Disintermediating the Interbank Market*

Tsz-Nga Wong† FRB Richmond

Mengbo Zhang‡ SUFE

This version: March 2025.

Abstract

We document a new *disintermediation* channel driving the decline in the interbank lending. In the early 2000s, over 80% of the interbank loans were intermediated, i.e., lent by banks who borrowed from others. By 2024 the number of intermediating banks has reduced by 85% and their borrowing by 93%. To explain this new channel, we devise an endogenous search model of divisible assets, solve the equilibrium in closed form, and derive testable implications from unconventional monetary policy (UMP) and regulation. Instrumented by granular IV and high-frequency monetary policy surprises, the bank-level data shows robust evidence supporting the model. We estimate the structural model and decompose the channel into five factors - (1) narrowing policy spread due to UMP, (2) abundant reserves injected by UMP, (3) Basel III regulations, (4) elimination of reserve requirements and (5) changes in reserve distribution. The disintermediation channel driven by (1) accounts for the largest portion of the decline in the interbank lending. The micro and macro reserve-demand elasticities (RDE) from our structural model are significantly different from the reduced-form estimates, highlighting the importance of our structural model for policy implications. Unlike typical search models, we prove that the equilibrium is constrained *inefficient* where banks actually intermediate *too often and too much*. Despite significant disintermediation, the welfare cost of implementing monetary policy in the decentralized interbank market has been much lower than before.

JEL Classification: G1, C78, D83, E44

Keywords: unconventional monetary policy, bank regulation, floor system, quantitative easing, wholesale funding, search and bargaining, constrained efficiency

*We are grateful to our discussants Semih Üslü, Shuo Liu, Sylvia Xiao and Gaston Navarro, and to Fernando Alvarez, Andy Atkeson, Huberto Ennis, Todd Keister, Anna Korver, Ricardo Lagos, Jeff Lacker, Guillaume Rocheteau, Pierre-Olivier Weill, Shengxing Zhang and participants of various seminars and conferences for helpful comments and suggestions. The views expressed here are those of the authors and do not necessarily reflect those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

†Research Department, Federal Reserve Bank of Richmond; Russell.Wong@rich.frb.org

‡School of Finance, Shanghai University of Finance and Economics; zhangmengbo@mail.sufe.edu.cn

1 Introduction

The Federal funds market, probably the most important market for monetary policy implementation, has been continuously shrinking since 2008, in spite of the fact that the US banks are larger than ever. The market volume of Federal funds by the end of 2019 (before the pandemic) has dropped to less than 20% of its peak in 2007, while the total assets of banks have expanded by 4.5 times during the same period. Although it has become a concern of policymakers, this secular decline of the Federal funds market is still largely a puzzle.

In this paper, we propose a new mechanism to explain the puzzle: the disintermediation channel. Specifically, the decline of Federal funds volume is driven by the decline of intermediated trading, where a Federal funds trade is intermediated if the reserve-purchasing bank is also selling reserves on the same day, i.e., the bank borrows to lend.[1] Our new mechanism is motivated by an observation that the Federal funds market was once heavily intermediated. Before 2008, typically more than 80% of the Federal funds were purchased by intermediary banks, i.e., the banks who do intermediated trading. This feature is distinct from the markets where the tradings of participants are usually one-sided. However, as illustrated in Figure 4, since the forth quarter of 2008, the decline in the Federal funds market is largely driven by the decline in the number of intermediary banks, as well as the volume of intermediated trades. We call the above phenomenon as the *disintermediation* of the Federal funds market.

To find out the sources of disintermediation, we notice that the timing of the disintermediation coincides with the time when the Federal Reserve implemented a series of unconventional monetary policies and regulation changes. The implementation started with the introduction of IOER, followed by three rounds of quantitative easing (QE hereafter) and the changes in regulation, such as the introduction of Basel III and widening of the basis of the FDIC's deposit insurance assessment fee. These policy and regulation changes greatly impacts the payoff and transaction cost of banks' Federal funds trading. In particular, the introduction of IOER changes the policy spread between the discount window rate and IOER, and the changes in regulation increase the balance sheet cost of holding reserves.

We argue that the adoption of unconventional monetary policies and changes in the transaction costs due to the regulation changes are among the main sources of disintermediation. While the disintermediation effect of transaction costs may seem straightforward, the disintermediation effect of the unconventional monetary policies calls for an explanation. It is puzzling since it is commonly thought that government sponsored enterprises (GSE) like Fannie Mae, Freddie Mac and Federal Home Loan Banks are not entitled to the IOER. It implies that under the unconventional monetary policies, there should be more Federal funds trades between GSEs and non-GSEs, and intermediated

[1] For readers not familiar with the jargons, Federal funds purchased are borrowing of reserves by buying the Federal funds today and selling back tomorrow. The price difference is the interest rate associated, i.e., the Federal funds rate. Similarly, Federal funds sold are lending of reserves.

loans in general, to earn the abitrage of the IOER. Our explanation is that unconventional monetary policies also amplify the disintermediation effect of transaction costs.

To illustrate the above mechanism of disintermediation, we develop a tractable continuous-time search-and-bargaining model of the over-the-counter Federal funds market. Our model extends the random search model in Afonso & Lagos (2015b) by allowing banks to choose their search intensities endogenously subject to transaction costs on holding reserves. This setup distinguishes our model from the standard random matching models, such as Afonso & Lagos (2015b), in multiple dimensions. First, on the effects of monetary policy and regulation on the Federal funds trading, the standard random matching models predict that any changes in the policy spread between the discount-window rate and IOER or the balance sheet cost do not affect the level of intermediation – all the effects are absorbed by the changes of the Federal funds rates. Moreover, the random matching models predict that the vast increase of reserves injected by QEs should have increased the level of intermediation instead. The reason is that, since matching is costless in Afonso & Lagos (2015b), banks always search for counterparties in the market, and they always trade to split their reserve holdings equally once they match with each other. It also implies that banks should trade more reserves when their average holding of reserves increases proportionally, ceteris paribus. Therefore, in the standard random matching models, the level of trades, along both the extensive and intensive margins, does not change even though the policy spread or the balance sheet cost changes the marginal value of holding reserves, as long as it is diminishing. On the contrary, our model shows that the endongenous search intensities and bilateral transaction volume respond endogenously to the gains of trade, which are determined by the policy spread, transaction cost and the aggregate reserve balances. Hence, our model predicts that the decline of policy spread and the reserves injection by QEs can reduce the level of intermediation and lower Federal funds trade on both the extensive and intensive margins. In addition, the increase of transaction cost raises the extensive margin of Federal funds trade and lowers the intensive margin.

Second, in the standard random matching models, the assumption of cost-free search implies that the constrained efficient allocation coincides with the equilibrium, where banks should always search and share the reserve holdings equally. However, our model shows that these features no longer hold when choosing search intensity becomes costly. Specifically, banks' search intensities are complementary in terms of bilateral matching, hence the equilibrium is constrained inefficient. But banks are not supposed to under-search due to the positive externality. Instead, they trade too much and search too much in the equilibrium compared with the constrained optimum.

To identify the theoretical predictions on the effects of unconventional monetary policies on the Federal funds trades and intermediated trades, we perform a series of instrumental variable regressions on a panel dataset of bank-level Federal funds trade volume. The dataset is collected from various sources, such as FFIEC Call Reports, Form FR-Y9C and SEC 10-Qs and 10-Ks. We find that the unconventional monetary policies significantly decrease the level of intermediated trades on both extensive and intensive margin, and also impede the allocation of Federal funds from net lenders to net

borrowers. These findings are robust to alternative specifications. Therefore, the disintermediation channel is empirically and economically significant.

We further calibrate our theoretical model with the empirical data via simulated method of moments, and conduct counterfactual analysis to evaluate the magnitudes of unconventional monetary policies and regulations on the disintermediation. We find that the disintermediation is mostly driven by the declining policy spread and the rising transaction cost, while the effect of excess reserve balances is small. In particular, in the year of 2018, the share of intermediation volume in total Federal funds volume doubles if we increase the policy spread to its 2006 level, and the share increases by four times if we decrease the estimated transaction costs to the 2006 level.

Literature. Our paper relates to several strands of literature. First, starting with Poole (1968), there has been a series of researches on the Federal funds market in partial equilibrium or general equilibrium models. Hamilton (1996) provides a partial equilibrium model to study the effects of transaction costs on the daily dynamics of the Federal funds rates. More recently, some studies focus on the monetary policy implementation and passthrough efficiency in the environment of excess reserves, such as Duffie & Krishnamurthy (2016), Bech & Keister (2017). In the meantime, other papers discuss the role of interbank markets and unconventional monetary policies on the aggregate outcome and welfare, such as Kashyap & Stein (2012), Ennis (2018), Williamson (2019), Bigio & Sannikov (2021) and Bianchi & Bigio (2022).

Another strand of literature focuses on capturing the over-the-counter (OTC) nature of the Federal funds markets and its implications. On the one hand, some reserches develop two-sided matching models to capture the search and matching frictions between lenders and borrowers, such as Berentsen & Monnet (2008), Bech & Monnet (2016), Afonso et al. (2019) and Chiu et al. (2020). These models are able to fit a number of aggregate empirical moments of the interbank markets in the U.S. and Europe and provide fruitful policy implications. However, the intermediation trades, which are important features of OTC markets, are missing in those models. On the other hand, people use continuous-time one-sided matching models to capture the intermediation feature of OTC markets. The one-sided matching models are pioneered by the seminal works of Afonso & Lagos (2015b) and Afonso & Lagos (2015a). Our model endogenizes the time-varying search intensity to study the disintermediation trades. The related papers include Duffie et al. (2005), Lagos & Rocheteau (2009), Trejos & Wright (2016), Farboodi et al. (2017), Lagos & Zhang (2019), Üslü (2019), Hugonnier et al. (2020) and Liu (2020).

There have been other papers that use network approach to study the interbank markets. For example, Bech & Atalay (2010) explores the network topology of the Federal funds market, and Gofman (2017) builds a network-based model of the interbank lending market and quantifies the efficiency-stability trade-offs of regulating large banks. Chang & Zhang (2018) develops a dynamic model that allows agents to endogenously choose counterparties and form network structure. They

find that some agents specialize in market making and become the core of the financial network, with the purpose of eliminating information frictions.

Outline. The remainder of the paper is as follows. Section 2 describes the institutional background and the aggregate empirical facts that motivate our paper. Section 3 presents the a search model of Federal funds market that allows for closed-form solutions and comparative statics. Section 4 documents the empirical evidence on the disintermediation effect of unconventional monetary policies at the individual bank level via testing the predictions of the theoretical model. Section 5 structurally estimates the theoretical model and quantitatively decomposes the effects of unconventional monetary policies on Federal funds intermediation. The final section, 6, concludes the paper.

2 The Landscape of the Federal Funds Market

This section introduces the institutional features, the policy and regulatory environment and the aggregate trade dynamics in the Federal funds market to motivate our estimation and theoretical model in the following sections. We will focus on the change of the landscape of this market before and after the Great Recession as the market has changed drastically since then. To measure the aggregate and composition of the Fed funds trade activity, we aggregate the data from a set of regulatory filings, including the quarterly Consolidated Report of Condition and Income for U.S. banks and branches (Call reports), the Consolidated Financial Statements (Form FR Y-9C) for bank holding companies (BHC) and SEC 10-Ks and 10-Qs for other eligible entities.

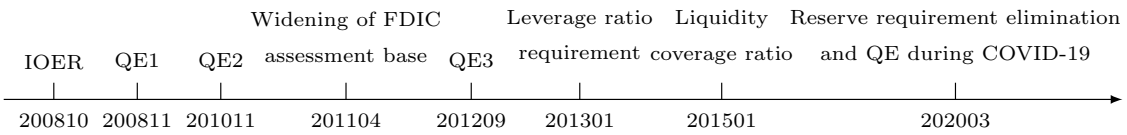
2.1 Institutional Background

The Federal funds market is a market for unsecured loans of dollar reserves held at the Federal Reserve Banks. The market interest rates on these loans are commonly referred to as the Federal funds rates. Most of the Federal funds transactions are overnight (99%). Financial institutions (FIs) rely on the Fed funds market for short-term liquidity needs: First, the Federal funds is not considered as the deposits to the borrower bank under Regulation D, thus it is useful for borrower banks to satisfy their reserve requirements and payments needs. Second, the lender FIs can lend excess reserves and earn overnight Fed funds rate. Regarding the market structure, the Federal funds market is an over-the-counter (OTC) market without centralized exchange. A borrower bank (Federal funds purchased) and a lender bank (Federal funds sold) meet and trade bilaterally, and the transfer of funds is completed through the Fed's reserve accounts.

The market of Federal Funds has been the epicenter where monetary policies are implemented. Before the Great Recession, the Federal Reserve adjusted the supply of reserve balances, by the purchase and sale of securities in the open market, so as to keep the Fed fund rates around the target of monetary policy. Since the Great Recession, the landscape of the market has changed drastically

due to a series of unconventional monetary policies and regulations. Figure 1 plots the timeline of these changes, which start with the introductin of interest on excess reserves (IOER), followed by three rounds of quantitative easing (QE) as well as changes in regulation, such as the widening of the basis for FDIC assessment fee and the introduction of Basel III regulations.

Figure 1: Timeline of unconventional monetary policy and regulation



Notes: This figure plots the timeline of unconventional monetary policies and regulations since the Great Recession. The numbers on the timeline represents the date (year-month) when the policy or regulation is introduced.

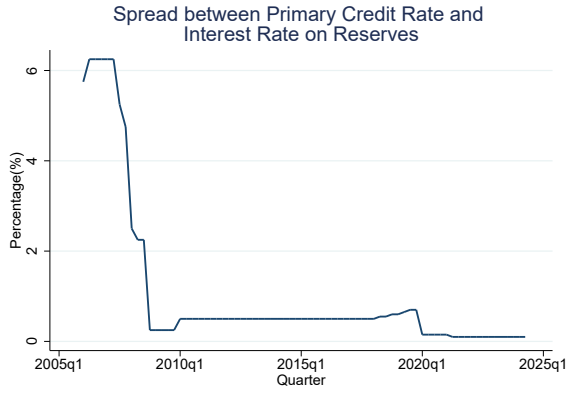
Due to the changes in policy and regulations, the Fed funds market has entered a stage with excess reserves, and the Federal Reserve relies on two new policy tools to implement its desired target range for the Federal funds rate: the IOER, which it offers to eligible depository institutions, is set at the top of the target ranges; and the rate of return at the overnight reverse repurchase (ON RRP) facility, which is available to an expanded set of counterparties including government-sponsored enterprises (GSEs) and some money market funds, is set at the bottom of the rage. Figure 2 shows the time series of the unconventional monetary policies. Panel (a) plots the path of IOER, which has been steadily increasing between 2008Q4 and 2018Q4. Panel (b) plots the mean and standard deviation of individual excess reserve balances in the same period , which has grown drastically since the Great Recession.

2.2 (Dis)intermediation in the Federal funds market

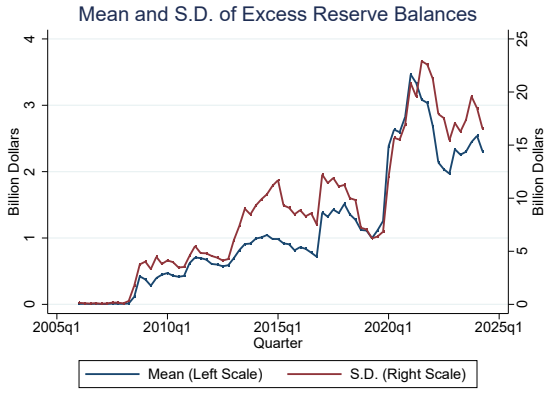
Due to the over-the-counter structure, the Fed funds trades involve a significant share of intermediation trading. A group of banks act as intermediaries by borrowing reserves from the lender banks and lending them to others on the same day. We find that the intermediary banks are responsible for most of the decline in Fed funds volume. Specifically, by consolidating the individual balance sheet data, we decompose the total Fed funds volume into three groups: intermediary banks, non-intermediary banks and government-sponsored enterprises (GSEs). As illustrated in Panel (a) of Figure 3, the decline of Fed funds purchased (borrowing) is entirely driven by intermediary banks, whose volume of borrowing sharply declined from the peak of $195 billion in 2007Q2 to an average of $22 billion in 2018. At the same time, the volume of borrowing by other groups stayed stable over time. Panel (b) of Figure 3 suggests that, on the supply side, the depository institutions account for most of the decline of Fed funds lending. In particular, the lendings by intermediary banks was more than $60 billion on average before 2008, but decreased sharply to almost zero right after the Great Recession.

Figure 2: IOER and Excess Reserves

(a) Spread between Primary Credit Rate and IOR



(b) Excess reserves



Notes: This figure plots the sequences of IOER, the mean and standard deviation of individual excess reserve balances from 2006Q1 to 2024Q2. Data source: FRED, Call reports, FR Y-9C.

The non-intermediary banks accounted for about $50 billion lending before the Great Recession, and shrank gradually to less than $5 billion over time in 2018.

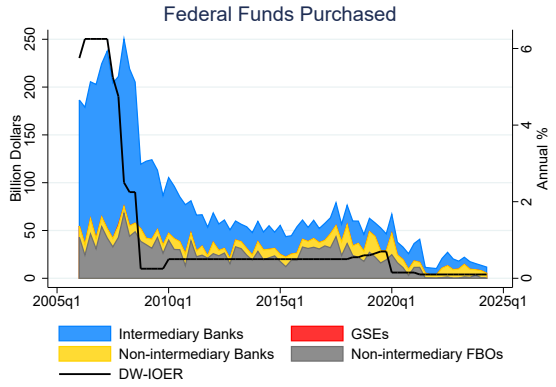
The decline of borrowing and lending by intermediaries imply the decline of Fed funds reallocation. We find that this decline occurs on both extensive and intensive margin. As plotted in Panel (a) of Figure 4, a significant share (more than 15%) of Fed funds volume is traded for intermediating purposes. However, since the financial crisis has declined by more than two thirds to less than 5%. Moreover, Panel (b) of Figure 4 shows that the number of intermediary banks has also decreased from 600 in 2006 to less than 100 at the end of 2018.

Why did disintermediation happen? Certainly, the Federal funds market has been going through a transition from the Great Recession, but it is worth noting that the timing of the disintermediation coincides with the changes in the monetary policies and regulations, as plotted in Figure 1. All these changes closely relate to banks' incentive to trade Fed funds. For example, the introduction of IOER raises the return of holding reserves, which lowers banks' lending incentives and raises their borrowing incentives. The QEs have left banks flush with excess reserves. As a result, the demand for borrowing reserves to meet the reserve requirement and payment needs has become rare. The regulation changes The widening of FDIC assessment base and Basel III regulations increase the balance sheet cost of holding reserves. For example, FDIC insurance premium is now charged according to the size of FI's assets (instead of the size of deposit), which is increasing in the Federal Funds borrowed. Furthermore, Basel III now imposes a cap on the FI's leverage ratio and a floor of the holding of liquid (and usually low-return) asset to cover potential cash outflow, increasing the regulation cost.

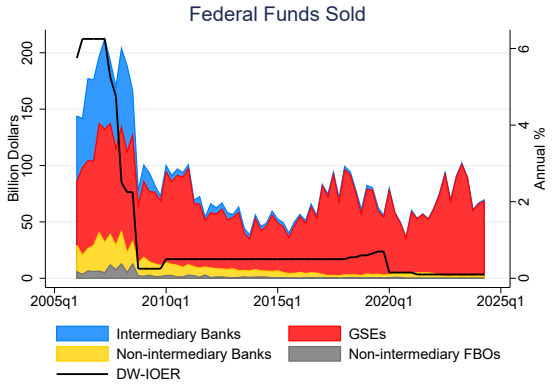
Our empirical facts about the disintermediation coincide with the existing literature. For example,

Figure 3: Decomposition of Federal funds volume

(a) Federal funds purchased



(b) Federal funds sold



Notes: This figure plots the decomposition of the aggregate Federal funds purchased and sold by groups from 2006Q1 to 2024Q2. Data source: Call reports, FR Y-9C, SEC 10-K and 10-Q.

Keating & Macchiavelli (2017) find that the proportion of intermediated funds declined sharply after the financial crisis. On the daily level, the domestic banks keep more than 99% percent of Fed funds borrowed and foreign banks keep more than 80%. These evidence document the importance of intermediation to the substantial decline of Fed funds volume. We will focus on examining banks' incentive to intermediate and its implications for the monetary policy implementation.

3 A Baseline Search Model

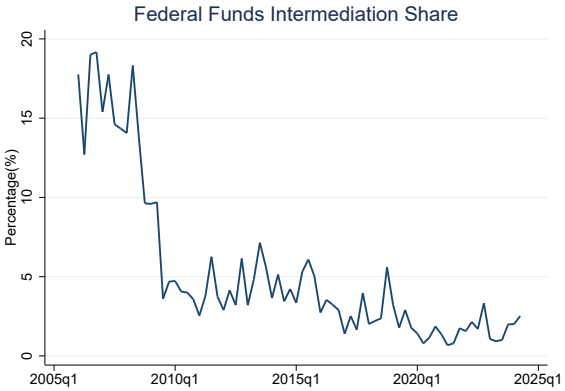
Overview. To highlight our new channel, our baseline model closely follows Afonso & Lagos (2015b) and Üslü (2019) but endogenizes search intensity.[2]

Time is continuous. A unit measure of agents (interpreted as banks) each holds a divisible amount k_t of a liquid asset (reserve balances). From these holdings, banks derive two sources of payoff. First, they earn a flow payoff $u(k_t)$ continuously from time 0 to T, capturing intraday *liquidity benefit* from holding balances for payments or other operations. Second, they receive a terminal payoff $U(k_T)$ at time T, reflecting the *portfolio benefit* from holding balances for their returns and meeting regulation requirements. Additionally, at time $T + \Delta$, a centralized market opens where banks trade illiquid assets and balances at a linear payoff, representing the rebalancing of their portfolios. We assume banks endow with sufficiently large amounts of illiquid assets, or can hold negative amounts, to ensure their rebalancing is never constrained.

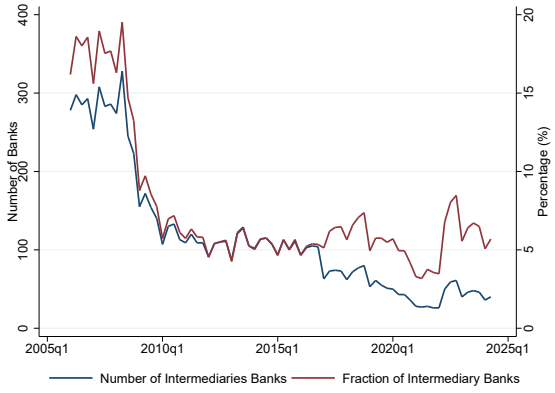
[2] Other differences from Afonso & Lagos (2015b) include divisible reserve balances (rather than discrete) and various extensions considered in Section XXX. Different from Üslü (2019), we solve the entire transitional dynamics rather than focusing on the steady state. Beyond the Federal funds market, the model can be applied to any decentralized markets of credit or of asset with delayed settlement.

Figure 4: Aggregate Fed funds intermediation

(a) Intermediation volume share



(b) Number of intermediary banks



Notes: This figure plots the intermediation volume share and number of intermediary banks. Data source: Call reports, FR Y-9C.

Federal funds market. Following Afonso & Lagos (2015b), banks begin with heterogeneous initial balances k_0 and can trade bilaterally in the federal funds market, which opens continuously from time 0 to T. Trading in the federal funds market is decentralized and subject to search frictions, so it takes time for each bank to find a counterparty. Consequently, a bank's balances evolve according to the jump process:

$$k_t = k_0 + \sum_{t_n \leq t} q_{t_n}, \tag{1}$$

where t_n denotes the moment the n-th counterparty is found, from whom q_{t_n} units of balances are purchased. These balances are repaid at time $T + \Delta$, using proceeds from the sale of illiquid assets. The paramenter $\Delta > 0$ captures the repayment delay in reality.

Unlike in Afonso & Lagos (2015b), banks can choose their trade frequency and counterparty-finding rate at time t by exerting higher search intensity $\varepsilon_t \in [0, 1]$, which is costly. As banks adjust their search intensities and execute trades, the distribution of balances evolves endogenously, denoted as $F_t(\cdot)$.

Search. For two banks exerting search intensities ε and ε', they find each other at the Poisson arrival rate $m(\varepsilon, \varepsilon')$.[3] We assume that the matching function is symmetric, increasing, supermodular, and additive in the counterparties' search intensities, satisfying

$$m\left[\alpha\varepsilon' + (1-\alpha)\varepsilon'', \varepsilon\right] = \alpha m\left(\varepsilon, \varepsilon'\right) + (1-\alpha) m\left(\varepsilon, \varepsilon''\right). \tag{2}$$

[3] For readers unfamilar with the Poisson model: if a bank follows a contingent plan of search intensity $\{\varepsilon_t\}_{t=0}^{T}$, then its probability of finding a counterparty bank within τ units of time is $\Pr\{t_1 \leq \tau\} = 1 - \exp\left\{-\int_0^\tau \int_{j\in[0,1]} m\left(\varepsilon_t, \varepsilon_t^j\right) djdt\right\}$.

Let $\varepsilon_t^*(k)$ denote the (equilibrium) search intensity strategy of a bank holding k units of balances (a "k-bank") at time t. We denote $\boldsymbol{\varepsilon}^* = \{\varepsilon_t^*(k)\}_{t\in[0,T],k\in\mathbb{R}}$ as the complete profile of all banks' strategies.

Preferences. The bank's objective is to choose a search intensity strategy $\boldsymbol{\varepsilon} = \{\varepsilon_t(k)\}_{t\in[0,T],k\in\mathbb{R}}$ that maximizes the expected discounted payoff (3):

$$\max_{\boldsymbol{\varepsilon}} \mathbb{E}^{\boldsymbol{\varepsilon}} \left\{ \int_0^T e^{-rt} r u(k_t)\, dt + e^{-rT} U(k_T) - \sum_{n=1,2,\ldots} \left[e^{-rt_n} \chi(\varepsilon_{t_n}, q_{t_n}) + e^{-r(T+\Delta)} \rho_{t_n} q_{t_n} \right] \right\}, \tag{3}$$

subject to (1).

Inside the integral and the terminal term, $u(k_t)$ and $U(k_T)$ capture the present value of the liquidity payoffs and the portfolio payoffs from holding balances. The last term of (3) reflects the two key costs of trading in the federal funds market: the cost $\chi(\varepsilon, q)$ of exerting search intensity ε to purchase (or sell, if negative) q units of balances, and the federal funds rate ρ applied to those transactions. These costs are incurred *ex post, i.e.,* only once counterparties are found and trades occur, making the model highly tractable.

To obtain closed-form solutions, we follow Üslü (2019) and Babus & Kondor (2018) and assume quadratic payoffs:

$$u(k) = -a\left(k - \bar{k}\right)^2, \; U(k) = -A\left(k - \tilde{k}\right)^2. \tag{4}$$

An interpretation is that $k = \bar{k}$ captures the bank's desirable level of balances for liquidity needs during the day, whereas $k = \tilde{k}$ captures the desirable level of balances for the portfolio purpose by the day's end. In general, these two targets differ from the bank's initial holdings, thus trading in the federal funds market allows banks to dynamically rebalance their holdings toward each target as time goes by. Although this quadratic assumption is stronger than strictly required for some results, we adopt it for clarity from the outset.

Similarly, we assume that the quadratic cost:

$$\chi(\varepsilon, q) = \kappa(\varepsilon) q^2.$$

The model in Afonso & Lagos (2015b) corresponds to the special case of $\chi(\varepsilon, q) = 0$. Beyond these particular choices for the payoffs and costs, we do not impose any additional functional-form assumptions on either the matching function $m(\varepsilon, \varepsilon')$ or the search cost function $\kappa(\varepsilon)$.

Bargaining. When two banks find each other, they negotiate the terms of trade (specifically q and ρ) via a Nash bargaining protocol, spliting their trade surplus. For a bank holding k unit of balances at time t, having exerted search intensity ε, and purchasing q additional units at the federal funds

rate ρ, the trade surplus is

$$B_t(q,\rho;k,\varepsilon) \equiv V_t(k+q) - e^{-r(T-t+\Delta)}\rho q - \chi(\varepsilon,q) - V_t(k),$$

where $V_t(k)$ denotes the maximal attainable continuation value of a k-bank at t. Similarly, the trade surplus of another bank with holdings k' and search intensity ε' is $B_t(-q,\rho;k',\varepsilon')$. The terms of trade q and ρ solve the following Nash bargaining problem:

$$\max_{q,\rho} B_t(q,\rho;k,\varepsilon)\, B_t\left(-q,\rho;k',\varepsilon'\right). \tag{5}$$

Let $q_t(k,k';\varepsilon,\varepsilon')$ and $\rho_t(k,k';\varepsilon,\varepsilon')$ be the solutions to (5). Given a strategy profile $\boldsymbol{\varepsilon}^*$ adopted by other banks, let $B_t^*(k,\varepsilon;k') \equiv B_t\left[q_t(k,k';\varepsilon,\varepsilon_t^*(k')),\rho_t(k,k';\varepsilon,\varepsilon_t^*(k'));k,\varepsilon\right]$ denote the corresponding trade suplus under the Nash bargaining solution.

