}\left[(\mathcal{H}f - I)\left(G\left(\frac{c_{NB}-\mathcal{H}f}{1-\mathcal{H}}\right) - G(f)\right) + (c_{NB}-c_B)\left(1 - G\left(\frac{c_{NB}-\mathcal{H}f}{1-\mathcal{H}}\right)\right)\right]. \quad (9)$$

2. *If $\phi(c_{NB}) > 0$, then*

   (a) *If $I \in (0, \mathcal{H}\phi(c_{NB}))$, there exists $c_1 \in \left(\frac{I}{\mathcal{H}}, c_{NB}\right)$ such that:*

      i. *If $c_B \in \left(\frac{I}{\mathcal{H}}, c_1\right)$, there is an* ***unconstrained bank dealer*** *equilibrium ($S^\star < c_{NB}$) that satisfies the following conditions:*

$$\phi(f^\star) = \frac{I}{\mathcal{H}},\ \phi(b^\star) = \frac{c_B - I}{1-\mathcal{H}},\ S^\star = \mathcal{H}f^\star + (1-\mathcal{H})\,b^\star; \quad (10)$$

      ii. *If $c_B \in [c_1, c_{NB}]$, there is a* ***constrained bank dealer*** *equilibrium ($S^\star = c_{NB}$), and $f^\star$ is the minimal solution of (9).*

   (b) *If $I \in [\mathcal{H}\phi(c_{NB}), \mathcal{H}c_{NB}]$ and $c_B \in \left(\frac{I}{\mathcal{H}}, c_{NB}\right]$, there is a* ***constrained bank dealer*** *equilibrium ($S^\star = c_{NB}$), and $f^\star$ is the minimal solution of (9).*

The proofs of all propositions are provided in the Appendix. Figure 1 provides a graphical illustration of the equilibria described in Proposition 1 in the space formed by I on the y-axis and by c_B on the x-axis. The left panel of the figure shows that when $\phi(c_{NB}) \leqslant 0$, only the constrained equilibrium exists for the entire range of $c_B \leq c_{NB}$. The right panel shows that when $\phi(c_{NB}) > 0$, constrained equilibria in region A exist for any level of the matchmaking cost (provided that the balance sheet cost is above a certain level). The unconstrained bank dealer equilibrium (region B) exists only if both the matchmaking cost and the bank dealer's balance sheet cost are low.[23]

[23] For example, if the need for immediacy of customers is exponentially distributed, then $\zeta(c_{NB}) = E(G)$ and $\phi(c_{NB}) = c_{NB} - E(G)$. The left panel then shows that whenever the balance sheet cost of the non-bank dealer is smaller than the mean value of immediacy in the population of customers, $c_{NB} \leq E(G)$, only the constrained bank dealer equilibrium exists. If $c_{NB} > E(G)$ and the matchmaking cost is in the range $[\mathcal{H}(c_{NB} - E(G)), \mathcal{H}c_{NB}]$, again only the constrained bank dealer equilibrium exists. If $c_{NB} > E(G)$ and the matchmaking cost is lower than $\mathcal{H}(c_{NB} - E(G))$, then there are two possible cases. If the balance sheet cost of the bank dealer is above a certain threshold, again only the constrained bank dealer equilibrium exists. Only if the balance sheet cost of the bank dealer is rather low (between $\frac{I}{\mathcal{H}}$ and $c_{NB} - E(G)$) does the unconstrained bank dealer equilibrium exist.

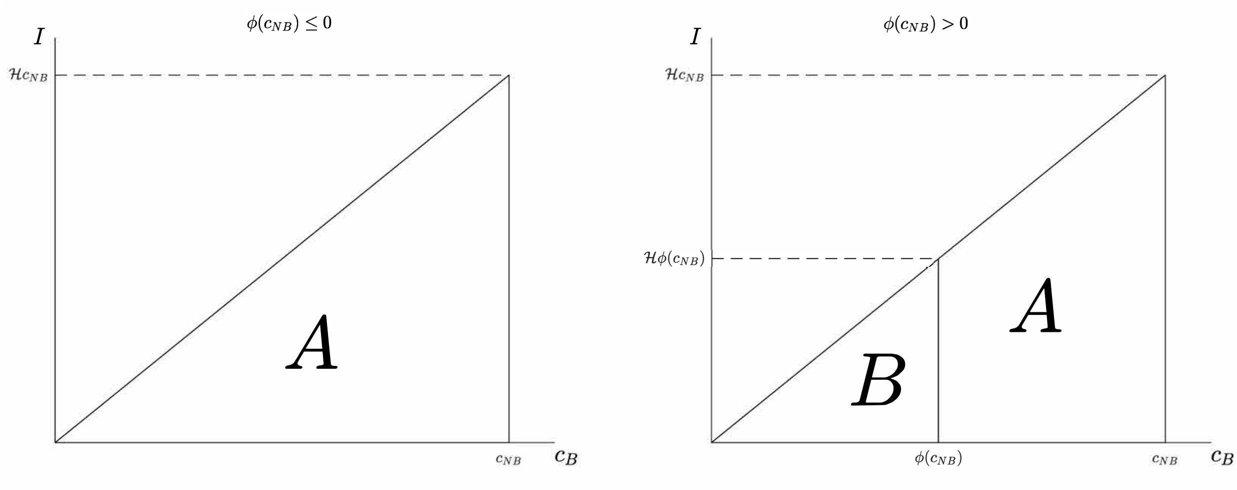

Figure 1: **Constrained and Unconstrained Bank Dealer Equilibria.** This figure provides a visual illustration of the equilibria in Proposition 1 in the space formed by I on the y-axis and by c_B on the x-axis. The left panel depicts the case where $\phi(c_{NB}) \leqslant 0$, while the right panel depicts the case where $\phi(c_{NB}) > 0$. The constrained bank dealer equilibrium ($S^\star = c_{NB}$) is represented by Region A, and the unconstrained bank dealer equilibrium ($S^\star < c_{NB}$) is represented by Region B.

Therefore, while Proposition 1 describes two types of equilibria, by far the more prevalent one appears to be the constrained bank dealer equilibrium. In this equilibrium, while the bank dealer would have preferred setting a monopolist market-making spread that is greater than c_{NB}, the non-bank dealer exerts competitive pressure and therefore the market-making spread needs to equal the non-bank dealer's balance sheet cost. The bank dealer then sets his matchmaking fee to maximize profit given this constraint.

The slope of the line that separates regions A and B in the right panel of the figure depends on the curvature of the inverse hazard rate (the Mills ratio) of the population distribution, $\zeta(x) = \frac{1-G(x)}{G'(x)}$. This decreasing function essentially describes the tradeoff between the population of customers who choose market making (in the numerator) and the rate at which these customers switch to the matchmaking mechanism (in the denominator). It is a measure of the propensity to switch to matchmaking at various points in the distribution of private benefits of immediacy.[24]

[24] A concave ζ means that, as we move from customers with low immediacy needs to those with high immediacy needs, the propensity to switch from market making to matchmaking increases at an increasing rate. A convex ζ means that the propensity of customers to switch from market making to matchmaking initially increases very quickly (when we are in the region of customers with low immediacy needs), but this propensity to switch increases only slowly when we get to customers with high immediacy needs.

A convex ζ, which characterizes many popular distributions such as the normal and uniform distributions, would result in a boundary that intersects the x-axis to the left of $\phi(c_{NB})$ and hence an even larger region A of the constrained equilibrium. A concave ζ would result in a boundary that intersects the x-axis to the right of $\phi(c_{NB})$, increasing the size of region B.

How does an increase in the bank dealer's regulatory costs impact the market? We can show that bank dealer profits are unequivocally lower when regulatory costs increase (or, put differently, as the bank dealer's implicit too-big-to-fail subsidy is reduced). This would explain why bank dealers opposed post-crisis regulations that increased their balance sheet costs. Less obvious is what happens to customer welfare and observable market outcome. To determine how an increase in bank regulatory costs impacts these quantities as c_B increases, we need to investigate how these attributes change within each of the two equilibrium regions.

Proposition 2 provides the comparative statics for the constrained bank dealer equilibrium. Specifically, we look at the following observable market outcomes: trading costs (market-making spread, matchmaking fee, and average transaction costs), and customers' choice of trading outcome and venue (i.e., volume and the market share of the two trading mechanisms). Most importantly, we examine how an increase in the bank dealer's balance sheet cost impacts (the unobservable) overall customer welfare.

Proposition 2. *When c_B increases in the* ***constrained bank dealer*** *equilibrium:*

1. *The spread is unchanged ($S^\star = c_{NB}$), the matchmaking fee $f^\star$ decreases, and average transaction costs decrease.*

2. *Trading volume increases, matchmaking increases, and market making decreases.*

3. *Overall customer welfare, π_c, increases.*

The comparative statics in Proposition 2 are all unambiguously signed. As competition from the non-bank dealer constrains the bank dealer's ability to pass the rising regulatory costs on to his market-making customers, the bank dealer seeks to extract higher profits from the matchmaking business. To this end, he increases overall trading volume (which is equivalent in our setting to increasing the fraction of customer types who trade, $1 - G(f^{\star})$) by lowering the matchmaking fee to attract customers with a low need for immediacy.

Shifting more trading to the less expensive matchmaking service lowers average transaction costs in the market, but even more notable is that overall customer welfare increases. The welfare gains come from three groups of customers. The first group refrains from trading when c_B is sufficiently low, but when the bank dealer lowers the matchmaking fee they begin trading and thus contribute to overall customer welfare. The second group consists of customers who trade via the matchmaking mechanism either way, but when c_B increases their welfare goes up because they pay a lower fee.

The third group are those customers who find it optimal to switch from market making to matchmaking when c_B increases because the matchmaking fee goes down. Given that the spread in this equilibrium region is constrained, their expected utility from using the market-making service does not change when c_B rises. Their choice to switch means that their expected utility considering both the lower fee and the expected waiting costs in the matchmaking service is higher (except for the marginal customer who is indifferent between the two trading mechanisms). Hence, all customers are either better off or no worse off as c_B goes up, which means that equilibria with higher bank regulatory costs Pareto-dominate those with lower bank regulatory costs in this region.[25]

It is in the constrained equilibrium that customers derive the most benefit from increased bank regulatory costs because the industrial organization angle (increased competition from the non-bank dealer) interacts with the market microstructure angle (increased utilization of a lower-cost alternative trading mechanism), leading to reduced bank dealer rents and increased overall customer welfare. We stress that if the bank dealer had operated only a market-making business, investor

[25] To simplify the model, the matching rate of customers in the matchmaking mechanism is directly determined by the bank dealer's search technology. We believe that the increase in overall customer welfare when bank regulatory costs go up in the constrained bank dealer equilibrium would hold in a more general search specification (e.g., Duffie, Gârleanu, and Pedersen (2005)) in which the matching intensity depends on the mass of customers who choose matchmaking. Specifically, Proposition 2 shows that the market share of matchmaking monotonically increases in c_B. Under an alternative structure in which the size of the pool of searching customers impacts the matching rate, this increase in market share would result in a higher matchmaking speed $\mathcal{H}$ and make the matchmaking service even more beneficial to customers. In this case, the switch to matchmaking would likely be more pronounced as c_B goes up, further improving overall customer welfare.

welfare would not have increased in this equilibrium. The change in equilibrium market structure is therefore at the core of the result that overall customer welfare increases.

While we can see from Proposition 1 that the constrained equilibrium is more prevalent in that it exists for the entire range of search costs $(0, \mathcal{H}\phi(c_{NB}))$ and irrespective of whether $\phi(c_{NB})$ is negative or positive, the unconstrained bank dealer equilibrium prevails if both the search cost and the balance sheet cost of the bank dealer are very low. The next proposition shows that the manner in which increased regulatory costs impacts customers and market outcomes in this region can differ considerably.

Proposition 3. *When c_B increases in the **unconstrained bank dealer** equilibrium:*

1. *The spread $S^\star$ increases, the matchmaking fee $f^\star$ is unchanged, and average transaction costs increase if $c_B < (1-\mathcal{H}) f^\star + I$ and decrease if $c_B \geqslant (1-\mathcal{H}) f^\star + I$.*

2. *Trading volume is unchanged, matchmaking increases, and market making decreases.*

3. *Overall customer welfare, π_c, decreases.*

Why are customers worse off in the unconstrained equilibrium? The bank dealer is a monopolist in both the market-making and matchmaking businesses. He passes any increase in regulatory costs to the market-making customers by increasing the spread, leaving his profit per trade in this trading mechanism unchanged. Given a particular distribution of customers' private value (or patience), a higher matchmaking fee increases compensation from each trade while decreasing the number of customers who choose to trade. This tradeoff results in a unique fee that maximizes the bank dealer's expected profit from matchmaking that depends only on the distribution of private values (and hence does not change as bank regulatory costs increase). As c_B increases, some customers stop paying the spread and opt instead to use the matchmaking mechanism, but the population of customers who refrain from trading (which depends on the magnitude of the matchmaking fee) does not change and hence volume is unchanged.

Average transaction costs can be computed as the weighted average of the market-making spread and the matchmaking fee with the respective populations of the trading customers in each of these mechanisms as weights. When regulatory costs increase, the spread increases but the fraction of customers choosing the market-making mechanism declines, and hence the direction of the change in average transaction costs can go either way. If the cost of balance sheet financing is lower than the bank dealer's matchmaking costs (the out-of-pocket search cost as well as the expected waiting cost to receive the fee), increasing the balance sheet cost will result in higher average transaction costs. Whether average transaction costs reach an interior maximum and begin declining before the unconstrained equilibrium changes to the constrained equilibrium depends on the specific distribution of customers' private values.

Irrespective of what happens to average transaction costs, however, overall customer welfare unequivocally decreases in the unconstrained equilibrium. This result is important in that it serves to demonstrate that the welfare of customers is not equivalent to measures that can be estimated from executed trades. All trading customers are worse off in this equilibrium when c_B increases, either because they pay a higher spread or because they are priced-out of the market-making service and they incur waiting costs when utilizing the matchmaking service. Furthermore, the population of customers who refrain from trading remains the same (because the matchmaking fee is unchanged), and hence overall customer welfare is lower when bank regulatory costs increase.

The unconstrained bank dealer equilibrium fits with views expressed by some market participants according to which a higher bank dealer balance sheet cost would negatively impact market liquidity and customer welfare. We stress that, in our model, these results arise because the bank dealer has an unconstrained monopoly on the provision of immediacy. In addition, the existence of this equilibrium critically depends on there being both a rather large too-big-to-fail subsidy (low c_B) and a rather low search cost in the matchmaking business. Considering the entire structure of equilibria when $c_B \leq c_{NB}$, though, a key insight that arises from our model is that counteracting the too-big-to-fail subsidy by increasing bank regulatory costs is not welfare-improving without market discipline. Namely, competition from non-bank dealers who stand ready to offer market-making services is crucial to attaining the welfare-improvement result.

4.1.2 When $c_B > c_{NB}$

If regulatory costs were to further increase, the bank dealer's balance sheet cost could potentially exceed the non-bank dealer's balance sheet cost, c_{NB}. In such a case, the bank dealer would not provide market-making services because the non-bank dealer could always charge $S_{NB} = c_B - \varepsilon$ and attract all customers who are willing to pay the spread to trade immediately. When $c_B = c_{NB}$, the only equilibrium spread possible is $S = c_B = c_{NB}$ and both bank and non-bank dealers make zero profits from market making. We include this boundary point as part of Proposition 1, but it can be added as easily to the case of $c_B > c_{NB}$ or we could also assume that customers who would like to use the market-making mechanism randomize between the bank and non-bank dealers.

