MONETARY POLICY OPERATIONS:
THEORY, EVIDENCE, AND TOOLS FOR QUANTITATIVE ANALYSIS

Ricardo Lagos
Gastón Navarro

WORKING PAPER 31370

NBER WORKING PAPER SERIES

MONETARY POLICY OPERATIONS: THEORY, EVIDENCE, AND TOOLS FOR QUANTITATIVE ANALYSIS

Ricardo Lagos
Gastón Navarro

Working Paper 31370
http://www.nber.org/papers/w31370

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2023

Lagos is employed by New York University. Navarro is employed by the Board of Governors of the Federal Reserve System. Neither has other material financial relationships that relate to this research. We thank Joshua Herman, Patrick Molligo, Maddie Penn, Charlotte Singer, and Kenji Wada for superb research assistance. We also thank Michele Cavallo, Jeff Huther, and Cindy Vojtech for useful comments and discussions. We are grateful to Heather Ford and Gina Sellito for helping us navigate the compliance requirements to access some of the data. The views expressed in the paper are those of the authors and are not necessarily reflective of views at the Federal Reserve Board, the Federal Reserve System, or the National Bureau of Economic Research.

Monetary Policy Operations: Theory, Evidence, and Tools for Quantitative Analysis
Ricardo Lagos and Gastón Navarro
NBER Working Paper No. 31370
June 2023
JEL No. C78,D83,E44,E49,G1,G18,G2,G21,G23,G28

ABSTRACT

We formulate a quantitative dynamic equilibrium theory of trade in the fed funds market, calibrate it to fit a comprehensive set of marketwide and micro-level cross-sectional observations, and use it to make two contributions to the operational side of monetary policy implementation. First, we produce global structural estimates of the aggregate demand for reserves—a crucial decision-making input for modern central banks. Second, we propose diagnostic tools to gauge the central bank's ability to track a given fed funds target, and the heterogeneous incidence of policy actions on the shadow cost of funding across banks.

Ricardo Lagos
Department of Economics
New York University
19 West Fourth Street
New York, NY 10012
and NBER
ricardo.lagos@nyu.edu

Gastón Navarro
Federal Reserve Board
Division of International Finance
20th St. and Constitution Ave. N.W.
Washington, DC 20551
gaston.m.navarro@frb.gov

Appendices are available at http://www.nber.org/data-appendix/w31370

1 Introduction

The Federal Reserve uses the fed funds rate to communicate and implement its monetary-policy stance. In each of the eight regularly scheduled meetings during the year, the Federal Open Market Committee (FOMC) chooses a fed funds rate target and issues implementation notes specifying the policy instruments that will be used to create market conditions for fed funds to trade at rates near the target. Two kinds of instruments are typically used to achieve this goal. The first, open-market operations, affect the market price of fed funds by changing the quantity of reserves. The second, a set of administered rates offered by standing facilities, such as the Discount-Window rate (DWR), the interest rate paid on reserves (IOR), and the offering rate on overnight reverse repurchase agreements (ONRRP), affect the market price of fed funds by changing banks' return from holding (or borrowing) reserves at the central bank. An *operating framework* for implementing monetary policy is a consistent usage of these instruments to implement the fed funds rate target. For example, an operating framework may rely mostly on managing *quantities* of reserves, and another on managing administered *rates*.

Figure 1 shows the stylized theoretical demand-and-supply model that policymakers use to think about how different operating frameworks can achieve a fed funds rate target.[1] Before the Great Financial Crisis of 2007-2008 (GFC), aggregate reserves were scarce, e.g., around a relatively low level such as Q_0 in the first panel of Figure 1. In this context, target rates like r_0^* were achieved by changing the quantity of reserves (represented by the vertical line in the figure) through open-market operations. This operating framework is known as a *corridor system* since it can implement any target rate inside the *corridor* defined by a *ceiling rate*, typically the Discount-Window rate, ι_w (possibly plus a stigma premium or other costs associated with borrowing from the central bank), and a *floor rate*, such as the interest that the central bank pays on bank reserves, ι_r (or a lower rate if not all fed funds participants can earn interest on reserves held at the central bank).[2]