Value and distribution. Given an equilibrium strategy profile $\varepsilon_t^*(\cdot)$ and a distribution $F_t(\cdot)$, the value function $V_t(k)$ satisfies following Hamilton-Jacob-Bellman (HJB) equation[4]

$$rV_t(k) = \dot{V}_t(k) + u(k) + \max_{\varepsilon_t(k)\in[0,1]} \int B_t^*\left[k,\varepsilon_t(k);k'\right] m\left[\varepsilon_t(k),\varepsilon_t^*\left(k'\right)\right] dF_t\left(k'\right), \tag{6}$$

subject to the terminal condition $V_T(k) = U(k)$.

The intuition behind the HJB equation is as follows. Following the equilibrium strategy profile $\varepsilon_t^*(\cdot)$, each k'-bank exerts search intensity $\varepsilon_t^*(k')$. A k-bank exerting any arbitary search intensity $\varepsilon_t(k)$ finds a k'-bank at the arrival rate $m[\varepsilon_t(k),\varepsilon_t^*(k')]$, from which it earns a trade surplus $B_t^*[k,\varepsilon_t(k);k']$. The expected gain from trading is thus given by integrating over all potential trading counterparties of k'-bank. In equilibrium, the strategy $\varepsilon_t(k) = \varepsilon_t^*(k)$ also maximizes (6) for all k and t.

Next, considering the inflow and outflow of banks' balances, the distribution $F_t(\cdot)$ evolves according to the following Kolmogorov forward equation (KFE)[5]

$$\dot{F}_t(k) = \begin{array}{l} \underbrace{-\int_{k^b\le k}\int m\left[\varepsilon_t^*\left(k^b\right),\varepsilon_t^*\left(k^s\right)\right] 1\left\{k^b + q_t\left(k^b,k^s;\varepsilon_t^*\left(k^b\right),\varepsilon_t^*\left(k^s\right)\right) > k\right\} dF_t\left(k^s\right) dF_t\left(k^b\right)}_{\text{outflow}} \\ \underbrace{+\int_{k^b>k}\int m\left[\varepsilon_t^*\left(k^b\right),\varepsilon_t^*\left(k^s\right)\right] 1\left\{k^b + q_t\left(k^b,k^s;\varepsilon_t^*\left(k^b\right),\varepsilon_t^*\left(k^s\right)\right) \le k\right\} dF_t\left(k^s\right) dF_t\left(k^b\right)}_{\text{inflow}} \end{array}, \tag{7}$$

given the terminal condition $F_0(k)$.

[4] For readers unfamiliar with the HJB equation, we derive (6) in the online Appendix C.2. The discretized version of (6) without search cost or transaction cost is Proposition 1 of Afonso & Lagos (2015b).

[5] For readers unfamiliar with the KFE, we derive (7) in the online Appendix C.2. When k is discrete, $F_t(k)$ is probability mass function shown in Proposition 2 of Afonso & Lagos (2015b).

Intuitively, $F_t(k)$ is the fraction of banks holding at most k units of balances at t. Balances-buying banks with at most k units of balances (the first term of outflow) may move out of the set $\{k^b + q_t \leq k\}$ after trading, while balances-buying banks with more than k units (the second term of inflow) may move into that set if their trades push them below k.

Equilibrium. The equilibrium dynamics are fully described by the backward PDE of the value function (6) and the forward PDE of the distribution function (7). Specifically, from the choice of search intensity in (6), the equilibrium search profile $\varepsilon_t^*(\cdot)$ is a fixed point function to the following functional:

$$\Gamma\left(\varepsilon_t^*\left(k'\right); V_t, F_t\right)(k) \equiv \arg\max_{\varepsilon_t(k)\in[0,1]} \left\{ \int B_t^*\left[k, \varepsilon_t(k); k'\right] m\left[\varepsilon_t(k), \varepsilon_t^*\left(k'\right)\right] dF_t\left(k'\right) \right\}. \tag{8}$$

Denote $\Omega(V_t, F_t) \subseteq [0,1]^{\mathbb{K}}$ as the set of fixed points to Γ, i.e., $\varepsilon_t^*(k) = \Gamma\left(\varepsilon_t^*(k'); V_t, F_t\right)(k)$ for all $\boldsymbol{\varepsilon}_t^* \in \Omega(V_t, F_t)$. The following lemma guarantees that the set of equilibrium strategies $\Omega(V_t, F_t)$ is never empty.

Lemma 1 *$\Omega(V_t, F_t)$ is not empty for any V_t and F_t.*

Having ruled out non-existence, as in many supermodular, multiple equilibria may nevertheless arise, necessitating an equilibrium-selection rule.[6] We impose an equilibrium-selection rule $\mathcal{E}(V_t, F_t) \in \Omega(V_t, F_t)$ that selects the equilibrium strategy with maximal aggregate joint surplus:

$$\mathcal{E}(V_t, F_t) \equiv \arg\max_{\boldsymbol{\varepsilon}_t^* \in \Omega(V_t, F_t)} \int \int \underbrace{\left\{ B_t^*\left[k, \varepsilon_t^*(k); k'\right] + B_t^*\left[k', \varepsilon_t^*\left(k'\right); k\right] \right\}}_{\text{joint surplus}} m\left[\varepsilon_t^*(k), \varepsilon_t^*\left(k'\right)\right] dF_t\left(k'\right) dF_t(k). \tag{10}$$

We are now ready to define a symmetric subgame-perfect equilibrium:

Definition 1 *An equilibrium consists of $\{\boldsymbol{\varepsilon}^*, \mathbf{V}, \mathbf{F}\} = \{\varepsilon_t^*(k), V_t(k), F_t(k)\}_{t\in[0,T], k\in\mathbb{R}}$ such that,*

*(a. **Value**) given $\boldsymbol{\varepsilon}^*$ and $\mathbf{F}$, $\mathbf{V}$ solves the bank's problem (6);*

*(b. **Distribution**) given $\boldsymbol{\varepsilon}^*$ and $\mathbf{V}$, $\mathbf{F}$ satisfies (7);*

*(c. **Search**) given $\mathbf{V}$ and $\mathbf{F}$, banks choose the strategy $\boldsymbol{\varepsilon}^* = \{\mathcal{E}(V_t, F_t)\}_{t\in[0,T]}$ of (10).*

[6] Because m and S are supermodular, coordination failure can lead to multiple equilibria. Furthermore, it is well-known that forward–backward differential equations can admit multiple solutions. For example, a simple system of forward-backward ODEs

$$\begin{aligned} \dot{y}(t) &= -x(t), \text{ where } y(2\pi) = 0, \\ \dot{x}(t) &= y(t), \text{ where } x(0) = 0, \end{aligned} \tag{9}$$

has a continuum of solutions $\{x(t) = A\sin t, y(t) = A\cos t\}$. In macroeconomics, the literature on equilibrium indeterminacy after the seminar work of Benhabib & Farmer (1994) shows how standard neoclassical growth models (which typically involve a forward-looking capital-accumulation equation and a backward-looking Euler equation) can yield multiple equilibria. Our setting features partial differential equations in which the state variable is the distribution of balances—an infinite-dimensional analog—which likewise permits multiplicity.

Proposition 1 *An equilibrium* $\{\boldsymbol{\varepsilon}^*, \mathbf{V}, \mathbf{F}\}$ *exists and is unique under the rule (10).*

The proposition establishes that, despite the possibility of multiple equilibria arising from supermodularities and the forward–backward structure of the model, a unique equilibrium emerges once we impose the equilibrium-selection rule (10). First, the Hamilton–Jacobi–Bellman (HJB) equation (6) determines the bank's optimal continuation value $V_t(k)$, given both the prevailing distribution of balances $F_t(k)$ and other banks' strategies. Meanwhile, the Kolmogorov forward equation (7) describes how $F_t(k)$ evolves in time. Supermodularity—where each bank's best response increases with the strategies of others—can generate multiple solutions to these equations. However, by selecting the strategy profile that yields the highest overall surplus, the equilibrium-selection rule narrows the set of equilibria to a single outcome: there is only one profile that simultaneously solves the HJB and KFE equations and maximizes total gains from trade at any point of time. Even so, the forward–backward nature of the model can, in theory, give rise to multiple solutions of the PDE system. Nonetheless, as shown in the proof of Proposition 1, a solution to the PDE system exists and is uniquely pinned down by this surplus-maximizing criterion.

3.1 Characterization

We guess and verify that the second derivative of the value function is degenerate, meaning there exists a time function H_t such that $V_t''(k) = -2H_t$ for all k. As in many models featuring Nash bargaining, the bargaining solution—and thus the equilibrium allocation of balances—depends on the marginal value of balances, reflecting the bilateral demand and supply curves. The model's tractability follows from the fact that this degenerate (yet time-varying) second derivative serves as a sufficient statistic for determining the marginal value of balances. Consequently, we need not track the entire distribution of the marginal value of balances over time.

Equilibrium federal funds trade. Given the premise $V_t''(k) = -2H_t$ and an equilibrium search intensity $\varepsilon_t^*(\cdot)$, the balances traded and federal funds rate under the Nash bargaining are given by

$$q_t^*(k,k') = \underbrace{\frac{H_t}{H_t + \frac{\kappa[\varepsilon_t^*(k)] + \kappa[\varepsilon_t^*(k')]}{2}}}_{\text{precaution–speed trade-off}} \times \underbrace{\frac{k'-k}{2}}_{\text{efficient bilateral trade without cost}}, \tag{11}$$

$$1 + \rho_t^*(k,k') = \underbrace{e^{r(T+\Delta-t)}}_{\text{time cost}} \left[\underbrace{\frac{V_t'(k) + V_t'(k')}{2}}_{\text{sharing marginal valuation}} + \underbrace{\frac{\kappa[\varepsilon_t^*(k')] - \kappa[\varepsilon_t^*(k)]}{2} q_t^*(k,k')}_{\text{speed premium (discount)}} \right]. \tag{12}$$

Afonso & Lagos (2015b) corresponds to the special case $\kappa(\varepsilon) = 0$. In this case, the two banks that meet exchange the efficient trade size $(k'-k)/2$, , leaving both with the same post-trade balances.

Because they have equal bargaining power, they agree on a price that equals the average of their marginal valuations of reserves. Once we introduce transaction costs and endogenous search intensity, however, the bilateral trade size falls below this efficient level. Higher κ, ε or ε' raise the marginal cost of transactions, reducing the equilibrium trade amount. This captures a precaution–speed trade-off: while a higher search intensity enables banks to find counterparties more quickly, it is also more costly, so they trade smaller orders in each match.

Endogenous search intensity likewise generates a speed premium or discount on the bilateral federal funds rate, mirroring a similar insight from Üslü (2019). The premium depends on both trade size and the difference in search intensities. Specifically, when $k' > k$ and $\varepsilon' > \varepsilon$, or $k' < k$ and $\varepsilon' < \varepsilon$, the seller bank searches more intensively than the buyer bank. This benefits the buyer, who faces less delay, while the seller bears a higher cost. Under Nash bargaining, some of that cost is shifted to the buyer via a premium on the federal funds rate. Conversely, when the buyer's search intensity exceeds the seller's (i.e.,$k' > k$ and $\varepsilon' < \varepsilon$, or $k' < k$ and $\varepsilon' > \varepsilon$), the seller faces less urgency, inducing a discount in the federal funds rate.

Trade frequency. Given the premise $V_t''(k) = -2H_t$, the search intensity profile chosen by the equilibrium-selection rule $\mathcal{E}(V_t, F_t)$ is also degenerate, as established by the following proposition.

Proposition 2 *(**Degeneracy**) Suppose $V_t''(k) = -2H_t$, then for any k and F_t we have $\varepsilon_t^*(k) = \varepsilon_t^*$ that solves*

$$\varepsilon_t^* = X(H_t) \equiv \frac{H_t^2}{4} \max_{\varepsilon \in [0,1]} \left\{ \frac{m(\varepsilon, \varepsilon)}{H_t + \kappa(\varepsilon)} \right\}, \tag{13}$$

The degeneracy rests on the complementarity of the matching function and the convexity of transaction costs. When two banks have different search intensities, their expected bilateral surplus is always greater if both adopt the average of their intensities. By doing so, the bilateral transaction cost decreases via Jensen's inequality, while the bilateral matching rate increases due to the complementarity property of the matching function.

Define $P_t(k)$ as the probability of a k-bank that will trade some balances in some future time $z \geq t$. Using Proposition 2, $P_t(k)$ has a closed-form solution:

$$P_t(k) = 1 - \exp\left(- \int_t^T m_s^* dz \right), \tag{14}$$

where $m_t^* \equiv m(\varepsilon_t^*, \varepsilon_t^*)$.

Net positions. Define the expected amount of net trades from time t to T of a bank holding k units of balances at time t as

$$L_t(k) \equiv \mathbb{E} \sum_{t_i \in [t,T]} q_{t_i}^* \left(k_{t_i}, k' \right) \text{ s.t. } k_t = k. \tag{15}$$

Using the equilibrium bargaining solution (11) and Proposition 2, the individual net trade admits a closed-form solution:

$$L_t(k) = \left\{1 - \exp\left(-\int_t^T \frac{m_z^*}{1+Y_z} dz\right)\right\}(K-k), \tag{16}$$

where $Y_t \equiv 1 + 2\kappa(\varepsilon_t^*)/H_t$.

Other closed-form solutions. With quadratic payoffs and degenerate equilibrium search intensity profile, there exists a unique equilibrium under the refinement. The following proposition further shows that the equilibrium admits an analytical solution.

Proposition 3 ***(Value function in closed-form)*** *In equilibrium, the value function is given by* $V_t(k) = -H_t k^2 + E_t k + D_t$, *where*

$$\dot{H}_t = rH_t - a + X(H_t), \quad \textit{where } H_T = A; \tag{17}$$

$$\dot{E}_t = rE_t - 2a\bar{k} + 2KX(H_t), \quad \textit{where } E_T = 2A\tilde{k}; \tag{18}$$

Having solved the time path of H_t, we are able to characterize the path of equilibrium reserve distribution as in the following lemma.

Proposition 4 ***(Moments of distribution in closed-form)*** *In equilibrium, the balances distribution solves the following PDE:*

$$\dot{F}_t(k) = m_t^* \left[\int F_t\left[(1+Y_t)k - Y_t k'\right] dF_t(k') - F_t(k)\right], \tag{19}$$

given the initial condition $F_0(k)$. *Denote the n-th moment of the balances distribution at time t as* $M_{n,t} \equiv \int k^n dF_t(k)$. *The moment function is given by the following ODE:*

$$\dot{M}_{n,t} = m_t^* \left[\sum_{i=0}^{n} C_n^i \frac{Y_t^i}{(1+Y_t)^n} M_{n-i,t} M_{i,t} - M_{n,t}\right], \tag{20}$$

with $M_{0,t} = 1$, $M_{1,t} = K$ *and*

$$M_{2,t} = K^2 + \left(M_{2,0} - K^2\right) \exp\left[-2\int_0^t \frac{m_z^* Y_z}{(1+Y_z)^2} dz\right]. \tag{21}$$

Thanks to Fourier transform, the model allows for an analytical expression for the paths of moments of the reserve distribution. In particular, equation (21) implies that the variance of the reserve distribution converges to zero at the speed of $2m_t^* Y_t/(1+Y_t)$, which is endogenously determined. In particular, a higher H_t implies a faster speed of convergence.

3.2 Constrained efficiency

In this section we examine the constrained efficiency of the federal funds market. Consider a social planner that dictates search decision $\{\varepsilon_t^p(k)\}$ and bilateral exchange of reserve balances $\{q_t^p(k,k')\}$ to maximize the following utilitarian welfare function.

Definition 2 *A constrained efficient allocation consists of* $\{\varepsilon_t^p(k), F_t^p(k), q_t^p(k,k')\}_{k,k'\in\mathbb{K},t\in[0,T]}$ *that solves*

$$\mathbb{W} = \max \left\{ \begin{array}{l} \int_0^T e^{-rt} \int u(k)\, dF_t^p(k)\, dt + e^{-rT} \int U(k)\, dF_T^p(k) \\ -\int_0^T \int\int e^{-rt} \chi\left[\varepsilon_t^p(k), q_t^p(k,k')\right] m\left[\varepsilon_t^p(k), \varepsilon_t^p(k')\right] dF_t^p(k')\, dF_t^p(k)\, dt \end{array} \right\} \tag{22}$$

subject to the law of motion of reserves

$$\dot{F}_t^p(k^w) = \left\{ \begin{array}{l} \int_{k>k^w} \int m\left[\varepsilon_t^p(k), \varepsilon_t^p(k')\right] 1\left\{k + q_t^p(k,k') \leq k^w\right\} dF_t^p(k')\, dF_t^p(k) \\ -\int_{k\leq k^w} \int m\left[\varepsilon_t^p(k), \varepsilon_t^p(k')\right] 1\left\{k + q_t^p(k,k') > k^w\right\} dF_t^p(k')\, dF_t^p(k) \end{array} \right\}, \tag{23}$$

given $F_0^p(k^w) = F_0(k^w)$.

The constrained efficient allocation $\{\varepsilon_t^p(k), q_t^p(k,k')\}$ maximizes the Hamiltonian. Denote $V_t^p(k)$ as the co-state to $dF_t^p(k)$, the Hamiltonian is given by

$$\begin{aligned} \mathcal{H}_t^p \equiv & \int u(k)\, dF_t^p(k) - \int\int \chi\left[\varepsilon_t^p(k), q_t^p(k,k')\right] m\left[\varepsilon_t^p(k), \varepsilon_t^p(k')\right] dF_t^p(k')\, dF_t^p(k) \\ & + \int\int m\left[\varepsilon_t^p(k), \varepsilon_t^p(k')\right] \left\{V_t^p\left[k + q_t^p(k,k')\right] - V_t^p(k)\right\} dF_t^p(k')\, dF_t^p(k) \\ & + \int\int \eta_t(k,k') \left[q_t^p(k,k') + q_t^p(k',k)\right] dF_t^p(k')\, dF_t^p(k), \end{aligned} \tag{24}$$

where $\eta_t(k,k')$ is the multiplier to the bilateral trade constraint $q_t^p(k,k') + q_t^p(k',k) = 0$. The evolution of the co-state solves[7]

$$\begin{aligned} rV_t^p(k) = & \dot{V}_t^p(k) + u(k) \\ & + \int \left\{ \begin{array}{l} V_t^p\left[k + q_t^p(k,k')\right] + V_t^p\left[k' - q_t^p(k,k')\right] - V_t^p(k) \\ -V_t^p(k') - \chi\left[\varepsilon_t^p(k), q_t^p(k,k')\right] - \chi\left[\varepsilon_t^p(k'), -q_t^p(k,k')\right] \end{array} \right\} m\left[\varepsilon_t^p(k), \varepsilon_t^p(k')\right] dF_t^p(k'), \end{aligned} \tag{25}$$

with $V_T^p(k) = U(k)$. The optimal allocation $\{q_t^p(k,k')\}_{k,k'\in\mathbb{K}}$ satisfies

$$q_t^p(k,k') = \arg\max_q \left\{V_t^p(k+q) + V_t^p(k'-q) - \chi\left(\varepsilon_t^p(k), q\right) - \chi\left(\varepsilon_t^p(k'), -q\right)\right\}$$

[7] We derive (25) in the Appendix C.9.

and the optimal search profile $\{\varepsilon_t^p(k)\}_{k\in\mathbb{K}}$ is

$$\{\varepsilon_t^p(k)\}_{k\in\mathbb{K}} \equiv \arg \max_{\{\varepsilon(k)\}_{k\in\mathbb{K}}\in[0,1]^{\mathbb{K}}} \int\int S_t^p\left[k,k',\varepsilon(k),\varepsilon\left(k'\right)\right] m\left[\varepsilon(k),\varepsilon\left(k'\right)\right] dF_t^p\left(k'\right) dF_t^p(k) \quad (26)$$

where

$$S_t^p\left(k,k',\varepsilon,\varepsilon'\right) \equiv \max_q \left\{ \begin{array}{c} V_t^p(k+q) - V_t^p(k) - \chi(\varepsilon,q) \\ +V_t^p(k'-q) - V_t^p(k') - \chi(\varepsilon',-q) \end{array} \right\}.$$

Note that the constrained-efficient search intensity $\varepsilon_t^p(k)$ of (26) also maximizes the joint surplus from reallocating balances. Like the equilibrium, the banks' search intensities are degenerate in the constrained optimal allocation and banks share the same path of optimal search intensity. However, observe that the equilibrium HJB (6) for $V_t(k)$ differs from the co-state HJB (25), because the bilateral gains from trade in the co-state HJB are *twice* those in the equilibrium HJB. As a result, the proposition below shows that banks *oversearch* in equilibrium relative to the constrained-efficient level. Therefore, the equilibrium allocation is generally not constrained-optimal: the welfare theorem is violated.

Proposition 5 ***(Inefficiency)*** *Equilibrium is not generically constrained optimal, and the equilibrium search intensity and balances traded exceeds the constrained-optimal level, i.e.,* $\varepsilon_t^p(k) \leq \varepsilon_t^*(k)$ *and* $|q_t^p(k,k')| \leq |q_t^*(k,k')|$. *It becomes constrained-optimal only if* $\chi = 0$.

Although the matching function implies complementarity in banks' search decisions, hence positive externality, banks do not undersearch in equilibrium. Instead, they are, in effect, over-trading along the extensive margin compared to the constrained optimum. This feature is different from the typical logic behind the Hosois condition. The reason is that, in the Federal funds market, banks rely on bilateral trades to reach their target level of balances. Because trading in an OTC market is opportunistic—owing to search frictions—banks tend to search more intensively to offset these frictions. If transaction costs are not sufficiently sensitive to increases in search intensity, banks may also trade more whenever an opportunity arises. In short, banks over-trade in equilibrium because of a precautionary motive amplified by search frictions.

Our results are novel. While Afonso & Lagos (2015b) show that the welfare theorem holds when banks are homogeneous (beyond initial balances), Proposition 5 shows it is no longer the case when there is search cost or transaction cost. Üslü (2019) shows that the welfare theorem does not hold when banks are ex-ante heterogeneous in payoff functions and exogenous counterparty-finding rates, because of the composition externality. Proposition 5 shows that even banks are ex-ante homogeneous, the welfare theorem still does not hold when counterparty-finding rates are endogenously chosen by banks via their search intensity.[8]

[8] Farboodi et al. (2017) obtains the similar finding for a different reason. The matching function in their model exhibits negative congestion externality, so agents oversearch in a steady state equilibrium. Our model has no congestion externality: matching function is increasing returns to scale, and the trading game is supermodular.

3.3 Testable Implications

This section provides the comparative statics of the closed-form solutions to policy and search cost parameters, which will serve as our testable implications for the later empirical investigation. Consider a regulation parameter κ_0 that measures how regulation makes trading balances costlier such that the trading cost is given by

$$\kappa(\varepsilon) = \kappa_0 + \tilde{\kappa}(\varepsilon).$$

Consider the discount window rate i^{DW} and the interest rate on excess reserve i^{ER}, where $i^{DW} > i^{ER}$ that enters the portfolio payoff of banks such that there exists k_+ and k_-, where $k_+ > k_-$, that the marginal terminal values of balances are given by

$$U'(k_- K) = 1 + i^{DW},\ U'(k_+ K) = 1 + i^{ER}.$$

In other words, banks will use the discount window when its balances fall to $k = k_- K$ so that the marginal terminal value of balances becomes $U'(k_- K) = 1 + i^{DW}$. It implies

$$A = \frac{i^{DW} - i^{ER}}{2K(k_+ - k_-)},\ \tilde{k} = \frac{\left(1 + i^{DW}\right) k_+ - \left(1 + i^{ER}\right) k_-}{i^{DW} - i^{ER}} K.$$

We are interested in how those parameters impact banks' search intensity, bilateral trade volume and net Federal funds borrowing. The following Proposition summarizes the comparative statics.

Proposition 6 *The comparative statics are*

	ε_t^*	$\lvert m_t^* q_t^* \rvert$	$L_0(k)$
$i^{DW} - i^{ER}$	$+$	$+$	$\text{sgn}(K - k)$
K	$-$	$-$	$+(-)$ for large (small) k
κ_0	$+$	$-$	$\text{sgn}(k - K)$

The above proposition characterizes the effects of policy variables ($i^{DW} - i^{ER}$, K and κ_0) on the search intensity ε_t, expected trade volumen flow $|m_t q_t|$ and net Federal funds purchase $L_0(k)$. The symbols $+$ and $-$ represent positive and negative effects, respectively. The function "sgn" is the sign function. The proposition has the following testable implications. First, the search intensity ε_t and expected flow of bilateral trade volume $|m_t q_t|$ increase in the policy spread $i^{DW} - i^{ER}$ and decrease in the aggregate excess reserves K. This is the disintermediation effect of policy spread and aggregate excess reserves: the bilateral gains of trade and intermediation decreases under a lower policy spread or a higher aggregate excess reserves, reducing banks' incentive to search and trade. Second, a higher transaction cost κ_0 raises the search intensity and lowers the expected flow of bilateral trade volume. This is the disintermediation effect of balance sheet cost: the bilateral transaction size q_t decreases in the transaction cost κ_0, thereby gives banks more incentive to search due to pre-cautionary savings.

However, the disintermediation effect on the intensive margin of κ_0 dominates, reducing the expected flow of bilateral trade volume. Finally, the disintermediation effects impact the net Federal funds borrowing of individual banks. A lower policy spread $i^{DW} - i^{ER}$, a higher aggregate excess reserves K, or a higher transaction cost κ_0 reduce the expected flow of bilateral trade $|m_t q_t|$, thereby impedes the reallocation of reserves from lender banks to borrower banks. This implies that the net borrowers will borrow less and the net lenders will lend less as well, giving rise to the comparative statics of $L_0(k)$.

3.4 Extensions and Robustness

Our closed-form model has focused on homogeneous banks except initial reserve balance so far. However, it allows for a set of extensions, in which we are still able to get closed-form solutions and conduct comparative statics analysis. In the appendix, we introduce four pieces of extensions separately to discuss the effects of other Federal funds market factors on the trade dynamics.

GSE and ON RRP. Our main extension is a heterogeneous-agent model, where we add peripheral traders, e.g. government-sponsored enterprises and other financial institutions without Fed Reserve accounts, to the existing group of banks. We assume the peripheral traders contact banks at a constant search intensity, and obtain closed-form solutions. Instead of conducting comparative statics, we estimate this extended model via simulated method of moments and evaluate the quantitative importance of the disintermediation effect of unconventional monetary policy. Section 5 describes the model setup and presents the quantitative analysis, while Appendix E provides the derivations for the closed-form solutions. ONRRP

FBO and balance-sheet cost. Our main extension is a heterogeneous-agent model, where we add peripheral traders, e.g. government-sponsored enterprises and other financial institutions without Fed Reserve accounts, to the existing group of banks. We assume the peripheral traders contact banks at a constant search intensity, and obtain closed-form solutions. Instead of conducting comparative statics, we estimate this extended model via simulated method of moments and evaluate the quantitative importance of the disintermediation effect of unconventional monetary policy. Section 5 describes the model setup and presents the quantitative analysis, while Appendix E provides the derivations for the closed-form solutions. ONRRP

Federal funds brokerage. Appendix F introduces Federal funds brokerage to the market to study how the unconventional monetary policies affect the size of brokerage. We assume the brokers compete for matchmaking services via free entry with non-zero entry cost. Thus the size of brokerage is endogenously determined. In particular, IOER has disintermediation effect on brokerage by lowering the equilibrium size of active brokers in the market.

Intraday payment shocks. Appendix G considers the effects of payment shocks on the market trade dynamics. We introduce both lumpy and continuous shocks to payment flows. In particular, we find that the payment shocks do not impact the equilibrium length of search and bilateral transaction size.

Counterparty risks. Appendix H discusses the effects of counterparty risk on the Federal funds trade. By counterparty risk, we assume both counterparties of a meeting could default on the trade independently with some constant probabilities. We find that the effects of higher counterparty risk are isomorphic to the effects of higher transaction costs or lower search intensity.

4 Empirical Evidence

In this section, we document the empirical relationship of Federal funds trades and the unconventional monetary policies. Our focus is to test the comparative statics predictions of Proposition 6 in the theoretical model. In particular, we test the following hypotheses using U.S. bank-level data described in Section 4.1.1:

Hypothesis 1 The number of intermediary banks and the individual bank's volume of Federal funds intermediation increase in the policy spread (the difference between Discount-window rate and IOER) and decrease in the aggregate excess reserves.