Given that Bessembinder, Jacobsen, Maxwell, and Venkataraman (2018) estimate bank dealers handle about 87% of principal trading even after the post-crisis regulatory reform, why should we care about the case of $c_B > c_{NB}$? The reason is that the same paper also presents evidence that non-bank dealers increased their market share in principal trading from about 3% in the pre-crisis period to about 13% in the post-regulatory-reform period. Our model is simplified in that we have a bang-bang solution: either the bank dealer or the non-bank dealer captures all market-making clients. The increase in principal trading by non-bank dealers documented by Bessembinder et al. may suggest that we are getting closer to the point where $c_B = c_{NB}$. It is therefore important to investigate what would happen to overall customer welfare if bank regulatory costs were to increase past this point.

As before, we focus only on equilibria in which the bank dealer engages in matchmaking. Specifically, given the bank dealer's choice of f, the non-bank dealer's problem is

$$\pi_{NB}(c_B) = \max_{c_{NB} \le S \le c_B} \Pi_{NB}(S) = \frac{2\mu}{r} \left[(S - c_{NB})(1 - G(b)) \right].$$

Here, we impose a tie-breaking rule according to which the bank dealer does not offer the market-making service if its profit would be zero. This implies that, when $S_{NB} = c_B$, the bank dealer does not operate the market-making business.

Given the non-bank dealer's choice of S, the bank dealer's problem is

$$\pi_B(c_B) = \max_{0 \leq f \leq S} \Pi_B(f) = \frac{2\mu}{r}(\mathcal{H}f - I)(G(b) - G(f)),$$

where $b = \frac{S - \mathcal{H}f}{1-\mathcal{H}}$. The following proposition establishes the existence of an equilibrium when the balance sheet cost of the bank dealer exceeds that of the non-bank dealer.

Proposition 4. *When $c_B > c_{NB}$ and $I < \mathcal{H}\min\{\tilde{c}_B, c_B\}$, there exists a unique equilibrium such that the non-bank dealer operates the market-making service and the bank dealer operates the matchmaking service, with $\tilde{c}_B$ as the unique solution of*

$$\xi(\tilde{c}_B) - \frac{\tilde{c}_B - c_{NB}}{1-\mathcal{H}} = 0,$$

provided that G is concave or G is convex with $G''' < 0$ and $\mathcal{H} < \frac{1}{2}$. In particular, there exists $c_2 > \max\{c_{NB}, \frac{I}{\mathcal{H}}\}$ such that,

1. *If $c_B \in (c_{NB}, c_2]$, the equilibrium is a **constrained non-bank dealer** equilibrium with $S^\star = c_B$.*
2. *If $c_B \in (c_2, \infty)$, the equilibrium is an **unconstrained non-bank dealer** equilibrium with $S^\star < c_B$.*

As before, all equilibria with bank dealer matchmaking require that the search cost is not overly high. In addition, the conditions on the curvature of the function G are required because under these conditions the bank dealer's best response is always continuous and unique, which can guarantee the existence of equilibrium. When c_B just exceeds c_{NB}, we enter the constrained non-bank dealer equilibrium region in which the non-bank dealer's spread, S_{NB}, is equal to the bank dealer's balance sheet cost, c_B. As c_B rises, the competitive pressure from the bank dealer eases, and the non-bank dealer can increase his market-making spread to extract more rents. Beyond a certain point, the bank dealer no longer exerts any competitive pressure and we reach an unconstrained non-bank dealer equilibrium, where S_{NB} is the non-bank dealer's monopoly spread. In the unconstrained equilibrium, further increases in the bank dealer's balance sheet cost no longer affect the equilibrium outcomes.

The unconstrained non-bank dealer equilibrium appears somewhat extreme and hence of limited interest. The constrained non-bank dealer equilibrium, on the other hand, can help us answer the question what happens to customer welfare and market outcomes when the balance sheet cost of the bank dealer just passes that of the non-bank dealer.

Proposition 5. *When c_B increases in the* ***constrained non-bank dealer*** *equilibrium:*

1. *$S^\star = c_B$ increases, $f^\star$ increases, and the change in average transaction costs is ambiguous.*

2. *Trading volume decreases, market making decreases, and the change in matchmaking is ambiguous.*

3. *Overall customer welfare, π_c, decreases.*

Increasing regulatory costs past the point at which the bank dealer's balance sheet cost is equal to c_{NB} weakens rather than strengthens competition. The non-bank dealer increases his spread, creating an opportunity for the bank dealer to increase his matchmaking fee as well. As a result, more customers forgo trading and hence trading volume decreases. Most importantly, overall customer welfare declines. The decline in welfare stems from the increase in market-making spread and matchmaking fee for those customers who trade as well as the increase in the number of customers who choose to refrain from trading because of the higher costs. At the same time, we can show that both dealers' profits go up unambiguously when bank regulatory costs increase as each dealer specializes in a different business and is able to use his market power to extract higher rents (the non-bank dealer from market making and the bank dealer from matchmaking).

This equilibrium demonstrates the peril of raising bank regulatory costs too much. While customers benefit when the increase in regulatory costs counters the too-big-to-fail subsidy and enhances competition, they fare worse if these costs are set so high that they push the bank dealer out of the market-making business.

4.2 What Happens when the Cost of Matchmaking Declines?

In this section we focus on the second important development of the past decade in the corporate bond market: technological advances that reduced the cost of matchmaking and therefore rendered the matchmaking mechanism more attractive.[26] This development is represented in our model by a reduction in the cost of search (or effort) required to effect a transaction in the matchmaking mechanism, I. The following proposition shows that, when competition between the bank and non-bank dealer constrains their strategies, a lower search cost unambiguously benefits customers.

Proposition 6. *When I decreases in either the* ***constrained bank dealer*** *equilibrium (section 4.1.1) or the* ***constrained non-bank dealer*** *equilibrium (section 4.1.2):*

1. *$S^\star$ is unchanged, $f^\star$ decreases, and average transaction costs decrease.*

2. *Trading volume increases, matchmaking increases, and market making decreases.*

3. *Overall customer welfare, π_c, increases.*

A lower matchmaking cost incentivizes the bank dealer to decrease the matchmaking fee to attract more customers to this trading mechanism. Competition between the bank and non-bank dealers in market-making services exerts enough pressure on the spread to essentially fix it to equal the higher of the two balance sheet costs irrespective of changes in the matchmaking cost. This leads to an unambiguous improvement in the welfare of customers who use the matchmaking mechanism, including those who optimally choose to switch from market making to matchmaking and those who start trading only after I declines. Higher volume in the matchmaking service also benefits the bank dealer, whose profits rise alongside the increase in overall customer welfare.[27]

The key to this unambiguous improvement in overall customer welfare, however, is binding competition in market-making services. We can clearly see the importance of competition by

[26] The report on electronic trading in fixed income markets issued by the Markets Committee of the Bank for International Settlements notes that growth in electronic trading has occurred primarily in the dealer-client segment, with technological innovations giving rise to new trading venues and protocols (BIS Markets Committee (2016)). Many of the systems in this segment are single-dealer platforms that are typically based on permutations of the request-for-quote (RFQ) protocol.

[27] Profits of the non-bank dealer, on the other hand, decline. As the bank dealer lowers the matchmaking fee, customers switch from market making to matchmaking, leaving the non-bank dealer with a smaller market share and hence lower profits.

comparing these results to the comparative statics obtained in the unconstrained bank dealer equilibrium in which the bank dealer operates as a monopolist in both trading mechanisms.

Proposition 7. *When I decreases in the* ***unconstrained bank dealer*** *equilibrium (section 4.1.1):*

1. *$S^{\star}$ increases (decreases) if ζ is convex (concave), $f^{\star}$ decreases, and the change in average transaction costs is ambiguous.*

2. *Trading volume increases, matchmaking increases, and market making decreases.*

3. *The change in overall customer welfare, π_c, is ambiguous if ζ is convex, but is positive if ζ is concave.*

While we can show that the bank dealer's profits unambiguously rise, the same cannot be said about overall customer welfare. We discussed the curvature of the Mills ratio, ζ, in Section 4.1.1, where it mattered for the shape of the boundary between regions A and B in Figure 1. Many familiar distributions (e.g., normal, uniform) have a convex ζ, although one could easily construct or specify distributions for which ζ is concave. Still, given the prevalence of distributions with convex ζ, it is important to note that for such distributions the change in overall customer welfare is ambiguous.

The reason for this ambiguity is that the monopolist bank dealer looks to increase profit margins from the market-making business by raising the spread. Customers with high immediacy needs who utilize the market-making mechanism are then worse off to a degree that can overwhelm improvements for customers who place a low value on immediacy. The surplus generated by a lower search cost in such a case can go mainly to the bank dealer, not to the customers. To ensure that customers unequivocally benefit from technological advances that lower search costs, one needs to be in the constrained equilibrium, which means that regulators need to increase the balance sheet cost of the bank dealer from the very low level that enables the unconstrained bank dealer equilibrium to prevail.

5 Did Post-Crisis Bank Regulations Harm Investors in the Corporate Bond Market?

We can use the model's insights to evaluate whether bank regulations following the financial crisis lowered overall investor welfare. Given the wide availability of empirical evidence regarding the US corporate bond market, we focus our discussion on this market, but the model's insights would apply to other markets that were similarly affected by post-crisis bank regulations. As we stress in the introduction, empirical estimates of transaction-based measures of liquidity cannot capture investors' loss of welfare when finding a trading counterparty takes more time or when investors forgo trading altogether. In contrast, our theoretical model takes these potential welfare losses into account alongside the out-of-pocket cost of trading. Our analysis shows that the implications of bank regulations for overall customer welfare depend on the manner in which competition between the bank and non-bank dealers shapes the equilibrium. Hence, to evaluate the impact of post-crisis regulations on welfare we need to tie the equilibrium regions in our model to the state of affairs in the corporate bond market around the enactment of the bank regulations. In this section we attempt to do so by mapping the implications of our model for observable market outcomes onto the findings reported in the empirical literature.

The first set of results involves the somewhat surprising empirical finding that average transaction costs have declined (Mizrach (2015), Adrian, Fleming, Shachar, and Vogt (2017), Anderson and Stulz (2017)) or have not changed (Bessembinder, Jacobsen, Maxwell, and Venkataraman (2018), Trebbi and Xiao (2019)) following the enactment of post-crisis regulations. In our model, the most prevalent equilibrium is the constrained bank dealer equilibrium in which average transaction costs unequivocally go down as c_B increases (region A in Figure 1). Still, if both the matchmaking cost and the bank dealer's balance sheet cost are low enough, the decline in average transaction costs could possibly start already in the right portion of the unconstrained bank equilibrium (region B in Figure 1).[28] Hence, the observation of declining average transaction costs is consistent with an increase in the balance sheet cost of the bank dealer starting from the right portion of region B

[28] This would happen if c_B increases beyond $(1 - \mathcal{H})\, f^{\star} + I$ before the equilibrium changes to the constrained bank equilibrium.

and across region A.

The second set of empirical results is that the cost of immediacy, or the cost of trading via the market-making mechanism, has risen (Bao, O'Hara, and Zhou (2018), Dick-Nielsen and Rossi (2018), Choi and Huh (2017)). In our model, this is also consistent with a movement that starts in region B and ends up in region A, although it could also be consistent with an increase in c_B that potentially overshoots and exceeds the balance sheet cost of the non-bank dealer in the constrained non-bank dealer equilibrium. The third set of empirical findings document an increase in matchmaking volume driven by bank dealers (Bao, O'Hara, and Zhou (2018), Choi and Huh (2017), Schultz (2017)) and a significant decline in capital commitment to market making by bank dealers without a sufficiently large offsetting increase in market making by non-bank dealers (Bao, O'Hara, and Zhou (2018), Bessembinder, Jacobsen, Maxwell, and Venkataraman (2018)). In our model, such a pattern can be observed throughout the constrained and unconstrained bank equilibria (regions A and B). This is not the pattern in the unconstrained non-bank dealer equilibrium, however, which suggests that it is unlikely that post-crisis regulations left us in the region where $c_B > c_{NB}$.

In the model, there is a clear distinction between the constrained bank equilibrium, where the bank dealer captures the entire market-making business, and the unconstrained non-bank equilibrium, where the non-bank dealer accommodates all customers who demand immediacy. We would expect a more gradual transition in reality. The empirical evidence of an increase (decline) in market-making activity by non-bank (bank) dealers following the enactment of post-crisis regulations suggests an increase in regulatory costs across region A that left us currently somewhere in the right portion of this region. However, the empirical evidence reported in Bessembinder, Jacobsen, Maxwell, and Venkataraman (2018), which shows that bank dealers completely dominate principal trading even after the post-crisis regulatory reform (with 87% of principal trading versus 13% of the non-bank dealers), suggests to us that c_B is still currently below c_{NB}.

The last set of empirical results concerns trading volume, and here we note that the findings appear to be somewhat nuanced. Overall trading volume in bonds has significantly increased, while turnover in each particular bond issue appears to have declined (increased) in more (less) active bonds (BIS Committee on the Global Financial System (2014), Mizrach (2015), Adrian, Fleming, Shachar, and Vogt (2017)). Turnover is computed as dollar trading volume divided by the value of the bond issue and is used by some market participants as yet another measure of volume alongside dollar trading volume. Given that volume is normalized by the size of the issue, though, patterns in issuance affect this measure. Some market observers note that the low interest-rate environment greatly boosted the attractiveness of bond financing, and business enterprises responded by issuing a record number of bonds. According to this explanation, abnormal issuance, not a decline in the desire to trade, reduced turnover in some bond issues (BIS Committee on the Global Financial System (2014)). In other words, the higher observed overall dollar volume could be driven by lower average transaction costs that attract more customers to trade bonds (as in our model), but the record number of new issues means that some bond issues would be traded less often. The only equilibrium region in our model in which dollar volume increases as it has in the empirical work is the constrained bank dealer equilibrium, suggesting again an increase in the balance sheet cost across region A.

Summarizing our discussion, the four sets of empirical findings associated with the implementation of bank regulations in the aftermath of the financial crisis all point to the particular role that the constrained bank dealer equilibrium plays in our model. It is reasonable to conjecture that we were in the unconstrained bank dealer equilibrium (region B) in the pre-crisis period and ended up in the constrained bank dealer equilibrium after bank regulations were imposed. The decline in average transaction costs and simultaneous increase in volume suggests that we have probably moved across a large portion of the c_B range in the constrained bank equilibrium getting closer to the point at which $c_B = c_{NB}$. With this in mind, we can now consider our model's predictions for overall customer welfare.

As Proposition 2 shows, overall customer welfare unequivocally increases as the balance sheet cost of the bank dealer increases throughout the constrained bank dealer equilibrium in region A. The clear mapping of the empirical findings to an increase in regulatory costs across region A in the model coupled with our welfare implications indicate that investors in the corporate bond market are likely better off under the current bank regulatory regime compared with the one that prevailed before the financial crisis. We cannot say whether a particular feature of the Basel framework is responsible for this result and our results do not imply that the Volcker Rule improved liquidity. Rather, considering the empirical findings pertaining to the impact of the entire suite of post-crisis regulations, our model would suggest that the increase in market-making costs benefits corporate bond investors even during normal times. We believe that this insight is an important contribution of our work to the debate over the merits of bank regulations that were imposed in the aftermath of the financial crisis.