[1] An individual bank's demand for reserves is typically thought of in terms of Poole (1968), who derives it from the static decision problem of an individual bank that chooses how much of its beginning-of-day reserves to lend (to earn a given market interest), and how much to hold (to insure against an otherwise uninsurable exogenous reduction in reserves, which would force the bank to engage in end-of-day borrowing at a central-bank discount rate higher than the market rate). This "Poole model" is the go-to framework in policy circles, see, e.g., Ennis and Keister (2008), Keister et al. (2008), Keister (2012), Afonso et al. (2020b), and Åberg et al. (2021).

[2] In a corridor system the target rate and the administered rates are typically chosen so that the target rate lies in the middle of the corridor, and the central bank does not use the standing facilities as instruments to manage the fed funds rate (although they are available for fed funds participants who wish to borrow from, or lend reserves to the central bank).

During, and in the aftermath of the GFC, the Federal Reserve undertook a series of large-scale asset purchase programs that increased banks' reserve balances to a very high level such as Q_1 in the top-right panel of Figure 1.[3] With a such a large supply of reserves, the Fed can no longer rely on routine open-market operations (which entail relatively small changes in the quantity of reserves) to instrument changes in the fed funds rate target. In this context, target rates like r_1^* are achieved by changing the administered rates, i.e., the DWR, the IOR, and the ONRRP. This operating framework is known as a *floor system*.[4]

To describe the levels of reserves compatible with these operating frameworks, policymakers often use "scarce reserves" to refer to the range for which the slope of the aggregate demand for reserves is "steep", "ample reserves" for the range for which it is "gentle", and "abundant reserves" for the range for which it is "flat", as illustrated in the top-right panel of Figure 1.[5] The Federal Reserve intends to continue operating a floor system in which an "ample" supply of reserves ensures that the fed funds rate is controlled by the administered rates, and in which "active management of the supply of reserves is not required."[6] In terms of the schematic in Figure 1, this operating framework seems easy to manage: the Fed just needs to ensure the supply of reserves remains "ample", i.e., at a level close to Q_1 in the top-right panel.

In practice, however, it is difficult to quantify how large the supply of reserves needs to be for it to be "ample". A supply of reserves larger than Q_1 in the top-right panel of Figure 1 would still allow the Fed to operate a floor system, but would entail unnecessary costs.[7] Conversely, a supply of reserves smaller than Q' would imply operating on the steep part of the aggregate demand for reserves, i.e., *de facto* abandoning the preferred operating framework for a corridor system that requires active day-to-day management of the supply of reserves to achieve the fed

[3] Total reserves were about $40 bn before the GFC through mid 2008, and reached $2.8 tn in 2014 (see Figure 16).

[4] In a floor system, the central bank actively operates two *standing facilities*: a lending facility that lends reserves to banks, and a deposit facility that enables qualifying institutions to lend reserves to the central bank. In the United States, for example, the Fed now sets three administered rates: DWR, IOR, and ONRRP. The IOR, which is regarded as the primary policy instrument, is the rate banks can earn by holding reserves in the deposit facility. The ONRRP, which is typically set lower than the IOR, is the rate that a broader set of financial institutions (including banks but also GSEs and money-market funds) can earn by holding reserves in the deposit facility. The logic is that the IOR acts a reservation price for lending banks, and the ONRRP acts as a reservation price for other (non-bank) lending institutions, so depending on the composition of trades, one of the two rates should act as a floor for negotiated rates.

[5] See Afonso et al. (2020b) and Afonso et al. (2022). The term "ample" has become standard language in FOMC press releases (see, e.g., Federal Reserve Board (2019c)).

[6] See, e.g., Federal Reserve Board (2019b).