Based on the facts shown in Figure 4 and the prediction of Proposition 6, the first hypothesis tests the causal effect of the policy spread and the aggregate excess reserves on the intermediation trading in Federal funds market. We examine the impact on both extensive margin and intensive margin. In addition to testing the impact on intermediation trades, we also investigate whether the disintermediation effect of IOER and aggregate excess reserves affect the allocation of reserves between the reserve net lenders and net borrowers.

Hypothesis 2 A lower policy spread and a higher aggregate excess reserves reduce the net Federal funds purchased by net borrowers (banks that have net borrowing of Federal funds) and the net Federal funds sold by net lenders (banks that have net lending of Federal funds).

This hypothesis examines whether borrower banks are less able to find lenders if the intermediation trades decrease. The following sections describe the data and estimation results.

4.1 Data

4.1.1 Bank-level data

The bank-level financial data are collected from various sources. We use the quarterly Consolidated Report of Condition and Income for U.S. banks and branches (commonly known as "Call reports")

and Consolidated Fiancial Statemennts (Form FR Y-9C) for Bank Holding Companies (BHCs).[9] The call reports and Form FR Y-9C are quarterly filed with the Federal Reserve by all U.S. banks and branches, and form FR Y-9C is filed by all U.S. holding companies with total consolidated assets of $1 billion or more (prior to 2015, this threshold was $500 million. Since September 2018, this number changes to $3 billion). These files report the balance sheet data of US banks at the end of each quarter, including the Federal funds purchased (Fed funds borrowing), Federal funds sold (Fed funds lending) and other balance sheet characteristics. Given the Fed funds are mostly overnight, the volume of Fed funds trade reported in these files measures banks' Fed funds borrowing and lending on the last business day of each quarter. Our data covers the period going from 2003Q1 to 2018Q4.[10] We measure each variable at the consolidated top holder level. Aggregating the variables to the top holder level not only avoids double counting, but also eliminates the bilateral trades between subsidiaries of a bank holding company that are not implemented in the Fed funds market.

For each top holder in each quarter, we construct the following variables: (1) Net volume of Fed funds purchased normalized by total assets (f^{net}), i.e.

$$f^{net} = \frac{\text{Fed funds purchased} - \text{Fed funds sold}}{\text{total assets}}.$$

It measures a bank's net borrowing of Fed funds as a share of bank assets. (2) Volume of Fed funds reallocation normalized by total assets (f^{int}), i.e.

$$f^{int} = \frac{\text{Fed funds purchased} + \text{Fed funds sold}}{\text{total assets}} - \left|f^{net}\right|.$$

This variable follows the definition of Fed funds reallocation in Afonso & Lagos (2015b), which is equal to the Fed funds trade in excess of the net borrowing. (3) Excess reserve balances before Federal funds trade normalized by total assets before Fed funds trade, i.e.

$$k = \frac{\text{excess reserve balances before Federal funds trade}}{\text{total assets}}.$$

The excess reserve balances before Federal funds trade represent a bank's holdings of Federal reserves balances in excess of its reserve requirement when it enters the Federal funds market. It captures individual heterogeneity of trade incentives in the Fed funds market. It is equal to a bank's excess reserve balances recorded in the bank balance sheets minus the net Federal funds purchased (Federal federal funds purchased minus Federal funds sold). Moreover, for individual controls, we include the following balance-sheet variables: (1) logged value of total assets; (2) total loans normalized by total assets; (3) total nonperforming loans normalized by total assets; (4) total high-quality liquid assets

[9] Appendix A describes the detailed data source and construction process.
[10] We also use the data in 2002Q4 as the lagged values of variables in 2003Q1.

normalized by total assets; (5) total equity normalized by total assets; (6) tier-1 leverage ratio; (7) ROA; (8) dummies of top holders' entity types.

4.1.2 Aggregate-level data

We use two sets of aggregate variables. The first set includes interest rate on reserves (i^{ER}), primary credit rate(i^{DW}), quarterly real GDP growth rate, quarterly unemployment rate, standard deviation of the Fed's general treasury account in a quarter. All these variables are measured at the end of a quarter. The interest rate on excess reserves and primary credit rate are the main regressors of monetary policy. They represent the outside return of holding reserves by lender banks and borrower banks at the end of a trading session, respectively. The other variables are the aggregate controls in regressions.

The second set of aggregate-level variables are obtained from bank-level data. For the cross section of top holders in each quarter, we construct the moments of excess reserve distribution: (1) aggregate excess reserves normalized by aggregate bank assets (K); (2) standard deviation of excess reserve balances normalized by the mean;[11] (3) skewness of excess reserve distribution. The aggregate excess reserves K is the third main regressor of monetary policy. It captures the effect of the Fed's total reserve balances on Fed funds trade. Meanwhile, we control the standard deviation and skewness to capture the effect of reserve distribution.

4.2 Effects on intermediation trade

Our first specification explores the impact of IOER and aggregate reserves on banks' intermediation trading. Note that in the data sample, only a fraction of banks are intermediaries, and the measure of individual bank's intermediation, $ffreallo_assets$, is non-negative. Thus we study how IOER and aggregate reserves impact both the probability of intermediation trades (extensive margin) and the volume of intermediation (intensive margin). In particular, we run probit and tobit regressions on the following specification on the sample of banks that hold positive total reserves at the Fed account and intermediate Federal funds at least once in the data sample:

$$\begin{aligned} y_{i,t} = a_0 + \beta_0 k_{it} + \beta_1 \left(i_t^{DW} - i_t^{ER}\right) + \beta_2 \left(i_t^{DW} - i_t^{ER}\right) \times k_{i,t} + \beta_3 K_t + \beta_4 K_t \times k_{i,t} \\ + \gamma \cdot controls_{i,t} + \varepsilon_{i,t}, \end{aligned} \tag{27}$$

where $y_{i,t} = \mathbf{1}\left\{f_{i,t}^{int} > 0\right\}$ in probit regression, and $y_{i,t} = f_{i,t}^{int}$ in tobit regressions. The variable $i_t^{DW} - i_t^{ER}$ is the policy spread between the primary credit rate and IOER. The term $controls_{i,t}$ includes individual-level characteristics and aggregate-level variables. The key coefficients are β_1 and β_3. According to Hypothesis 4, we expect a positive β_1 and a negative β_3. Moreover, by adding the

[11] Using standard deviation of excess reserves normalized by average assets produces similar results.

interaction between the policy variables and individual excess reserve balances, we also investigate the potential heterogeneous effects of the unconventional monetary policies across banks.

The probit and tobit estimation assumes exogeneity of the regressors. However, the Fed funds trade volume could depend on unobserved factors that correlate with the main regressors. For example, a bank's Federal funds trade volume and excess reserve balances could be driven by some common unobserved factors such as sophistication of balance sheet management. Moreover, a bank's incentive to trade Federal funds could be driven by some unobserved aggregate shocks that are correlated with the changes in IOER, primary credit rate and aggregate excess reserves. Thus we augment the estimation with instrumental-variable probit and tobit regressions to examine the potential endogeneity of excess reserves, aggregate policies and Federal funds trades. First, the instruments for the policy spread are the cumulative monetary policy shocks (policy news shocks) over past 4 quarters, which are obtained from Nakamura & Steinsson (2018).[12] Second, the instrument for the aggregate excess reserves is the one-period lag of 4-quarter change in aggregate excess reserves to aggregate bank assets ratio. Third, the instrument for individual excess reserves is one-period lag of individual excess reserves. For the instruments of interaction terms, we use the interactions between the corresponding instruments mentioned above.

The results of probit regressions are shown in Table 3, where we report three groups of estmation: column (1) and (2) reports the standard Probit estimation, Column (3) and (4) report the estimation of a random-effect panel probit model, and column (5) and (6) report the estimation of the instrumental-variable probit model. In all columns, the probability of intermediation trade increase in the policy spread and decrease in the aggregate excess reserves. Except the coefficient of the policy spread in column (5), all the other coefficients of policy spread and aggregate excess reserves are significant and robust. This implies that the unconventional monetary policies have strong disintermediation effect on the extensive margin. Moreover, by adding the interaction terms, we find that the impact of policy spread and aggregate excess reserves on the probability of intermediation trade can be heterogeneous across banks, but the signs of the coefficients for the interaction terms are not consistent and robust across the columns.

The results of tobit regressions are reported in Table 4, where we also have three groups of estimation. The main results of tobit regressions are similar to those of probit regressions. On average, under a lower value of policy spread or a higher value of aggregate excess reserves, banks are less likely to do intermediation trades. The signs of coefficients are of expectation, and the values of the coefficients are significant and robust across columns. In summary, the estimation results of probit and tobit regressions imply significantly and consistently negative effect of unconventional monetary policies on intermediation trade, which reveals a strong disintermediation channel.

[12] The original sample period of the policy shocks end in 2014, and Acosta & Saia (2020) update the shocks to 2019. We use the later in our estimation.

4.3 Effects on Net Borrowing of Fed Funds

Our second specification relates the net Fed funds borrowing to a bank's excess reserve balances, IOER and aggregate reserve balances. We estimate the following equation on the sample of banks that hold positive total reserves at the Fed account and trade Federal funds at least once in the data sample:

$$f_{i,t}^{net} = \alpha_i + \beta_0 k_{i,t} + \beta_1 \left(i_t^{DW} - i_t^{ER}\right) + \beta_2 \left(i_t^{DW} - i_t^{ER}\right) \times k_{i,t} + \beta_3 K_t + \beta_4 K_t \times k_{i,t} + \gamma \cdot controls_{i,t} + \varepsilon_{i,t}, \tag{28}$$

where i represents a bank and t denotes the last business day of a quarter. The parameters α_i represent the bank fixed effects. The control variables $controls_{i,t}$ include both the bank-level controls and the aggregate controls mentioned above. This regression examines how the level of policy spread and aggregate excess reserves impact individual banks' net Fed funds borrowing. The interaction terms allow us to investigate the heterogeneous effects of the monetary policies across banks. According to Hypothesis 4, we expect a negative β_2 and a positive β_4. Note that the dependent variable is the net borrowing of Federal funds. Thus a negative β_2 means a higher policy spread increases the net borrowing of a net borrower (small k) and the net lending of a net lender (large k). A positive β_4 means a higher aggregate excess reserves reduces the net borrowing of a net borrower and the net lending of a net lender.

Colums (1) to (3) of Table 5 report the results of OLS estimation. Column (1) does not include the interaction terms, thus estimates the average effect of the monetary policies on banks' net Fed funds borrowing. Column (2) reports the estimation of our baseline specification (28), while Column (3) additionally controls the time fixed effects. We have the following findings. First, the coefficient of individual excess reserves, β_0, is significantly negative across all the columns. It implies that banks with more excess reserves borrow less Federal funds. Second, the OLS estimation shows significant and robust heterogeneous effects of monetary policies on net Fed funds borrowing. In particular, the coefficient of the interaction between policy spread and individual excess reserves, β_2, is significantly negative. The coefficient of the interaction between the aggregate excess reserves and individual excess reserves, β_4, is significantly positive. Therefore, the OLS estimation results are consistent with our theoretical predictions.

Column (4) to (6) of Table 5 report the results of 2SLS estimation, where the specification of each column corresponds to Column (1) to (3). The results are consistent with the OLS estimation. In Column (4), we find that banks net Fed funds borrowing increases in policy spread and decreases in aggregate excess reserves on average. In Column (5) and (6), the coefficients of all interaction terms are significant and consistent with the OLS estimation. Thus our estimation documents robust

positive effect of policy spread and negative effect of aggregate excess reserves on Federal funds reallocation. This verifies our second hypothesis.

5 Quantitative Analysis

This section provides a quantitative evaluation for the effects of unconventional monetary policy on disintermediation. The evaluation is based on an extended model that captures the main institutional features of the Federal funds market. The setup is as follows. There are two groups of agents: a unit continuum of banks as in the baseline model, and a continuum of peripheral traders that have no Federal reserve accounts. The peripheral traders represent government-sponsored enterprises and other financial institutions that participate in the Federal funds market but have no access to IOER. The mass of peripheral traders is ϑ. We assume a peripheral trader only contacts banks at a constant arrival rate φ. Moreover, the banks choose search intensity ε in the contact with other banks, at an arrival rate $m(\varepsilon, \varepsilon')$. The bargaining power of banks in the meeting with peripheral traders is $\theta \in (0,1)$. Each peripheral trader is endowed with some reserve balances $\tilde{k}$, and we denote the distribution of peripheral traders' reserve balances as $\tilde{F}_t\left(\tilde{k}\right)$, with $\tilde{F}_0\left(\tilde{k}\right)$ given.[13] We assume the peripheral traders have no flow payoff of reserve holdings, but only enjoy the end-of-period payoff from the overnight reverse repurchase facility (ON RRP), i.e. $\tilde{U}\left(\tilde{k}\right) = \left(1 + i^{RRP}\right)\tilde{k}$.

For quantitative motivation, we assume the transaction cost of a bank in a meeting is $\chi_t(\varepsilon, q) = (\kappa_{0,t} + \kappa_1 \varepsilon) q^2$, where $\kappa_{0,t}$ is a time-varying parameter that captures the regulatory changes on bank balance sheet. The peripheral traders are not subject to balance sheet regulations, thus their transaction cost is assumed to be 0. Since banks do not choose search intensity in contacting peripheral traders, their transaction costs in such contacts is $\kappa_{0,t} q^2$. This extended model has closed-form solutions and Appendix E presents the derivations. In particular, we find the banks' value functions are still quadratic and the peripheral traders' value functions are linear in their reserve balances.

To capture the change in the regulatory requirement on bank balance sheet and the opportunity cost of liquidity, we allow for time-varying transaction cost and liquidity benefits. Specifically, we assume κ_0 and γ change over years in the following form:

$$\begin{aligned} \kappa_{0,yr} &= \kappa_{0,2006} \times \exp\left[g_{\kappa_0}\left(yr - 2006\right)\right], \\ \gamma_{yr} &= \gamma_{2006} \times \exp\left[g_{\gamma}\left(yr - 2006\right)\right], \end{aligned} \tag{29}$$

where yr denotes a year and takes values from 2006 to 2018. In our estimation, we set 2006 as the first year and 2018 as the last year of the sample. Therefore, instead of estimating g_{κ_0} and g_γ, we estimate $\kappa_{0,2018}$ and γ_{2018}.

[13] As is shown in Appendix E, the distribution $\tilde{F}_t\left(\tilde{k}\right)$ is redundant in equilibrium.

5.1 Estimation

Instead of calibrating the deterministic theoretical model, we conduct a simulated method of moments estimation on a discretized version of the model to pin down the parameters. In the discretized version, we assume the reserve distribution is atomic (so there is a finite number of banks) and given by the empirical distribution of reserve balances in the data. The outcome of the discretized model is random since each bank faces idiosyncratic random meetings. We estimate the model parameters via simulated method of moments. The Appendix I describes the algorithm of simulation and estimation.

In the current version of estimation, we first normalize $r = a = 0$, and set $T = 2.5/24$ to represent the 2.5 hr trading session of the daily Federal funds market. Second, we normalize the size of peripheral traders $\vartheta = 1$ since it cannot be identified separately from the contact rate φ. Third, the individual excess reserves are the quarterly bank-level data (Call reports and Form FR Y9-C) of individual excess reserves before Federal funds trade divided by bank assets. The data of IOER, primary credit rate and ON RRP are obtained from FRED. We conduct the simulated method of moments based on the data over 2006Q1-2018Q4 to estimate the following parameters

$$\{\lambda, \lambda_0, k_+, k_-, \gamma_{2006}, \gamma_{2018}, \theta, \varphi, \kappa_1, \kappa_{0,2006}, \kappa_{0,2018}\}.$$

The moments for estimation are (1) the regression coefficients of $i^{ER} \times k$ and $K \times k$ in the Federal funds net purchase regressions 28; (2) the banks' aggregate share of intermediation volume in 2006 and 2018; (3) the aggregate Fed funds sold by intermediaries normalized by aggregate bank assets in 2006 and 2018; (4) the aggregate Fed funds purchased by intermediaries normalized by aggregate bank assets in 2006 and 2018; (5) the aggregate fraction of trading banks in 2006 and 2018; (6) the average effective Fed funds rates in 2006 and 2018. The parameter estimation results are listed in Table 6. The simulated moments are listed in Table 7 and 8.

We find that the estimated transaction cost κ_0 increases from 2006 to 2008, while the liquidity benefit γ decreases in the same period. This implies the rise of bank balance sheet cost due to stronger regulations, and the declined liquidity benefit due to the increasing aggregate excess reserves. The moments produced by our estimation are close to the targets. In particular, the simulated regression coefficients have the correct signs and similar magnitudes, and the fraction of trading banks and effective Federal funds rates are almost exactly calibrated.

5.2 Counterfactual Analysis

Given the estimation we conduct counterfactual analysis to evaluate the quantitative importance of unconventional monetary policies and regulations to the disintermediation channel. In particular, we consider the following exercises and examine how the level of intermediation in 2018 changes: (1) Change the paths of IOER, primary credit rate and ON RRP in 2018 to the paths in 2006. This

exercise investigates how the level of intermediation changes in 2018 if the Federal Reserve recovers the policy spread in 2006. (2) Proportionally change individual banks' reserve balances in 2018, such that the average individual reserve balances are equal to the levels in 2006. This exercise examines the effect of aggregate excess reserves on disintermediation. (3) Change $\kappa_{0,2018}$ to $\kappa_{0,2006}$. This exercise evaluates the impact of rising transaction cost on disintermediation.

Table 9 reports the results of counterfactuals. We find that recovering the policy spread in 2006 doubles the intermediation volume share in 2018, while reducing the transaction cost can increase the level of intermediation by about 4 times. However, the effect of aggregate excess reserves on disintermediation is small, since the intermediation share almost doesn't change in the counterfactual analysis.

6 Conclusion

This paper proposes a new channel of monetary policy and regulation on the monetary policy implementation, the disintermediation channel. When the policy spread between discount-window rate and interest rate on excess reserves decreases or the balance sheet cost rises, the intermediation trades by banks decline in the Federal funds market. We rationalize this channel in a continuous-time search-and-bargaining model of divisible funds and endogenous search intensity, which nests the matching model of Afonso & Lagos (2015b) and the transaction model of Hamilton (1996). A lower policy spread decreases the spread of marginal value of reserves, and balance sheet cost increases the marginal cost of holding reserves, both of which lower the gains of intermediation. We find that the equilibrium is constrained inefficient as banks trade too frequently. The disintermediation channel is both empirically and quantitatively important. Empirically, it significantly impede the reallocation of reserves from lender banks to borrower banks. Quantitatively, eliminating IOER and reducing the balance sheet cost can greatly raise the level of intermediation during the period after the Great Recession. For further research, we will focus on the investigating how the disintermediation channel impacts the effects of current monetary policy framework on the Federal funds rate and real economy, as well as characterizing the optimal monetary policy and regulation via quantitative analysis.

References

Acosta, M. & Saia, J. (2020). Estimating the effects of monetary policy via high frequency factors. Columbia University working paper.

Adda, J. & Cooper, R. (2003). *Dynamic economics: quantitative methods and applications*. MIT press.

Afonso, G., Armenter, R., & Lester, B. (2019). A model of the federal funds market: yesterday, today, and tomorrow. *Review of Economic Dynamics*, 33, 177–204.

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

Afonso, G. & Lagos, R. (2015a). The over-the-counter theory of the fed funds market: A primer. *Journal of Money, Credit and Banking*, 47(S2), 127–154.

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

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

Bech, M. & Keister, T. (2017). Liquidity regulation and the implementation of monetary policy. *Journal of Monetary Economics*, 92, 64–77.

Bech, M. & Monnet, C. (2016). A search-based model of the interbank money market and monetary policy implementation. *Journal of Economic Theory*, 164, 32–67.

Bech, M. L. & Atalay, E. (2010). The topology of the federal funds market. *Physica A: Statistical Mechanics and its Applications*, 389(22), 5223–5246.

Benhabib, J. & Farmer, R. E. (1994). Indeterminacy and increasing returns. *Journal of Economic Theory*, 63(1), 19–41.

Berentsen, A. & Monnet, C. (2008). Monetary policy in a channel system. *Journal of Monetary Economics*, 55(6), 1067–1080.

Bianchi, J. & Bigio, S. (2022). Banks, liquidity management, and monetary policy. *Econometrica*, 90(1), 391–454.

Bigio, S. & Sannikov, Y. (2021). *A Model of Credit, Money, Interest, and Prices*. Working Paper 28540, National Bureau of Economic Research.

Bracewell, R. N. (2000). *The Fourier transform and its applications*. McGraw-Hill New York.

Chang, B. & Zhang, S. (2018). Endogenous market making and network formation. Available at SSRN 2600242.

Chiu, J., Eisenschmidt, J., & Monnet, C. (2020). Relationships in the interbank market. *Review of Economic Dynamics*, 35, 170–191.

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

Duffie, D. & Krishnamurthy, A. (2016). Passthrough efficiency in the fed's new monetary policy setting. In *Designing Resilient Monetary Policy Frameworks for the Future. Federal Reserve Bank of Kansas City, Jackson Hole Symposium* (pp. 1815–1847).

Ennis, H. M. (2018). A simple general equilibrium model of large excess reserves. *Journal of Monetary Economics*, 98, 50–65.

Farboodi, M., Jarosch, G., & Shimer, R. (2017). *The emergence of market structure*. Working Paper 23234, National Bureau of Economic Research.

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

Hamilton, J. D. (1996). The daily market for federal funds. *Journal of Political Economy*, 104(1), 26–56.

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

Ihrig, J. E., Vojtech, C. M., & Weinbach, G. C. (2019). How have banks been managing the composition of high-quality liquid assets? *Review*, 101(3), 177–201.

Kashyap, A. K. & Stein, J. C. (2012). The optimal conduct of monetary policy with interest on reserves. *American Economic Journal: Macroeconomics*, 4(1), 266–82.

Keating, T. & Macchiavelli, M. (2017). Interest on reserves and arbitrage in post-crisis money markets. FEDS Working Paper No. 2017-124.

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

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

Lagos, R. & Zhang, S. (2019). A monetary model of bilateral over-the-counter markets. *Review of Economic Dynamics*, 33, 205–227.

Liu, S. (2020). Dealers' search intensity in us corporate bond markets. Available at SSRN 3644132.

Milgrom, P. & Shannon, C. (1994). Monotone comparative statics. *Econometrica*, 62(1), 157–180.

Nakamura, E. & Steinsson, J. (2018). High-frequency identification of monetary non-neutrality: the information effect. *The Quarterly Journal of Economics*, 133(3), 1283–1330.

Poole, W. (1968). Commercial bank reserve management in a stochastic model: implications for monetary policy. *The Journal of finance*, 23(5), 769–791.

Sigman, K. (2007). Poisson processes, and compound (batch) poisson processes. Lecture notes, Columbia University, http://www.columbia.edu/ ks20/4703-Sigman/4703-07-Notes-PP-NSPP.pdf.

Trejos, A. & Wright, R. (2016). Search-based models of money and finance: An integrated approach. *Journal of Economic Theory*, 164, 10–31.

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

van Imhoff, E. (1982). *Optimal economic growth and non-stable population*. Springer-Verlag, Berlin, Germany.

Williamson, S. D. (2019). Interest on reserves, interbank lending, and monetary policy. *Journal of Monetary Economics*, 101, 14–30.

Appendices

A Details of Data and Measurement

In this section, we describe how we collect the data and construct various measurement we used for the summary statistics and estimation.

A.1 Sources

Financial data of the Federal funds market participants come from the following:

- **Call Reports**. This is the source of the subsidiary-level data. In particular, we use form FFIEC 031 for banks with both domestic and foreign offices, form FFIEC 041 for banks with domestic offices only, and form FFIEC 002 for U.S. branches and agencies of foreign banks (FBO). These forms are available for download at the Federal Financial Institutions Examination Council (FFIEC) and the Federal Reserve Bank of Chicago.[14]

- **FR Y-9C**. This is the source of the consolidated data at the level of holding companies (for bank holding companies, savings and loan holding companies, and intermediate holding companies) with total consolidated assets of $1 billion or more (prior to 2015, this threshold was just $500 million). This is available for download at the Federal Reserve Bank of Chicago and National Information Center (NIC).[15]

- **Attributes, relationships, and transformations tables**. This is the source of the ownership structure of holding companies upon their subsidiaries. They are available for download at National Information Center (NIC).[16]

- **10Q and 10K**. This is the source of government sponsored enterprises (GSE) data. These forms are available for download at the Security Examination Commission (SEC).[17] The GSE data is fully available since 2006Q1.

- **H.4.1**. This is the source of the balance sheet of the Federal Reserve System and factors affecting reserve balances of depository institutions. This is available for download at the Board of Governors of the Federal Reserve System.[18]

[14] https://cdr.ffiec.gov/public/ and https://www.chicagofed.org/banking/financial-institution-reports/commercial-bank-structure-data

[15] https://www.chicagofed.org/banking/financial-institution-reports/bhc-data and https://www.ffiec.gov/npw/FinancialReport/FinancialDataDownload?selectedyear=2021

[16] https://www.ffiec.gov/npw/FinancialReport/DataDownload

[17] https://www.sec.gov/edgar/searchedgar/companysearch.html

[18] https://www.federalreserve.gov/releases/h41/

- **Time series of the economy**. It is available for download at the Federal Reserve Bank of St Louis (FRED).[19]

A.2 Consolidated sample

Whenever possible, we always measure variables at the holding-company level. We think that holding companies are desirable sample unit because first, usually the subsidiaries' reserves, which are not directly observable in the Call reports, are corresponded by their holding company's master accounts in the Federal Reserve Banks, which are observable. Second, sometimes the decision of Federal Funds trading is delegated to the holding company. Third, it avoids double-counting the intra-holding-company Federal Funds trades, which are different from those normal interbank transactions.

Consolidation is done by referring to items filed in FR Y-9C. For the holding companies not eligible to file FR Y-9C, or items not available from FR Y-9C, we directly consolidate the Call report items from the subsidiary level up to the topmost holding-company level, based on the relationships table from NIC. In this appendix, we always refer i as the index for holding companies and j as the index for i's subsidiaries. We focus on banks that have positive amounts of asset and total reserve balances, and trade at least once in the Federal funds market in the data sample.

A.3 Excess Reserves

The formula to measure excess reserves bank i holds at the Federal Reserve account at the end of quarter t is given by

$$\textit{Excess Reserves}_{it} = \textit{Total Reserves}_{it} - \left\{\sum\nolimits_j \textit{Required Reserves}_{jt} - \textit{Vault Cash}_{it}\right\}_+ .$$

$\textit{Total Reserves}_{it}$ is measured by item RCFD0090 in FR Y-9C ("Balances due from Federal Reserve Banks"). $\textit{Vault Cash}_{it}$ is approximated by item RCON0080 in FR Y-9C ("Currency and coin"). The formula to calculate $\textit{Required Reserves}_{jt}$ is based on subsidiary j's net transaction accounts. For example, the formula of reserve requirement in 2010 is given by the following table:

Table 1: Reserve requirement in 2010

Net transaction accounts	% required
\$0 to \$10.7 million	0
More than \$0.7 million to \$55.2 million	3
More than \$55.2 million	10

[19] https://fred.stlouisfed.org/

The table is updated every year.[20] To estimate net transaction acccounts, we substract item RCON 2215 of j's Call Report ("Total Transaction Accounts") from the sum of item RCFD 0083 ("Balances due from depository institutions in the U.S.: U.S. branches and agencies of foreign banks (including their IBFs)"), item RCFD 0085 ("Balances due from depository institutions in the U.S.: Other depository institutions in the U.S. (including their IBFs)") and item RCON 0020 ("Cash items in process of collection and unposted debit"). Then we apply the historical reserve requirement formulas on net trans accounts to calculate $Required\ Reserves_{jt}$.

To measure the excess reserves bank i holds before entering the Federal funds market, we subtract the net Federal funds purchase from $Excess\ Reserves_{it}$. Thus the pre-trade excess reserves is given by

$$\begin{aligned} Excess\ Reserves\ pre\text{-}trade_{it} &= Excess\ Reserves_{\ it} - Federal\ funds\ purchased_{it} \\ &\quad + Federal\ funds\ sold_{it}. \end{aligned} \tag{A.1}$$

By dividing the pre-trade excess reserves by bank assets, we obtain the measure $exres_assets$ in the regressions.