Lastly, our observation that post-crisis bank regulations likely caused a shift from an equilibrium in region B to an equilibrium in region A is also important when considering technological innovations that decrease search costs in the matchmaking business. In our model, the bank dealer earns higher profits by adopting innovations that lower search costs in either equilibrium region. However, the impact of such innovations on overall customer welfare is ambiguous in region B, while Proposition 6 demonstrates that overall customer welfare unambiguously increases when search costs decline in region A. In the equilibrium that we believe is most likely prevailing in the market today, therefore, the surplus from lower search costs is shared by both customers and bank dealers. As such, regulations that help facilitate (or remove obstacles that could hamper) the adoption of such innovations by bank dealers would make investors in the corporate bond market better off.

6 Robustness

In this section, we study the robustness of our main results. Section 6.1 presents a variant of the model in which multiple differentiated bank dealers compete. Section 6.2 examines an alternative measure of welfare.

6.1 An Oligopoly Model of Bank Dealers

Our baseline model has a single representative bank dealer. To some readers, this monopoly setting may appear stark. In this section, we study an oligopoly model in which multiple differentiated bank dealers compete for customers. We show that, as in the representative bank dealer model, there is a parameter region in which overall customer welfare increases in the bank dealers' balance sheet cost. We further show that this region would shrink to zero only if bank dealers were perfectly competitive (i.e., completely undifferentiated). Thus, as long as bank dealers have even a small amount of market power, our main result—raising the balance sheet cost of bank dealers can increase overall customer welfare—is robust.

Model setup

There are $N + 1$ dealers, consisting of N bank dealers and one non-bank dealer. The N bank dealers, indexed by j, have identical balance sheet cost c_B, search cost I, and matchmaking speed $\mathcal{H}$. The non-bank dealer has balance sheet cost c_{NB}. Because the most empirically relevant case is when the bank dealers have a balance sheet cost advantage over the non-bank dealer, we focus in this section on the parameter region $c_B \leq c_{NB}$.

There is a static mass $\frac{2N\mu}{r}$ of customers, indexed by i, and customer i has a "taste" $\epsilon_{ij} \sim F(\sigma)$ for being a customer of bank j. The taste ϵ_{ij} can be interpreted as customer i's private value for using bank j. Customers have zero taste for the non-bank dealer. The customer taste shocks ϵ_{ij} generate differentiation among bank dealers, which limits customer substitution and allows bank dealers to charge positive markups. The variance of the taste shocks, σ, captures the degree of differentiation and therefore competition between bank dealers. If taste shocks have a low variance, bank dealers are more homogeneous, customers' choice between them is primarily determined by their prices, and bank dealers compete intensely by choosing lower market-making spreads and matchmaking fees. If, on the other hand, taste shocks have a large variance, bank dealers are more differentiated, customers put less weight on prices, and bank dealers compete less intensely by choosing higher S and f. Our use of the variance of the taste shocks to model the degree of differentiation and competition follows from Perloff and Salop (1985).

The timing of the model is as follows:

Stage 1. Each bank dealer j chooses (f_j, S_j), and the non-bank dealer chooses S_{NB}, all simultaneously.

Stage 2. Each customer i receives the ϵ_{ij} taste shock and chooses to affiliate with bank dealer j. Once affiliated, the customer can use only the services (matchmaking or market making) of bank dealer j.

Stage 3. Each customer i observes her private benefit for trading immediately, $x \sim G$, and chooses between:

(a) using bank dealer j's matchmaking service with the fee f_j,

(b) using bank dealer j's market-making service with the spread S_j,

(c) using the non-bank dealer's market-making service with the spread S_{NB}, and

(d) not trading.

In this setup, customers can substitute between market making, matchmaking, and not trading in stage 3, and substitute between bank dealers in stage 2. The customer's choice set in stage 3 is identical to that in the baseline model, and the customer's choice between multiple bank dealers in stage 2 can be viewed as an "overlay" stage of the game whose outcome is taken as given in our baseline model. As a result, if we take the limit as the variance of the taste shocks goes to infinity, we effectively recover our baseline monopolist bank dealer model. On the other hand, if we take the limit as the variance of the taste shocks goes to zero we end up in an economy in which bank dealers are homogeneous and therefore engage in a Bertrand competition. Varying the variance of the taste shocks, therefore, enables us to examine how the degree of market power bank dealers wield influences our conclusions.[29]

Customers' decisions

As usual, we use backward induction to solve this model. In stage 3, customer i is already locked in to a single bank dealer. Facing (f_j, S_j, S_{NB}), customer i's decision is the same as that of the customer in our baseline model.

In stage 2, before observing her private value for trading, customer i's expected utility from

[29]This particular oligopoly setup achieves consistency with the baseline model and at the same time builds in the customer's choice between bank dealers in a tractable framework. For this reason, although the setup rules out substitution between bank dealers after the customers observe their private values, we view it as a reasonable compromise.

affiliating with bank dealer j is

$$\pi_{cji}(S_j, f_j) = \underbrace{\int_{x=f}^{b} (x - f_j)\mathcal{H} dG(x) + \int_{x=b}^{\infty} (x - S_j) dG(x)}_{\bar{\pi}_{cj}(S_j, f_j)} + \epsilon_{ij}, \tag{11}$$

where $b = \frac{S_j - \mathcal{H} f_j}{1 - \mathcal{H}}$, and we have labeled the portion of utility not related to taste as $\bar{\pi}_{cj}(S_j, f_j)$. The customer's decision problem in stage 2 is therefore

$$\max_j \pi_{ci}(S_j, f_j). \tag{12}$$

For tractability, we assume that $\{\epsilon_{ij}\}$ are independent and identically distributed logistic random variables with variance $\frac{\pi^2}{6}\sigma^2$. This parametrization implies that the share of customers who choose bank j is

$$s_j(S_j, f_j, S_{-j}, f_{-j}) = \frac{\exp(\frac{1}{\sigma}\bar{\pi}_{cj}(S_j, f_j))}{\sum_{j'} \exp(\frac{1}{\sigma}\bar{\pi}_{cj'}(S_{j'}, f_{j'}))}, \tag{13}$$

where $S_{-j} = (S_{j'})_{j' \neq j}$ and $f_{-j} = (f_{j'})_{j' \neq j}$ are the other bank dealers' prices.

Bank dealers' decisions

In stage 1, each bank dealer sets prices to maximize his profits. The bank dealer's expected per-capita profits is

$$\pi_j = (\mathcal{H} f_j - I)(G(b) - G(f_j)) + (S_j - c_B)(1 - G(b)). \tag{14}$$

Then, bank dealer j's problem is

$$\max_{0 \leq f_j \leq S_j \leq c_{NB}} \Pi_j \equiv \frac{2N\mu}{r} \times s_j(S_j, f_j, S_{-j}, f_{-j}) \times \pi_j. \tag{15}$$

We focus on symmetric strategies, that is, for all j and j', we have $(f_j, S_j) = (f_{j'}, S_{j'})$. Assuming that such a symmetric equilibrium exists, each bank dealer's market share would be $s_j = 1/N$.

Taking the first-order condition of Π_j and substituting in $s_j = 1/N$, we obtain

$$\frac{\partial \Pi_j}{\partial f_j} = \frac{2N\mu}{r}\left[\frac{N-1}{N^2}\frac{1}{\sigma}\frac{\partial \bar{\pi}_{cj}}{\partial f_j}\pi_j + \frac{1}{N}\frac{\partial \pi_j}{\partial f_j}\right] = \frac{2\mu}{r}\left[\frac{N-1}{N}\frac{1}{\sigma}\frac{\partial \bar{\pi}_{cj}}{\partial f_j}\pi_j + \frac{\partial \pi_j}{\partial f_j}\right], \tag{16}$$

$$\frac{\partial \Pi_j}{\partial S_j} = \frac{2N\mu}{r}\left[\frac{N-1}{N^2}\frac{1}{\sigma}\frac{\partial \bar{\pi}_{cj}}{\partial S_j}\pi_j + \frac{1}{N}\frac{\partial \pi_j}{\partial S_j}\right] = \frac{2\mu}{r}\left[\frac{N-1}{N}\frac{1}{\sigma}\frac{\partial \bar{\pi}_{cj}}{\partial S_j}\pi_j + \frac{\partial \pi_j}{\partial S_j}\right]. \tag{17}$$

From this expression, it is clear that, as $\sigma \to \infty$, the two partial derivatives converge to $\frac{2\mu}{r}\frac{\partial \pi_j}{\partial f_j}$ and $\frac{2\mu}{r}\frac{\partial \pi_j}{\partial f_j}$, respectively, which are the same as the first-order conditions in the baseline model. Therefore, if bank dealers are sufficiently differentiated, they will have a stable set of customers based on taste and would no longer compete for market share.

Proposition 8. *As $\sigma \to \infty$, the equilibrium spread and fee charged by each bank dealer converge to those in the baseline model.*

On the other hand, as $\sigma \to 0$, bank dealers are no longer differentiated and can attract customers only by offering low prices. The following proposition shows that raising the balance sheet cost of perfectly competitive bank dealers always decreases overall customer welfare.

Proposition 9. *As $\sigma \to 0$, in the limiting symmetric equilibrium, bank dealers earn zero profits, and increasing c_B results in lower overall customer welfare.*

These two limiting cases illustrate the flexibility of this oligopoly model: it spans the monopoly case at one extreme and the competitive case at the other. While the two limiting cases are relatively simple, the generic solution for any positive but finite σ is involved. We therefore resort to numerical solutions, and demonstrate the most salient result in Figure 2. Bessembinder, Jacobsen, Maxwell, and Venkataraman (2018) report that "The number of dealers that comprise 70% of market share each year ranges between 10 and 12." Given that trading in corporate bonds is concentrated among a small number of very large bank dealers, we use $N = 10$ for the number of bank dealers in our example. Case I, Case II, and Case III in the figure represent combinations of I, $\mathcal{H}$, and c_{NB} that correspond to Part 1, Part 2(a), and Part 2(b) of Proposition 1.

Our main result in the baseline model is that there is a region over which raising the balance sheet cost of the bank dealer increases overall customer welfare. Our goal in this robustness section is to demonstrate how the size of this region changes with the degree of market power. Therefore, Figure 2 showcases c_B on the y-axis and σ on the x-axis, with market power increasing as we move to the right. For each σ, the lighter orange area represents the range of c_B for which customer welfare is increasing in c_B, and the darker blue area represents the range for which welfare is decreasing in c_B.

A salient theme across all three cases is that, as long as there is some degree of differentiation between banks (i.e., a modest positive σ), overall customer welfare increases in c_B in a nontrivial range of the parameter. Equally interestingly, this region seems to expand rapidly as σ increases. Only if $\sigma = 0$ does customer welfare always decrease in c_B. While we have not been able to derive the analytical characterizations for these regions, the numerical solutions nonetheless show that our main result—overall customer welfare increasing in the bank dealers' balance sheet cost—is not an artifact of the baseline monopoly model. This insight is rather robust as long as bank dealers possess any positive amount of market power.

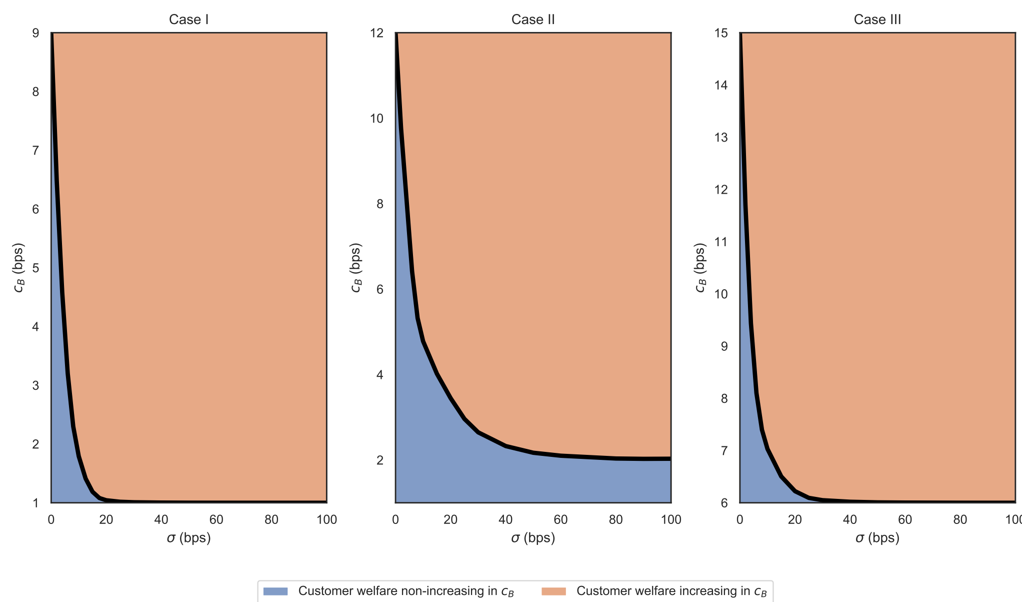

Figure 2: **Regions of (σ, c_B) in which overall customer welfare increases or decreases in c_B.** In all three cases, $N = 10$, $\mathcal{H} = 0.5$, and we use the exponential distribution for G, with $E(G) = 10$ bps. In Case I, $c_{NB} = 9$ bps and $I = 0.5$ bp. In Case II, $c_{NB} = 12$ bps and $I = 0.5$ bp. In Case III, $c_{NB} = 15$ bps and $I = 3$.

6.2 An Alternative Welfare Measure

Throughout the paper we focus on customer welfare as the metric with which to evaluate the ultimate benefit or cost of post-crisis regulations. This quantity, defined in (5), reflects the gains from trade of customers in the model (who represent the investor population in our economy). For us, this is a natural benchmark for regulators to consider.

One may argue that our customer welfare measure misses two components. First, the too-big-to-fail subsidy the bank dealer receives—reflected in the difference in balance sheet costs between the bank and non-bank dealers—ultimately comes out of the pocket of investors (or the public in general), so we may want to include this adjustment. Second, both the bank and non-bank dealers are agents in the economy, and one may want to include them in the definition of welfare as well.

Therefore, we consider an alternative welfare measure as the sum of customer welfare and the two dealers' profits, while adjusting for the implicit too-big-to-fail subsidy of the bank dealer:

$$W = \pi_c + \pi_B + \pi_{NB} - \underbrace{\frac{2\mu}{r}\left(1 - G\left(b\right)\right)\left(c_{NB} - c_B\right)\mathbb{I}_{S=S_B}}_{\text{Too-big-to-fail subsidy of the bank dealer, if } c_B < c_{NB}} . \tag{18}$$

The first three terms of this expression are defined in the model section, but the last term requires additional discussion. Our assumption is that, without giving an implicit public subsidy to the bank dealer, the balance sheet costs of the bank dealer would be similar to those of the non-bank dealer—that is, the same market-making activity would require the same cost of capital regardless of who provides it.[30] Hence, the difference between the two balance sheet costs, multiplied by the number of customers who trade using the market-making services of the bank dealer, represents the welfare loss to investors associated with granting this subsidy to the bank dealer. As post-crisis regulations increase the market-making costs of the bank dealer, the likelihood that the dealer would need to be rescued decreases, and with it the implicit subsidy. As c_B approaches c_{NB}, the subsidy given to the bank dealer's balance sheet cost disappears. The indicator function at the end of the term is there to ensure that the subsidy is considered only when the bank dealer makes

[30] The cost of capital in this context means the cost of capital of the activity, not of the entire institution that may have many subsidiaries. In the data, the cost of capital for each business line might be difficult to observe.

markets.