[7] See Bernanke and Kohn (2016) and Ireland (2018) for discussions of the economic and political risks associated with an operating framework that involves paying interest on a large stock of reserves.

funds rate target. Notice that the operational difficulty goes beyond resolving the arbitrariness involved in specifying numerical thresholds for the slope of the reserve demand to be considered "steep", "gentle", or "flat". Even if we agreed on a precise definition of "gentle slope", the main difficulty is to find the associated quantity of reserves, which requires reliable estimates of the slope of the aggregate demand for reserves for a wide range of the aggregate quantity of reserves. In other words, running a floor system like the one the Federal Reserve has adopted, requires *global* estimates of the slope of the aggregate demand for reserves. This presents a significant challenge because existing state-of-the-art empirical methods only deliver *local* estimates of the slope of the demand for reserves, i.e., estimates obtained based on instrumented variation around a relatively narrow range of the aggregate supply of reserves.[8]

In this paper we develop a quantitative model of the fed funds market, calibrate it to match key micro and macro statistics that characterize fed funds trading in the United States—including available empirical estimates of the *local* slope of the aggregate demand for reserves—and use it to bridge the *local-global gap*. Specifically, we use the relationship between the aggregate supply of reserves and the equilibrium fed funds rate implied by the theory, to infer the *global* shape of the actual aggregate demand for reserves.

The theory incorporates search and bilateral bargaining to represent the well-documented over-the-counter microstructure of the fed funds market. The theory also accounts for relevant institutional considerations, such as the differential regulatory treatment of the reserve balances held by Government Sponsored Enterprises (GSEs) vis á vis depository institutions, and incorporates the array of policy instruments and regulations that affect participants' demands for reserves, such as the administered policy rates (DWR, IOR, ONRRP), the regulatory requirements on reserve holdings, and the aggregate quantity of reserves supplied to the system. The theory also accommodates the large degree of heterogeneity among fed funds participants across several dimensions, such as: market power in bilateral negotiations, frequency and size distribution of idiosyncratic payment shocks originated by forces outside the fed funds trading motives, measures of trading activity (frequency of trade, number of counterparties, participation rate in aggregate volume of trade), and degree of centrality in the endogenous market-making activity that reallocates reserves across the trading network.

[8]The empirical challenge is illustrated in the bottom panels of Figure 1, which show situations in which structural parameters are Π_i at the time the quantity-price pair (Q_i, r_i^*) is observed, for $i \in \{0,1\}$. Without theoretical guidance to identify the structural parameters whose variation, e.g., from Π_0 to Π_1, shift the demand for reserves, one may be led to believe that the observations $\{(Q_i, r_i^*)\}_{i\in\{0,1\}}$ lie on a single demand curve, and therefore overestimate (bottom-left panel) or underestimate (bottom-right panel) the relevant slope.

We calibrate the parameters of the theory that govern the heterogeneity in payment and trading activities using high-frequency micro-level transaction data from Fedwire, and find that the model is able to fit the targeted observations well, e.g., as in the data, a small number of very active banks account for the majority of loans, and carry out most of the intermediation. The calibration strategy also ensures that—at the current level of total reserves—the magnitude of the variation in the equilibrium fed funds rate induced by exogenous variation in the supply of reserves is in line with standard reduced-form empirical estimates of the "liquidity effect". The calibrated model is also broadly consistent with empirical observations not targeted in the calibration, such as the cross-sectional distribution of bilateral interest rates, the distribution of bid-ask spreads, and the intraday flow of reserves and supporting interest rates between pairs of banks with different trading activities.

We use the quantitative theory to make two practical contributions to the operational side of monetary policy implementation. First, we use the calibrated model—disciplined and validated by micro data—to deliver global structural estimates of the aggregate demand for reserves, which should be useful to central banks that wish to operate floor systems. Second, we use the calibrated model as the basis for two "navigational instruments" for monetary policy implementation. The first, which we term *Monetary Confidence Band* (MCB), is a hybrid of theory and data: it is a simple procedure that uses the empirical distribution of daily reserve-draining shocks to construct a confidence band around the aggregate demand for reserves that, for each outstanding quantity of reserves, will contain the equilibrium fed funds rate with a desired degree of confidence, e.g., 95%. The second is the cross-sectional distribution of banks' shadow cost of procuring funding in the fed funds market implied by the theory.