A.4 Federal Funds Trades and Intermediation

We compute the net Federal funds borrowed by substracting item BHDM B993 in FR Y-9C ("Federal funds purchased in domestic offices") from item BHDM B987 ("Federal funds sold in domestic offices"). We measure bank's intermediation by $Reallocated\ Funds_{it}$:

$$\begin{aligned} Reallocated\ Funds_{it} &= Federal\ funds\ purchased_{it} + Federal\ funds\ sold_{it} \\ &\quad - \left|Federal\ funds\ purchased_{it} - Federal\ funds\ sold_{it}\right|. \end{aligned} \tag{A.2}$$

By dividing the net Federal funds borrowed and *Reallocated Funds* by bank assets respectively, we obtain the measure $ffnet_assets$ and $ffreallo_assets$ in the regressions.

A.5 Bank-level Controls

We use the following items from Call report to measure various attributes of banks.

- Size and scope
  - logarithm of assets (item RCFD 2170 "Total assets").
  - bank equity (item RCFD 3210 "Total bank equity capital") over bank assets.
- Marginal benefit of liquidity

[20] The historical reserve requirement can be found on https://www.federalreserve.gov/monetarypolicy/reservereq.htm

  - ROA
  - High-quality liquid assets (HQLA) over total assets (Ihrig et al., 2019)

- Risk
  - ratio of non-performing loan (sum of items 1 through 8.b of Column B and C in Schedule RC-N) over bank assets, as in Afonso et al. (2011)
  - ratio of loan (item RCFD 2122 "Total loans and leases held for investment and held for sale") over bank assets

- Regulation
  - Tier-1 leverage ratio (item RCFA 7204 "Tier 1 leverage ratio")

- Other indicators
  - bank entity type (in the NIC attributes table)
  - Fed District dummy (in the NIC attributes table)

A.6 Economy-wide Controls

- quarterly real GDP growth rate (available from FRED)
- quarterly unemployment rate (available from FRED)
- standard deviation of the Fed's general treasury account in a quarter (available from H.4.1)

B Tables

B.1 Summary statistics

Table 2: Summary statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
Net Fed funds purchase/Assets	92,785	-0.0068	0.0347	-0.8742	0.6852
Ex. res. pre-trade/Assets	92,785	0.0348	0.0618	-0.6622	4.1827
log (Assets)	92,785	13.7081	1.2849	4.6728	21.6874
Dummy: reallocation	92,785	0.0932	0.2907	0	1
Fed funds reallocation/Assets	92,785	0.0006	0.0033	0	0.0259
IOER (%)	64	0.3602	0.5470	0	2.4
Primary credit rate (%)	64	2.0781	0.1804	0.5	6.25
Agg. ex. res./Agg. assets	64	0.0428	0.0410	-0.0084	0.1070

Notes: This table presents the summary statistics of key variables. The observations for the first 5 variables are bank-quarter. "Net Fed funds purchase/Assets" is a bank's net Federal funds purchase divided by bank assets. "Ex. res. pre-trade/Assets" is a bank's excess reserve balances before Federal funds trade divided by bank assets. "log(Assets)" is the log value of bank assets. "Dummy: reallocation" is equal to 1 if a bank intermediates Federal funds on a day, and equal to 0 otherwise. "Fed funds reallocation/Assets" is a bank's volume of Federal funds reallocation divided by bank assets. "Agg. ex. res/Agg. assets" is the aggregate excess reserve balances before Federal funds trade divided by the aggregate bank assets. The sample consists of U.S. banks that hold positive total reserves at the Fed account and trade Federal funds at least once in the data sample. The sample period is from 2003Q1 to 2018Q4.

B.2 Regression results

Table 3: Probit on Reallocation

Dep. Var.	Dummy: Intermediation or not					
	Probit (Pooled)		Panel Probit (RE)		IV Probit	
	(1)	(2)	(3)	(4)	(5)	(6)
$i^{DW} - i^{ER}$	0.0256***	0.0229***	0.0524***	0.0452***	0.0445	0.0483*
	(0.0067)	(0.0072)	(0.0124)	(0.0126)	(0.0275)	(0.0282)
$\left(i^{DW} - i^{ER}\right) \times k$		-0.3939		-0.8649**		0.6663
		(0.2678)		(0.3784)		(0.9171)
K	-7.4060***	-7.2581***	-13.1188***	-13.8352***	-6.3299***	-5.7380**
	(0.7000)	(0.7508)	(1.2348)	(1.3013)	(2.3488)	(2.4261)
$K \times k$		23.0237		-64.0178**		76.4506
		(24.2872)		(27.0229)		(69.9103)
k	-4.5381***	-4.9683***	-11.0011***	-6.8335***	-4.5948***	-9.3636*
	(1.1315)	(1.2377)	(1.0379)	(1.6515)	(0.1754)	(5.0332)
Bank controls	Y	Y	Y	Y	Y	Y
Agg. controls	Y	Y	Y	Y	Y	Y
Specification tests						
Wald test of exogeneity						
χ^2 stat					1.27	4.51
p-value					[0.5312]	[0.3419]
Weak instrument test						
χ^2 stat					224.51	220.82
p-value					[0.0000]	[0.0000]
Pseudo R^2	0.1822	0.1840				
Wald χ^2			2412.84	2408.41	6789.89	6637.70
Number of observations	81,255	81,255	81,255	81,255	77,506	77,506
Number of banks	3,022	3,022	3,022	3,022	3,000	3,000

Notes: This table presents the estimation results on the Probit regression of Federal funds intermediation (27). The sample consists of U.S. banks that hold positive total reserves at the Fed account and intermediate Federal funds at least once in the data sample. The sample period is from 2003Q1 to 2018Q4. Standard errors clustered by banks are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 4: Tobit on Reallocation

Dep. Var.	Intermediation/Assets					
	Tobit (Pooled)		Panel Tobit (RE)		IV Tobit	
	(1)	(2)	(3)	(4)	(5)	(6)
$i^{DW} - i^{ER}$	0.0004***	0.0003***	0.0003***	0.0003***	0.0007*	0.0007*
	(0.0001)	(0.001)	(0.0001)	(0.0001)	(0.0004)	(0.0004)
$\left(i^{DW} - i^{ER}\right) \times k$		-0.0027		0.0013		0.0105
		(0.0037)		(0.0010)		(0.0121)
K	-0.1144***	-0.1107***	-0.1119***	-0.1132***	-0.0965***	-0.0861**
	(0.0107)	(0.0112)	(0.0056)	(0.0056)	(0.0341)	(0.0346)
$K \times k$		0.4710		-0.1863***		1.3220
		(0.3575)		(0.0640)		(0.9272)
k	-0.0586***	-0.0706***	-0.0343***	-0.0303***	-0.0597***	-0.1357**
	(0.0151)	(0.0165)	(0.0020)	(0.0047)	(0.0023)	(0.0667)
Bank controls	Y	Y	Y	Y	Y	Y
Agg. controls	Y	Y	Y	Y	Y	Y
Specification tests						
Wald test of exogeneity						
χ^2 stat					0.89	2.59
p-value					[0.6422]	[0.6280]
Weak instrument test						
χ^2 stat					247.15	248.02
p-value					[0.0000]	[0.0000]
Pseudo R^2	-1.0076	-1.0249				
Wald χ^2			3029.94	3028.46	4116.20	3902.30
Number of observations	81,255	81,255	81,255	81,255	77,506	77,506
Number of banks	3,022	3,022	3,022	3,022	3,000	3,000

Notes: This table presents the estimation results on the Tobit regression of Federal funds intermediation (27). The sample consists of U.S. banks that hold positive total reserves at the Fed account and intermediate Federal funds at least once in the data sample. The sample period is from 2003Q1 to 2018Q4. Standard errors clustered by banks are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 5: Effects of IOER and aggregate excess reserves on net Federal funds purchased

Dep. Var.	Net Fed Funds Purchased/Assets					
	OLS			2SLS		
	(1)	(2)	(3)	(4)	(5)	(6)
$i^{DW} - i^{ER}$	-0.0007***	-0.0008***		0.0005*	0.0005**	
	(0.0001)	(0.0001)		(0.0003)	(0.0002)	
$\left(i^{DW} - i^{ER}\right) \times k$		-0.0893***	-0.0893***		-0.0615**	-0.0623**
		(0.0102)	(0.0103)		(0.0310)	(0.0309)
K	0.0747***	0.0784***		-0.0815***	-0.0637***	
	(0.0118)	(0.0090)		(0.0212)	(0.0185)	
$K \times k$		2.5696***	2.5568***		4.3948**	4.2226**
		(0.7597)	(0.7579)		(1.7435)	(1.7294)
k	-0.3127***	-0.3719***	-0.3725***	-0.3145***	-0.4685***	-0.4682***
	(0.0320)	(0.0494)	(0.0495)	(0.0324)	(0.1065)	(0.1062)
Bank FE	Y	Y	Y	Y	Y	Y
Quarter FE	N	N	Y	N	N	Y
Bank controls	Y	Y	Y	Y	Y	Y
Agg. controls	Y	Y	N	Y	Y	N
Specification tests						
Underidentification test						
χ^2 stat				1800.8	28.92	29.05
p-value				[0.0000]	[0.0000]	[0.0000]
Weak Instrument test						
F stat				5366.2	5.435	10.90
Relative OLS bias>10% (p-value)				[0.0000]	[0.0873]	[0.0000]
Relative OLS bias>30% (p-value)				[0.0000]	[0.0060]	[0.0000]
Adj. R^2	0.664	0.786	0.789	0.286	0.535	0.546
Number of observations	81,240	81,240	81,240	77,486	77,486	77,486
Number of banks	3,022	3,022	3,022	3,000	3,000	3,000

Notes: This table presents the estimation results on the net Federal funds purchased regression (28). The sample consists of U.S. banks that hold positive total reserves at the Fed account and trade Federal funds at least once in the data sample. The sample period is from 2003Q1 to 2018Q4. Standard errors clustered by banks are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

B.3 Tables in Quantitative Analysis

Table 6: Parameter estimation

Parameter	λ	λ_0/λ_0	k_+	k_-	θ	ρ
Estimated Value	20.1987	0.5605	2.9480	−0.0596	0.7005	0.2000
Standard deviation	0.0007	3.6×10^{-5}	0.0038	0.0048	0.0043	1.3×10^{-5}
Parameter	κ_1	$\kappa_{0,2006}$	$\kappa_{0,2018}$	γ_{2006}	γ_{2018}	
Estimated Value	0.00568	0.00001	0.000705	0.00035	0.00028	
Standard deviation	0.0054	0.0038	0.0024	0.0062	0.0019	

Notes: This table lists the estimated values and standard deviations of the model parameters from simulated method of moments.

Table 7: Simulated regression coefficients

Moments	Target	Simulation	95% CI
Coef of ind. ex. res.	-0.468	-0.199	[-0.252,-0.151]
Coef of ind. ex. res× (dw-ioer)†	-0.0623	-0.0107	[-0.0152,-0.006]
Coef of ind. ex. res×agg. ex. res.†	4.223	2.2829	[1.548,3.069]

Notes: This table presents the simulated coefficients of Federal funds net purchase regressions under the estimated parameters. The column "Target" lists the estimated coefficients from the original regressions. The column "Simulation" lists the simulated coefficients. The column "95% CI" lists the 95% confidence interval of the simulated coefficients. The sign † represents the target is used in estimation. "ind. ex. res." is the individual excess reserves divided by individual bank assets. "ioer" is the interest rate on excess reserves. "dw" is the primary credit rate. "agg. ex. res." is the aggregate excess reserves divided by aggregate bank assets.

Table 8: Simulated moments

Year	2006		2018	
	Target	Simulation	Target	Simulation
Intermediation volume share	0.2150	0.1715	0.0663	0.0726
FF sold by intermediary	0.0045	0.0034	0.0002	0.0009
FF purchased by intermediary	0.0107	0.0062	0.0013	0.0031
Fraction of trading banks	0.8894	0.8805	0.6896	0.6985
Effective Federal funds rate	0.0514	0.0511	0.0204	0.0207

Notes: This table presents the simulated moments under the estimated parameters. The column "Target" lists the moments from the data. The column "Simulation" lists the simulated moments. All the targets are used in estimation. "Intermediation volume share" is the share of Federal funds reallocation in total Federal funds volume. "FF sold by intermediary" is the volume of Federal funds sold by intermediary banks as a share of aggregate bank assets. "FF purchased by intermediary" is the volume of Federal funds purchased by intermediary banks as a share of aggregate bank assets. "Fraction of trading banks" is the fraction of banks that trade in the total number of banks. All the moments are average values across quarters within each year.

Table 9: Counterfactual analysis

			Counterfactual analysis		
			(1)	(2)	(3)
Moments in 2018	Target	Simulation	IOER	Agg ex res	Transct cost
Intermediation Volume Share	0.0663	0.0726	0.1328	0.0654	0.3025
FF sold by intermediary	0.0002	0.0009	0.0019	0.0006	0.0166
FF purchased by intermediary	0.0013	0.0031	0.0041	0.0029	0.0382
Fraction of trading banks	0.6896	0.6985	0.8802	0.6985	0.6985
Effective Federal funds rate	0.0204	0.0207	0.0318	0.0203	0.0331

Notes: This table presents the simulated counterfactual analysis under the estimated parameters. The column "Target" lists the moments from the data. The column "Simulation" lists the simulated moments of the estimated model. The columns under "Counterfactual analysis" lists the simulated moments. under the corresponding counterfactual exercise. "IOER" represents the exercise that changes the values of IOER, primary credit rate and ON RRP from 2018 to 2006. "Agg ex res" represents the exercise that changes the aggregate excess reserves from 2018 to 2006 by proportionaly scaling individual excess reserves. "Transct cost" represents the exercise that changes the transaction parameter κ_0 from the 2018 value to 2006 value. "Intermediation volume share" is the share of Federal funds reallocation in total Federal funds volume. "FF sold by intermediary" is the volume of Federal funds sold by intermediary banks as a share of aggregate bank assets. "FF purchased by intermediary" is the volume of Federal funds purchased by intermediary banks as a share of aggregate bank assets. "Fraction of trading banks" is the fraction of banks that trade in the total number of banks. All the moments are average values across quarters within each year.

C Proofs and Derivations

C.1 Derivation of the general form of $m(\varepsilon, \varepsilon')$

For any $\varepsilon, \varepsilon' \in [0,1]$, equation (2) implies that

$$\begin{aligned} m\left(\varepsilon, \varepsilon'\right) &= \varepsilon' m\left(\varepsilon, 1\right) + \left(1-\varepsilon'\right) m\left(\varepsilon, 0\right) \\ &= \left[m\left(\varepsilon, 1\right) - m\left(\varepsilon, 0\right)\right] \varepsilon' + m\left(\varepsilon, 0\right). \end{aligned} \tag{C.3}$$

By symmetry we have

$$\begin{aligned} m\left(\varepsilon, 1\right) = m\left(1, \varepsilon\right) &= \left[m\left(1,1\right) - m\left(1,0\right)\right] \varepsilon + m\left(1,0\right), \\ m\left(\varepsilon, 0\right) = m\left(0, \varepsilon\right) &= \left[m\left(0,1\right) - m\left(0,0\right)\right] \varepsilon + m\left(0,0\right). \end{aligned} \tag{C.4}$$

Thus we can get

$$
\begin{aligned}
m\left(\varepsilon,\varepsilon'\right) &= \left[m\left(\varepsilon,1\right)-m\left(\varepsilon,0\right)\right]\varepsilon'+m\left(\varepsilon,0\right) && \text{(C.5)}\\
&= \left\{\left[m\left(1,1\right)-m\left(1,0\right)\right]\varepsilon+m\left(1,0\right)-\left[m\left(0,1\right)-m\left(0,0\right)\right]\varepsilon-m\left(0,0\right)\right\}\varepsilon' && \text{(C.6)}\\
&\quad+\left[m\left(0,1\right)-m\left(0,0\right)\right]\varepsilon+m\left(0,0\right) && \text{(C.7)}\\
&= \left[m\left(1,1\right)-m\left(1,0\right)-m\left(0,1\right)+m\left(0,0\right)\right]\varepsilon\varepsilon'+\left[m\left(0,1\right)-m\left(0,0\right)\right]\varepsilon && \text{(C.8)}\\
&\quad+\left[m\left(1,0\right)-m\left(0,0\right)\right]\varepsilon'+m\left(0,0\right) && \text{(C.9)}\\
&= \left(\lambda-2\lambda_1+\lambda_0\right)\varepsilon\varepsilon'+\left(\lambda_1-\lambda_0\right)\left(\varepsilon+\varepsilon'\right)+\lambda_0.
\end{aligned}
$$

C.2 Derivation of HJB (6) and KFE (7)

For all $k\in\mathbb{K}$ and $t\in[0,T]$, the value function is given by

$$
V_t\left(k\right)=\mathbb{E}^{\boldsymbol{\epsilon}}\left\{\begin{array}{l}\int_0^{\min\{t_{+1},T\}-t}e^{-r\tau}u\left(k\right)d\tau+1_{t_{+1}>T}e^{-r(T-t)}U\left(k\right)\\ +1_{t_{+1}\leq T}e^{-r(t_{+1}-t)}\int\left\{\begin{array}{l}V_{t_{+1}}\left[k+q_{t_{+1}}\left[k,k^a,\varepsilon_{t_{+1}},\varepsilon^a_{t_{+1}}\left(k^a\right)\right]\right]\\ -\chi\left[\varepsilon_{t_{+1}},q_{t_{+1}}\left[k,k^a,\varepsilon_{t_{+1}},\varepsilon^a_{t_{+1}}\left(k^a\right)\right]\right]\\ -e^{-r(T+\Delta-t_{+1})}R_{t_{+1}}\left[k,k^a,\varepsilon_{t_{+1}},\varepsilon^a_{t_{+1}}\left(k^a\right)\right]\end{array}\right\}\frac{m\left[\varepsilon_{t_{+1}},\varepsilon^a_{t_{+1}}\left(k^a\right)\right]}{m\left(\varepsilon_{t_{+1}},\bar{\varepsilon}^a_{t_{+1}}\right)}dF_{t_{+1}}\left(k^a\right)\end{array}\right\}, \tag{C.10}
$$

where

$$
\begin{aligned}
q_t\left(k,k^a,\varepsilon,\varepsilon^a\right) &= \arg\max_q\left\{V_t\left(k+q\right)+V_t\left(k^a-q\right)-\chi\left(\varepsilon,q\right)-\chi\left(\varepsilon^a,-q\right)\right\}, && \text{(C.11)}\\
e^{-r(T+\Delta-t)}R_t\left(k,k^a,\varepsilon,\varepsilon^a\right) &= \frac{1}{2}\left\{\begin{array}{l}V_t\left[k+q_t\left(k,k^a,\varepsilon,\varepsilon^a\right)\right]-V_t\left(k\right)-\chi\left[\varepsilon,q_t\left(k,k^a,\varepsilon,\varepsilon^a\right)\right]\\ V_t\left(k^a\right)-V_t\left[k^a-q_t\left(k,k^a,\varepsilon,\varepsilon^a\right)\right]+\chi\left[\varepsilon^a,-q_t\left(k,k^a,\varepsilon,\varepsilon^a\right)\right]\end{array}\right\},
\end{aligned}
$$

and t_{+1} is the random time of matching the next counterparty, arriving at the rate $m\left(\varepsilon_t,\bar{\varepsilon}^a_t\right)$. The costs of search intensities, $\chi\left(\varepsilon,q\right)$ and $\chi\left(\varepsilon',q\right)$, are shared in the bargaining; it creates the cost shifting effect.

By the property of Poisson process, the equation (C.10) for value function $V_t\left(k\right)$ can be rewritten as

$$
\begin{aligned}
&V_t\left(k\right) && \text{(C.12)}\\
&=\max_{\{\varepsilon_z\}_{z\in[t,T]}\in[0,1]^{[t,T]}}\left\{\begin{array}{l}\int_t^T e^{-\int_t^z[r+m(\varepsilon_s,\bar{\varepsilon}_s)]ds}\left\{\begin{array}{l}u\left(k\right)+\int_{k'}\left\{\begin{array}{l}V_z\left[k+q_z\left(k,k',\varepsilon_z,\varepsilon_z\left(k'\right)\right)\right]\\ -\chi\left[\varepsilon_z,q_z\left(k,k',\varepsilon_z,\varepsilon_z\left(k'\right)\right)\right]\\ -e^{-r(T+\Delta-z)}R_z\left(k,k',\varepsilon_z,\varepsilon_z\left(k'\right)\right)\end{array}\right\}\\ \times m\left(\varepsilon_z,\varepsilon_z\left(k'\right)\right)dF_z\left(k'\right)\end{array}\right\}dz\\ +e^{-\int_t^T[r+m(\varepsilon_s,\bar{\varepsilon}_s)]ds}U\left(k\right)\end{array}\right\}
\end{aligned}
$$

Denote $\varepsilon^*_t\left(k\right)$ as one equilibrium search profile. By taking the first-order derivative of $V_t\left(k\right)$ w.r.t. t

and plugging in the solution to $e^{-r(T+\Delta-z)}R_z\left(k,k',\varepsilon_z,\varepsilon_z\left(k'\right)\right)$, we can obtain

$$rV_t\left(k\right)=\dot{V}_t\left(k\right)+u\left(k\right)+\int\frac{1}{2}S_t\left[k,k',\varepsilon_t^*\left(k\right),\varepsilon_t^*\left(k'\right)\right]m\left[\varepsilon_t^*\left(k\right),\varepsilon_t^*\left(k'\right)\right]dF_t\left(k'\right).$$

To derive the optimality condition for $\varepsilon_t^*\left(k\right)$, let $\mathbf{B}$ denote the space of bounded real-valued functions defined on $\mathbb{K}\times[0,T]$. Define a mapping $\mathcal{M}$ on $\mathbf{B}$ as follows:

$$\left(\mathcal{M}w\right)\left(k,t\right) \tag{C.13}$$
$$=\max_{\{\varepsilon_z\}_{z\in[t,T]}\in[0,1]^{[t,T]}}\left\{\begin{array}{l}\int_t^T e^{-\int_t^z[r+m(\varepsilon_s,\bar{\varepsilon}_s)]ds}\left\{\begin{array}{l}u\left(k\right)+\int_{k'}\left\{\begin{array}{l}w\left[k+b_z\left(k,k',\varepsilon_z,\varepsilon_z\left(k'\right)\right),z\right]\\-\chi\left[\varepsilon_z,b_z\left(k,k',\varepsilon_z,\varepsilon_z\left(k'\right)\right)\right]\\-e^{-r(T+\Delta-z)}Y_z\left(k,k',\varepsilon_z,\varepsilon_z\left(k'\right)\right)\end{array}\right\}\\\times m\left(\varepsilon_z,\varepsilon_z\left(k'\right)\right)dF_z\left(k'\right)\end{array}\right\}dz\\+e^{-\int_t^T[r+m(\varepsilon_s,\bar{\varepsilon}_s)]ds}U\left(k\right)\end{array}\right\}$$

where

$$b_t\left(k,k',\varepsilon,\varepsilon\prime\right)\in\arg\max_b\left\{\begin{array}{c}w\left(k+b,t\right)-w\left(k,t\right)-\chi\left(\varepsilon,b\right)\\+w\left(k'-b,t\right)-w\left(k',t\right)-\chi\left(\varepsilon',-b\right)\end{array}\right\}$$

and

$$e^{-r(T+\Delta-t)}Y_t\left(k,k',\varepsilon,\varepsilon'\right)=\frac{1}{2}\left\{\begin{array}{c}w\left(k+b_t\left(k,k',\varepsilon,\varepsilon'\right),t\right)-w\left(k,t\right)-\chi\left(\varepsilon,b_t\left(k,k',\varepsilon,\varepsilon'\right)\right)\\+w\left(k',t\right)-w\left(k'-b_t\left(k,k',\varepsilon,\varepsilon'\right),t\right)+\chi\left(\varepsilon',-b_t\left(k,k',\varepsilon,\varepsilon'\right)\right)\end{array}\right\}.$$

It is clear that the solution $V_t\left(k\right)$ to the HJB (6) is a fixed point of the mapping $\mathcal{M}$. Therefore, $\varepsilon_t^*\left(k\right)$ must be the solution to the right-hand side of $\left(\mathcal{M}w\right)\left(k,t\right)$ if we replace w with V. Note that since the time variable t is continuous, we have a continuum of control variables. We follow the heuristic approach in van Imhoff (1982) to derive the condition for $\varepsilon_t^*\left(k\right)$. This approach relies on interpreting the integral in $\left(\mathcal{M}w\right)\left(k,t\right)$ as a summation of discrete variables over intervals with widths dz and dt. Then the Lebesgue dominated convergence theorem guarantees that the summation converges to the original integral as the widths of intervals approach 0. Then the terms in $\left(\mathcal{M}w\right)\left(k,t\right)$ which are

related to $\varepsilon_t(k)$ can be written as

$$e^{-\int_t^{t+dt}[r+m(\varepsilon_t(k),\bar{\varepsilon}_t)]ds}\left\{\begin{array}{l} u(k)+\int_{k'}\left\{\begin{array}{l} w\left[k+b_t\left(k,k',\varepsilon_t(k),\varepsilon_t(k')\right),t\right] \\ -\chi\left[\varepsilon_t(k),b_t\left(k,k',\varepsilon_t(k),\varepsilon_t(k')\right)\right] \\ -e^{-r(T+\Delta-t)}Y_t\left(k,k',\varepsilon_t(k),\varepsilon_t(k')\right)\end{array}\right\} \\ \times m\left(\varepsilon_t(k),\varepsilon_t(k')\right)dF_t(k')\end{array}\right\}dt \tag{C.14}$$

$$+e^{-\int_t^{t+dt}[r+m(\varepsilon_t(k),\bar{\varepsilon}_t)]ds}w(k,t-dt) \tag{C.15}$$

$$=(1-rdt)\,w(k,t-dt)+o(|dt|)+\left\{\begin{array}{l} u(k)+\int_{k'}\left\{\begin{array}{l} w\left[k+b_t\left(k,k',\varepsilon_t(k),\varepsilon_t(k')\right),t\right] \\ -w(k,t-dt) \\ -\chi\left[\varepsilon_t,b_t\left(k,k',\varepsilon_t(k),\varepsilon_t(k')\right)\right] \\ -e^{-r(T+\Delta-t)}Y_t\left(k,k',\varepsilon_t(k),\varepsilon_t(k')\right)\end{array}\right\} \\ \times m\left(\varepsilon_t(k),\varepsilon_t(k')\right)dF_t(k')\end{array}\right\}dt.$$

Thus the maximizer of $\varepsilon_t(k)$ to the above equation when $dt\to 0$ is given by

$$\varepsilon_t(k)\in\arg\max_{\varepsilon\in[0,1]}\left\{\begin{array}{l}\int_{k'}\frac{1}{2}\left[\begin{array}{l} w\left[k+b_t\left(k,k',\varepsilon,\varepsilon_t(k')\right),t\right]-w(k,t)-\chi\left[\varepsilon_t,b_t\left(k,k',\varepsilon,\varepsilon_t(k')\right)\right] \\ +w\left[k'-b_t\left(k,k',\varepsilon,\varepsilon_t(k')\right),t\right]-w(k',t)-\chi\left[\varepsilon_t(k'),-b_t\left(k,k',\varepsilon,\varepsilon_t(k')\right)\right]\end{array}\right] \\ \times m\left(\varepsilon_t(k),\varepsilon_t(k')\right)dF_t(k')\end{array}\right\}$$

where we plug in the solution to $e^{-r(T+\Delta-t)}Y_t\left(k,k',\varepsilon,\varepsilon_t(k')\right)$. This gives the HJB (6).