Because the subsidy to the bank dealer is relevant if and only if the bank dealer makes markets, we can make the intuition in (18) more transparent by expanding it in the region $c_B < c_{NB}$ (hence $\pi_{NB} = 0$):

$$W = \frac{2\mu}{r}\left[\int_{x=f}^{b}(\mathcal{H}x - I)dG(x) + \int_{x=b}^{\infty}(x - c_{NB})dG(x)\right], \text{ if } c_B < c_{NB}. \tag{19}$$

This expression essentially measures investors' gains from trades, after adjusting for the search cost I and the non-bank dealer's balance sheet cost c_{NB}, which is taken to be the market-determined, risk-based balance sheet cost of market-making activity. While c_B does not show up explicitly in (19), an increase in c_B changes the equilibrium values of f, b, and $\mathcal{H}$. Overall, W reflects the allocative efficiency of the market. Its comparative statics with respect to c_B are shown in the following proposition.

Proposition 10. *In the equilibrium of Proposition 1, the welfare measure W is increasing in c_B.*

Proposition 10 shows that, once the subsidy of the bank dealer is explicitly accounted for, market-wide welfare increases in the balance sheet cost of the bank dealer. While this result is unambiguous, there are good reasons to focus on overall customer welfare, π_c, rather than on W. In particular, the dealers' profits π_B and π_{NB} arise from their market power, and it is hard to imagine a situation in which regulators justify their actions by appealing to the need to maximize banks' market power rents. As such, we think that overall customer welfare remains the most natural metric in our framework with which to judge the impact of post-crisis regulations on the corporate bond market.

7 Concluding Remarks

Our paper highlights the complex and multifaceted consequences that post-crisis bank regulations have for market liquidity and investor welfare. While the explicit goal of those regulations was to enhance financial stability in times of stress, our work shows that these regulations can improve market efficiency even during normal times by prompting a change in the nature of liquidity provision. In particular, post-crisis regulations eliminate obstacles to competition in the most profitable business (market making) and incentivize bank dealers to reprice their services to attract customers to an alternative market structure (matchmaking) that is better at serving the needs of many customers. The industrial organization angle combines with the market microstructure angle to deliver this positive outcome.

The key insight we offer in this paper is that an increase in regulatory costs imposed on banks can make customers in the corporate bond market better off. Specifically, our focus is on balance sheet costs that were imposed on banks following the financial crisis to deter proprietary trading and as a result increased the costs of market making. A notable aspect of our work is that we do not focus on whether these costs reduce the likelihood or severity of a crisis, which are presumably the stated goals of the regulation. Rather, we show that the increase in regulatory costs can lower average transaction costs during normal times and, more importantly, increase overall customer welfare. The driving forces behind our results are three elements that characterize the corporate bond market: the coexistence of two distinct trading mechanisms (market making and matchmaking), the market power enjoyed by bank dealers, and potential market-making competition from non-bank dealers. Although our paper is motivated by and specifically addresses observations in the corporate bond market, the model we present can be applied to other over-the-counter markets that feature these three elements.

How can customers be made better off by regulations that deter market making? The rationale is that while market making is a very profitable form of intermediation for bank dealers, it can be an expensive form of intermediation for customers. When regulation increases the cost of market making for bank dealers and competition prevents them from transferring the cost increase to their customers, bank dealers are incentivized to lower fees in their matchmaking business in order to raise the volume of trading. This increases overall customer welfare because more customers can now be better served by a menu of trading mechanisms that cater to their needs. If market making were the only trading mechanism, raising regulatory costs would not have made customers better off. The change in equilibrium market structure is integral to the result that overall customer welfare increases. The welfare gains are even more pronounced when combined with the second important development of the past decade: technological advances that reduce search costs in the matchmaking mechanism.

Regulators always worry about the unintended consequences of their regulatory interventions. While increasing the costs of market making was meant to enhance bank dealers' resilience in times of market stress, we believe that one (perhaps) "unintended" benefit of these bank regulations was to push bank dealers to enhance their matchmaking service. This can improve overall customer welfare during normal times and therefore materially changes the supposed tradeoff between resilience in times of stress and day-to-day liquidity. The market microstructure angle in this case is crucial to delivering this result. It is also important to emphasize that the improvement in customer welfare in our model likely understates the extent of the true effect. Our model intentionally shuts off one of the routes through which customers are made better off when they switch to matchmaking. Specifically, the rate at which customers are matched in a more general search model could rise as more customers switch to trading through the matchmaking mechanism, reducing waiting costs and further enhancing overall customer welfare.

Raising regulatory costs beyond a certain level would inevitably worsen welfare. A natural question is whether we can help regulators identify the optimal level of such costs. We show that overall customer welfare increases as the cost differential between bank and non-bank dealers shrinks. Bank regulators can therefore consider the market-making costs of non-bank dealers to help them judge the level they should impose on bank dealers to ensure sufficient competition.

Five regulatory agencies—the Fed, the FDIC, the Office of the Comptroller of the Currency, the SEC, and the CFTC—have recently adopted the revised Volcker Rule.[31] The revision was intended, among other things, to "provide banking entities with clarity about what activities are prohibited." While our model is not about the Volcker Rule per se, we stress that simply lowering the regulatory costs of bank dealers need not help liquidity in the corporate bond market nor would it necessarily improve customer welfare. Our analysis suggests that, in addition to using the market-making costs of non-bank dealers as a benchmark, regulators could seek to incentivize greater utilization of matchmaking to help ensure that customers continue to benefit even as bank regulatory costs decline.

[31] See https://www.sec.gov/rules/final/2019/bhca-7.pdf.

All theoretical models employ simplifications in the process of creating the appropriate structure and deriving the results. Our model is no different. Aiming for robustness and clarity necessitates abstracting from some features of the economic environment. One example is that we do not model the matchmaking business of non-bank dealers. Several other such features—intermediation chains, dealer inventory, the nature of dealer funding, and the moral hazard problem that arises when a customer employs an agent for searching—have recently been discussed in other papers (An, Song, and Zhang (2017), An and Zheng (2017), Li and Li (2017), and Cimon and Garriott (2019)). Partially for that reason we forgo incorporating them in our model (e.g., we abstract from the inventory consideration by using the same arrival rates for buyers and sellers). Instead, we focus on dimensions we deem critical to understanding how customer welfare in the corporate bond market changes when regulation increases the costs of market making by bank dealers: the dual trading mechanisms, dealer market power, and the nature of competition between bank and non-bank dealers. The resulting model provides insights that we hope capture the more salient trade-offs.

Crucial to our analysis are customers who optimize over several key choices—from the basic decision of whether to trade to the tradeoff between search time and the cost of execution in the two trading mechanisms—but we recognize that we could not capture all aspects of customer behavior. For example, increased market-making costs could incentivize customers to shift their trading to newly issued securities to minimize transaction costs, and such coping strategies may limit portfolio flexibility. Similarly, customers may choose to reduce their average trade size in response to the higher cost of market making. We believe that these aspects of customer behavior (choosing newer issues or a smaller trade size) play a lesser role in determining customer welfare than the pronounced shift from market making to matchmaking, but nonetheless they suggest a complexity to the manner in which customers respond to changes in their economic environment.

The evolving regulatory frameworks and the breathtaking pace at which technology impacts securities markets continue to dominate the agendas of regulators, practitioners, and academics. We hope that our work would serve to both highlight important tradeoffs and spur additional work on the changing nature of our securities markets.

References

Adrian, T., M. Fleming, O. Shachar, and E. Vogt. 2017. Market Liquidity After the Financial Crisis. *Annual Review of Financial Economics* 9:43–83.

Allahrakha, M., J. Centina, B. Munyan, and S. Watugala. 2019. The Effects of the Volcker Rule on Corporate Bond Trading: Evidence from the Underwriting Exemption. Working Paper.

An, Y., Y. Song, and X. Zhang. 2017. The Intermediary Rat Race. Working Paper.

An, Y., and Z. Zheng. 2017. Conflicted Immediacy Provision. Working Paper.

Anand, A., C. Jotikasthira, and K. Venkataraman. 2020. Do Buy-Side Institutions Supply Liquidity in Bond Markets? Evidence from Mutual Funds. Forthcoming, *Review of Financial Studies*.

Anderson, M., and R. Stulz. 2017. Is Post-Crisis Bond Liquidity Lower? Working Paper.

Bagnoli, M., and T. Bergstrom. 2005. Log-concave probability and its applications. *Economic Theory* 26:445–469.

Bao, J., M. O'Hara, and X. (Alex) Zhou. 2018. The Volcker Rule and corporate bond market making in times of stress. *Journal of Financial Economics* 130:95–113.

Bessembinder, H., S. Jacobsen, W. Maxwell, and K. Venkataraman. 2018. Capital Commitment and Illiquidity in Corporate Bonds. *Journal of Finance* 73:1615–61.

BIS Committee on the Global Financial System. 2014. Market-making and proprietary trading: industry trends, drivers and policy implications. Report.

———. 2016. Fixed income market liquidity. Report.

BIS Markets Committee. 2016. Electronic trading in fixed income markets. Report.

Bulow, J., and P. Klemperer. 2002. Prices and the Winner's Curse. *The RAND Journal of Economics* 33:1–21.

Chang, B., and S. Zhang. 2020. OTC Market Structure under Reforms. Working Paper.

Chen, Y., and M. Riordan. 2008. Price-Increasing Competition. *The RAND Journal of Economics* 39:1042–1058.

Chernenko, S., and A. Sunderam. 2018. Corporate Bond Liquidity: A Revealed Preference Approach. Working Paper, Harvard Business School.

Choi, J., and Y. Huh. 2017. Customer Liquidity Provision: Implications for Corporate Bond Transaction Costs. Working Paper.

Cimon, D. A., and C. Garriott. 2019. Banking Regulation and Market Making. Working Paper.

Davis Polk. 2014. Supplementary Leverage Ratio. Memo.

Dick-Nielsen, J., and M. Rossi. 2018. The Cost of Immediacy for Corporate Bonds. *The Review of Financial Studies* 32:1–41.

Duffie, D. 2012. Market Making Under the Proposed Volcker Rule. Working Paper.

Duffie, D. 2018. Post-Crisis Bank Regulations and Financial Market Liquidity. Baffi Lecture.

Duffie, D., N. Gârleanu, and L. H. Pedersen. 2005. Over-the-Counter Markets. *Econometrica* 73:1815–47.

Ederington, L. H., W. Guan, and P. K. Yadav. 2014. Dealer Spreads in the Corporate Bond Market: Agent vs. Market-Making Roles. Working Paper.

Goldstein, M. A., and E. S. Hotchkiss. 2020. Providing Liquidity in an Illiquid Market: Dealer Behavior in U.S. Corporate Bonds. *Journal of Financial Economics* 135:16–40.

Greenwood, R., S. G. Hanson, J. C. Stein, and A. Sunderam. 2017. Strengthening and Streamlining Bank Capital Regulation. *Brookings Papers on Economic Activity*.

Fudenberg, D., and J. Tirole. 1991. *Game Theory*. MIT Press.

Johnson, J.P., and D.P. Myatt. 2013. Multiproduct Quality Competition: Fighting Brands and Product Line Pruning. *The American Economic Review* 93:748–774.

Li, J., and W. Li. 2017. Agency Trading and Principal Trading. Working Paper.

Katz. M. 1984. Firm-Specific Differentiation and Competition Among Multiproduct Firms. *Journal of Business* 57:149–166.

Mussa, M., and S. Rosen. 1978. Monopoly and Product Quality. *Journal of Economic Theory* 18:301–317.

Mizrach, B. 2015. Analysis of Corporate Bond Liquidity. Working Paper.

Nocke, V., and N. Schutz. 2018. Multiproduct-Firm Oligopoly: An Aggregative Games Approach. *Econometrica* 86:523–557.

O'Hara, M. 1995. Market Microstructure Theory. *John Wiley & Sons*.

O'Hara, M., and X. (Alex) Zhou. 2020. The Electronic Evolution of Corporate Bond Dealers. Forthcoming, *Journal of Financial Economics*.

Perloff, J.M., and S.C. Salop. 1985. Equilibrium with product differentiation. *Review of Economic Studies* 52:107-120.

Randall, O. 2015. How Do Inventory Costs Affect Dealer Behavior in the US Corporate Bond Market? Working Paper.

Schultz, P. 2001. Corporate bond trading costs: A peek behind the curtain. *Journal of Finance* 56:677–98.

———. 2017. Inventory Management by Corporate Bond Dealers. Working Paper.

Trebbi, F., and K. Xiao. 2019. Regulation and Market Liquidity. *Management Science* 65:1949–1968.

Weyl, G., and M. Fabinger. 2013. Pass-Through as an Economic Tool: Principles of Incidence under Imperfect Competition. *Journal of Political Economy* 121:528–583.

Appendix

A List of Variables

Table 1: List of Model Variables

Variables	Explanations
v	Expected fundamental value of the asset
r	Discount rate
$G(\cdot)$	Cumulative distribution function of customer's private value of trading immediately
μ	Arrival rate of new buyers (and new sellers)
$\zeta(\cdot)$	$\zeta(x) = \frac{1-G(x)}{G'(x)}$
S_B	Bank dealer's market-making spread
S_{NB}	Non-bank dealer's market-making spread
c_B	Bank dealer's balance sheet cost
c_{NB}	Non-bank dealer's balance sheet cost
I	Matchmaking cost
H	Search intensity in the matchmaking service
$\mathcal{H}$	Speed of matchmaking
f	Matchmaking fee
b	Private value of the marginal customer who is indifferent between matchmaking and market-making
π_c	Overall customer welfare
π_B	Bank dealer's profit
π_{NB}	Non-bank dealer's profit
$\phi(\cdot)$	$\phi(x) = x - \zeta(x)$

B Proof of Propositions

B.1 Proof of Proposition 1

The bank dealer's problem is

$$\pi_B\left(c_B\right) = \max_{0 \le f \le S \le c_{NB}} \Pi_B\left(c_B\right)$$

where

$$\Pi_B\left(c_B\right) = \frac{2\mu}{r}\left[\left(\mathcal{H}f - I\right)\left(G\left(\frac{S-\mathcal{H}f}{1-\mathcal{H}}\right) - G\left(f\right)\right) + \left(S - c_B\right)\left(1 - G\left(\frac{S-\mathcal{H}f}{1-\mathcal{H}}\right)\right)\right].$$

Consider the following change of variables,

$$b = \frac{S-\mathcal{H}f}{1-\mathcal{H}};$$

then

$$\Pi_B\left(c_B\right) = \frac{2\mu}{r}\left[\left(\mathcal{H}f - I\right)\left(G\left(b\right) - G\left(f\right)\right) + \left(\left(1-\mathcal{H}\right)b + \mathcal{H}f - c_B\right)\left(1 - G\left(b\right)\right)\right], \tag{20}$$

and the domain $\{(f,S)\,|0 \le f \le S \le c_{NB}\}$ becomes $\{(f,b)\,|0 \le f \le b; (1-\mathcal{H})\,b + \mathcal{H}f \le c_{NB}\}$, which is a triangle in the (f,b) space.

The first-order derivatives of (20) are

$$\frac{\partial \Pi_B}{\partial f} = \frac{2\mu}{r}\mathcal{H}G'\left(f\right)\left[\frac{1-G\left(f\right)}{G'\left(f\right)} - f + \frac{I}{\mathcal{H}}\right],$$

$$\frac{\partial \Pi_B}{\partial b} = \frac{2\mu}{r}\left(1-\mathcal{H}\right)G'\left(b\right)\left(\frac{1-G\left(b\right)}{G'\left(b\right)} - b + \frac{c_B - I}{1-\mathcal{H}}\right).$$

Since $\phi\left(x\right) = x - \frac{1-G(x)}{G'(x)}$ is an increasing function, $\Pi_B\left(f,b\right)$ is a unimodal function of f with maximum satisfying $\phi\left(f_{max}\right) = \frac{I}{\mathcal{H}}$, and $\Pi_B\left(f,b\right)$ is a unimodal function of b with maximum satisfying $\phi\left(b_{max}\right) = \frac{c_B - I}{1-\mathcal{H}}$. One immediate observation is that

$$f_{max} < b_{max} \Longleftrightarrow \frac{I}{\mathcal{H}} < \frac{c_B - I}{1-\mathcal{H}} \Longleftrightarrow c_B > \frac{I}{\mathcal{H}}.$$

When $\phi\left(c_{NB}\right) \le 0$, we must have $\phi\left(c_{NB}\right) < \frac{I}{\mathcal{H}}$, which implies $f_{max} > c_{NB}$. Then, the solution $(f^\star, b^\star)$ either satisfies $f^\star = b^\star$, so there is no matchmaking, or satisfies the constraint $\left(1-\mathcal{H}\right)b^\star + \mathcal{H}f^\star = c_{NB}$. When $c_B > \frac{I}{\mathcal{H}}$, suppose the bank dealer chooses $f = b = S$. It is straightforward to obtain

$$\frac{\partial \Pi_B}{\partial(-f)}|_{f=b=S} = \frac{2\mu}{r}\frac{\mathcal{H}}{1-\mathcal{H}}G'\left(f\right)\left(c_B - \frac{I}{\mathcal{H}}\right) > 0.$$

In this case, the bank dealer always has an incentive to lower the fee and therefore matchmaking must exist. The equilibrium must satisfy $\left(1-\mathcal{H}\right)b^\star + \mathcal{H}f^\star = c_{NB}$, which is equivalent to $S^\star = c_{NB}$.

The equilibrium fee $f^\star$ is the minimal solution (by our equilibrium refinement) of

$$f^\star = \arg\max_f \frac{2\mu}{r}\left[(\mathcal{H}f - I)\left(G\left(\frac{c_{NB}-\mathcal{H}f}{1-\mathcal{H}}\right) - G(f)\right) + (c_{NB}-c_B)\left(1-G\left(\frac{c_{NB}-\mathcal{H}f}{1-\mathcal{H}}\right)\right)\right].$$

This is the constrained bank dealer equilibrium.

When $\phi(c_{NB}) > 0$, if $\frac{I}{\mathcal{H}} < \phi(c_{NB})$, we have $f_{max} < c_{NB}$. When $c_B > \frac{I}{\mathcal{H}}$, the optimal solution $(f^\star, b^\star)$ is either (f_{max}, b_{max}) or it satisfies the constraint $(1-\mathcal{H})\,b^\star + \mathcal{H}f^\star = c_{NB}$. Define c_1 as the solution of

$$\frac{1-G\left(\frac{c_{NB}-\mathcal{H}f_{max}}{1-\mathcal{H}}\right)}{G'\left(\frac{c_{NB}-\mathcal{H}f_{max}}{1-\mathcal{H}}\right)} - \frac{c_{NB}-\mathcal{H}f_{max}}{1-\mathcal{H}} + \frac{c_1 - I}{1-\mathcal{H}} = 0.$$

It is easy to show that $c_1 \in \left(\frac{I}{\mathcal{H}}, c_{NB}\right)$. When $c_B < c_1$, the constraint $(1-\mathcal{H})\,b + \mathcal{H}f \le c_{NB}$ is not binding, which means that $(f^\star = f_{max}, b^\star = b_{max})$ must be the equilibrium. This is the unconstrained bank dealer equilibrium. When $c_B \ge c_1$, the constraint $(1-\mathcal{H})\,b + \mathcal{H}f \le c_{NB}$ is binding, which implies $S^\star = c_{NB}$ and $f^\star$ is the minimal solution (by our equilibrium refinement) of

$$f^\star = \arg\max_f \frac{2\mu}{r}\left[(\mathcal{H}f - I)\left(G\left(\frac{c_{NB}-\mathcal{H}f}{1-\mathcal{H}}\right) - G(f)\right) + (c_{NB}-c_B)\left(1-G\left(\frac{c_{NB}-\mathcal{H}f}{1-\mathcal{H}}\right)\right)\right].$$

This is also the constrained bank dealer equilibrium.

If $\frac{I}{\mathcal{H}} \ge \phi(c_{NB})$, we have $f_{max} \ge c_{NB}$. Then, the solution $(f^\star, b^\star)$ either satisfies $f^\star = b^\star$, so there is no matchmaking or satisfies the constraint $(1-\mathcal{H})\,b^\star + \mathcal{H}f^\star = c_{NB}$. When $c_B > \frac{I}{\mathcal{H}}$, suppose the bank dealer chooses $f = b = S$. In this case, we have

$$\frac{\partial \Pi_B}{\partial(-f)}|_{f=b=S} = \frac{2\mu}{r}\frac{\mathcal{H}}{1-\mathcal{H}}G'(f)\left(c_B - \frac{I}{\mathcal{H}}\right) > 0,$$

so the bank dealer has an incentive to lower the fee and therefore matchmaking must exist and the equilibrium must be constrained, $S^\star = c_{NB}$. The equilibrium fee $f^\star$ is the minimal solution (by our equilibrium refinement) of

$$f^\star = \arg\max_f \frac{2\mu}{r}\left[(\mathcal{H}f - I)\left(G\left(\frac{c_{NB}-\mathcal{H}f}{1-\mathcal{H}}\right) - G(f)\right) + (c_{NB}-c_B)\left(1-G\left(\frac{c_{NB}-\mathcal{H}f}{1-\mathcal{H}}\right)\right)\right].$$

This is the constrained bank dealer equilibrium.

From this proposition, we know that when $c_B \in \left(\frac{I}{\mathcal{H}}, c_{NB}\right]$, the bank dealer always operates the matchmaking business. This implies that, if one starts with an equilibrium in which the parameters satisfy $c_B \in \left(\frac{I}{\mathcal{H}}, c_{NB}\right)$, then increasing the balance sheet cost locally or decreasing the matchmaking cost locally will still make the condition $c_B \in \left(\frac{I}{\mathcal{H}}, c_{NB}\right]$ hold, which means that matchmaking will exist in the new equilibrium as well. If one starts with an equilibrium in which the parameters satisfy $c_B = c_{NB}$ and $\frac{I}{\mathcal{H}} < c_{NB}$, then decreasing the matchmaking cost locally will still make the condition $c_B \in \left(\frac{I}{\mathcal{H}}, c_{NB}\right]$ hold, and in the proof of Proposition 4 we will show that increasing the balance sheet cost locally in this case will also lead to equilibrium with matchmaking.

B.2 Proof of Proposition 2

In this equilibrium, obviously $S^\star = c_{NB}$ is unchanged. $f^\star$ is the minimal solution of the following problem

$$f^\star = \arg\max_f \frac{2\mu}{r}\left[(\mathcal{H}f - I)\left(G\left(\frac{c_{NB} - \mathcal{H}f}{1-\mathcal{H}}\right) - G(f)\right) + (c_{NB} - c_B)\left(1 - G\left(\frac{c_{NB} - \mathcal{H}f}{1-\mathcal{H}}\right)\right)\right].$$

Since

$$\frac{\partial^2\left[(\mathcal{H}f - I)\left(G\left(\frac{c_{NB} - \mathcal{H}f}{1-\mathcal{H}}\right) - G(f)\right) + (c_{NB} - c_B)\left(1 - G\left(\frac{c_{NB} - \mathcal{H}f}{1-\mathcal{H}}\right)\right)\right]}{\partial c_B \partial f} < 0,$$

by Topkis's theorem, when c_B increases, $f^\star$ must decrease. Therefore, $b^\star = \frac{c_{NB} - \mathcal{H}f^\star}{1-\mathcal{H}}$ must increase.

Average transaction costs are

$$\begin{aligned} AC &= \frac{(G(b^\star) - G(f^\star)) f^\star + (1 - G(b^\star)) S^\star}{1 - G(f^\star)} \\ &= S^\star - \frac{G(b^\star) - G(f^\star)}{1 - G(f^\star)} (S^\star - f^\star). \end{aligned} \tag{21}$$

When c_B increases, we know that $f^\star$ decreases and $b^\star$ increases, thus

$$\frac{G(b^\star) - G(f^\star)}{1 - G(f^\star)} = \frac{G(b^\star) - G(f^\star)}{1 - G(b^\star) + G(b^\star) - G(f^\star)} = \frac{1}{1 + \frac{1 - G(b^\star)}{G(b^\star) - G(f^\star)}}$$

increases. We also know that $(S^\star - f^\star)$ increases, which means that AC decreases.

The trading volume $(1 - G(f^\star))$ must increase because $f^\star$ decreases. Matchmaking $(G(b^\star) - G(f^\star))$ must increase because $b^\star$ increases, which also implies that market making $(1 - G(b^\star))$ decreases.

Overall customer welfare is

$$\pi_c = \int_{f^\star}^{b^\star} \mathcal{H}(x - f^\star)\, dG(x) + \int_{b^\star}^{\infty} (x - S^\star)\, dG(x),$$

so

$$\frac{d\pi_c}{dc_B} = \int_{f^\star}^{b^\star} \mathcal{H}\frac{-df^\star}{dc_B} dG(x) > 0.$$

B.3 Proof of Proposition 3

The equilibrium fee $f^\star$ can be found from $\phi(f^\star) = \frac{I}{\mathcal{H}}$, so it is independent of c_B. $b^\star$ is the solution to $\phi(b^\star) = \frac{c_B - I}{1-\mathcal{H}}$, so $b^\star$ is increasing in c_B. Thus, $S^\star = (1-\mathcal{H}) b^\star + \mathcal{H}f^\star$ is increasing in c_B. Average transaction costs are

$$\begin{aligned} AC &= \frac{(G(b^\star) - G(f^\star)) f^\star + (1 - G(b^\star)) S^\star}{1 - G(f^\star)} \\ &= \frac{(G(b^\star) - G(f^\star)) f^\star + (1 - G(b^\star)) (\mathcal{H}f^\star + (1-\mathcal{H}) b^\star)}{1 - G(f^\star)}. \end{aligned}$$

Then,

$$
\begin{aligned}
\frac{dAC}{dc_B} &= \frac{db^\star}{dc_B} \cdot \frac{G'\left(b^\star\right) f^\star + \left(-G'\left(b^\star\right)\right)\left(\mathcal{H} f^\star + \left(1-\mathcal{H}\right) b^\star\right) + \left(1-G\left(b^\star\right)\right)\left(1-\mathcal{H}\right)}{1-G\left(f^\star\right)} \\
&= \frac{db^\star}{dc_B} \cdot \frac{G'\left(b^\star\right)\left(1-\mathcal{H}\right)}{1-G\left(f^\star\right)} \left(f^\star - b^\star + \frac{1-G\left(b^\star\right)}{G'\left(b^\star\right)}\right) \\
&= \frac{db^\star}{dc_B} \cdot \frac{G'\left(b^\star\right)\left(1-\mathcal{H}\right)}{1-G\left(f^\star\right)} \left(f^\star - \frac{c_B - I}{1-\mathcal{H}}\right).
\end{aligned}
$$

When $c_B < \left(1-\mathcal{H}\right) f^\star + I$, average transaction costs are increasing in c_B; when $c_B \geq \left(1-\mathcal{H}\right) f^\star + I$, average transaction costs are decreasing in c_B.

Since $f^\star$ is unchanged and $b^\star$ increases, trading volume $\left(1-G\left(f^\star\right)\right)$ must be unchanged. Matchmaking $\left(G\left(b^\star\right) - G\left(f^\star\right)\right)$ increases, and market making $\left(1-G\left(b^\star\right)\right)$ decreases.

Overall customer welfare is

$$
\pi_c = \frac{2\mu}{r}\left[\int_{f^\star}^{b^\star} \mathcal{H}\left(x-f^\star\right) dG\left(x\right) + \int_{b^\star}^{\infty} \left(x-S^\star\right) dG\left(x\right)\right],
$$

so

$$
\frac{d\pi_c}{dc_B} = \frac{2\mu}{r}\int_{b^\star}^{\infty} -\frac{dS^\star}{dc_B} dG\left(x\right) < 0.
$$

B.4 Proof of Proposition 4

The non-bank dealer's profit function is

$$
\Pi_{NB}\left(S\right) = \frac{2\mu}{r}\left[\left(S-c_{NB}\right)\left(1-G\left(b\right)\right)\right].
$$

To establish the existence of an equilibrium and find the equilibrium, we first conjecture that the equilibrium exists, and then verify it.

Suppose the equilibrium is $\left(f^\star, S^\star\right)$. The first observation is that, for bank dealer, $f < \frac{I}{\mathcal{H}}$ is strictly dominated by $f = \frac{I}{\mathcal{H}}$. Then, given any $f \geq \frac{I}{\mathcal{H}}$, let us consider the properties of the non-bank dealer's best response function without considering the constraint $S \leq c_B$.

The first-order derivative of the non-bank dealer's profit is

$$
\begin{aligned}
\frac{d\pi_{NB}}{dS} &= \frac{2\mu}{r}\left[1-G\left(b\right) - \left(S-c_{NB}\right) G'\left(b\right) \frac{1}{1-\mathcal{H}}\right] \\
&= \frac{2\pi}{r} G'\left(b\right)\left[\frac{1-G\left(b\right)}{G'\left(b\right)} - \frac{S-c_{NB}}{1-\mathcal{H}}\right].
\end{aligned}
$$

We first verify that matchmaking services must be offered in equilibrium. To see this, if there is no matchmaking in equilibrium, then the non-bank dealer's best response is $S_{BR} = \tilde{c}_B$. Since in this case $\frac{I}{\mathcal{H}} < \tilde{c}_B$, the bank dealer can always choose $f = \tilde{c}_B - \epsilon$ to obtain positive revenues. Thus, in equilibrium (if it exists), matchmaking services must be offered.