Next we take a heuristic approach to derive the KFE. Let Δ be a small time interval that is close to 0. Then by definition of $F_t(k)$, we have

$$F_{t+\Delta}(k^w)=\left[1-\Delta\cdot m\left(\varepsilon_t(k),\bar{\varepsilon}_t\right)\right]F_t(k^w) \tag{C.16}$$

$$+\int_{k\le k^w}\int_{k'}\Delta\cdot m\left(\varepsilon_t(k),\varepsilon_t\left(k'\right)\right)1\left\{k+q_t\left(k,k'\right)\le k^w\right\}dF_t\left(k'\right)dF_t(k) \tag{C.17}$$

$$+\int_{k>k^w}\int_{k'}\Delta\cdot m\left(\varepsilon_t(k),\varepsilon_t\left(k'\right)\right)1\left\{k+q_t\left(k,k'\right)\le k^w\right\}dF_t\left(k'\right)dF_t(k).$$

On the right-hand side, the first term represents the mass of banks that do not meet counterparties during $[t,t+\Delta]$. The second term represents the banks that have meetings during $[t,t+\Delta]$ and hold reserves no more than k^w both before and after the meeting. The third term represents the banks that have meetings during $[t,t+\Delta]$ and hold reserves more than k^w before meeting and no more than k^w after the meeting. These three groups of banks constitute the mass of banks with reserves

no more than k^w at $t+\Delta$. By rearranging terms, we can get

$$\begin{aligned}
\frac{F_{t+\Delta}\left(k^w\right)-F_t\left(k^w\right)}{\Delta} &= -m\left(\varepsilon_t\left(k\right),\bar{\varepsilon}_t\right)F_t\left(k^w\right) && \text{(C.18)}\\
&\quad+\int_{k\leq k^w}\int_{k'} m\left(\varepsilon_t\left(k\right),\varepsilon_t\left(k'\right)\right)1\left\{k+q_t\left(k,k'\right)\leq k^w\right\}dF_t\left(k'\right)dF_t\left(k\right) && \text{(C.19)}\\
&\quad+\int_{k>k^w}\int_{k'} m\left(\varepsilon_t\left(k\right),\varepsilon_t\left(k'\right)\right)1\left\{k+q_t\left(k,k'\right)\leq k^w\right\}dF_t\left(k'\right)dF_t\left(k\right) && \text{(C.20)}\\
&= -\int_{k\leq k^w}\int_{k'} m\left(\varepsilon_t\left(k\right),\varepsilon_t\left(k'\right)\right)1\left\{k+q_t\left(k,k'\right)> k^w\right\}dF_t\left(k'\right)dF_t\left(k\right) && \text{(C.21)}\\
&\quad+\int_{k>k^w}\int_{k'} m\left(\varepsilon_t\left(k\right),\varepsilon_t\left(k'\right)\right)1\left\{k+q_t\left(k,k'\right)\leq k^w\right\}dF_t\left(k'\right)dF_t\left(k\right),
\end{aligned}$$

where in the second equality we expand $m\left(\varepsilon_t\left(k\right),\bar{\varepsilon}_t\right)F_t\left(k^w\right)$ to

$$\int_{k\leq k^w}\int_{k'} m\left(\varepsilon_t\left(k\right),\varepsilon_t\left(k'\right)\right)dF_t\left(k'\right)dF_t\left(k\right),$$

and combine it with $\int_{k\leq k^w}\int_{k'} m\left(\varepsilon_t\left(k\right),\varepsilon_t\left(k'\right)\right)1\left\{k+q_t\left(k,k'\right)\leq k^w\right\}dF_t\left(k'\right)dF_t\left(k\right)$. Then we can take $\Delta\to 0$ and obtain the KFE (7).

C.3 Proof of Lemma 1

Proof. Define

$$\mathcal{E}^*\left(W,G\right)\equiv\arg\max_{\mathbf{x}\in[0,1]^{\mathbb{K}}}\int\int S\left[k,k',x\left(k\right),x\left(k'\right);W\right]m\left[x\left(k\right),x\left(k'\right)\right]dG\left(k'\right)dG\left(k\right). \qquad \text{(C.22)}$$

Obviously, $\mathcal{E}^*\left(W,G\right)$ exists. The first order condition of (C.22) with respect to $\mathbf{x}$ is

$$\begin{array}{l}\int\frac{d}{dx}\left\{S\left[k,k',x,\mathcal{E}^*\left(W,G\right)\left(k'\right);W\right]m\left[x,\mathcal{E}^*\left(W,G\right)\left(k'\right)\right]\right\}dG\left(k'\right)\\ +\int\frac{d}{dx}\left\{S\left[k',k,\mathcal{E}^*\left(W,G\right)\left(k'\right),x;W\right]m\left[\mathcal{E}^*\left(W,G\right)\left(k'\right),x\right]\right\}dG\left(k'\right)\end{array}\Bigg|_{x=\omega^*(W,G)(k)}\begin{array}{l}>0 \text{ if } \mathcal{E}^*\left(W,G\right)\left(k\right)=1\\ =0 \text{ if } \mathcal{E}^*\left(W,G\right)\left(k\right)\in(0,1)\\ <0 \text{ if } \mathcal{E}^*\left(W,G\right)\left(k\right)=0\end{array} \qquad \text{(C.23)}$$

Obviously, if $\mathcal{E}^*\left(W,G\right)$ is an equilibrium search profile, it maximizes the aggregate joint surplus. So what we need to prove is that $\mathcal{E}^*\left(W,G\right)$ is an equilibrium search profile. Suppose not, hence the following first order condition of (10) with respect to $\mathbf{x}$ is is violated

$$\int\frac{d}{dx}\left\{S\left[k,k',x,\mathcal{E}^*\left(W,G\right)\left(k'\right);W\right]m\left[x,\mathcal{E}\left(W,G\right)\left(k'\right)\right]\right\}dG\left(k'\right)\Big|_{x=\mathcal{E}^*(W,G)(k)}\begin{array}{l}<0 \text{ if } \mathcal{E}^*\left(W,G\right)\left(k\right)=1\\ =0 \text{ if } \mathcal{E}^*\left(W,G\right)\left(k\right)\in(0,1)\\ >0 \text{ if } \mathcal{E}^*\left(W,G\right)\left(k\right)=0\end{array} \qquad \text{(C.24)}$$

Since $S\left[k,k',x\left(k\right),x\left(k'\right);W\right]m\left[x\left(k\right),x\left(k'\right)\right]$ is symmetric in k and k', we have

$$\frac{d}{dx}\left\{S\left[k,k',\mathcal{E}^{*}\left(W,G\right)\left(k\right),x;W\right]m\left[\mathcal{E}\left(W,G\right)\left(k\right),x\right]\right\} \tag{C.25}$$

$$=\frac{d}{dx}\left\{S\left[k',k,x,\mathcal{E}^{*}\left(W,G\right)\left(k'\right);W\right]m\left[x,\mathcal{E}\left(W,G\right)\left(k'\right)\right]\right\} \tag{C.26}$$

$$\Leftrightarrow\int\frac{d}{dx}\left\{S\left[k,k',\mathcal{E}^{*}\left(W,G\right)\left(k\right),x;W\right]m\left[\mathcal{E}\left(W,G\right)\left(k\right),x\right]\right\}dG\left(k'\right) \tag{C.27}$$

$$=\int\frac{d}{dx}\left\{S\left[k',k,x,\mathcal{E}^{*}\left(W,G\right)\left(k\right);W\right]m\left[x,\mathcal{E}\left(W,G\right)\left(k\right)\right]\right\}dG\left(k'\right)$$

Thus, (C.23) implies

$$\begin{array}{l}\int\frac{d}{dx}\left\{S\left[k,k',x,\mathcal{E}^{*}\left(W,G\right)\left(k'\right);W\right]m\left[x,\mathcal{E}\left(W,G\right)\left(k'\right)\right]\right\}dG\left(k'\right)\\+\int\frac{d}{dx}\left\{S\left[k',k,\mathcal{E}^{*}\left(W,G\right)\left(k'\right),x;W\right]m\left[\mathcal{E}\left(W,G\right)\left(k'\right),x\right]\right\}dG\left(k'\right)\end{array}\Bigg|_{x=\mathcal{E}^{*}(W,G)(k)} \tag{C.28}$$

$$=\int\frac{d}{dx}\left\{S\left[k,k',x,\mathcal{E}^{*}\left(W,G\right)\left(k'\right);W\right]m\left[x,\mathcal{E}\left(W,G\right)\left(k'\right)\right]\right\}dG\left(k'\right)\Big|_{x=\mathcal{E}^{*}(W,G)(k)} \tag{C.29}$$

$$\begin{array}{l}>0\text{ if }\mathcal{E}^{*}\left(W,G\right)\left(k\right)=1\\=0\text{ if }\mathcal{E}^{*}\left(W,G\right)\left(k\right)\in\left(0,1\right)\\<0\text{ if }\mathcal{E}^{*}\left(W,G\right)\left(k\right)=0\end{array}$$

which leads to contradiction. **Q.E.D.**

C.4 Proof of Lemma 1

Proof. Given $\mathbf{x}^{a}=\left\{x_{t}^{a}\left(k\right)\right\}_{k\in\mathbb{K},t\in[0,T]}$ and $\mathbf{G}=\left\{G_{t}\left(k\right)\right\}_{k\in\mathbb{K},t\in[0,T]}$, let $\mathcal{B}$ denote the space of bounded real-valued functions defined on $\mathbb{K}\times[0,T]$. Define a functional $\mathcal{M}$ of $\mathbf{W}=\left\{W_{t}\left(k\right)\right\}_{k\in\mathbb{K},t\in[0,T]}\in\mathcal{B}$ as

$$\mathcal{M}\left(\mathbf{W};\mathbf{x}^{a},\mathbf{G}\right)\left(k,t\right) \tag{C.30}$$

$$\equiv u\left(k\right)\left[1-e^{-r(T-t)}\right]+e^{-r(T-t)}U\left(k\right) \tag{C.31}$$

$$+\frac{1}{2}\int_{t}^{T}\left\{e^{-(r+\lambda)(z-t)}\max_{x_{z}(k)}\int S\left[k,k',x_{z}\left(k\right),x_{z}^{a}\left(k'\right);W_{z}\right]m\left[x_{z}\left(k\right),x_{z}^{a}\left(k'\right)\right]dG_{z}\left(k'\right)\right\}dz$$

A value function $\mathbf{V}=\left\{V_{t}\left(k\right)\right\}_{k\in\mathbb{K},t\in[0,T]}$ solving the bank's maximization problem (6) is the fixed point of $\mathcal{M}$ given $\mathbf{F}=\left\{F_{t}\left(k\right)\right\}_{k\in\mathbb{K},t\in[0,T]}$, i.e., $\mathbf{V}=\mathcal{M}\left(\mathbf{V};\boldsymbol{\varepsilon}^{a},\mathbf{F}\right)$. Given $\mathbf{G}=\left\{G_{t}\left(k\right)\right\}_{k\in\mathbb{K},t\in[0,T]}$, define the metric $D:\mathcal{B}\times\mathcal{B}\to\mathbb{R}_{+}$ as

$$D\left(\mathbf{W},\mathbf{W}'\right)\equiv\sup_{k,t}\left\{e^{-\beta t}\left|W_{t}\left(k\right)-W_{t}'\left(k\right)\right|\right\},$$

where β is a constant satisfies

$$\max\left\{0,\lambda-r\right\}<\beta<\infty.$$

The metric space $(\mathcal{B}, D)$ is complete. $\mathcal{M}$ is a contraction mapping on the complete metric space $(\mathcal{B}, D)$. Thus, there exists unique $\mathbf{V}$ solving (6). **Q.E.D.**

C.5 Proof of Proposition 7

Proof. Denote v_t^w as the co-state to a_t, the Hamiltonian is thus given by

$$\mathcal{H}_t^w \equiv u\left(\frac{a_t}{1+\rho_t^w}\right) - e^{-r(T+\Delta-t)} d\delta_t + v_t^w \left(\frac{\dot{\rho}_t^w}{1+\rho_t^w} a_t + d\delta_t\right). \tag{C.32}$$

The evolution of costate is given by $r v_t^w - \dot{v}_t^w = \frac{\partial \mathcal{H}_t^w}{\partial a_t}$, i.e.

$$\dot{v}_t^w = r v_t^w - \frac{1}{1+\rho_t^w} u'\left(\frac{a_t}{1+\rho_t^w}\right) - v_t^w \frac{\dot{\rho}_t^w}{1+\rho_t^w}. \tag{C.33}$$

The first order condition with respect to $d\delta_t$ is

$$v_t^w = e^{-r(T+\Delta-t)}. \tag{C.34}$$

Since the first order condition is independent to a_t and δ_t, all banks must have the same value of costate. But since the evolution of costate, C.33, depends on a_t, the only possibility is that all banks have the same a_t for all $t > 0$. This implies $\delta_t(a)$ is given by result (b), such that they hold K units of reserve balance for all $t > 0$. Substituting (C.34) to the evolution of costate, (C.33), we have

$$\dot{\rho}_t^w = -e^{r(T+\Delta-t)} u'(K).$$

The solution to the above ODE is

$$\rho_t^w = \rho_T^w + e^{r\Delta}\left[e^{r(T-t)} - 1\right] \frac{u'(K)}{r}.$$

Notice that at T the bank problem is

$$\max_{q_T} \left\{U(k+q_T) - e^{-r\Delta}(1+\rho_T^w) q_T\right\}.$$

To yield $k + q_T = K$, we have

$$\rho_T^w = e^{r\Delta} U'(K) - 1.$$

Q.E.D.

C.6 Proof of Lemma 2

Proof. Suppose $W = -Hk^2 + Ek + D$, then we have

$$S\left(k, k', \varepsilon, \varepsilon'; W\right) m\left(\varepsilon, \varepsilon'\right) = \left[H\left(k' - k\right)\right]^2 \frac{m\left(\varepsilon, \varepsilon'\right)}{2H + \kappa\left(\varepsilon\right) + \kappa\left(\varepsilon'\right)}.$$

Using Lemma 1 we have

$$\begin{aligned}
&\max_{\mathbf{x}\in\Omega(W,G)} \int\int S\left[k, k', x\left(k\right), x\left(k'\right); W\right] m\left[x\left(k\right), x\left(k'\right)\right] dG\left(k'\right) dG\left(k\right) \\
&= H^2 \max_{\mathbf{x}\in\Omega(W,G)} \int\int \frac{m\left[x\left(k\right), x\left(k'\right)\right]}{2H + \kappa\left[x\left(k\right)\right] + \kappa\left[x\left(k'\right)\right]} \left(k' - k\right)^2 dG\left(k'\right) dG\left(k\right) \\
&\leq \max_{\varepsilon,\varepsilon'\in[0,1]} \left\{\frac{m\left(\varepsilon, \varepsilon'\right)}{2H + \kappa\left(\varepsilon\right) + \kappa\left(\varepsilon'\right)}\right\} H^2 \int\int \left(k' - k\right)^2 dG\left(k'\right) dG\left(k\right) \qquad \text{(C.35)}
\end{aligned}$$

We want to show $\varepsilon = \varepsilon'$. Suppose not, i.e., $\varepsilon > \varepsilon'$ maximize (C.35). Consider a deviation $\varepsilon'' = \left(\varepsilon + \varepsilon'\right)/2$. Note that convexity of κ implies that $2\kappa\left(\varepsilon''\right) < \kappa\left(\varepsilon\right) + \kappa\left(\varepsilon'\right)$. Additivity of m implies

$$m\left(\varepsilon'', \varepsilon''\right) - m\left(\varepsilon, \varepsilon'\right) = \left(\lambda - 2\lambda_1 + \lambda_0\right)\left[\left(\varepsilon''\right)^2 - \varepsilon\varepsilon'\right] > 0$$

where the last inequality follows the fact that the arithmetric mean is always greater than the geometric mean. Thus we have contradiction as

$$\frac{m\left(\varepsilon'', \varepsilon''\right)}{2H + 2\kappa\left(\varepsilon''\right)} > \frac{m\left(\varepsilon, \varepsilon'\right)}{2H + \kappa\left(\varepsilon\right) + \kappa\left(\varepsilon'\right)}.$$

Finally, we want to show $x\left(k\right) \equiv \varepsilon$ that maximizes (C.35), i.e. $x\left(k\right) \equiv \tilde{\omega}\left(H\right) \in \arg\max_{\varepsilon\in[0,1]} \frac{m(\varepsilon,\varepsilon)}{H+\kappa(\varepsilon)}$, is an equilibrium. Given $x\left(k'\right) = \varepsilon$ for all $k' \neq k$, the search intensity of a k-bank is

$$\begin{aligned}
&\arg\max_{x(k)} \int S\left[k, k', x\left(k\right), x\left(k'\right); W\right] m\left[x\left(k\right), x\left(k'\right)\right] dG\left(k'\right) \qquad \text{(C.36)} \\
&= \arg\max_{x(k)} \int \frac{m\left[x\left(k\right), \varepsilon\right]}{2H + \kappa\left[x\left(k\right)\right] + \kappa\left(\varepsilon\right)} \left(k' - k\right)^2 dG\left(k'\right) \qquad \text{(C.37)} \\
&= \arg\max_{x(k)} \frac{m\left[x\left(k\right), \varepsilon\right]}{2H + \kappa\left[x\left(k\right)\right] + \kappa\left(\varepsilon\right)}
\end{aligned}$$

with the following first-order condition:

$$\text{F.O.C.:}\ \frac{\left[\left(\lambda - 2\lambda_1 + \lambda_0\right)x + \left(\lambda_1 - \lambda_0\right)\right]\left[2H + \kappa\left[x\left(k\right)\right] + \kappa\left(x\right)\right] - m\left[x\left(k\right), x\right]\kappa'\left(x\left(k\right)\right)}{\left[2H + \kappa\left[x\left(k\right)\right] + \kappa\left(x\right)\right]^2}. \qquad \text{(C.38)}$$

Note that the numerator of the above equation is decreasing in $x\left(k\right)$ since its derivative is

$$-m\left[x\left(k\right), x\right]\kappa''\left(x\left(k\right)\right) < 0,$$

which implies that the individual optimal $x(k)$ is unique. The first-order condition for $\tilde{\omega}(H) \in \arg\max_{x\in[0,1]} \frac{m(x,x)}{H+\kappa(x)}$ is

$$\text{F.O.C.: } \frac{2\left[(\lambda - 2\lambda_1 + \lambda_0)\, x + (\lambda_1 - \lambda_0)\right]\left[H + \kappa(x)\right] - m(x,x)\,\kappa'(x)}{\left[H + \kappa(x)\right]^2}. \tag{C.39}$$

Note that the FOC (C.39) is proportional to the FOC (C.38) when $x(k) = x$, thus ε and ε' that maximize (C.35) is an equilibrium. **Q.E.D.**

C.7 Proof of Proposition 3

Proof. Define $\tau = T - t$ and $v_\tau(k) \equiv V_{t-\tau}(k)$. We prove the proposition by real induction. We need to establish the following three conditions: (a) $v_0(k)$ is quadratic; (b) if $v_a(k)$ is quadratic, then there exist $b > a$ such that $v_\tau(k)$ is quadratic for any $\tau \in [a, b)$; (c) if $v_\tau(k)$ is quadratic for any $\tau \in [0, b)$, then $v_b(k)$ is also quadratic. If condition (a), (b) and (c) are satisfied, then by the principle of real induction $v_\tau(k)$ is quadratic for all $\tau \in [0, T]$.

Condition (a) is automatic as $v_0(k) = U(k)$ is quadratic. Suppose $v_a(k)$ is quadratic. Taking the Taylor expansion along the time dimension around a, for all $\tau \in [a, b)$ we have $\frac{\partial^n}{\partial k^n} v_\tau(k)$ given by

$$\frac{\partial^n}{\partial k^n} v_\tau(k) = \frac{\partial^n}{\partial k^n} v_a(k) + \sum_{i=1}^{\infty} \frac{\partial^i}{\partial \tau^i} \frac{\partial^n}{\partial k^n} v_\tau(k)\Big|_{\tau=a} \frac{(\tau - a)^i}{i!}.$$

Note that in the equilibrium, we have

$$\frac{\partial}{\partial \tau} v_\tau(k) = -r v_\tau(k) + r u(k) + \frac{1}{2} \int S\left[k, k', \varepsilon_\tau^a(k), \varepsilon_\tau^a(k'); v_\tau\right] m\left[\varepsilon_\tau^a(k), \varepsilon_\tau^a(k')\right] dF_\tau(k'). \tag{C.40}$$

where ε_τ^a satisfies (10). Note that given that $v_a(k)$ is quadratic, the RHS of (C.40) is quadratic at $\tau = a$. Differentiating (C.40) with respect to k for n times, $n \geq 3$, at $\tau = a$, we have

$$\frac{\partial}{\partial \tau} \frac{\partial^n}{\partial k^n} v_a(k) = 0.$$

Thus we have $\frac{\partial^i}{\partial \tau^i} \frac{\partial^n}{\partial k^n} v_\tau(k) = 0$ for any $n \geq 3$ and $i \geq 1$, i.e., $v_\tau(k)$ is quadratic for any $\tau \in [a, b)$. We establish condition (b).

Suppose $v_\tau(k)$ is quadratic for any $\tau \in [0, b)$ but $v_b(k)$ is not quadratic. Integrating (C.40) from

any $\tau \in [0, b)$ to b, we have

$$v_\tau(k) = v_b(k) - \int \dot{v}_z(k)\, dz \tag{C.41}$$

$$= v_b(k) - ru(k)(b-\tau) + r\int_\tau^b v_z(k)\, dz \tag{C.42}$$

$$-\frac{1}{2}\int_\tau^b \int S\left[k, k', \varepsilon_z^a(k), \varepsilon_z^a(k'); v_z\right] m\left[\varepsilon_z^a(k), \varepsilon_z^a(k')\right] dF_z(k')\, dz.$$

Since $v_b(k)$ is not quadratic, although all other terms are, we have $v_\tau(k)$ is not quadratic for any $\tau \in [0, b)$. It contradicts the premise and we establish condition (c). **Q.E.D.**

C.8 Proof of Lemma 4

Proof. Plug the closed-form solution (**??**) and $\varepsilon_t(k) \equiv \varepsilon_t$ into the KFE (7), we can get

$$\dot{F}_t(k^w) = m(\varepsilon_t, \varepsilon_t) \left\{ \begin{array}{c} \int_{k>k^w} \int 1\left\{k + \frac{H_t(k'-k)}{2\kappa(\varepsilon_t)+2H_t} \leq k^w\right\} dF_t(k')\, dF_t(k) \\ -\int_{k\leq k^w} \int 1\left\{k + \frac{H_t(k'-k)}{2\kappa(\varepsilon_t)+2H_t} > k^w\right\} dF_t(k')\, dF_t(k) \end{array} \right\} \tag{C.43}$$

$$= m(\varepsilon_t, \varepsilon_t) \left\{ \begin{array}{c} \int_{k>k^w} F_t\left[2\left(1+\frac{\kappa(\varepsilon_t)}{H_t}\right)k^w - \left(1+\frac{2\kappa(\varepsilon_t)}{H_t}\right)k\right] dF_t(k) \\ -\int_{k\leq k^w}\left[1 - F_t\left[2\left(1+\frac{\kappa(\varepsilon_t)}{H_t}\right)k^w - \left(1+\frac{2\kappa(\varepsilon_t)}{H_t}\right)k\right]\right] dF_t(k) \end{array} \right\} \tag{C.44}$$

$$= m(\varepsilon_t, \varepsilon_t)\left[\int F_t\left[2\left(1+\frac{\kappa(\varepsilon_t)}{H_t}\right)k - \left(1+\frac{2\kappa(\varepsilon_t)}{H_t}\right)k'\right] dF_t(k') - F_t(k)\right].$$

Then the probability density function solves the following PDE:

$$\dot{f}_t(k) = m(\varepsilon_t, \varepsilon_t)\left[2\left(1+\frac{\kappa(\varepsilon_t)}{H_t}\right)\int f_t\left[2\left(1+\frac{\kappa(\varepsilon_t)}{H_t}\right)k - \left(1+\frac{2\kappa(\varepsilon_t)}{H_t}\right)k'\right] f_t(k')\, dk' - f_t(k)\right]. \tag{C.45}$$

To characterize the dynamics of moment function, we take advantage of the Fourier transform. We follow the definition of Bracewell (2000) for the Fourier transform:

$$h^*(\nu) = \int e^{-i2\pi\nu x} h(x)\, dx,$$

where $h^*(\cdot)$ is the Fourier transform of the function $h(\cdot)$.

Let $f_t^*(\cdot)$ be the Fourier transform of the equilibrium pdf $f_t(\cdot)$. Then the Fourier transform of equation (C.45) is

$$\dot{f}_t^*(\nu) = m(\varepsilon_t, \varepsilon_t)\left[f_t^*\left(\frac{H_t}{2(H_t+\kappa(\varepsilon_t))}\nu\right) f_t^*\left(\frac{H_t+2\kappa(\varepsilon_t)}{2(H_t+\kappa(\varepsilon_t))}\nu\right) - f_t^*(\nu)\right]. \tag{C.46}$$

The PDE (C.46) cannot be solved in closed form. However, it facilitates the calculation fo the moment function which is the derivative of the transform, with respect to ν, at $\nu = 0$. Let us denote $f_t^{*(n)}(\nu)$ be the n-th derivative of $f_t^*(\nu)$ with respect to ν. By taking n-th derivative with respect to ν to both

sides of (C.46), we can obtain

$$\dot{f}_t^{*(n)}(\nu) = m(\varepsilon_t, \varepsilon_t)\left[\sum_{i=0}^{n} C_n^i \frac{(H_t)^{n-i}(H_t + 2\kappa(\varepsilon_t))^i}{2^n (H_t + \kappa(\varepsilon_t))^n} f_t^{*(n-i)}\left(\frac{H_t}{2(H_t + \kappa(\varepsilon_t))}\nu\right) f_t^{*(i)}\left(\frac{H_t + 2\kappa(\varepsilon_t)}{2(H_t + \kappa(\varepsilon_t))}\nu\right) - f_t^{*(n)}(\nu)\right] \tag{C.47}$$

Evaluating the above equation at $\nu = 0$, we can get

$$\dot{M}_{n,t} = m(\varepsilon_t, \varepsilon_t)\left[\sum_{i=0}^{n} C_n^i \frac{(H_t)^{n-i}(H_t + 2\kappa(\varepsilon_t))^i}{2^n (H_t + \kappa(\varepsilon_t))^n} M_{n-i,t} M_{i,t} - M_{n,t}\right].$$

In particular, by definition we have $M_{0,t} = \int f_t(k)\, dk = 1$ and $M_{1,t} = \int k f_t(k)\, dk = K$. Moreover, the second moment of reserve distribution satisfies

$$\dot{M}_{2,t} = m(\varepsilon_t, \varepsilon_t)\left[-\frac{H_t(H_t + 2\kappa(\varepsilon_t))}{2(H_t + \kappa(\varepsilon_t))^2} M_{2,t} + \frac{H_t(H_t + 2\kappa(\varepsilon_t))}{2(H_t + \kappa(\varepsilon_t))^2} K^2\right].$$

Solving this first-order ODE gives rise to the solution (21). **Q.E.D.**

C.9 Derivation of Equation (25)

Following Üslü (2019), the planner's current-value Hamiltonian can be written as

$$\begin{aligned}\mathcal{H}_t^p \equiv & \int u(k)\, dF_t^p(k) - \int\int \chi\left[\varepsilon_t^p(k), q_t^p(k,k')\right] m\left[\varepsilon_t^p(k), \varepsilon_t^p(k')\right] dF_t^p(k')\, dF_t^p(k) \\ & + \int\int m\left[\varepsilon_t^p(k), \varepsilon_t^p(k')\right]\left\{V_t^p\left[k + q_t^p(k,k')\right] - V_t^p(k)\right\} dF_t^p(k')\, dF_t^p(k) \\ & + \int\int \eta_t(k,k')\left[q_t^p(k,k') + q_t^p(k',k)\right] dF_t^p(k')\, dF_t^p(k).\end{aligned} \tag{C.48}$$

Applying the feasibility condition of asset reallocation,

$$q_t^p(k,k') + q_t^p(k',k) = 0,$$

and the symmetry of matching function,

$$m(\varepsilon, \varepsilon') = m(\varepsilon', \varepsilon),$$

we can rewrite the Hamiltonian as

$$\begin{aligned}\mathcal{H}_t^p \equiv & \int u(k)\, dF_t^p(k) - \frac{1}{2}\int\int \chi\left[\varepsilon_t^p(k), q_t^p\left(k,k'\right)\right] m\left[\varepsilon_t^p(k), \varepsilon_t^p\left(k'\right)\right] dF_t^p\left(k'\right) dF_t^p(k) \\ & -\frac{1}{2}\int\int \chi\left[\varepsilon_t^p\left(k'\right), -q_t^p\left(k,k'\right)\right] m\left[\varepsilon_t^p\left(k'\right), \varepsilon_t^p(k)\right] dF_t^p\left(k'\right) dF_t^p(k) \\ & +\frac{1}{2}\int\int m\left[\varepsilon_t^p(k), \varepsilon_t^p\left(k'\right)\right]\left\{V_t^p\left[k+q_t^p\left(k,k'\right)\right] - V_t^p(k)\right\} dF_t^p\left(k'\right) dF_t^p(k) \\ & +\frac{1}{2}\int\int m\left[\varepsilon_t^p\left(k'\right), \varepsilon_t^p(k)\right]\left\{V_t^p\left[k'-q_t^p\left(k',k\right)\right] - V_t^p\left(k'\right)\right\} dF_t^p\left(k'\right) dF_t^p(k) \\ = & \int u(k)\, dF_t^p(k) + \frac{1}{2}\int\int S_t^p\left(k,k',\varepsilon_t^p(k),\varepsilon_t^p\left(k'\right)\right) m\left[\varepsilon_t^p(k),\varepsilon_t^p\left(k'\right)\right] dF_t^p\left(k'\right) dF_t^p(k) \end{aligned} \tag{C.49}$$

where

$$\begin{aligned} S_t^p\left(k,k',\varepsilon,\varepsilon'\right) = & V_t^p\left[k+q_t^p\left(k,k'\right)\right] - V_t^p(k) - \chi\left[\varepsilon, q_t^p\left(k,k'\right)\right] \\ & + V_t^p\left[k'-q_t^p\left(k,k'\right)\right] - V_t^p\left(k'\right) - \chi\left[\varepsilon', -q_t^p\left(k,k'\right)\right]. \end{aligned} \tag{C.50}$$