Given any $f \geq \frac{I}{\mathcal{H}}$, we can show

$$\frac{1-G(b)}{G'(b)} - \frac{S-c_{NB}}{1-\mathcal{H}}|_{S=c_{NB}} > 0$$

$$\frac{1-G(b)}{G'(b)} - \frac{S-c_{NB}}{1-\mathcal{H}}|_{S=\infty} < 0.$$

It is easy to see that $\frac{1-G(b)}{G'(b)} - \frac{S-c_{NB}}{1-\mathcal{H}}$ is strictly decreasing in S. Define $M(S,f) = \frac{1-G\left(\frac{S-\mathcal{H}f}{1-\mathcal{H}}\right)}{G'\left(\frac{S-\mathcal{H}f}{1-\mathcal{H}}\right)} - \frac{S-c_{NB}}{1-\mathcal{H}}$. Then, the best response $S_{BR}(f)$ is unique and is solved by

$$M(S_{BR}, f) = 0.$$

We want to find a uniform upper bound for the non-bank dealer's equilibrium strategy to help us use the Kakutani Fixed-Point Theorem later. The best response is solved by

$$M(S_{BR}, f) = 0 \Longleftrightarrow \frac{S_{BR}(f) - c_{NB}}{1-\mathcal{H}} = \frac{1-G\left(\frac{S_{BR}(f)-\mathcal{H}f}{1-\mathcal{H}}\right)}{G'\left(\frac{S_{BR}(f)-\mathcal{H}f}{1-\mathcal{H}}\right)} \leq \frac{1-G(0)}{G'(0)}$$
$$\Longleftrightarrow S_{BR}(f) \leq c_{NB} + (1-\mathcal{H})\frac{1-G(0)}{G'(0)}.$$

So WLOG, we just focus on the non-bank dealer's strategy space $S \in \left[c_{NB}, c_{NB} + (1-\mathcal{H})\frac{1-G(0)}{G'(0)}\right]$.

Since M is continuously differentiable in both S and f, we know that the unique best response function $S_{BR}(f)$ is also continuous. Besides,

$$\frac{\partial M}{\partial S} dS_{BR} + \frac{\partial M}{\partial f} df = 0 \Longrightarrow \frac{dS_{BR}}{df} = -\frac{\frac{\partial M}{\partial f}}{\frac{\partial M}{\partial S}}.$$

It is obvious that

$$\frac{\partial M}{\partial f} > 0, \frac{\partial M}{\partial S} < 0,$$

so we must have

$$\frac{dS_{BR}}{df} > 0.$$

Hence, $S_{BR}(f)$ is strictly increasing in f.

Consider the bank dealer's best response function. Given the non-bank dealer's choice

$$S \in \left[c_{NB}, c_{NB} + (1-\mathcal{H})\frac{1-G(0)}{G'(0)}\right],$$

the first-order derivative of the bank dealer's profit function is

$$\begin{aligned}\frac{r}{2\mu}\frac{\partial\pi_B}{\partial f} &= \mathcal{H}\left(G\left(b\right)-G\left(f\right)\right)+\left(\mathcal{H}f-I\right)\left(-\frac{\mathcal{H}}{1-\mathcal{H}}G'\left(b\right)-G'\left(f\right)\right)\\ &= \mathcal{H}G'\left(f\right)\left[\frac{1-G\left(f\right)}{G'\left(f\right)}-f+\frac{I}{\mathcal{H}}\right]-\mathcal{H}G'\left(b\right)\left(\frac{\mathcal{H}f-I}{1-\mathcal{H}}+\frac{1-G\left(b\right)}{G'\left(b\right)}\right)\\ &= \mathcal{H}G'\left(f\right)\left[\frac{1-G\left(f\right)}{G'\left(f\right)}-\frac{\mathcal{H}f-I}{\mathcal{H}}\right]-\mathcal{H}G'\left(b\right)\left(\frac{1-G\left(b\right)}{G'\left(b\right)}+\frac{\mathcal{H}f-I}{1-\mathcal{H}}\right).\end{aligned}$$

It is clear that $\frac{\partial\pi_B}{\partial f}|_{f=\frac{I}{\mathcal{H}}}>0$ and $\frac{\partial\pi_B}{\partial f}|_{f=S}<0$, so the best response $f_{BR}\left(S\right)$ must be interior. Thus, the FOC must be satisfied and we have

$$\frac{1-G\left(f_{BR}\right)}{G'\left(f_{BR}\right)}-f_{BR}+\frac{I}{\mathcal{H}}>0\Longleftrightarrow\phi\left(f_{BR}\right)<\frac{I}{\mathcal{H}}.$$

Remember that the best response must satisfy $f_{BR}\in\left[\frac{I}{\mathcal{H}},c_B\right]$, so we have $f_{BR}\in\left[\frac{I}{\mathcal{H}},\min\{\phi^{-1}\left(\frac{I}{\mathcal{H}}\right),c_B\}\right]$. In region $f_{BR}\in\left[\frac{I}{\mathcal{H}},\min\{\phi^{-1}\left(\frac{I}{\mathcal{H}}\right),c_B\}\right]$:

1. if G is concave, then $G'\left(f\right)$ is decreasing in f. Moreover, $\frac{1-G(f)}{G'(f)}-f+\frac{I}{\mathcal{H}}$ is positive and decreasing in f, $G'\left(b\right)$ is increasing in f, $\frac{\mathcal{H}f-I}{1-\mathcal{H}}+\frac{1-G(b)}{G'(b)}$ is increasing in f, and therefore $\frac{\partial\pi_B}{\partial f}$ is decreasing in $f_{BR}\in\left[\frac{I}{\mathcal{H}},\min\{\phi^{-1}\left(\frac{I}{\mathcal{H}}\right),c_B\}\right]$. This implies that the best response must be unique, and we can show that the best response function is continuous by the implicit function theorem.

2. if G is convex, $G'''\leq 0$ and $\mathcal{H}\leq\frac{1}{2}$, we can rewrite $\frac{\partial\pi_B}{\partial f}$ as

$$\frac{r}{2\mu}\frac{\partial\pi_B}{\partial f}=\mathcal{H}\left(G\left(b\right)-G\left(f\right)\right)-\frac{\left(\mathcal{H}f-I\right)}{1-\mathcal{H}}\left(\mathcal{H}G'\left(b\right)+\left(1-\mathcal{H}\right)G'\left(f\right)\right).$$

When $f\geq\frac{I}{\mathcal{H}}$, we know $\mathcal{H}\left(G\left(b\right)-G\left(f\right)\right)$ is strictly decreasing in f, and $\frac{(\mathcal{H}f-I)}{1-\mathcal{H}}$ is positive and strictly increasing in f. If we can show that $\left(\mathcal{H}G'\left(b\right)+\left(1-\mathcal{H}\right)G'\left(f\right)\right)$ is increasing in f, then $\frac{\partial\pi_B}{\partial f}$ would be strictly decreasing in f, implying that the the best response is unique. The following claim confirms this result.

Claim 1. If $\mathcal{H}<\frac{1}{2}$ and $G'''\left(x\right)<0$ for all x, then $\left(\mathcal{H}G'\left(\frac{S-\mathcal{H}f}{1-\mathcal{H}}\right)+\left(1-\mathcal{H}\right)G'\left(f\right)\right)$ is increasing in f.

Proof. $\mathcal{H}b+\left(1-\mathcal{H}\right)f=f+\mathcal{H}\left(b-f\right)=f+\frac{\mathcal{H}}{1-\mathcal{H}}\left(S-f\right)=\frac{\mathcal{H}}{1-\mathcal{H}}S+\frac{1-2\mathcal{H}}{1-\mathcal{H}}f$. Consider any f_1 and f_2 that satisfy $\frac{I}{\mathcal{H}}\leq f_1<f_2\leq S$. Define $\delta f=f_2-f_1>0$, $b_1=\frac{S-\mathcal{H}f_1}{1-\mathcal{H}}$, $b_2=\frac{S-\mathcal{H}f_2}{1-\mathcal{H}}$, and $\delta b=b_2-b_1=-\frac{\mathcal{H}}{1-\mathcal{H}}\delta f<0$. Then,

$$\left(\mathcal{H}G'\left(b_2\right)+\left(1-\mathcal{H}\right)G'\left(f_2\right)\right)\geq\left(\mathcal{H}G'\left(b_2-\frac{1-2\mathcal{H}}{1-\mathcal{H}}\delta f\right)+\left(1-\mathcal{H}\right)G'\left(f_2-\frac{1-2\mathcal{H}}{1-\mathcal{H}}\delta f\right)\right)$$

because $G''>0$. And

$$\left(\mathcal{H}G'\left(b_2-\frac{1-2\mathcal{H}}{1-\mathcal{H}}\delta f\right)+\left(1-\mathcal{H}\right)G'\left(f_2-\frac{1-2\mathcal{H}}{1-\mathcal{H}}\delta f\right)\right)\geq\left(\mathcal{H}G'\left(b_1\right)+\left(1-\mathcal{H}\right)G'\left(f_1\right)\right)$$

because G' is concave, and

$$\mathcal{H}\left(b_2 - \frac{1-2\mathcal{H}}{1-\mathcal{H}}\delta f\right) + (1-\mathcal{H})\left(f_2 - \frac{1-2\mathcal{H}}{1-\mathcal{H}}\delta f\right) = \mathcal{H}b_1 + (1-\mathcal{H})\, f_1.$$

□

Since f_{BR} is unique and continuous on $\left[\frac{I}{\mathcal{H}}, \min\{\phi^{-1}\left(\frac{I}{\mathcal{H}}\right), c_B\}\right]$ and S_{BR} is unique and continuous on $\left[c_{NB}, c_{NB} + (1-\mathcal{H})\frac{1-G(0)}{G'(0)}\right]$, by the Kakutani Fixed Point Theorem the equilibrium must exist, and it satisfies

$$\frac{1-G(b)}{G'(b)} - \frac{S-c_{NB}}{1-\mathcal{H}} = 0, \tag{22}$$

$$\mathcal{H}G'(f)\left[\frac{1-G(f)}{G'(f)} - \frac{\mathcal{H}f - I}{\mathcal{H}}\right] - \mathcal{H}G'(b)\left(\frac{1-G(b)}{G'(b)} + \frac{\mathcal{H}f-I}{1-\mathcal{H}}\right) = 0. \tag{23}$$

We now show that there is a unique solution satisfying (22) and (23). Suppose there exist two solutions of the above two conditions: (f_1, S_1) and (f_2, S_2). WLOG we can assume $f_1 < f_2$. Then, we must have $S_1 < S_2$ because $S_{BR}(f)$ is strictly increasing in f. Using (22) and $S_1 < S_2$, we can show that $b_1 > b_2$.

1. If G is concave, since (f_1, S_1) satisfies (23), we know that

$$\mathcal{H}G'(f_1)\left[\frac{1-G(f_1)}{G'(f_1)} - \frac{\mathcal{H}f_1 - I}{\mathcal{H}}\right] = \mathcal{H}G'(b_1)\left(\frac{1-G(b_1)}{G'(b_1)} + \frac{\mathcal{H}f_1-I}{1-\mathcal{H}}\right). \tag{24}$$

Since $f_1 < f_2$, and $\frac{1-G(f_i)}{G'(f_i)} - \frac{\mathcal{H}f_i-I}{\mathcal{H}} > 0$ is satisfied for both $i = 1, 2$, we have

$$\frac{1-G(f_1)}{G'(f_1)} - \frac{\mathcal{H}f_1 - I}{\mathcal{H}} > \frac{1-G(f_2)}{G'(f_2)} - \frac{\mathcal{H}f_2 - I}{\mathcal{H}} > 0.$$

Besides, we have $G'(f_1) > G'(f_2)$. So

$$\mathcal{H}G'(f_1)\left[\frac{1-G(f_1)}{G'(f_1)} - \frac{\mathcal{H}f_1 - I}{\mathcal{H}}\right] > \mathcal{H}G'(f_2)\left[\frac{1-G(f_2)}{G'(f_2)} - \frac{\mathcal{H}f_2 - I}{\mathcal{H}}\right]. \tag{25}$$

Similarly, since $b_1 > b_2$ and $f_1 < f_2$, we have

$$\mathcal{H}G'(b_1)\left(\frac{1-G(b_1)}{G'(b_1)} + \frac{\mathcal{H}f_1-I}{1-\mathcal{H}}\right) < \mathcal{H}G'(b_2)\left(\frac{1-G(b_2)}{G'(b_2)} + \frac{\mathcal{H}f_2-I}{1-\mathcal{H}}\right). \tag{26}$$

Equations (24), (25), and (26) imply that

$$\mathcal{H}G'(f_2)\left[\frac{1-G(f_2)}{G'(f_2)} - \frac{\mathcal{H}f_2 - I}{\mathcal{H}}\right] < \mathcal{H}G'(b_2)\left(\frac{1-G(b_2)}{G'(b_2)} + \frac{\mathcal{H}f_2-I}{1-\mathcal{H}}\right).$$

But this is impossible because, by the definition of (f_2, S_2), we know that

$$\mathcal{H}G'(f_2)\left[\frac{1-G(f_2)}{G'(f_2)} - \frac{\mathcal{H}f_2 - I}{\mathcal{H}}\right] = \mathcal{H}G'(b_2)\left(\frac{1-G(b_2)}{G'(b_2)} + \frac{\mathcal{H}f_2-I}{1-\mathcal{H}}\right).$$

Hence, there must be a unique solution satisfying (22) and (23).

2. If G is convex, $\mathcal{H} < \frac{1}{2}$ and $G''' < 0$, we know that

$$\mathcal{H}\left(G\left(b_1\right) - G\left(f_1\right)\right) > \mathcal{H}\left(G\left(b_2\right) - G\left(f_2\right)\right).$$

By Claim 1, we know that

$$\begin{aligned}\left(\mathcal{H}G'\left(\frac{S_1 - \mathcal{H}f_1}{1-\mathcal{H}}\right) + (1-\mathcal{H})\,G'\left(f_1\right)\right) &< \left(\mathcal{H}G'\left(\frac{S_1 - \mathcal{H}f_2}{1-\mathcal{H}}\right) + (1-\mathcal{H})\,G'\left(f_2\right)\right) \\ &< \left(\mathcal{H}G'\left(\frac{S_2 - \mathcal{H}f_2}{1-\mathcal{H}}\right) + (1-\mathcal{H})\,G'\left(f_2\right)\right).\end{aligned}$$

Since

$$\mathcal{H}\left(G\left(b_1\right) - G\left(f_1\right)\right) - \left(\mathcal{H}G'\left(b_1\right) + (1-\mathcal{H})\,G'\left(f_1\right)\right) = 0,$$

we must have

$$\mathcal{H}\left(G\left(b_2\right) - G\left(f_2\right)\right) - \left(\mathcal{H}G'\left(b_2\right) + (1-\mathcal{H})\,G'\left(f_2\right)\right) < 0,$$

which violates the FOC. Hence, there must be a unique solution satisfying (22) and (23).

Denote the equilibrium as $\left(\bar{f}^\star, \bar{S}^\star\right)$, and define $c_2 = \bar{S}^\star$. When $c_B > c_2$, the unique equilibrium is $\left(\bar{f}^\star, \bar{S}^\star\right)$, and this is the unconstrained equilibrium. When $c_B \leq c_2$, in equilibrium we must have $S^\star = c_B$ (because otherwise it will be the unconstrained equilibrium). The bank dealer's unique best response $f^\star$ is solved by

$$\mathcal{H}G'\left(f^\star\right)\left[\frac{1 - G\left(f^\star\right)}{G'\left(f^\star\right)} - \frac{\mathcal{H}f^\star - I}{\mathcal{H}}\right] - \mathcal{H}G'\left(\frac{c_B - \mathcal{H}f^\star}{1-\mathcal{H}}\right)\left(\frac{1 - G\left(\frac{c_B - \mathcal{H}f^\star}{1-\mathcal{H}}\right)}{G'\left(\frac{c_B - \mathcal{H}f^\star}{1-\mathcal{H}}\right)} + \frac{\mathcal{H}f^\star - I}{1-\mathcal{H}}\right) = 0.$$

Because the non-bank dealer's best response is increasing in f, $S^\star = c_B$ must be his best response, which implies that $(f^\star, c_B)$ is the unique equilibrium.

From this proposition, we know that when $c_B > c_{NB}$ and $I < \mathcal{H}\min\{\tilde{c}_B, c_B\}$, the bank dealer always operates the matchmaking business. This implies that, if one starts with an equilibrium in which the parameters satisfy $c_B > c_{NB}$ and $I < \mathcal{H}\min\{\tilde{c}_B, c_B\}$, then increasing the balance sheet cost locally or decreasing the matchmaking cost locally will still make the condition $c_B > c_{NB}$ and $I < \mathcal{H}\min\{\tilde{c}_B, c_B\}$ hold, which means that matchmaking will still exist in the new equilibrium. If one starts with an equilibrium in which the parameters satisfy $c_B = c_{NB}$ and $\frac{I}{\mathcal{H}} < c_{NB}$, from Proposition 1 we know that matchmaking exists in this equilibrium. Increasing the balance sheet cost locally in this case will make the conditions $c_B > c_{NB}$ and $I < \mathcal{H}\min\{\tilde{c}_B, c_B\}$ hold, and thus the matchmaking service will still exist in the new equilibrium.

B.5 Proof of Proposition 5

When establishing the existence of an equilibrium in this case, we already imposed the following additional assumptions on G:

- $G(x)$ is a concave function, or
- $G(x)$ is a convex function, and $\mathcal{H} < \frac{1}{2}$, $G'''(x) < 0$ for all x.

In the constrained non-bank dealer equilibrium, the bank dealer's profit is

$$\Pi_B = \frac{2\mu}{r}(\mathcal{H}f - I)\left(G\left(\frac{c_B - \mathcal{H}f}{1-\mathcal{H}}\right) - G(f)\right).$$

In equilibrium, we have

$$\frac{\partial^2 \pi_B}{\partial f \partial c_B} = \frac{2\mu}{r}\frac{\mathcal{H}}{1-\mathcal{H}}\left(G'(b^\star) - \frac{\mathcal{H}f^\star - I}{1-\mathcal{H}}G''(b^\star)\right).$$

When G is concave, it is clear that $\frac{\partial^2 \pi_B}{\partial f \partial S} > 0$. When G is convex, in equilibrium the FOC is satisfied:

$$\begin{aligned}
&\mathcal{H}(G(b^\star) - G(f^\star)) + (\mathcal{H}f^\star - I)\left(-\frac{\mathcal{H}}{1-\mathcal{H}}G'(b^\star) - G'(f^\star)\right) = 0\\
\Longleftrightarrow &G(b^\star) - G(f^\star) = \frac{\mathcal{H}f^\star - I}{\mathcal{H}(1-\mathcal{H})}\left(\mathcal{H}G'(b^\star) + (1-\mathcal{H})G'(f^\star)\right).
\end{aligned}$$

Since G is convex, we have $G(b^\star) - G(f^\star) \le (b^\star - f^\star)G'(b^\star)$, thus

$$\begin{aligned}
&\frac{\mathcal{H}f^\star - I}{\mathcal{H}(1-\mathcal{H})}\left(\mathcal{H}G'(b^\star) + (1-\mathcal{H})G'(f^\star)\right) \le (b^\star - f^\star)G'(b^\star)\\
\Longleftrightarrow &\left(f^\star - \frac{I}{\mathcal{H}}\right)\left(\mathcal{H}G'(b^\star) + (1-\mathcal{H})G'(f^\star)\right) \le (c_B - f^\star)G'(b^\star)\\
\Longleftrightarrow &\left(f^\star - \frac{I}{\mathcal{H}}\right)(1-\mathcal{H})G'(f^\star) \le (c_B - f^\star - (\mathcal{H}f^\star - I))G'(b^\star).
\end{aligned}$$

This implies that $c_B - f^\star - (\mathcal{H}f^\star - I) > 0$. Moreover, when G' is concave, we have $G''(b^\star) \le \frac{G'(b^\star) - G'(f^\star)}{b^\star - f^\star} < \frac{G'(b^\star)}{b^\star - f^\star}$. Then,

$$\begin{aligned}
G'(b^\star) - \frac{\mathcal{H}f^\star - I}{1-\mathcal{H}}G''(b^\star) &> G'(b^\star) - \frac{\mathcal{H}f^\star - I}{1-\mathcal{H}}\frac{G'(b^\star)}{b^\star - f^\star}\\
&= G'(b^\star)\frac{c_B - f^\star - (\mathcal{H}f^\star - I)}{c_B - f^\star}\\
&> 0.
\end{aligned}$$

Hence, in both cases we have $\frac{\partial^2 \pi_B}{\partial f \partial S}|_{f^\star} > 0$, and by Topkis's theorem $f^\star$ must be increasing in c_B. To consider the change in $b^\star$,

$$\begin{aligned}
\frac{\partial \Pi_B}{\partial f} &= \frac{2\mu}{r}\left[\mathcal{H}(G(b^\star) - G(f^\star)) + (\mathcal{H}f^\star - I)\left(-\frac{\mathcal{H}}{1-\mathcal{H}}G'(b^\star) - G'(f^\star)\right)\right]\\
&= \frac{2\mu}{r}\left[\mathcal{H}G'(f)\left(\frac{1-G(f)}{G'(f)} - f + \frac{I}{\mathcal{H}}\right) - \mathcal{H}(1-G(b)) - (\mathcal{H}f - I)\frac{\mathcal{H}}{1-\mathcal{H}}G'(b)\right].
\end{aligned}$$

The optimal solution $f^\star$ must satisfy $f^\star \in \left(\frac{I}{\mathcal{H}}, \phi^{-1}\left(\frac{I}{\mathcal{H}}\right)\right)$. In the proof of the equilibrium existence, we already showed that $\Pi_B(f)$ is unimodal in $f^\star \in \left(\frac{I}{\mathcal{H}}, \phi^{-1}\left(\frac{I}{\mathcal{H}}\right)\right)$ with the additional assumptions on G. This implies that if $\frac{\partial \Pi_B}{\partial f}(x; c_B) < 0$, then $\frac{\partial \Pi_B}{\partial f}(f; c_B) < 0$ for all $f \in \left(x, \phi^{-1}\left(\frac{I}{\mathcal{H}}\right)\right)$. Suppose that, when $c_B = c_{B1}$, the equilibrium is $f^\star = f_1$. When c_B increases to c_{B2}, denote the equilibrium as f_2. Suppose f increases to $\tilde{f}_2$ in this case such that

$$b_1 = \frac{c_{B1} - \mathcal{H} f_1}{1 - \mathcal{H}} = \frac{c_{B2} - \mathcal{H} \tilde{f}_2}{1 - \mathcal{H}}.$$

Then,

$$\frac{\partial \Pi_B}{\partial f}\left(\tilde{f}_2; c_{B2}\right) = \frac{2\mu}{r}\left[\mathcal{H} G'\left(\tilde{f}_2\right)\left(\frac{1 - G\left(\tilde{f}_2\right)}{G'\left(\tilde{f}_2\right)} - \tilde{f}_2 + \frac{I}{\mathcal{H}}\right) - \mathcal{H}\left(1 - G\left(b_1\right)\right) - \left(\mathcal{H}\tilde{f}_2 - I\right)\frac{\mathcal{H}}{1 - \mathcal{H}} G'\left(b_1\right)\right].$$

The optimality of f_1 implies that

$$\frac{\partial \Pi_B}{\partial f}\left(f_1; c_{B1}\right) = \frac{2\mu}{r}\left[\mathcal{H} G'\left(f_1\right)\left(\frac{1 - G\left(f_1\right)}{G'\left(f_1\right)} - f_1 + \frac{I}{\mathcal{H}}\right) - \mathcal{H}\left(1 - G\left(b_1\right)\right) - \left(\mathcal{H} f_1 - I\right)\frac{\mathcal{H}}{1 - \mathcal{H}} G'\left(b_1\right)\right] = 0.$$

If G is concave, it is straightforward to show that $\frac{\partial \Pi_B}{\partial f}\left(\tilde{f}_2; c_{B2}\right) < \frac{\partial \Pi_B}{\partial f}\left(f_1; c_{B1}\right) = 0$, so $\frac{\partial \Pi_B}{\partial f}\left(f; c_{B2}\right) < 0$ for all $f \in \left(\tilde{f}_2, \phi^{-1}\left(\frac{I}{\mathcal{H}}\right)\right)$, and we must have $f_2 < \tilde{f}_2$, which implies that $b_2 = \frac{c_{B2} - \mathcal{H} f_2}{1 - \mathcal{H}} > b_1$.

Similarly, if G is convex,

$$\frac{\partial \Pi_B}{\partial f}\left(\tilde{f}_2; c_{B2}\right) = \frac{2\mu}{r}\left[\mathcal{H}\left(1 - G\left(\tilde{f}_2\right) - \left(\tilde{f}_2 - \frac{I}{\mathcal{H}}\right) G'\left(\tilde{f}_2\right)\right) - \mathcal{H}\left(1 - G\left(b_1\right)\right) - \left(\mathcal{H}\tilde{f}_2 - I\right)\frac{\mathcal{H}}{1 - \mathcal{H}} G'\left(b_1\right)\right].$$

We know that

$$\frac{\partial \Pi_B}{\partial f}\left(f_1; c_{B1}\right) = \frac{2\mu}{r}\left[\mathcal{H}\left(1 - G\left(f_1\right) - \left(f_1 - \frac{I}{\mathcal{H}}\right) G'\left(f_1\right)\right) - \mathcal{H}\left(1 - G\left(b_1\right)\right) - \left(\mathcal{H} f_1 - I\right)\frac{\mathcal{H}}{1 - \mathcal{H}} G'\left(b_1\right)\right] = 0.$$

Moreover, when f_1 increases (and we keep b_1 unchanged), $1 - G\left(f_1\right) - \left(f_1 - \frac{I}{\mathcal{H}}\right) G'\left(f_1\right)$ must be decreasing when $f_1 \geq \frac{I}{\mathcal{H}}$. Therefore, the RHS of the above expression is decreasing in f_1, which implies $\frac{\partial \Pi_B}{\partial f}\left(\tilde{f}_2; c_{B2}\right) < \frac{\partial \Pi_B}{\partial f}\left(f_1; c_{B1}\right) = 0$, so $\frac{\partial \Pi_B}{\partial f}\left(f; c_{B2}\right) < 0$ for all $f \in \left(\tilde{f}_2, \phi^{-1}\left(\frac{I}{\mathcal{H}}\right)\right)$. We therefore must have $f_2 < \tilde{f}_2$ and as a result also $b_2 = \frac{c_{B2} - \mathcal{H} f_2}{1 - \mathcal{H}} > b_1$. Hence, we can conclude that $b^\star$ is increasing in c_B.

From above analysis we can show that trading volume $\left(1 - G\left(f^\star\right)\right)$ is decreasing in c_B and market making $\left(1 - G\left(b^\star\right)\right)$ is decreasing in c_B.

Overall customer welfare is

$$\pi_c = \int_{f^\star}^{b^\star} \mathcal{H}\left(x - f^\star\right) dG\left(x\right) + \int_{b^\star}^{\infty}\left(x - S^\star\right) dG\left(x\right).$$

Hence,

$$\frac{d\pi_c}{dc_B} = \int_{f^\star}^{b^\star} \mathcal{H}\left(-\frac{df^\star}{dc_B}\right) dG\left(x\right) + \int_{b^\star}^{\infty}\left(-\frac{dS^\star}{dx}\right) dG\left(x\right) < 0.$$

B.6 Proof of Proposition 6

When I decreases, it's obvious that $S^\star$ is unchanged in both the constrained bank dealer equilibrium and the constrained non-bank dealer equilibrium.

In the constrained bank dealer equilibrium, $f^\star$ is the minimal solution (by our equilibrium refinement) of

$$f^\star = \arg\max_f \frac{2\mu}{r}\left[(\mathcal{H}f - I)\left(G\left(\frac{c_{NB}-\mathcal{H}f}{1-\mathcal{H}}\right) - G(f)\right) + (c_{NB}-c_B)\left(1 - G\left(\frac{c_{NB}-\mathcal{H}f}{1-\mathcal{H}}\right)\right)\right].$$

It is straightforward to show that

$$\frac{\partial^2\left[(\mathcal{H}f - I)\left(G\left(\frac{c_{NB}-\mathcal{H}f}{1-\mathcal{H}}\right) - G(f)\right) + (c_{NB}-c_B)\left(1 - G\left(\frac{c_{NB}-\mathcal{H}f}{1-\mathcal{H}}\right)\right)\right]}{\partial I \partial f} > 0,$$

and therefore

$$\frac{df^\star}{dI} > 0.$$

Hence, when I decreases, $f^\star$ decreases as well.

In the constrained non-bank dealer equilibrium, $f^\star$ is solved by

$$f^\star = \arg\max_f (\mathcal{H}f - I)\left(G\left(\frac{c_B-\mathcal{H}f}{1-\mathcal{H}}\right) - G(f)\right).$$

It is clear that the objective function has an increasing difference in (I, f). Hence, when I decreases, $f^\star$ decreases as well.

Therefore, in both cases we have that $b^\star$ increases because $f^\star$ decreases and $S^\star$ is unchanged. Thus, $G(f^\star)$ decreases, $G(b^\star) - G(f^\star)$ increases, and $1 - G(b^\star)$ decreases.

Average transaction costs are

$$\begin{aligned} AC &= \frac{(G(b^\star) - G(f^\star))\, f^\star + (1 - G(b^\star))\, S^\star}{1 - G(f^\star)} \\ &= S^\star - \frac{G(b^\star) - G(f^\star)}{1 - G(f^\star)}(S^\star - f^\star). \end{aligned} \tag{27}$$

When I decreases, we know that $f^\star$ decreases and $b^\star$ increases, and thus

$$\frac{G(b^\star) - G(f^\star)}{1 - G(f^\star)} = \frac{G(b^\star) - G(f^\star)}{1 - G(b^\star) + G(b^\star) - G(f^\star)} = \frac{1}{1 + \frac{1-G(b^\star)}{G(b^\star)-G(f^\star)}}$$

increases. We also know that $(S^\star - f^\star)$ increases. Hence, AC decreases.

Overall customer welfare is

$$\pi_c = \frac{2\mu}{r}\left[\int_{f^\star}^{b^\star} \mathcal{H}(x - f^\star)\, dG(x) + \int_{b^\star}^{\infty} (x - S^\star)\, dG(x)\right],$$

We can obtain

$$\frac{d\pi_c}{dI} \propto \int_{f^\star}^{b^\star} \mathcal{H}\left(-\frac{df^\star}{dI}\right) dG(x) < 0.$$

Hence, when I decreases, overall customer welfare increases.