Optimality condition for $\{\varepsilon_t^p(k)\}_{k\in\mathbb{K}}$ and $\{q_t^p(k,k')\}_{k,k'\in\mathbb{K}}$. Since the planner solution maximizes the Hamiltonian, the optimal search intensity profile must maximize the second term of (C.49), i.e.

$$\{\varepsilon_t^p(k)\}_{k\in\mathbb{K}} \equiv \arg\max_{\{\varepsilon(k)\}_{k\in\mathbb{K}}\in[0,1]^{\mathbb{K}}} \int\int S_t^p\left[k,k',\varepsilon(k),\varepsilon\left(k'\right)\right] m\left[\varepsilon(k),\varepsilon\left(k'\right)\right] dF_t^p\left(k'\right) dF_t^p(k).$$

Moreover, due to the feasiblity condition of asset reallocation, the optimal asset reallocation profile must maximize the bilateral joint surplus, i.e.

$$q_t^p\left(k,k'\right) = \arg\max_q \left\{V_t^p(k+q) + V_t^p\left(k'-q\right) - \chi\left(\varepsilon_t^e(k), q\right) - \chi\left(\varepsilon_t^e\left(k'\right), -q\right)\right\}.$$

ODE for co-state variables. Following Üslü (2019), the co-state variables must satisfy the following ODEs in an optimum:

$$\nabla_{n_t(k)}\mathcal{H}_t^p\left(F_t^p\right) = rV_t^p(k) - \dot{V}_t^p(k),$$

where $n_t(k)$ is the degenerate measure with all the probability on bank k, and ∇_n denotes the Gâteaux differential in the direction of measure n:

$$\nabla_n\mathcal{H}_t^p\left(F_t^p\right) = \lim_{\alpha\to 0}\frac{\mathcal{H}_t^p\left(F_t^p+\alpha\cdot n\right) - \mathcal{H}_t^p\left(F_t^p\right)}{\alpha}.$$

For small α, we obtain up to second-order terms:

$$\mathcal{H}_t^p\left(F_t^p+\alpha\cdot n\right)-\mathcal{H}_t^p\left(F_t^p\right) \tag{C.51}$$

$$=\alpha\int u\left(k\right)dn_t\left(k\right)+\frac{\alpha}{2}\int\int S_t^p\left(k,k',\varepsilon_t^p\left(k\right),\varepsilon_t^p\left(k'\right)\right)m\left[\varepsilon_t^p\left(k\right),\varepsilon_t^p\left(k'\right)\right]dF_t^p\left(k'\right)dn_t\left(k\right) \tag{C.52}$$

$$+\frac{\alpha}{2}\int\int S_t^p\left(k,k',\varepsilon_t^p\left(k\right),\varepsilon_t^p\left(k'\right)\right)m\left[\varepsilon_t^p\left(k\right),\varepsilon_t^p\left(k'\right)\right]dn_t\left(k'\right)dF_t^p\left(k\right),$$

Thus, by applying the symmetry of $S_t^p\left(k,k',\varepsilon_t^p\left(k\right),\varepsilon_t^p\left(k'\right)\right)$ and $m\left[\varepsilon_t^p\left(k\right),\varepsilon_t^p\left(k'\right)\right]$ w.r.t. k and k',

$$\nabla_{n_t(k)}\mathcal{H}_t^p\left(F_t^p\right)=u\left(k\right)+\int S_t^p\left(k,k',\varepsilon_t^p\left(k\right),\varepsilon_t^p\left(k'\right)\right)m\left[\varepsilon_t^p\left(k\right),\varepsilon_t^p\left(k'\right)\right]dF_t^p\left(k'\right).$$

Therefore, the ODE for the co-state variables in an optimum is

$$rV_t^p\left(k\right)=\dot{V}_t^p\left(k\right)+u\left(k\right)+\int S_t^p\left(k,k',\varepsilon_t^p\left(k\right),\varepsilon_t^p\left(k'\right)\right)m\left[\varepsilon_t^p\left(k\right),\varepsilon_t^p\left(k'\right)\right]dF_t^p\left(k'\right).$$

First-order conditions. First, take any optimal q_t^e and

$$\begin{aligned}\hat{q}_t\left(k,k'\right)&=q_t^e\left(k,k'\right)+\alpha_q\mathbf{1}\left\{V_t^e\left(k\right)>V_t^e\left(k'\right)\right\}-\alpha_q\mathbf{1}\left\{V_t^e\left(k\right)<V_t^e\left(k'\right)\right\}\\&=q_t^e\left(k,k'\right)+\alpha_q\Delta_t\left(k,k'\right),\end{aligned} \tag{C.53}$$

where α_q is an arbitrary scalar. Second, take any optimal $\varepsilon_t^e\left(k\right)$, an arbitrary admissible deviation $\delta_t\left(k\right)$ and a scalar α_ε, let

$$\hat{\varepsilon}_t\left(k\right)=\varepsilon_t^e\left(k\right)+\alpha_\varepsilon\cdot\delta_t\left(k\right).$$

For small α_q and α_ε, we obtain up to second-order terms:

$$\mathcal{H}_t^p\left(\hat{\varepsilon}_t,\hat{q}_t\right)-\mathcal{H}_t^p\left(\varepsilon_t^e,q_t^e\right) \tag{C.54}$$

$$=-\alpha_\varepsilon\int\int\left\{\begin{array}{c}\chi_1\left[\varepsilon_t^e(k),q_t^e(k,k')\right]m\left[\varepsilon_t^e(k),\varepsilon_t^e(k')\right]\delta_t(k)\\+\chi\left[\varepsilon_t^e(k),q_t^e(k,k')\right]m_1\left[\varepsilon_t^e(k),\varepsilon_t^e(k')\right]\delta_t(k)\\+\chi\left[\varepsilon_t^e(k),q_t^e(k,k')\right]m_2\left[\varepsilon_t^e(k),\varepsilon_t^e(k')\right]\delta_t(k')\end{array}\right\}dF_t^p\left(k'\right)dF_t^p(k) \tag{C.55}$$

$$+\alpha_\varepsilon\int\int\left\{\begin{array}{c}m_1\left[\varepsilon_t^e(k),\varepsilon_t^e(k')\right]\delta_t(k)\\+m_2\left[\varepsilon_t^e(k),\varepsilon_t^e(k')\right]\delta_t(k')\end{array}\right\}\left\{V_t^p\left[k+q_t^e\left(k,k'\right)\right]-V_t^p(k)\right\}dF_t^p\left(k'\right)dF_t^p(k) \tag{C.56}$$

$$-\alpha_q\int\int\chi_2\left[\varepsilon_t^e(k),q_t^e\left(k,k'\right)\right]m\left[\varepsilon_t^e(k),\varepsilon_t^e\left(k'\right)\right]\Delta_t\left(k,k'\right)dF_t^p\left(k'\right)dF_t^p(k) \tag{C.57}$$

$$+\alpha_q\int\int m\left[\varepsilon_t^e(k),\varepsilon_t^e\left(k'\right)\right]V_t^{p\prime}\left[k+q_t^e\left(k,k'\right)\right]\Delta_t\left(k,k'\right)dF_t^p\left(k'\right)dF_t^p(k) \tag{C.58}$$

$$+\alpha_q\int\int\eta_t\left(k,k'\right)\left[\Delta_t\left(k,k'\right)+\Delta_t\left(k',k\right)\right]dF_t^p\left(k'\right)dF_t^p(k) \tag{C.59}$$

$$=\alpha_\varepsilon\int\int\left\{\begin{array}{c}m_1\left[\varepsilon_t^e(k),\varepsilon_t^e(k')\right]\left\{V_t^p\left[k+q_t^e(k,k')\right]-V_t^p(k)\right\}\\+m_2\left[\varepsilon_t^e(k'),\varepsilon_t^e(k)\right]\left\{V_t^p\left[k'-q_t^e(k,k')\right]-V_t^p(k')\right\}\\-\chi_1\left[\varepsilon_t^e(k),q_t^e(k,k')\right]m\left[\varepsilon_t^e(k),\varepsilon_t^e(k')\right]\\-\chi\left[\varepsilon_t^e(k),q_t^e(k,k')\right]m_1\left[\varepsilon_t^e(k),\varepsilon_t^e(k')\right]\\-\chi\left[\varepsilon_t^e(k'),-q_t^e(k,k')\right]m_2\left[\varepsilon_t^e(k'),\varepsilon_t^e(k)\right]\end{array}\right\}\delta_t(k)\,dF_t^p\left(k'\right)dF_t^p(k) \tag{C.60}$$

$$+\frac{\alpha_q}{2}\int\int m\left[\varepsilon_t^e(k),\varepsilon_t^e\left(k'\right)\right]\left\{V_t^{p\prime}\left[k+q_t^e\left(k,k'\right)\right]-\chi_2\left[\varepsilon_t^e(k),q_t^e\left(k,k'\right)\right]\right\}\Delta_t\left(k,k'\right)dF_t^p\left(k'\right)dF_t^p(k) \tag{C.61}$$

$$+\frac{\alpha_q}{2}\int\int m\left[\varepsilon_t^e(k),\varepsilon_t^e\left(k'\right)\right]\left\{V_t^{p\prime}\left[k'+q_t^e\left(k',k\right)\right]-\chi_2\left[\varepsilon_t^e\left(k'\right),q_t^e\left(k',k\right)\right]\right\}\Delta_t\left(k',k\right)dF_t^p\left(k'\right)dF_t^p(k) \tag{C.62}$$

$$=\alpha_\varepsilon\int\int\left\{\begin{array}{c}m_1\left[\varepsilon_t^e(k),\varepsilon_t^e(k')\right]\left\{V_t^p\left[k+q_t^p(k,k')\right]-V_t^p(k)\right\}\\+m_1\left[\varepsilon_t^e(k),\varepsilon_t^e(k')\right]\left\{V_t^p\left[k'-q_t^p(k,k')\right]-V_t^p(k')\right\}\\-\chi_1\left[\varepsilon_t^e(k),q_t^p(k,k')\right]m\left[\varepsilon_t^e(k),\varepsilon_t^e(k')\right]\\-\chi\left[\varepsilon_t^e(k),q_t^p(k,k')\right]m_1\left[\varepsilon_t^e(k),\varepsilon_t^e(k')\right]\\-\chi\left[\varepsilon_t^e(k'),-q_t^p(k,k')\right]m_1\left[\varepsilon_t^e(k),\varepsilon_t^e(k')\right]\end{array}\right\}\delta_t(k)\,dF_t^p\left(k'\right)dF_t^p(k) \tag{C.63}$$

$$+\frac{\alpha_q}{2}\int\int m\left[\varepsilon_t^e(k),\varepsilon_t^e\left(k'\right)\right]\left\{\begin{array}{c}V_t^{p\prime}\left[k+q_t^e(k,k')\right]-V_t^{p'}\left[k'-q_t^e(k,k')\right]\\-\chi_2\left[\varepsilon_t^e(k),q_t^e(k,k')\right]+\chi_2\left[\varepsilon_t^e(k'),-q_t^e(k,k')\right]\end{array}\right\}\Delta_t\left(k,k'\right)dF_t^p\left(k'\right)dF_t^p(k),$$

where we apply $\Delta_t(k,k')+\Delta_t(k',k)=0$ in the second and third equality and $q_t^e(k,k')+q_t^e(k',k)=0$ in the third equality.

If $\{\varepsilon_t^e,q_t^e\}$ is optimal, this must be negative. Thus the integrand in the second term must be zero everywhere. Then the FOC for $q_t^e(k,k')$ becomes

$$V_t^{p\prime}\left[k+q_t^e\left(k,k'\right)\right]-V_t^{p'}\left[k'-q_t^e\left(k,k'\right)\right]-\chi_2\left[\varepsilon_t^e(k),q_t^e\left(k,k'\right)\right]+\chi_2\left[\varepsilon_t^e\left(k'\right),-q_t^e\left(k,k'\right)\right]=0.$$

In other words, $q_t^p(k,k')$ is the solution to

$$q_t^p\left(k,k'\right)=\arg\max_q\left\{V_t^p\left(k+q\right)+V_t^p\left(k'-q\right)-\chi\left(\varepsilon_t^e\left(k\right),q\right)-\chi\left(\varepsilon_t^e\left(k'\right),-q\right)\right\}.$$

Moreover, for the FOC of ε_t^e, since $\delta_t(k)$ is an arbitrary admissible deviation, we must have

$$m_1\left[\varepsilon_t^e\left(k\right),\varepsilon_t^e\left(k'\right)\right]\left\{\begin{array}{c}V_t^p\left[k+q_t^p\left(k,k'\right)\right]-V_t^p\left(k\right)-\chi\left[\varepsilon_t^e\left(k\right),q_t^p\left(k,k'\right)\right]\\+V_t^p\left[k'-q_t^p\left(k,k'\right)\right]-V_t^p\left(k'\right)-\chi\left[\varepsilon_t^e\left(k'\right),-q_t^p\left(k,k'\right)\right]\end{array}\right\} \tag{C.64}$$

$$-\chi_1\left[\varepsilon_t^e\left(k\right),q_t^p\left(k,k'\right)\right]m\left[\varepsilon_t^e\left(k\right),\varepsilon_t^e\left(k'\right)\right] \tag{C.65}$$

$$\left\{\begin{array}{ll}\leq 0, & \text{if } \varepsilon_t^e\left(k\right)=0,\\ =0, & \text{if } \varepsilon_t^e\left(k\right)\in(0,1),\ ,\\ \geq 0, & \text{if } \varepsilon_t^e\left(k\right)=1.\end{array}\right.$$

Thus the constrained efficiency solution of ε_t^p must satisfy

$$\Gamma_t^p\left(\varepsilon_t^p\right)(k)\equiv\arg\max_{\varepsilon\in[0,1]}\left\{\int S_t^p\left(k,k',\varepsilon,\varepsilon_t^p\left(k'\right)\right)m\left[\varepsilon,\varepsilon_t^p\left(k'\right)\right]dF_t\left(k'\right)\right\},$$

where

$$\begin{aligned}S_t^p\left(k,k',\varepsilon,\varepsilon'\right)=V_t^p\left[k+q_t^p\left(k,k'\right)\right]-V_t^p\left(k\right)-\chi\left[\varepsilon,q_t^p\left(k,k'\right)\right]\\+V_t^p\left[k'-q_t^p\left(k,k'\right)\right]-V_t^p\left(k'\right)-\chi\left[\varepsilon',-q_t^p\left(k,k'\right)\right].\end{aligned} \tag{C.66}$$

C.10 Derivations of positive measures of liquidity

Price impact. Note that the terms of trade between k and k' are

$$1+\rho_t\left(k,k'\right)=e^{r(T+\Delta-t)}\left[E_t-H_t\left(k+k'\right)\right],$$

$$q_t\left(k,k'\right)=\frac{H_t\left(k'-k\right)}{2\left(\kappa\left(\varepsilon_t\right)+H_t\right)}\Rightarrow k'=k+\frac{2\left(\kappa\left(\varepsilon_t\right)+H_t\right)}{H_t}q_t\left(k,k'\right).$$

Therefore, given k and q, we can infer the reserve holding of the counterparty $k'(k,q)$. Thus the Federal funds rate of a bank k that trades reserves q is given by

$$\log(1+\rho_t(k,q)) = r(T+\Delta-t) + \log\left[E_t - H_t\left(k+k'(k,q)\right)\right] \tag{C.67}$$

$$= r(T+\Delta-t) + \log\left[E_t - 2kH_t\right] + \log\left[1 - \frac{2(\kappa(\varepsilon_t)+H_t)}{E_t-2kH_t}q\right] \tag{C.68}$$

$$\approx r(T+\Delta-t) + \log\left[V_t'(k)\right] - \frac{2(\kappa(\varepsilon_t)+H_t)}{V_t'(k)}q \tag{C.69}$$

$$= r(T+\Delta-t) + \log\left[V_t'(k)\right] - \frac{2(\kappa(\varepsilon_t)+H_t)}{-2H_t}\frac{q}{k}\frac{kV_t''(k)}{V_t'(k)} \tag{C.70}$$

$$= r(T+\Delta-t) + \log\left[V_t'(k)\right] + \frac{V_t''(k)}{V_t'(k)}\cdot q \cdot \frac{1}{1-\left(1-\frac{\bar{V}''}{2\kappa(\varepsilon_t)}\right)^{-1}}$$

Denote $\theta_{V,t}(k) \equiv -\frac{V_t''(k)}{V_t'(k)}$ and $\omega_t \equiv \left(1-\frac{\bar{V}''}{2\kappa(\varepsilon_t)}\right)^{-1}$, we get equation (148).

Return reversal. The average Federal funds rate is

$$1+\varrho_t = e^{r(T+\Delta-t)}\left[E_t - 2H_tK\right], \tag{C.71}$$

then the difference between individual Federal funds rate and the average Federal funds rate is

$$\rho_t\left(k,k'\right) - \varrho_t = e^{r(T+\Delta-t)}\left(2K-k-k'\right)H_t.$$

Differentiating the rates with respect to time:

$$\dot{\varrho}_t = e^{r(T+\Delta-t)}\left(\dot{E}_t - 2K\dot{H}_t\right) - r(1+\varrho_t) = e^{r(T+\Delta-t)}\left(2a_2K - a_1\right), \tag{C.72}$$

$$\dot{\rho}_t\left(k,k'\right) = e^{r(T+\Delta-t)}\left[\dot{E}_t - \dot{H}_t\left(k+k'\right)\right] - r\left(1+\rho_t\left(k,k'\right)\right) \tag{C.73}$$

$$= e^{r(T+\Delta-t)}\left[-a_1 + \left(k+k'\right)a_2 + \frac{2K-k-k'}{4}\frac{H_t^2 m(\varepsilon_t,\varepsilon_t)}{\kappa(\varepsilon_t)+H_t}\right].$$

This implies

$$\frac{d}{dt}\left[\rho_t\left(k,k'\right) - \varrho_t\right] = e^{r(T+\Delta-t)}\left(2K-k-k'\right)\left[-a_2 + \frac{1}{4}\frac{H_t^2 m(\varepsilon_t,\varepsilon_t)}{\kappa\varepsilon_t+H_t}\right] \tag{C.74}$$

$$= -\left[\frac{a_2}{H_t} - \frac{1}{4}\frac{H_t m(\varepsilon_t,\varepsilon_t)}{\kappa\varepsilon_t+H_t}\right]\left[\rho_t\left(k,k'\right) - \varrho_t\right].$$

Price dispersion. The standard deviation of the bilateral Federal funds rates is

$$\sigma_{\rho,t} = \left\{ \int \int \left[\rho_t \left(k, k' \right) - \underline{\rho}_t \right]^2 dF_t \left(k' \right) dF_t \left(k \right) \right\}^{1/2} \tag{C.75}$$

$$= e^{r(T+\Delta-t)} H_t \left[\int \int \left(2K - k - k' \right)^2 dF_t \left(k' \right) dF_t \left(k \right) \right]^{1/2} \tag{C.76}$$

$$= e^{r(T+\Delta-t)} H_t \left\{ \int \int \left[(K-k)^2 + \left(K - k' \right)^2 + 2 \left(K - k \right) \left(K - k' \right) \right] dF_t \left(k' \right) dF_t \left(k \right) \right\}^{1/2} \tag{C.77}$$

$$= e^{r(T+\Delta-t)} H_t \cdot \sqrt{2} \sigma_{k,t}.$$

This gives our measure of price dispersion.

Intermediation markup. By definition, the rate spread is

$$\Delta_{\rho,t} \left(k, q \right) \tag{C.78}$$

$$\equiv \int \rho_t \left(k + q, k' \right) dF_t \left(k' \right) - \rho_t \left(k, q \right) \tag{C.79}$$

$$= \int e^{r(T+\Delta-t)} \left[E_t - H_t \left(k + q + k' \right) \right] dF_t \left(k' \right) - e^{r(T+\Delta-t)} \left[E_t - H_t \left(k + k + \frac{2 \left(\kappa \left(\varepsilon_t \right) + H_t \right)}{H_t} q \right) \right] \tag{C.80}$$

$$= e^{r(T+\Delta-t)} \left[-H_t \left(k + q + K \right) + 2kH_t + 2 \left(\kappa \left(\varepsilon_t \right) + H_t \right) q \right] \tag{C.81}$$

$$= e^{r(T+\Delta-t)} \left[-H_t \left(K - k \right) + \left(2\kappa \left(\varepsilon_t \right) + H_t \right) q \right].$$

Thus the intermediation markup is given by taking $\Delta_{\rho,t} \left(k, q \right)$ differentiation with respect to q.

Unilization rate of trade opportunities. By definition,

$$UR_t = \frac{\int_k \int_{k' \geq k} m \left(\varepsilon_t, \varepsilon_t \right) q_t \left(k, k' \right) dF_t \left(k' \right) dF_t \left(k \right)}{TO_t} \tag{C.82}$$

$$= \frac{\int_k \int_{k' \geq k} \frac{H_t (k' - k)}{2(\kappa(\varepsilon_t) + H_t)} m \left(\varepsilon_t, \varepsilon_t \right) dF_t \left(k' \right) dF_t \left(k \right)}{TO_t} \tag{C.83}$$

$$= \frac{H_t m \left(\varepsilon_t, \varepsilon_t \right)}{2 \left(\kappa \left(\varepsilon_t \right) + H_t \right)} \frac{\int_k \int_{k' \geq k} \left(k' - k \right) dF_t \left(k' \right) dF_t \left(k \right)}{TO_t} \tag{C.84}$$

$$= \frac{H_t m \left(\varepsilon_t, \varepsilon_t \right)}{\kappa \left(\varepsilon_t \right) + H_t}.$$

Extensive margins. We provide a heuristic approach to derive the dynamics of the extensive margins. Let Δ be a small time length, and denote $m_t \equiv m \left(\varepsilon_t, \varepsilon_t \right)$ as the equilibrium matching rate. Then by definition,

$$1 - P_t^{0\prime} \left(k \right) = \left(1 - \Delta \cdot m_t \right) \left[1 - P_{t+\Delta}^{0\prime} \left(k \right) \right] + \Delta \cdot m_t \cdot 0,$$

where $1-P_t^{0\prime}(k)$ denotes the probability of no trade over $[t,T]$ conditional on $k_t=k$, $1-\Delta\cdot m_t$ represents the probability of no meetings during $[t,t+\Delta]$, and 0 means the probability of no trade is 0 given a meeting arrives at t. Take $\Delta\to 0$, we can obtain

$$\dot{P}_t^{0\prime}(k)=\lim_{\Delta\to 0}\frac{P_{t+\Delta}^{0\prime}(k)-P_t^{0\prime}(k)}{\Delta}=-m_t\left[1-P_t^{0\prime}(k)\right].$$

The evolution of $P_t^b(k)$ and $P_t^s(k)$ can be derived similarly as follows.

$$1-P_t^b(k)=(1-\Delta\cdot m_t)\left[1-P_{t+\Delta}^b(k)\right]+\Delta\cdot m_t\int_{k'\le k}\left[1-P_{t+\Delta}^b\left(k+q_t\left(k,k'\right)\right)\right]dF_t\left(k'\right),$$

$$1-P_t^s(k)=(1-\Delta\cdot m_t)\left[1-P_{t+\Delta}^s(k)\right]+\Delta\cdot m_t\int_{k'\ge k}\left[1-P_{t+\Delta}^s\left(k+q_t\left(k,k'\right)\right)\right]dF_t\left(k'\right).$$

Take $\Delta\to 0$ gives

$$\dot{P}_t^b(k)=-m_t\left[1-F_t(k)\right]\left[1-P_t^b(k)\right]-m_t\int_{k'\le k}\left[P_t^b\left(k+q_t\left(k,k'\right)\right)-P_t^b(k)\right]dF_t\left(k'\right),$$

$$\dot{P}_t^s(k)=-m_tF_t(k)\left[1-P_t^s(k)\right]-m_t\int_{k'\ge k}\left[P_t^s\left(k+q_t\left(k,k'\right)\right)-P_t^s(k)\right]dF_t\left(k'\right).$$

Then the evolution of $P_t^{int}(k)$ is

$$\dot{P}_t^{int}(k)=\dot{P}_t^b(k)+\dot{P}_t^s(k)-\dot{P}_t^{0\prime}(k) \tag{C.85}$$

$$=-m_t\int_{k'\le k}\left[P_t^b\left(k+q_t\left(k,k'\right)\right)-P_t^{int}(k)\right]dF_t\left(k'\right) \tag{C.86}$$

$$-m_t\int_{k'\ge k}\left[P_t^s\left(k+q_t\left(k,k'\right)\right)-P_t^{int}(k)\right]dF_t\left(k'\right).$$

Intensive margins. We provide an heuristic derivation of the absolute trades and net rades. First, for the individual absolute trades, let Δ be an infinitesimal time period. Then by the property of Poisson process,

$$Q_t(k)=\Delta\cdot m(\varepsilon_t,\varepsilon_t)\cdot\left[\int_{k'}\left|q_t\left(k,k'\right)\right|dF_t\left(k'\right)+\int_{k'}Q_{t+\Delta}\left(k+q_t\left(k,k'\right)\right)dF_t\left(k'\right)\right] \tag{C.87}$$

$$+\left[1-\Delta\cdot m(\varepsilon_t,\varepsilon_t)\right]Q_{t+\Delta}(k).$$

Thus the aggregate absolute trades is given by

$$
\begin{aligned}
Q_t &= \int Q_t(k)\, dF_t(k) && \text{(C.88)}\\
&= \Delta \cdot m(\varepsilon_t, \varepsilon_t) \cdot \int_k \int_{k'} \left| q_t\left(k, k'\right) \right| dF_t\left(k'\right) dF_t(k) && \text{(C.89)}\\
&\quad + \int_k \left\{ \Delta \cdot m(\varepsilon_t, \varepsilon_t) \cdot \int_{k'} Q_{t+\Delta}\left(k + q_t\left(k, k'\right)\right) dF_t\left(k'\right) + \left[1 - \Delta \cdot m(\varepsilon_t, \varepsilon_t)\right] Q_{t+\Delta}(k) \right\} dF_t(k) && \text{(C.90)}\\
&= \Delta \cdot m(\varepsilon_t, \varepsilon_t) \cdot \int_k \int_{k'} \left| q_t\left(k, k'\right) \right| dF_t\left(k'\right) dF_t(k) + Q_{t+\Delta}, && \text{(C.91)}
\end{aligned}
$$

where the last equality is given by the definition of $Q_t(k)$ and Q_t. Taking $\Delta \to 0$, we can obtain the following ODEs for $Q_t(k)$ and Q_t:

$$
\begin{aligned}
\dot{Q}_t(k) &= \lim_{\Delta \to 0} \frac{Q_{t+\Delta}(k) - Q_t(k)}{\Delta} && \text{(C.92)}\\
&= -m(\varepsilon_t, \varepsilon_t) \cdot \left[\int_{k'} \left| q_t\left(k, k'\right) \right| dF_t\left(k'\right) + \int_{k'} Q_{t+\Delta}\left(k + q_t\left(k, k'\right)\right) dF_t\left(k'\right) \right] + m(\varepsilon_t, \varepsilon_t)\, Q_t(k), && \text{(C.93)}
\end{aligned}
$$

and

$$
\begin{aligned}
\dot{Q}_t &= \lim_{\Delta \to 0} \frac{Q_{t+\Delta} - Q_t}{\Delta} && \text{(C.94)}\\
&= -m(\varepsilon_t, \varepsilon_t) \cdot \int_k \int_{k'} \left| q_t\left(k, k'\right) \right| dF_t\left(k'\right) dF_t(k) && \text{(C.95)}\\
&= -m(\varepsilon_t, \varepsilon_t) \frac{H_t}{2(\kappa\varepsilon_t + H_t)} \int_k \int_{k'} \left| k' - k \right| dF_t\left(k'\right) dF_t(k). && \text{(C.96)}
\end{aligned}
$$

This implies

$$
Q = \int_0^T m(\varepsilon_t, \varepsilon_t) \frac{H_t}{2(\kappa\varepsilon_t + H_t)} \left(\int_k \int_{k'} \left| k' - k \right| dF_t\left(k'\right) dF_t(k) \right) dt. \tag{C.97}
$$

Second, for the individual net Federal funds purchase, it satisfies

$$
\begin{aligned}
L_t(k) &= \Delta \cdot m_t \int \frac{H_t(k' - k)}{2(\kappa\varepsilon_t + H_t)} dF_t\left(k'\right) + \Delta \cdot m_t \int L_{t+\Delta}\left(k + q_t\left(k, k'\right)\right) dF_t\left(k'\right) && \text{(C.98)}\\
&\quad + (1 - \Delta \cdot m_t)\, L_{t+\Delta}(k) && \text{(C.99)}\\
&= \Delta \cdot m_t \frac{H_t(K - k)}{2(\kappa\varepsilon_t + H_t)} + \Delta \cdot m_t \int L_{t+\Delta}\left(k + q_t\left(k, k'\right)\right) dF_t\left(k'\right) && \text{(C.100)}\\
&\quad + (1 - \Delta \cdot m_t)\, L_{t+\Delta}(k). && \text{(C.101)}
\end{aligned}
$$

We guess and verify that $L_t(k) = \Theta_{1,t} - \Theta_{2,t} k$. Plugging the guessed formula into the above equation and matching the coefficients, we can get

$$
\frac{\dot{\Theta}_{1,t}}{K} = \dot{\Theta}_{2,t} = \frac{m_t H_t}{2(\kappa\varepsilon_t + H_t)} (\Theta_{2,t} - 1).
$$

With terminal condition $\Theta_{1,T} = \Theta_{2,T} = 0$, we have the following closed-form solution:

$$\begin{aligned}\Theta_{2,t} &= 1 - \exp\left[-\int_t^T \frac{m_z H_z}{2\left(\kappa\varepsilon_z + H_z\right)} dz\right], \\ \Theta_{1,t} &= K \cdot \Theta_{2,t}.\end{aligned} \tag{C.98}$$

Thus the individual net trades is given by

$$L_t(k) = \left\{1 - \exp\left[-\int_t^T \frac{m_z H_z}{2\left(\kappa\varepsilon_z + H_z\right)} dz\right]\right\}(K - k),$$

and the aggregate net trades is

$$L = \int |L_0(k)| \, dF_0(k) = \left\{1 - \exp\left[-\int_0^T \frac{m_z H_z}{2\left(\kappa\varepsilon_z + H_z\right)} dz\right]\right\} \int |K - k| \, dF_0(k). \tag{C.99}$$

Federal funds rate. The average Federal funds rate at t is given by equation (C.71). It satisfies the ODE (C.72) with terminal condition $1 + \varrho_T = e^{r\Delta}\left[A_1 - 2A_2 K\right]$, which has the following closed-form solution:

$$1 + \varrho_t = e^{r\Delta}\left(A_1 - 2A_2 K\right) - \frac{2a_2 K - a_1}{r}\left[e^{r(T+\Delta-t)} - e^{r\Delta}\right] \tag{C.100}$$

$$= e^{r\Delta}\left[1 + \gamma + \frac{\left(k_+ - 1\right) i^{DW} - \left(k_- - 1\right) i^{ER}}{k_+ - k_-}\right] - \frac{2a_2 K - a_1}{r}\left[e^{r(T+\Delta-t)} - e^{r\Delta}\right] \tag{C.101}$$

$$= e^{r\Delta}\left[1 + \gamma + i^{ER} + \frac{k_+ - 1}{k_+ - k_-}\Delta i\right] - \frac{2a_2 K - a_1}{r}\left[e^{r(T+\Delta-t)} - e^{r\Delta}\right].$$

When r is closed to zero, the spread between the federal funds rate and IOER is

$$\varrho_t - i^{ER} = \gamma + \frac{k_+ - 1}{k_+ - k_-}\Delta i - \left(2a_2 K - a_1\right)(T - t)$$

C.11 Proofs on comparative statics

To prove Proposition 6, we first present the following lemmas. Define

$$\Upsilon\left(H, \varepsilon, \varepsilon'\right) = \frac{Hm\left(\varepsilon, \varepsilon'\right)}{2H + \kappa(\varepsilon) + \kappa\left(\varepsilon'\right)}$$

and

$$\Upsilon^*(H) = \max_{\varepsilon, \varepsilon' \in [0,1]} \frac{Hm\left(\varepsilon, \varepsilon'\right)}{2H + \kappa(\varepsilon) + \kappa\left(\varepsilon'\right)}.$$

Next, we conduct comparative statics of model parameters, including $H, \lambda_C, \lambda_I, \lambda_0, \kappa(\cdot)$, on the equilibrium search intensity. Since the set $\arg\max_{\varepsilon \in [0,1]} \Upsilon(H, \varepsilon, \varepsilon)$ may not be a singleton, we apply monotone comparative statics.

Lemma 2 *The set* $\arg\max_{\varepsilon\in[0,1]}\Upsilon(H,\varepsilon,\varepsilon)$ *is increasing in* H, λ_C, λ_I, *and decreasing in* λ_0. *Moreover, if the search cost function belongs to a family* $\{\kappa(\cdot;s)\}$ *that is increasing in* s *and obeys single crossing differences, then* $\arg\max_{\varepsilon\in[0,1]}\Upsilon(H,\varepsilon,\varepsilon)$ *is decreasing in* $\kappa(\cdot)$.

Proof. According to Milgrom & Shannon (1994), it suffices to show that the family of functions $\left\{\frac{m(x,x)}{H+\kappa(x)}\right\}$ obeys single crossing differences in H, m and κ. Note that for all $\varepsilon'>\varepsilon$ and $H'>H$, we have

$$\frac{m(x',x')}{H+\kappa(x')}-\frac{m(x,x)}{H+\kappa(x)}\geq 0\Rightarrow\frac{m(x',x')}{m(x,x)}\geq\frac{H+\kappa(x')}{H+\kappa(x)}>\frac{H'+\kappa(x')}{H'+\kappa(x)}\Rightarrow\frac{m(x',x')}{H'+\kappa(x')}-\frac{m(x,x)}{H'+\kappa(x)}>0.$$

Thus $\frac{m(x,x)}{H+\kappa(x)}$ obeys single crossing differences in H.

Moreover, taking derivatives with respect to λ_C, λ_I and λ_0 implies that $\frac{m(x',x')}{m(x,x)}$ is increasing in λ_C, λ_I and $1/\lambda_0$ for all $x'>x$, it follows that

$$\frac{m(x',x')}{H+\kappa(x')}-\frac{m(x,x)}{H+\kappa(x)}\geq 0\Rightarrow\frac{H+\kappa(x')}{H+\kappa(x)}\leq\frac{m(x',x')}{m(x,x)}<\frac{\tilde{m}(x',x')}{\tilde{m}(x,x)}\Rightarrow\frac{\tilde{m}(x',x')}{H+\kappa(x')}-\frac{\tilde{m}(x,x)}{H+\kappa(x)}>0.$$

Thus $\frac{m(x,x)}{H+\kappa(x)}$ obeys single crossing differences in $(\lambda_C,\lambda_I,1/\lambda_0)$.

Finally, for any two $\kappa(\cdot;s)$ and $\kappa(\cdot;s')$ such that $s'>s$, and any $x'>x$, a sufficient condition for $\frac{m(x',x')}{H+\kappa(x';s)}-\frac{m(x,x)}{H+\kappa(x;s)}\geq 0\Rightarrow\frac{m(x',x')}{H+\kappa(x';s')}-\frac{m(x,x)}{H+\kappa(x;s')}\geq 0$ is that

$$m(x,x)\,\kappa(x';s)-m(x',x')\,\kappa(x;s)\geq m(x,x)\,\kappa(x';s')-m(x',x')\,\kappa(x;s'),$$

which is equivalent to

$$m(x',x')\left[\kappa(x;s')-\kappa(x;s)\right]\geq m(x,x)\left[\kappa(x';s')-\kappa(x';s)\right].$$

The above condition holds if

$$\kappa(x;s)\ \text{ increases in } s$$

and

$$\kappa(x';s)-\kappa(x;s)\geq\kappa(x';s')-\kappa(x;s').$$

If $\kappa(\varepsilon)=\kappa_0+\tilde{\kappa}(\varepsilon)$, then an increase in κ_0 is equivalent to an increase in H, thus $\arg\max_{\varepsilon\in[0,1]}\Upsilon(H,\varepsilon,\varepsilon)$ is increasing in κ_0.

Q.E.D.

Therefore, if the set of refined equilibria is not unique, we focus on the one with the largest search intensity, i.e. $\max\left\{\arg\max_{\varepsilon\in[0,1]}\Upsilon(H,\varepsilon,\varepsilon)\right\}$. The comparative statics of the largest search intensity follows the above lemma.

Next, we show that the time paths of H_t and ε_t are monotone, which is implied by the following lemma.

Lemma 3 *$\Upsilon^*(H)$ is continuous and increasing in H.*

Proof. By the maximum theoreom, $\Upsilon^*(H)$ is continuous in H. For the monotonicity, note that for any $\varepsilon, \varepsilon' \in [0,1]$, $\Upsilon(H, \varepsilon, \varepsilon')$ is continuously differentiable and concave in H. This guarantees that $\Upsilon(H, \varepsilon, \varepsilon')$ is absolutely continuous in H for all $\varepsilon, \varepsilon' \in [0,1]$. Moreover, note that

$$\begin{aligned} \left|\Upsilon_H\left(H, \varepsilon, \varepsilon'\right)\right| &= \left|\frac{m(\varepsilon, \varepsilon')\left[\kappa(\varepsilon) + \kappa(\varepsilon')\right]}{\left[2H + \kappa(\varepsilon) + \kappa(\varepsilon')\right]^2}\right| \\ &\leq \left|\frac{m(1,1)}{2}\frac{\kappa(1)}{H\left[H + \kappa(1)\right]}\right|, \end{aligned} \tag{C.102}$$

where $\frac{m(1,1)}{2}\frac{\kappa(1)}{H[H+\kappa(1)]}$ is integrable over an arbitrary interval $H \in \left[\underline{H}, \overline{H}\right]$. Then by Theorem 2 in Milgrom and Segal (2002), $\Upsilon^*(H)$ is absolutely continuous with the derivative

$$\Upsilon^*_H(H) = \Upsilon_H\left(H, \varepsilon^*, \varepsilon'^*\right) \geq 0$$

almost everywhere. Thus $\Upsilon^*(H)$ is increasing in H. **Q.E.D.**

Thus the right-hand side of (**??**) is monotonically increasing in H with a unique positive zero point. We also have the comparative statics of $\Upsilon^*(H)$ on function $m(\cdot,\cdot)$ and $\kappa(\cdot)$.

Lemma 4 *For any $m(\cdot,\cdot)$ and $\tilde{m}(\cdot,\cdot)$ such that $m(\varepsilon, \varepsilon') \leq \tilde{m}(\varepsilon, \varepsilon')$ for all $\varepsilon, \varepsilon' \in [0,1]$, we have $\Upsilon^*(H; m(\cdot,\cdot)) \leq \Upsilon^*(H; \tilde{m}(\cdot,\cdot))$. For any $\kappa(\cdot)$ and $\tilde{\kappa}(\cdot)$ such that $\kappa(\varepsilon) \leq \tilde{\kappa}(\varepsilon)$ for all $\varepsilon \in [0,1]$, we have $\Upsilon^*(H; \kappa(\cdot)) \geq \Upsilon^*(H; \tilde{\kappa}(\cdot))$.*

Proof. For $m(\cdot,\cdot)$, denote the solution of search intensity as $\varepsilon^*(m(\cdot,\cdot))$ and $\varepsilon'^*(m(\cdot,\cdot))$. By definition,

$$\begin{aligned} \Upsilon^*(H; m(\cdot,\cdot)) &= \Upsilon\left(H, \varepsilon^*(m(\cdot,\cdot)), \varepsilon'^*(m(\cdot,\cdot)); m(\cdot,\cdot)\right) && \text{(C.103)} \\ &\leq \Upsilon\left(H, \varepsilon^*(m(\cdot,\cdot)), \varepsilon'^*(m(\cdot,\cdot)); \tilde{m}(\cdot,\cdot)\right) && \text{(C.104)} \\ &\leq \Upsilon^*(H; \tilde{m}(\cdot,\cdot)). \end{aligned}$$

For κ, denote the solution of search intensity as $\varepsilon^*(\kappa(\cdot))$ and $\varepsilon'^*(\kappa(\cdot))$. By definition,

$$\begin{aligned} \Upsilon^*(H; \tilde{\kappa}(\cdot)) &= \Upsilon\left(H, \varepsilon^*(\tilde{\kappa}(\cdot)), \varepsilon'^*(\tilde{\kappa}(\cdot)); \tilde{\kappa}(\cdot)\right) && \text{(C.105)} \\ &\leq \Upsilon\left(H, \varepsilon^*(\tilde{\kappa}(\cdot)), \varepsilon'^*(\tilde{\kappa}(\cdot)); \kappa(\cdot)\right) && \text{(C.106)} \\ &\leq \Upsilon^*(H; \kappa(\cdot)). \end{aligned}$$

Q.E.D.

The above lemmas imply

Lemma 5 *H_t increases in A_2 and $\kappa(\cdot)$, and decreases in $m(\cdot,\cdot)$.*

Proof. Note that

$$\dot{H}_t = rH_t - a + \frac{H_t}{2}\Upsilon^*(H_t; m, \kappa).$$

By 4, $\Upsilon^*(H_t; m, \kappa)$ increases in m and decreases in κ, it follows that H_t decreases in m and increases in κ. For the former argument, note that $\dot{H}_T$ increases in m, thus in the neighborhood of T, we have H_t decreases in m. If there is a sufficiently small t such that H_t increases in m, then the two curves must interact with $\dot{H}_t$ decreases in m, contradiction. Similar arguments apply for the latter. **Q.E.D.**

Finally, we prove Proposition 6 as follows.

Proof of Proposition 6. [ε_t] For the comparative statics of ε_t w.r.t. $i^{DW} - i^{ER}$ and K, it sufficies to show that ε_t is increasing in A_2. This is established by Lemma 2 and 5, which show that H_t increases in A_2 and ε_t increases in H_t. For comparative statics of ε_t w.r.t. κ_0, note that the monotone comparative statics shows that $\arg\max_{x\in[0,1]}\left\{\frac{m(x,x)}{H+\kappa(x)}\right\}$ increases in H and κ_0. Since H also increases in κ_0, it follows that ε_t increases in κ_0.

[$|m_t q_t|$] Note that $|m_t q_t| = \max_\varepsilon\left\{\frac{H_t m(\varepsilon,\varepsilon)}{H_t+\kappa(\varepsilon)}\right\}\left|\frac{k'-k}{2}\right|$. The envelope theorem implies that $|m_t q_t|$ increases in H_t. Since H_t increases in A_2, it follows that $|m_t q_t|$ increases in $i^{DW} - i^{ER}$ and decreases in K. For the effect of κ_0 on $|m_t q_t|$, rewrite (17) as

$$\max_\varepsilon\left\{\frac{H_t m(\varepsilon,\varepsilon)}{H_t+\kappa(\varepsilon)}\right\} = 4\left(\frac{a+\dot{H}_t}{H_t} - r\right).$$

Since H_t increases in κ_0, $\dot{H}_t$ decreases in κ_0, and $a + \dot{H}_t > 0$, it follows that $\frac{a+\dot{H}_t}{H_t}$ decreases in κ_0. This implies $|m_t q_t|$ decreases in κ_0.

[$L_0(k)$] By Lemma 3 and 5 we have

$$\frac{\partial\Upsilon^*(H_t)}{\partial A_2} = \frac{\partial\Upsilon^*(H_t)}{\partial H_t}\frac{\partial H_t}{\partial A_2} > 0 \Rightarrow \frac{\partial\left\{1-\exp\left[-\int_t^T \Upsilon^*(H_z)\,dz\right]\right\}}{\partial A_2} > 0$$

Since $\frac{\partial A_2}{\partial i^{ER}} < 0$, $\frac{\partial A_2}{\partial i^{DW}} > 0$ and $\frac{\partial A_2}{\partial K} < 0$, we can get

$$sgn\left(\frac{\partial L_0(k)}{\partial i^{ER}}\right) = sgn\left(\frac{\partial A_2}{\partial i^{ER}}(K-k)\right) = sgn(k-K),$$

$$sgn\left(\frac{\partial L_0(k)}{\partial i^{DW}}\right) = sgn\left(\frac{\partial A_2}{\partial i^{DW}}(K-k)\right) = sgn(K-k).$$

For the comparative statics w.r.t. K, note that

$$\frac{\partial L_0(k)}{\partial K} = \frac{\partial\left\{1-\exp\left[-\int_t^T \Upsilon^*(H_z)\,dz\right]\right\}}{\partial A_2}\frac{\partial A_2}{\partial K}(K-k) + \left\{1-\exp\left[-\int_t^T \Upsilon^*(H_z)\,dz\right]\right\}$$

whic is positive iff

$$k > K + \frac{1-\exp\left[-\int_t^T \Upsilon^*(H_z)\,dz\right]}{\frac{\partial\left\{1-\exp\left[-\int_t^T \Upsilon^*(H_z)dz\right]\right\}}{\partial A_2}\frac{\partial A_2}{\partial K}}.$$

Next, for the comparative statics w.r.t. κ_0, it suffices to show that the equilibrium $\int_t^T \Upsilon^*(H_z)\,dz$ is increasing in m and decreasing in κ. Note that

$$\dot{H}_t = rH_t - a + \frac{H_t}{2}\Upsilon^*(H_t; m, \kappa),$$

which implies

$$\int_t^T \Upsilon^*(H_z)\,dz = 2\int_t^T \frac{\dot{H}_z - rH_z + a}{H_z}dz = 2\left\{\ln A_2 - \ln H_t - r(T-t) + \int_t^T \frac{a}{H_z}dz\right\}.$$

Since Lemma 5 shows that H_t increasing in κ_0, this gives comparative statics of $L_0(k)$. **Q.E.D.**

D Walrasian benckmark

To see the role of search intensity, consider the Walrasian benchmark where there is no search fricton ($\lambda_0 = \infty$) and trades are organized in a competitive market. Banks are free to trade at any $t \in [0, T]$, taking the competitive Federal funds rates ρ_t^w as given. It will be useful to express the bank's problem in term of its value of reserve balances, $a_t \equiv (1+\rho_t^w)k_t$. The evolution of a_t is thus given by

$$da_t = \frac{\dot{\rho}_t^w}{1+\rho_t^w}a_t dt + d\delta_t, \tag{D.107}$$

where the first term is the appreciation of the reserve value due to the appreciation of Federal funds rate and the second term, δ_t, is the value of Federal fund purchased up to t. Notice that we allow $d\delta_t$ to be infinitesimal or lumpy. At $T+\Delta$ the bank will settle the accumulated Federal funds purchased, which is δ_T. Similar to (3), given the path of competitive Federal funds rates $\{\rho_t^w\}$, the bank problem is given by

$$\max_{\{\delta_t\}} \mathbb{E}\left\{\int_0^T e^{-rt}u\left(\frac{a_t}{1+\rho_t^w}\right)dt + e^{-rT}U\left(\frac{a_T}{1+\rho_T^w}\right) - e^{-r(T+\Delta)}\delta_T\right\}, \text{ s.t. (D.107)}. \tag{D.108}$$

Denote $\delta_t(a_0)$ as the solution chosen at t by a bank that holds a_0 units of reserve value at $t=0$. In the competitive equilibrium, ρ_t^w clears the market clearing such that for all t

$$0 = \int \delta_t\left[\left(1+\rho_0^w\right)k\right] dF_0(k). \tag{D.109}$$

Proposition 7 *In the competitive equilibrium, we have*
(a) $\rho_t^w = e^{r\Delta}\left\{U'(K) + \left[e^{r(T-t)}-1\right]\frac{u'(K)}{r}\right\} - 1$;
(b) $\delta_t(a) = \left(1+\rho_0^w\right)K - a$ *for all* $t \in [0,T]$.

In the Walrasian benchmark, banks trade instantaneously at $t=0$ such that every bank maintains K units of reserve balances throughtout the horizon. In the competitive equilibrium, the Federal funds rate is decreasing over time, in order to compensate for the utility from holding reserve. Also, notice that the Walrasian benchmark is the first-best allocation.

E Extension: Heterogeneous agents with peripheral traders

We guess and verify the closed-form solutions. First, we guess the banks' value function is $V_t(k) = -H_t k^2 + E_t k + D_t$, and the peripheral trader's value function is $\tilde{V}_t\left(\tilde{k}\right) = -\tilde{H}_t\tilde{k}^2 + \tilde{E}_t\tilde{k} + \tilde{D}_t$. Since the matching between banks is independent of the matching between a bank and a peripheral trader, banks are only choosing the search intensities in contacting other banks, we can impose the same refinement and obtain the same equilibrium solution to the bank meetings as the baseline model.

For the meetings between a bank and a peripheral trader, they solve

$$\begin{aligned}&\max_{R,q}\left[V_t(k+q) - e^{-r(T-t+\Delta)}R - V_t(k) - \chi(0,q)\right]^{\theta} \\ &\times\left[\tilde{V}_t\left(\tilde{k}-q\right) + e^{-r(T-t+\Delta)}R - \tilde{V}_t\left(\tilde{k}\right)\right]^{1-\theta}.\end{aligned} \tag{E.110}$$

The maximized surplus and optimal trade size are given by

$$\tilde{S}_t\left(k,\tilde{k}\right) = \frac{\left[E_t - \tilde{E}_t + 2\left(\tilde{H}_t\tilde{k} - H_t k\right)\right]^2}{4\left[H_t + \tilde{H}_t + \kappa_0\right]}, \tag{E.111}$$

$$\tilde{q}_t\left(k,\tilde{k}\right) = \frac{E_t - \tilde{E}_t + 2\left(\tilde{H}_t\tilde{k} - H_t k\right)}{2\left[H_t + \tilde{H}_t + \kappa_0\right]}. \tag{E.112}$$

Therefore, the HJB for peripheral traders is

$$r\tilde{V}_t\left(\tilde{k}\right) = \dot{\tilde{V}}_t\left(\tilde{k}\right) + (1-\theta)\varphi\int \tilde{S}_t\left(k,\tilde{k}\right) dF_t(k).$$

By matching coefficients we can obtain

$$\dot{\tilde{H}}_t = r\tilde{H}_t + \frac{(1-\theta)\,\varphi\tilde{H}_t^2}{H_t + \tilde{H}_t + \kappa_0}, \text{ with } \tilde{H}_T = 0; \tag{E.113}$$

$$\dot{\tilde{E}}_t = r\tilde{E}_t - (1-\theta)\,\varphi\tilde{H}_t\frac{E_t - \tilde{E}_t - 2H_tK_t}{H_t + \tilde{H}_t + \kappa_0}, \text{ with } \tilde{E}_T = 1 + i^{RRP}; \tag{E.114}$$

$$\dot{\tilde{D}}_t = r\tilde{D}_t - (1-\theta)\,\varphi\int\frac{\left[E_t - \tilde{E}_t - 2H_tk\right]^2}{4\left[H_t + \tilde{H}_t + \kappa_0\right]}dF_t(k), \text{ with } \tilde{D}_t = 0.$$

Given $\tilde{H}_T = 0$, $\tilde{E}_T = 1 + i^{RRP}$, we can get that $\tilde{H}_t \equiv 0$ and $\tilde{E}_t = \left(1 + i^{RRP}\right)e^{-r(T-t)}$. Thus the bilateral Federal funds rate in a meeting between bank and peripheral trader is

$$\begin{aligned} 1 + \tilde{\rho}_t\left(k, \tilde{k}\right) &= e^{r(T+\Delta-t)}\left[\frac{1-\theta}{2}\left(E_t - \tilde{E}_t - 2H_tk\right) + \tilde{E}_t\right] \\ &= e^{r(T+\Delta-t)}\left[(1-\theta)\,(H_t + \kappa_0)\,\tilde{q}_t\left(k, \tilde{k}\right) + \tilde{E}_t\right] \end{aligned} \tag{E.115}$$

On the other hand, the HJB for banks is

$$rV_t(k) = \dot{V}_t(k) + u(k) + \int\frac{1}{2}S_t\left(k, k'\right)m\left(\varepsilon_t, \varepsilon_t\right)dF_t\left(k'\right) + \theta\varphi\vartheta\int\tilde{S}_t\left(k, \tilde{k}\right)d\tilde{F}_t\left(\tilde{k}\right),$$

which implies

$$\dot{H}_t = rH_t - a_2 + \frac{H^2}{4}\max_{x\in[0,1]}\left\{\frac{m(x,x)}{H_t + \kappa(x)}\right\} + \frac{\theta\varphi\vartheta H_t^2}{H_t + \kappa(0)}, \tag{E.116}$$

$$\dot{E}_t = rE_t - a_1 + K_t\frac{H^2}{2}\max_{x\in[0,1]}\left\{\frac{m(x,x)}{H_t + \kappa(x)}\right\} + \theta\varphi\vartheta H_t\frac{E_t - \tilde{E}_t}{H_t + \kappa(0)}, \tag{E.117}$$

$$\dot{D}_t = rD_t - \max_{x\in[0,1]}\left\{\frac{m(x,x)}{H_t + \kappa(x)}\right\}\int k'^2dF_t\left(k'\right) - \theta\varphi\vartheta\frac{\left[E_t - \tilde{E}_t\right]^2}{4\left[H_t + \kappa(0)\right]}. \tag{E.118}$$

It follows that

$$\frac{d\left(E_t-\tilde{E}_t\right)}{dt}=\left(r+\frac{\theta\varphi\vartheta H_t}{H_t+\kappa\left(0\right)}\right)\left(E_t-\tilde{E}_t\right)-a_1+K_t\frac{H^2}{2}\max_{x\in[0,1]}\left\{\frac{m\left(x,x\right)}{H_t+\kappa\left(x\right)}\right\} \tag{E.119}$$

$$\frac{d\left(E_t-\tilde{E}_t-2H_tK_t\right)}{dt}=\left(r+\frac{\theta\varphi\vartheta H_t}{H_t+\kappa\left(0\right)}\right)\left(E_t-\tilde{E}_t\right)-a_1+K_t\frac{H^2}{2}\max_{x\in[0,1]}\left\{\frac{m\left(x,x\right)}{H_t+\kappa\left(x\right)}\right\} \tag{E.120}$$

$$-2K_t\left[rH_t-a_2+\frac{H^2}{4}\max_{x\in[0,1]}\left\{\frac{m\left(x,x\right)}{H_t+\kappa\left(x\right)}\right\}+\frac{\theta\varphi\vartheta H_t^2}{H_t+\kappa\left(0\right)}\right] \tag{E.121}$$

$$-2H_t\varphi\vartheta\frac{E_t-\tilde{E}_t-2H_tK_t}{2\left[H_t+\kappa\left(0\right)\right]} \tag{E.122}$$

$$=\left(r-\frac{\left(1-\theta\right)\varphi\vartheta H_t}{H_t+\kappa\left(0\right)}\right)\left(E_t-\tilde{E}_t-2H_tK_t\right)-a_1+2a_2K_t \tag{E.123}$$

and

$$\dot{K}_t=\varphi\vartheta\int\frac{E_t-\tilde{E}_t+2\left(\tilde{H}_t\tilde{k}-H_tk\right)}{2\left[H_t+\tilde{H}_t+\kappa\left(0\right)\right]}dF_t\left(k\right)=\varphi\vartheta\frac{E_t-\tilde{E}_t-2H_tK_t}{2\left[H_t+\kappa\left(0\right)\right]}, \tag{E.124}$$

with the boundary condition $K_0=K$ and $E_T-\tilde{E}_T=A_1-1-i^{RRP}$. We use the above equations in numerical simulations.

F Extension: Federal Funds Brokerage

In this section we model the brokerage of Federal funds following Lagos & Rocheteau (2007). In practice, Federal funds brokers reach out their banks' contact for matchmaking. Consider the following timing of actions. Having secured a pair of banks for potential Federal funds trading, the broker negotiates with each banks about its brokerage fee. In this stage, the broker does not reveal the identities of counterparties but informs the banks about the reserve balances held by their counterparties (the sufficient information banks need to know to initiate Federal funds trade in this model). This prevents the side-trading between the counterparty banks circumventing the broker's fee. Having determined the brokerage fees, the identities are revealed and the two banks negotitate the terms of trade like any bilateral Federal funds trades we described before. The brokerage fee is settled in numéraire at $T+\Delta$. We assume the matching rate between a broker and the bank counterparties is α, thus the contact rate of banks with a broker is $\alpha\nu$, where ν is the measure of active brokers. Brokers are free entry with entry cost ψ per broker.