B.7 Proof of Proposition 7

In the unconstrained bank dealer equilibrium, the equilibrium $(S^\star, f^\star, b^\star)$ is solved using the following conditions:

$$\frac{1-G(f^\star)}{G'(f^\star)} - f^\star + \frac{I}{\mathcal{H}} = 0,$$
$$\frac{1-G(b^\star)}{G'(b^\star)} - b^\star + \frac{c_B - I}{1-\mathcal{H}} = 0;$$
$$S^\star = \mathcal{H} f^\star + (1-\mathcal{H}) b^\star.$$

Recall that

$$\phi(x) = x - \frac{1-G(x)}{G'(x)} = x - \zeta(x)$$

is increasing in x. From the above equilibrium conditions we have

$$\frac{1}{\mathcal{H}} = \phi'(f^\star)\frac{df^\star}{dI}, \tag{28}$$
$$\frac{1}{1-\mathcal{H}} = \phi'(b^\star)\left(-\frac{db^\star}{dI}\right), \tag{29}$$
$$\frac{dS^\star}{dI} = \mathcal{H}\frac{df^\star}{dI} + (1-\mathcal{H})\frac{db^\star}{dI}.$$

Therefore, $\frac{df^\star}{dI} > 0$ and $-\frac{db^\star}{dI} > 0$. Then, when I decreases, $f^\star$ decreases and $b^\star$ increases. It is also straightforward to show that

$$\begin{aligned}\frac{dS^\star}{dI} &= \mathcal{H}\frac{df^\star}{dI} + (1-\mathcal{H})\frac{db^\star}{dI} \\ &= \frac{1}{\phi'(f^\star)} - \frac{1}{\phi'(b^\star)}.\end{aligned}$$

If $\phi'(\cdot)$ is an increasing function, i.e., ζ is concave, then $\frac{1}{\phi'(f^\star)} - \frac{1}{\phi'(b^\star)} \geq 0$, and thus $S^\star$ decreases when I decreases. Similarly, when ζ is convex, we can show that $S^\star$ increases when I decreases.

From our results regarding $f^\star$ and $b^\star$, we know that trading volume $(1-G(f^\star))$ increases, matchmaking $(G(b^\star) - G(f^\star))$ increases, and market making $(1-G(b^\star))$ decreases.

Overall customer welfare is

$$\pi_c = \frac{2\mu}{r}\left[\int_{f^\star}^{b^\star} \mathcal{H}(x-f^\star)\,dG(x) + \int_{b^\star}^{\infty}(x-S^\star)\,dG(x)\right].$$

Then,

$$\begin{aligned}
\frac{d\pi_c}{dI} &= \frac{2\mu}{r}\left[\int_{f^\star}^{b^\star}\mathcal{H}\left(-\frac{df^\star}{dI}\right)dG\left(x\right)+\int_{b^\star}^{\infty}\left(-\frac{dS^\star}{dI}\right)dG\left(x\right)\right]\\
&= \frac{2\mu}{r}\left[-\mathcal{H}\frac{df^\star}{dI}\left(G\left(b^\star\right)-G\left(f^\star\right)\right)-\left(1-G\left(b^\star\right)\right)\frac{dS^\star}{dI}\right]\\
&= \frac{2\mu}{r}\left[-\mathcal{H}\frac{df^\star}{dI}\left(G\left(b^\star\right)-G\left(f^\star\right)\right)-\left(1-G\left(b^\star\right)\right)\left(\mathcal{H}\frac{df^\star}{dI}+\left(1-\mathcal{H}\right)\frac{db^\star}{dI}\right)\right]\\
&= \frac{2\mu}{r}\left[-\mathcal{H}\frac{df^\star}{dI}\left(1-G\left(f^\star\right)\right)-\left(1-G\left(b^\star\right)\right)\left(1-\mathcal{H}\right)\frac{db^\star}{dI}\right]
\end{aligned}$$

Thus,

$$\begin{aligned}
\frac{d\pi_c}{dI}\leq 0 &\iff -\mathcal{H}\frac{df^\star}{dI}\left(1-G\left(f^\star\right)\right)-\left(1-G\left(b^\star\right)\right)\left(1-\mathcal{H}\right)\frac{db^\star}{dI}\leq 0\\
&\iff \left(1-G\left(b^\star\right)\right)\left(1-\mathcal{H}\right)\left(-\frac{db^\star}{dI}\right)\leq\mathcal{H}\frac{df^\star}{dI}\left(1-G\left(f^\star\right)\right)\\
&\iff \frac{\left(1-\mathcal{H}\right)\left(-\frac{db^\star}{dI}\right)}{\mathcal{H}\frac{df^\star}{dI}}\leq\frac{1-G\left(f^\star\right)}{1-G\left(b^\star\right)}.
\end{aligned}$$

Substituting (28) and (29) into the above equation, we obtain

$$\begin{aligned}
\frac{d\pi_c}{dI}\leq 0 &\iff \frac{\phi'\left(f^\star\right)}{\phi'\left(b^\star\right)}\leq\frac{1-G\left(f^\star\right)}{1-G\left(b^\star\right)}\\
&\iff \frac{\phi'\left(f^\star\right)}{1-G\left(f^\star\right)}\leq\frac{\phi'\left(b^\star\right)}{1-G\left(b^\star\right)}.
\end{aligned}$$

When $\zeta\left(x\right)$ is concave, $\phi\left(x\right)=x-\zeta\left(x\right)$ must be convex, and thus $\phi'\left(x\right)$ is increasing in x. As a result, $\frac{\phi'(x)}{1-G(x)}$ must be increasing in x. We know that in equilibrium, $f^\star<b^\star$, so we must have

$$\frac{\phi'\left(f^\star\right)}{1-G\left(f^\star\right)}<\frac{\phi'\left(b^\star\right)}{1-G\left(b^\star\right)}.$$

This implies $\frac{d\pi_c}{dI}<0$. Therefore, if ζ is concave, when I decreases overall customer welfare, π_c, increases.

B.8 Proof of Proposition 9

With σ as the exogenous parameter, denote $\left(f_\sigma,S_\sigma\right)$ as the equilibrium strategy. Then, the bank dealer's expected per-capita payoff is

$$\pi_B^\sigma=\pi_j\left(S_\sigma,f_\sigma\right),$$

and a customer's equilibrium expected utility without the taste shock is

$$\bar{\pi}_c^\sigma=\bar{\pi}_{cj}\left(S_\sigma,f_\sigma\right).$$

Denote

$$(f_0, S_0) = \lim_{\sigma\to 0} (f_\sigma, S_\sigma),$$

$$\pi_B^0 = \lim_{\sigma\to 0} \pi_B^\sigma,$$

$$\bar{\pi}_c^0 = \lim_{\sigma\to 0} \bar{\pi}_c^\sigma.$$

First, we can show that $S_0 = c_B$ and $f_0 = \frac{I}{\mathcal{H}}$ in two steps. In Step 1, we show that as $\sigma \to 0$, the equilibrium profit that each bank dealer can earn is zero, i.e., $\pi_B^0 = 0$. In Step 2, we show that, as $\sigma \to 0$, the bank dealers make zero profit in both the matchmaking business and the market-making business, i.e., $f_0 = \frac{I}{\mathcal{H}}$ and $S_0 = c_B$.

Step 1:

To show the first result, notice first that, in equilibrium, the market share that bank dealer j can obtain is

$$s_j = \frac{\exp\left(\frac{1}{\sigma}\bar{\pi}_{cj}(S_j, f_j)\right)}{\sum_{j'} \exp\left(\frac{1}{\sigma}\bar{\pi}_{cj'}\left(S_{j'}, f_{j'}\right)\right)} = \frac{1}{(N-1)\exp\left(\frac{1}{\sigma}\left(\bar{\pi}_c^\sigma - \bar{\pi}_{cj}(S_j, f_j)\right)\right) + 1}.$$

Consider the case when $\sigma \to 0$. If $\pi_B^0 > 0$, then we must have $S_0 > c_B$ or $f_0 > \frac{I}{\mathcal{H}}$.

1. If $S_0 > c_B$, bank dealer j's equilibrium payoff (without deviation) is

   $$\frac{1}{N}\pi_B^0.$$

   Now, let bank dealer j deviate to $S_j = S_0 - \epsilon$. Then, $\lim_{\sigma\to 0} \bar{\pi}_{cj}(S_j, f_j) > \bar{\pi}_c^0$, and therefore bank dealer j's market share becomes

   $$\lim_{\sigma\to 0} \frac{1}{(N-1)\exp\left(\frac{1}{\sigma}\left(\bar{\pi}_c^\sigma - \bar{\pi}_{cj}(S_j, f_j)\right)\right) + 1} = 1.$$

   Bank dealer j's payoff after the deviation is

   $$1 \cdot \pi_j(S_0 - \epsilon, f_0),$$

   which is greater than $\frac{1}{N}\pi_B^0$ if ϵ is small enough.

2. If $f_0 > \frac{I}{\mathcal{H}}$, a similar proof can be used to show that bank dealer j always wants to deviate to $f = f_0 - \epsilon$.

The above contradiction implies that, in equilibrium, we must have $\pi_B^0 = 0$.

Step 2:

To show the second step, we need to establish the following result:

Claim 2. $\left(S = c_B, f = \frac{I}{\mathcal{H}}\right) = \arg\max_{(S_j, f_j)} \bar{\pi}_{cj}(S_j, f_j) + \pi_j(S_j, f_j)$.

Proof. We know that

$$\bar{\pi}_{cj}(S_j, f_j) + \pi_j(S_j, f_j) = \int_{f_j}^{b_j} (\mathcal{H}x - I)\, dG(x) + \int_{b_j}^{\infty} (x - c_B)\, dG(x).$$

The FOCs are

$$\frac{d\left[\bar{\pi}_{cj}\left(S_j,f_j\right)+\pi_j\left(S_j,f_j\right)\right]}{dS_j}=\left(\mathcal{H}b_j-I-\left(b_j-c_B\right)\right)G'\left(b_j\right)\frac{\partial b_j}{\partial S_j}=0,$$

$$\frac{d\left[\bar{\pi}_{cj}\left(S_j,f_j\right)+\pi_j\left(S_j,f_j\right)\right]}{df_j}=\left(\mathcal{H}b_j-I-\left(b_j-c_B\right)\right)G'\left(b_j\right)\frac{\partial b_j}{\partial f_j}-\left(\mathcal{H}f_j-I\right)=0.$$

The solution of the above FOCs is $\left(S=c_B,f=\frac{I}{\mathcal{H}}\right)$. □

Suppose that in the limit equilibrium $S_0>c_B$ and $f_0<\frac{I}{\mathcal{H}}$, or $S_0<c_B$ and $f_0>\frac{I}{\mathcal{H}}$. Based on the above Claim, we have

$$\bar{\pi}_{cj}\left(S_0,f_0\right)+\pi_j\left(S_0,f_0\right)<\bar{\pi}_{cj}\left(c_B,\frac{I}{\mathcal{H}}\right)+\pi_j\left(c_B,\frac{I}{\mathcal{H}}\right).$$

Since we have already shown that $\pi_j\left(S_0,f_0\right)=\pi_B^0=0$, and we know that $\pi_j\left(c_B,\frac{I}{\mathcal{H}}\right)=0$, we must have

$$\bar{\pi}_{cj}\left(S_0,f_0\right)<\bar{\pi}_{cj}\left(c_B,\frac{I}{\mathcal{H}}\right).$$

By the continuity of $\bar{\pi}_{cj}$ and π_j, there must exist $\epsilon>0$ such that

$$\bar{\pi}_{cj}\left(S_0,f_0\right)<\bar{\pi}_{cj}\left(c_B+\epsilon,\frac{I}{\mathcal{H}}\right).$$

It is clear that

$$\pi_j\left(c_B+\epsilon,\frac{I}{\mathcal{H}}\right)>0,$$

and so if bank dealer j deviates to $\left(c_B+\epsilon,\frac{I}{\mathcal{H}}\right)$, it can offer higher expected utility to all customers, thus obtaining a larger market share. At the same time, his expected per-capita profit is now positive, so total profits increase for bank dealer j, and therefore this cannot be an equilibrium. Thus, in equilibrium we must have $S_0=c_B$ and $f_0=\frac{I}{\mathcal{H}}$.

From the above analysis, we know that the customer's welfare is equal to

$$\max_{\left(S_j,f_j\right)}\bar{\pi}_{cj}\left(S_j,f_j\right)+\pi_j\left(S_j,f_j\right),$$

this is because the solution of the above problem is $S_j=c_B$ and $f_j=\frac{I}{\mathcal{H}}$, and we also know that $\pi_j\left(c_B,\frac{I}{\mathcal{H}}\right)=0$. Hence,

$$\bar{\pi}_{cj}\left(S_j,f_j\right)+\pi_j\left(S_j,f_j\right)=\int_{f_j}^{b_j}\left(\mathcal{H}x-I\right)dG\left(x\right)+\int_{b_j}^{\infty}\left(x-c_B\right)dG\left(x\right).$$

By the envelope theorem, we know that the customer's welfare is decreasing in c_B.

B.9 Proof of Proposition 10

When $c_B < c_{NB}$, we know that $\pi_{NB} = 0$. Hence, it is straightforward to obtain

$$\begin{aligned} W &= \pi_c + \pi_B + \pi_{NB} - \frac{2\mu}{r}\left(1 - G\left(b\right)\right)\left(c_{NB} - c_B\right) \\ &= \int_{f^\star}^{b^\star} \mathcal{H} x dG\left(x\right) + \int_{b^\star}^{\infty} \left(x - c_{NB}\right) dG\left(x\right), \end{aligned}$$

$$d\pi_c = \frac{2\mu}{r}\left[-\mathcal{H}\left(G\left(b^\star\right) - G\left(f^\star\right)\right) df^\star - \left(1 - G\left(b^\star\right)\right) dS^\star\right],$$

$$d\pi_B = \frac{2\mu}{r}\left[-\left(1 - G\left(b^\star\right)\right) dc_B\right],$$

$$d\left(-\frac{2\mu}{r}\left(1 - G\left(b\right)\right)\left(c_{NB} - c_B\right)\right) = \frac{2\mu}{r}\left[\left(c_{NB} - c_B\right) G'\left(b^\star\right) db^\star + \left(1 - G\left(b^\star\right)\right) dc_B\right].$$

In the unconstrained equilibrium, we know that $f^\star$ is constant, and thus

$$dW = \left(\mathcal{H} b^\star - \left(b^\star - c_{NB}\right)\right) db^\star.$$

We know in this case that $db^\star > 0$. Then

$$\begin{aligned} & dW > 0 \\ \Longleftrightarrow & \mathcal{H} b^\star > \left(b^\star - c_{NB}\right) \\ \Longleftrightarrow & c_{NB} > \left(1 - \mathcal{H}\right) b^\star = S^\star - \mathcal{H} f^\star \\ \Longleftrightarrow & c_{NB} + \mathcal{H} f^\star > S^\star \end{aligned}$$

We also know in this case that $S^\star < c_{NB}$, and hence the above inequality holds.

In the constrained equilibrium, we know that $S^\star$ is constant, $f^\star$ is decreasing in c_B, and $b^\star$ is increasing in c_B. Therefore,

$$\begin{aligned} dW &= d\pi_c + d\pi_B + d\left(-\frac{2\mu}{r}\left(1 - G\left(b\right)\right)\left(c_{NB} - c_B\right)\right) \\ &= \frac{2\mu}{r}\left[-\mathcal{H}\left(G\left(b^\star\right) - G\left(f^\star\right)\right) df^\star + \left(c_{NB} - c_B\right) G'\left(b^\star\right) db^\star\right] \\ &> 0 \end{aligned}$$

Hence we obtain that W is increasing in c_B.