We solve the outcome backward. Consider that a broker has identified a k-bank and a k'-bank at t. Each bank anticipates their trade surplus from trading with the arranged counterparties as $0.5S_t\left(k,k'\right)$. Denote $Y_t\left(k,k'\right)$ as the brokerage fee paid by k-bank for arranging the match with k'-bank; vice versa for the brokerage fee $Y_t\left(k',k\right)$ paid by k'-bank. To the k-bank, the surplus of brokerage is $0.5S_t\left(k,k'\right)-Y_t\left(k,k'\right)$. To the broker, the surplus of brokeraging the side of k-bank is

simply $Y_t(k,k')$. Thus, the brokerage fee solves the following Nash bargaining problem:

$$Y_t\left(k,k'\right)=\arg\max_y\left\{y\left[0.5S_t\left(k,k'\right)-y\right]\right\}.$$

Hence the bargaining solution is

$$Y_t\left(k,k'\right)=Y_t\left(k',k\right)=0.25S_t\left(k,k'\right).$$

The value of the broker, J_t, solves the following HJB equation

$$rJ_t=\dot{J}_t+\alpha\int\int\left[Y_t\left(k,k'\right)+Y_t\left(k',k\right)\right]dF_t\left(k'\right)dF_t\left(k\right),\text{ where }J_T=0.$$

Denote the dependence of J_t on the broker size ν as $J_t(\nu)$. In the equilibrium, ν is determined by the free-entry condition to the brokers:

$$\psi=J_0(\nu).$$

The bank's HJB is

$$rV_t(k)=\dot{V}_t(k)+u(k)+\max_{\varepsilon_t\in[0,1]}\int\frac{1}{2}S_t\left(k,k',\varepsilon_t,\varepsilon_t\left(k'\right)\right)m\left(\varepsilon_t,\varepsilon_t\left(k'\right)\right)dF_t\left(k'\right)+\alpha\nu\int\frac{1}{4}S_t\left(k,k',0,0\right)dF_t\left(k'\right)$$

With quadratic utility function, we guess and verify $V_t(k)=-H_tk^2+E_tk+D_t$, the solution is

$$\begin{aligned}rV_t(k)=\dot{V}_t(k)+u(k)+\frac{1}{2}\frac{\left(H_t\right)^2m(\varepsilon,\varepsilon_t)}{\kappa(\varepsilon)+\kappa(\varepsilon_t)+2H_t}\int\left(k'-k\right)^2dF_t\left(k'\right)\\+\frac{\alpha\nu}{8}H_t\int\left(k'-k\right)^2dF_t\left(k'\right)\end{aligned}\tag{F.125}$$

By matching coefficients we obtain

$$\left(r+\frac{\alpha\nu}{8}\right)H_t=\dot{H}_t+a_2-\frac{1}{4}\frac{H_t^2m(\varepsilon_t,\varepsilon_t)}{\kappa(\varepsilon_t)+H_t},\tag{F.126}$$

thus $\alpha\nu$ changes the discount rate to the banks. The surplus function is

$$S_t\left(k,k',\varepsilon_t\right)=\frac{\left[H_t\left(k'-k\right)\right]^2}{2\left(\kappa\left(\varepsilon_t\right)+H_t\right)}$$

and the broker's HJB is

$$\begin{aligned}rJ_t&=\dot{J}_t+\frac{\alpha}{2}\int\int S_t\left(k,k',0\right)dF_t\left(k'\right)dF_t\left(k\right)\\&=\dot{J}_t+\frac{\alpha}{4}H_t\left[\int k^2dF_t(k)-K^2\right]\end{aligned}\tag{F.127}$$

The solution is

$$J_0(\nu) = \frac{\alpha}{4}\int_0^T e^{-rt} H_t \left[\int k^2 dF_t(k) - K^2\right] dt, \tag{F.128}$$

where

$$\int k^2 dF_t(k) \tag{F.129}$$

$$= \int k^2 dF_0(k) \exp\left\{-\int_0^t m(\varepsilon_z, \varepsilon_z) \frac{H_z(H_z + 2\kappa(\varepsilon_z))}{2(H_z + \kappa(\varepsilon_z))^2} dz - \frac{\alpha\nu}{2} t\right\} \tag{F.130}$$

$$+K^2 \int_0^t \exp\left\{-\int_z^t m(\varepsilon_s, \varepsilon_s) \frac{H_s(H_s + 2\kappa(\varepsilon_s))}{2(H_s + \kappa(\varepsilon_s))^2} ds - \frac{\alpha\nu}{2}(t - z)\right\} \left[m(\varepsilon_z, \varepsilon_z) \frac{H_z(H_z + 2\kappa(\varepsilon_z))}{2(H_z + \kappa(\varepsilon_z))^2} + \frac{\alpha\nu}{2}\right] dz \tag{F.131}$$

$$= K^2 + \left[\int k^2 dF_0(k) - K^2\right] \exp\left\{-\int_0^t m(\varepsilon_z, \varepsilon_z) \frac{H_z(H_z + 2\kappa(\varepsilon_z))}{2(H_z + \kappa(\varepsilon_z))^2} dz - \frac{\alpha\nu}{2} t\right\}$$

Thus the solution to $J_0(\alpha)$ can be written as

$$J_0(\nu) = \frac{\alpha}{4}\left[\int k^2 dF_0(k) - K^2\right] \int_0^T e^{-rt} H_t \exp\left\{-\int_0^t m(\varepsilon_z, \varepsilon_z) \frac{H_z(H_z + 2\kappa(\varepsilon_z))}{2(H_z + \kappa(\varepsilon_z))^2} dz - \frac{\alpha\nu}{2} t\right\} dt.$$

Thus the equilibrium matchmaking is

$$\psi = \frac{\alpha}{4}\left[\int k^2 dF_0(k) - K^2\right] \int_0^T e^{-rt} H_t \exp\left\{-\int_0^t m(\varepsilon_z, \varepsilon_z) \frac{H_z(H_z + 2\kappa(\varepsilon_z))}{2(H_z + \kappa(\varepsilon_z))^2} dz - \frac{\alpha\nu}{2} t\right\} dt. \tag{F.132}$$

The following proposition characterizes the comparative statics of the equilibrium measure of brokers with respect to policy and technology parameters.

Proposition 8 *Suppose κ is sufficiently small. The comparative statics of ν are*

	i^{ER}	i^{DW}	K
ν	$-$	$+$	$-$

Proof. Note that $J_0(\infty) = 0$ and $J_0(0) > 0$. For the existence of equilibrium we assume $\psi < J_0(0)$. Due to free entry, we focus on the equilibrium ν^* with $J_0'(\nu^*) < 0$. We define

$$M_t \equiv e^{-rt} H_t \exp\left\{-\int_0^t m(\varepsilon_z, \varepsilon_z) \frac{H_z(H_z + 2\kappa(\varepsilon_z))}{2(H_z + \kappa(\varepsilon_z))^2} dz - \frac{\alpha\nu}{2} t\right\},$$

which implies

$$\frac{\dot{M}_t}{M_t} = -\frac{3}{8}\alpha\nu - \frac{a_2}{H_t} - \frac{m(\varepsilon_t, \varepsilon_t)}{4} \frac{H_t}{H_t + \kappa\varepsilon_t}\left(\frac{2\kappa(\varepsilon_t)}{H_t + \kappa(\varepsilon_t)} + 1\right) < 0,$$

and

$$\frac{\partial}{\partial H_t}\left(\frac{\dot{M}_t}{M_t}\right) = \frac{a_2}{H_t} - \frac{m(\varepsilon_t, \varepsilon_t)}{4} \frac{\kappa(\varepsilon_t)}{(H_t + \kappa(\varepsilon_t))^2}\left(\frac{4\kappa(\varepsilon_t)}{H_t + \kappa(\varepsilon_t)} - 1\right).$$

Thus a sufficient condition for $\frac{\partial}{\partial H_t}\left(\frac{\dot{M}_t}{M_t}\right) > 0$ is $\frac{\kappa(\varepsilon_t)}{H_t+\kappa(\varepsilon_t)} < \frac{1}{4}$, which requires a sufficiently small κ. Given this condition and note that $M_0 = H_0$, we can obtain that the path of M_t shifts upward if the path of H_t shifts upward. Combining with the result that $\frac{\partial H_t}{\partial A_2} > 0$, we can obtain that

$$\frac{\partial J_0(\nu)}{\partial A_2} = \int_0^T \frac{\partial J_0(\nu)}{\partial M_t}\frac{\partial M_t}{\partial A_2}dt > 0.$$

By implicit function theorem, we can obtain $\frac{\partial \nu^*}{\partial A_2} > 0$. Given $\frac{\partial A_2}{\partial i^{ER}} < 0$, $\frac{\partial A_2}{\partial i^{DW}} > 0$ and $\frac{\partial A_2}{\partial K} < 0$, this establishes our proposition. **Q.E.D.**

G Extension: Payment Shocks

Since Poole (1968) there has been a long history of analyzing the effects of payment flow on the Federal funds market. In this extension we study the role of payment on disintermediation. Suppose that banks are receiving and sending exogenous and stochastic payment flows of reserve balances. There are two types of payment flows: lumpy or continuous. Lumpy payments occur occasionally at the arrival rate ζ, with the amount w (negative value means outflow of reserve balances) drawn from a symmetric distribution G with mean 0 and standard deviation σ_L. Continuous payments occur continuously that follows a Brownian motion with mean μ and volatility σ_C. Thus the aggregate inflow of reserve balances from payment flow is μ. The HJB equation becomes

$$\begin{aligned} rV_t(k) = \dot{V}_t(k) + u(k) + \max_{\varepsilon\in[0,1]} \int \frac{1}{2} S_t\left(k, k', \varepsilon, \varepsilon_t\left(k'\right)\right) m\left(\varepsilon, \varepsilon_t\left(k'\right)\right) dF_t\left(k'\right) \\ +\zeta \int \left[V_t(k+w) - V_t(k)\right] dG(w) + \mu \frac{\partial}{\partial k} V_t(k) + \frac{\sigma_C^2}{2}\frac{\partial^2}{\partial k^2} V_t(k). \end{aligned} \tag{G.133}$$

Given $\{F_t\}$, the value function in an equilibrium is given by

$$V_t(k) = -H_t k^2 + E_t k + D_t, \tag{G.134}$$

where H_t, E_t and D_t are given by the solutions to the following initial-value ODE problems

$$\dot{H}_t = rH_t - a_2 + \frac{1}{4}\frac{H_t^2 m(\varepsilon_t, \varepsilon_t)}{\kappa(\varepsilon_t) + H_t}, \tag{G.135}$$

$$\dot{E}_t = rE_t - a_1 + \frac{K_t}{2}\frac{H_t^2 m(\varepsilon_t, \varepsilon_t)}{\kappa(\varepsilon_t) + H_t} + 2\mu H_t, \tag{G.136}$$

$$\dot{D}_t = rD_t - \frac{1}{4}\frac{H_t^2 m(\varepsilon_t, \varepsilon_t)}{\kappa(\varepsilon_t) + H_t}\int k'^2 dF_t\left(k'\right) + \left(\zeta\sigma_L^2 + \sigma_C^2\right) H_t - \mu E_t, \tag{G.137}$$

where $H_T = A_2$, $E_T = A_1$ and $D_T = 0$. The equilibrium search profile of $\Omega(S_t, F_t)$ is given by

$$\varepsilon_t(k) = \arg\max_{\varepsilon} \left\{ \frac{H_t^2 m(\varepsilon, \varepsilon)}{\kappa(\varepsilon) + H_t} \right\} \tag{G.138}$$

The Federal funds purchased $q_t(k, k')$ and the Federal funds rate $\rho_t(k, k')$ are given by

$$q_t(k, k') = \frac{H_t(k' - k)}{2(\kappa(\varepsilon_t) + H_t)}, \tag{G.139}$$

$$\rho_t(k, k') = e^{r(T+\Delta-t)} \left[E_t - H_t(k + k') \right]. \tag{G.140}$$

Note that H_t does not depend on the payment shocks, while E_t is only affected by μ. The following Proposition summarize the grid-locking effect of payment shocks.

Proposition 9 *The comparative statics of the length of search, $\bar{\tau}$, the amount of Federal funds purchased, $q_t(k, k')$ and the Federal fund rates, $\rho_t(k, k')$ are given by the following table*

	$\bar{\tau}$	$q(k, k', \tau)$	$\rho(k, k', \tau)$
ζ	0	0	0
μ	0	0	$-$
σ	0	0	0

Proof. Since H_t is independent of the payment shocks, the comparative statics of $\bar{\tau}$ and q_t over payment shock parameters are zero. For ρ_t, the comparative statics is non-zero only for μ. Note that a higher μ means a higher K_t and a larger $2\mu H_t$. This implies a larger $\dot{E}_t$. Since E_T is given, it means E_t decreases in μ. Thus ρ_t decreases in μ. **Q.E.D.**

Intuitively, a larger μ means the excess reserves increase faster. This implies a lower marginal value of holding reserves, leading to lower Federal funds rates.

H Extension: Counterparty Risk

Afonso et al. (2011) documents the importance of counterparty risk in explaining the rise of Federal funds rate and decline Federal funds trade during the crisis. Our model can be extended to incorporate two kinds of counterparty risk. Consider that there is probability $1 - p_L$ that, after the terms of trade is determined, the Federal funds lender cannot deliver the corresponding reserves to the borrower and the trade has to be cancelled. Also, there is a probability $1 - p_B$ that the Federal funds borrower cannot repay R when it is due. The borrower's surplus is thus given by

$$p_L \left[V_t(k + q) - p_B e^{-r(T+\Delta-t)} R \right] - p_L V_t(k) - \chi(\varepsilon, q).$$

The lender's surplus is given by

$$p_L\left[V_t\left(k'-q\right)+p_B e^{-r(T+\Delta-t)}R\right]-p_L V_t\left(k'\right)-\chi\left(\varepsilon',-q\right).$$

The solution to Nash bargaining problem becomes

$$q_t\left(k,k',\varepsilon,\varepsilon'\right)=\frac{H_t\left(k'-k\right)}{\left(\kappa\left(\varepsilon\right)+\kappa\left(\varepsilon'\right)\right)/p_L+2H_t}, \tag{H.141}$$

$$R_t\left(k,k',\varepsilon,\varepsilon'\right)=\frac{e^{r(T+\Delta-t)}}{p_B}\left[E_t-H_t\left(k+k'\right)-\frac{\kappa\left(\varepsilon\right)-\kappa\left(\varepsilon'\right)}{2p_L}q_t\left(k,k',\varepsilon,\varepsilon'\right)\right]q_t\left(k,k',\varepsilon,\varepsilon'\right), \tag{H.142}$$

$$\rho\left(k,k',\tau\right)=\frac{R\left(k,k',\tau\right)}{q\left(k,k,\tau\right)}=\frac{e^{r(T+\Delta-t)}}{p_B}\left[E_t-H_t\left(k+k'\right)-\frac{\kappa\left(\varepsilon\right)-\kappa\left(\varepsilon'\right)}{2p_L}q_t\left(k,k',\varepsilon,\varepsilon'\right)\right], \tag{H.143}$$

$$S_t\left(k,k',\varepsilon,\varepsilon'\right)=\frac{p_L\left[H_t\left(k'-k\right)\right]^2}{\left(\kappa\left(\varepsilon\right)+\kappa\left(\varepsilon'\right)\right)/p_L+2H_t}. \tag{H.144}$$

The optimal search intensity in the refined equilibrium is

$$\varepsilon_t=\arg\max_{\varepsilon}\left\{\frac{H_t^2 m\left(\varepsilon,\varepsilon\right)}{\kappa\left(\varepsilon\right)/p_L+H_t}\right\}.$$

The solution to the value function is that $V_t\left(k\right)=-H_t k^2+E_t k+D_t$, where

$$\dot{H}_t=rH_t-a_2+\frac{p_L}{4}\frac{H_t^2 m\left(\varepsilon_t,\varepsilon_t\right)}{\kappa\left(\varepsilon_t\right)/p_L+H_t},\ \text{where } H_T=A_2; \tag{H.145}$$

$$\dot{E}_t=rE_t-a_1+\frac{Kp_L}{2}\frac{H_t^2 m\left(\varepsilon_t,\varepsilon_t\right)}{\kappa\left(\varepsilon_t\right)/p_L+H_t},\ \text{where } E_T=A_1; \tag{H.146}$$

$$\dot{D}_t=rD_t-\frac{p_L}{4}\frac{H_t^2 m\left(\varepsilon_t,\varepsilon_t\right)}{\kappa\left(\varepsilon_t\right)/p_L+H_t}\int k'^2 dF_t\left(k'\right),\ \text{where } D_T=0. \tag{H.147}$$

Overall, the effects of higher counterparty risk (a higer $1-p_L$) are isomorphic to the effects of higher transaction cost κ and lower matching rate λ and λ_0.

I Algorithm of Simulation and Estimation

Simulation. Let us denote N as the number of banks, $i\in\{1,2,...,N\}$ as the index of individual banks. Since the size of peripheral traders is redundant for simulation, we assume there is only one peripheral trader and denote it as $i=N+1$. We also denote $m\in\mathbb{N}$ as the index for bilateral meetings, where a smaller m means an earlier meeting. Since the number of banks is finite, the total number of meetings is also finite. Moreover, denote $k_0\left(i\right)$ as the initial reserve balances of bank i before entering the Federal funds market, and $k_m\left(i\right)$ as the reserve balances of bank i after meeting m takes place. Note that $k_0\left(i\right)$ is given by banks' empirical excess reserves divided by

bank assets, and $k_m(i) \neq k_{m-1}(i)$ only if bank i is one of the counterparties in meeting m. It is important to note that the mass of an individual bank is normalized to 1, and the search intensity λ, λ_0 and φ represent the search intensity for an individual bank. Thus the total mass of banks is N, and the contact rate for a bank with another bank is $\frac{m(\varepsilon_t,\varepsilon_t)}{N}$. There are $\frac{N(N-1)}{2}$ pairs of bilateral meetings between banks and N pairs of bilateral meetings between a bank and a peripheral trader. All these meetings are independent Poisson process. Thus the sum of all these meetings follows a Poisson process with intensity $\frac{N(N-1)}{2}\frac{m(\varepsilon_t,\varepsilon_t)}{N} + N\varphi = \frac{N-1}{2}m(\varepsilon_t,\varepsilon_t) + N\varphi$. $\frac{N(N-1)}{2}\frac{m\left(\varepsilon_t^{b,b},\varepsilon_t^{b,b}\right)}{N} + N\mu_g m\left(\varepsilon_t^{b,g},\varepsilon_t^{g,b}\right) + N\mu_f m\left(\varepsilon_t^{b,f},\varepsilon_t^{f,b}\right) + \mu_g\mu_f m\left(\varepsilon_t^{g,f},\varepsilon_t^{f,g}\right)$We simulate the discretized version of the model via the following algorithm.

1. Given the model parameters and policy parameters, we numerically solve the paths of H_t, ε_t via Proposition **??** and solve the paths of E_t and K_t via the ODEs (E.119) and (E.124).

2. Given the path of ε_t, simulate a Poisson process for bilateral meetings up to time T via a thinning algorithm:[21]

   (a) Set a sufficiently large $\lambda_{\max}$ (such that $\lambda_{\max} > \lambda$). Generate a random integer $\hat{M}$ distributed as Poisson with mean $\left(\frac{N-1}{2}\lambda_{\max} + N\varphi\right)T$. $\left(\frac{N-1}{2} + N\mu_g + N\mu_f + \mu_g\mu_f\right)\lambda_{\max}T$ If $\hat{M} = 0$ stop.

   (b) Generate $\hat{M}$ random numbers distributed as i.i.d. uniforms on $(0,1)$, i.e. $U_1, ..., U_{\hat{M}}$, and reset $U_m = T \cdot U_m$, $m \in \left\{1, ..., \hat{M}\right\}$.

   (c) Place the U_m in ascending order to obtain the order statistics $U_{(1)} < U_{(2)} < ... < U_{(\hat{M})}$.

   (d) Set $\hat{t}_m = U_{(m)}$.

   (e) For each $\hat{t}_m$, generate an i.i.d. uniform on $(0,1)$, $\hat{U}_m$. If

   $$\hat{U}_m \leq \frac{\frac{N-1}{2}m\left(\varepsilon_{\hat{t}_m},\varepsilon_{\hat{t}_m}\right) + N\varphi}{\frac{N-1}{2}\lambda_{\max} + N\varphi},$$

   then keep $\hat{t}_m$. Otherwise, drop it.

   (f) For each kept $\hat{t}_m$, draw a pair of integers $\hat{p}_m = \{i,j\}$ with $1 \leq i < j \leq N+1$ from the weighted distribution j

   $$\Pr(i,j) = \begin{cases} \frac{\frac{N-1}{2}m\left(\varepsilon_{\hat{t}_m},\varepsilon_{\hat{t}_m}\right)}{\frac{N-1}{2}m\left(\varepsilon_{\hat{t}_m},\varepsilon_{\hat{t}_m}\right)+N\varphi}\frac{2}{N(N-1)}, & \text{if } i,j \leq N, \\ \frac{N\varphi}{\frac{N-1}{2}m\left(\varepsilon_{\hat{t}_m},\varepsilon_{\hat{t}_m}\right)+N\varphi}\frac{1}{N}, & \text{if } j = N+1. \end{cases}$$

   (g) For each kept $\hat{t}_m$ and $\hat{p}_m$, we relabel them with $\{t_m, p_m\}_{m=1}^{M}$, where $t_m < t_{m+1}$ and M is the number of kept $\hat{t}_m$. The sequence of $\{t_m, p_m\}_{m=1}^{M}$ is the Poisson process for bilateral

[21] See Sigman (2007) for a detailed description and proof of the thinning algorithm.

meetings for our simulation. and denote the number of kept trade according to the rule. Update $k_n(i)$ and $k_n(j)$.

3. Update individual reserve balances and bilateral terms of trade: denote $k_m(i)$, $q_m(i)$ and ρ_m as bank i's reserve balances after meeting m, bank i's cumulative absolute Federal funds trade after meeting m, and the bilateral Federal funds rate in meeting m. We start with the data $k_0(i)$ and set $q_0(i) = 0$ by definition. For each meeting m, if $i \in p_m$, then update $k_m(i)$, $q_m(i)$ and ρ_m according to the theoretical formulae. For any $i \notin p_m$, do not update $k_m(i)$ and $q_m(i)$.

4. Use the sequence $\{k_m(i), q_m(i), \rho_m\}$ to calculate the aggregate moments and regression coefficients.

Estimation. The simulated method of moments estimation follows a standard two-step procedure.[22] For each quarter, we simulate the model for $S = 2,000$ times.

6.1 Positive Implications on Liquidity

The closed-form solution allows us to obtain a set of measures on liquidity in analytical form. We list these measures in this section for possible quantitative analysis. The derivations of all the measures are provided in the Appendix C.10.

Price impact. The price impact of a trade measures how much the Federal fund rate changes in response to a given Federal fund purchased. The higher the price impact, the more expensive to borrow reserve balances, reflecting lower liquidity. In the Walrasian benchmark, the price impact is always zero. Log-linearizing (12), the Federal fund rate of a k-bank borrowing q units of reserves is

$$\rho_t(k,q) \cong \underbrace{r(T+\Delta-t)}_{\text{time effect}} + \underbrace{\log V_t'(k)}_{\text{bank fixed effect}} - \underbrace{\frac{\theta_{V,t}(k)}{1-\omega_t}}_{\text{price impact}} q, \tag{148}$$

where $\theta_{V,t}(k)$ is the elasticity of value function and ω_t is the equilibrium precaution-speed trade-off:

$$\begin{aligned} \theta_{V,t}(k) &\equiv \frac{H_t}{V_t'(k)}, \\ \omega_t &\equiv \left[1 + \frac{H_t}{\kappa(\varepsilon_t)}\right]^{-1}. \end{aligned} \tag{149}$$

The price impact depends on the ratio between the elasticity of value function and the precaution-speed trade-off.

[22] See Adda & Cooper (2003) for the reference on simulated method of moments.

Return reversal. If the Federal funds market is liquid, the price impact is transistory and the Federal fund rate will swiftly reverse to the mean. The return reversal measures how swift the Federal fund rate stablizes disturbances. In the Walrasian benchmark, the return reversal is always infinity. In our model, the dynamics of the Federal fund rate is given by

$$\frac{d}{dt}\log\left[\left|\rho_t\left(k,k'\right)-\varrho_t\right|\right]=-\underbrace{\frac{a-X\left(H_t\right)}{H_t}}_{\text{return reversal}},$$

where ϱ_t is the average Federal funds rate defined by $\varrho_t \equiv \int\int \rho_t\left(k,k'\right)dF_t\left(k'\right)dF_t\left(k\right)$. Note that the value of $V_t''\left(k\right)$ and the search intensity both control the speed of return reversal.

Price dispersion. The law of one price tends to apply when the Federal fund market is extremely liquid. The price dispersion measures the prevalence of arbitrage opportunity arise of the search friction. In the Walrasian benchmark, the price dispersion is always zero. In our model, the price dispersion is given by

$$\underbrace{\sigma_{\rho,t}}_{\text{price dispersion}}=\sqrt{2}e^{r(T+\Delta-t)}H_t\sigma_{k,t},$$

where $\sigma_{\rho,t}$ is the standard deviation of Federal fund rate and $\sigma_{k,t}$ is the standard deviation of reserve balances. Since the Federal fund rates are more dispersed when banks hold more dispersed reserve balances, we normalize the price dispersion with the standard deviation of reserve balances.

Intermediation markup. Recall that banks intermediate by purchasing Federal funds to sell. Intermediation is not risk-free as the bank exposes itself to the risk of selling Federal funds at a lower price than the purchasing price. The rate spread is the between the expected Federal fund rate of the selling leg and the realized Federal fund rate of the purchasing leg:

$$\Delta_{\rho,t}\left(k,q\right)\equiv\int\rho_t\left(k+q,k'\right)dF_t\left(k'\right)-\rho_t\left(k,q\right).$$

The intermediation markup measures the change in the rate spread in response to the size of the intermediation trade. In our model, the intermediation markup is given by

$$\underbrace{\frac{\partial\Delta_{\rho,t}\left(k,q\right)}{\partial q}}_{\text{intermediation markup}}=e^{r(T+\Delta-t)}\left[2\kappa\left(\varepsilon_t\right)+H_t\right].$$

Utilization rate of trade opportunities. The total trade opportunities in this economy is

$$TO_t=\int_k\int_{k'\geq k}\frac{k'-k}{2}dF_t\left(k'\right)dF_t\left(k\right).$$

The utilitzation rate of trade opportunities measure how fast the trade opportunities are realized. In Afonso & Lagos (2015b), the utilization rate is the exogeneous matching rate. In our model, the utilization rate is

$$UR_t = \frac{\int_k \int_{k' \geq k} m\left(\varepsilon_t, \varepsilon_t\right) q_t\left(k, k', \varepsilon_t, \varepsilon_t\right) dF_t\left(k'\right) dF_t\left(k\right)}{TO_t} = \frac{H_t m\left(\varepsilon_t, \varepsilon_t\right)}{\kappa\left(\varepsilon_t\right) + H_t}.$$

Thanks to the closed-form solution, we also derive the comparative statics of $L_t\left(k\right)$ on policy parameters in Section 3.3. Given the individual net trade volume, the aggregate volume of the absolute net trade is

$$L \equiv \int \left|L_0\left(k\right)\right| dF_0\left(k\right) = \left\{1 - \exp\left[-\int_0^T \frac{m\left(\varepsilon_t, \varepsilon_t\right) H_t}{2\left[\kappa\left(\varepsilon_t\right) + H_t\right]} dt\right]\right\} \int \left|K - k\right| dF_0\left(k\right).$$

Given the aggregate volume of absolute trade and net trade, we define the level of intermediation and fraction of intermediation as

$$\begin{aligned} \text{Int} &= Q - L, \\ \text{IntR} &= \frac{Q - L}{Q}. \end{aligned} \tag{150}$$

Federal fund rate. The average Federal fund rate at t is given by

$$\begin{aligned} 1 + \varrho_t &= \int \int \left[1 + \rho_t\left(k, k'\right)\right] dF_t\left(k'\right) dF_t\left(k\right) = e^{r(T+\Delta-t)}\left[E_t - 2H_t K\right] \\ &= e^{r\Delta}\left[\gamma + 1 + i^{ER} + \frac{k_+ - 1}{k_+ - k_-}\Delta i\right] - \frac{2a\left(K - \bar{k}\right)}{r}\left[e^{r(T+\Delta-t)} - e^{r\Delta}\right], \end{aligned} \tag{151}$$

$$\begin{aligned} \varrho &= \gamma + \frac{1 - k_-}{k_+ - k_-} i^{ER} + \frac{k_+ - 1}{k_+ - k_-}\Delta i - a\left(K\right)\left(K - \bar{k}\right) T \\ \varrho_t - i^{ER} &= \left(1 + i^{ER}\right)\left(e^{r\Delta} - 1\right) + \end{aligned} \tag{152}$$

where $\Delta i = i^{DW} - i^{ER}$ is the policy rate spread. Compared with the frictionless Federal fund rate, the liquidity effect is weaker with seach frictions. The range of the Federal funds rates is given by $1 + \rho_t\left(k, k'\right) \in \left[1 + \rho_t^{\min}, 1 + \rho_t^{\max}\right]$, where

$$\begin{aligned} 1 + \rho_t^{\min} &= e^{r(T+\Delta-t)}\left[E_t - 2H_t k_{\max}\right], \\ 1 + \rho_t^{\max} &= e^{r(T+\Delta-t)}\left[E_t - 2H_t k_{\min}\right]. \end{aligned} \tag{153}$$