This paper contributes to the large empirical and theoretical literature that studies the fed funds market, e.g., Hamilton (1996), Ashcraft and Duffie (2007), Bech and Atalay (2010), Afonso et al. (2011), Bech and Klee (2011), Afonso and Lagos (2012, 2015a,b), Ennis and Weinberg (2013), Armenter and Lester (2017), Afonso et al. (2019), Beltran et al. (2021), Ennis (2019), Chiu et al. (2020), and Afonso et al. (2022). Methodologically, we build on the strand of the finance microstructure literature that uses search theory to model over-the-counter markets, e.g., Duffie et al. (2005), Lagos and Rocheteau (2007, 2009), Weill (2007), Lagos et al. (2011), Üslü (2019), and Hugonnier et al. (2020). Specifically, our model builds on Afonso and Lagos (2015b), which we generalize along several dimensions to make it a serviceable quantitative tool for monetary policy implementation.

The rest of the paper is organized as follows. Section 2 presents the model, discusses the main assumptions, and defines equilibrium. Section 3 documents the key statistics that will guide the quantitative implementation of the theory, e.g., bank-level measures of fed funds trading activity (participation and intermediation), frequency and size distribution of micro-level intraday payments between banks, and the typical beginning-of-day cross-sectional distribution of reserve balances (net of predictable payments and regulatory requirements). Section 3 also reports empirical estimates of the distribution of aggregate daily reserve-draining shocks for the fed funds market since the GFC, and of the "liquidity effect" associated with exogenous variation in the aggregate supply of reserves. Section 4 discusses the calibration strategy. Section 5 reports how well the quantitative model fits empirical price and quantity observations not targeted by the calibration. Section 6 analyzes the aggregate demand for reserves generated by the model, and uses it as the basis for a quantitative-theoretic estimation of the aggregate demand for reserves in the United States. Section 7 proposes two "navigational instruments" to guide the routine monetary policy operations necessary to implement a fed-funds rate target. Section 8 uses the quantitative theory to rationalize the well-known fed-funds rate spikes of September 2019. Section 9 concludes. The appendices contain supplementary material.

2 Theory

There is a unit measure of *banks* that are heterogeneous along several dimensions. We represent this heterogeneity with a finite set $\mathbb{N}$ of bank types, and let $n_i \in [0,1]$ represent the proportion of banks of type $i \in \mathbb{N}$, with $\sum_{i\in\mathbb{N}} n_i = 1$. Banks hold an asset we interpret as (claims to) *reserve balances* that can be traded with other banks during the time interval $\mathbb{T} = [0,T]$. The reserve balance that a bank holds at a given time is represented by a real number, e.g., $a \in \mathbb{R}$. The cumulative distribution function of reserve balances across all banks at time $t \in \mathbb{T}$ is denoted $F_t(a) = \sum_{i\in\mathbb{N}} n_i F_t^i(a)$, where $F_t^i(a) : \mathbb{R}\times\mathbb{T} \to [0,1]$ is the cumulative distribution of balances across banks of type i at time t. The initial distribution, $\left\{F_0^i(\cdot)\right\}_{i\in\mathbb{N}}$, is given, and so is the aggregate supply of reserve balances throughout the trading session, denoted $Q \equiv \int a dF_0(a)$.

Banks trade reserves with other banks in a bilateral over-the-counter market where a bank of type $i \in \mathbb{N}$ contacts another bank at random times generated by a Poisson process with arrival rate $\beta_i \in \mathbb{R}_+$. Conditional on a meeting, the counterparty is a random (uniform) draw from the population of banks. Once two banks have made contact, they bargain over the size of the loan and the quantity of reserve balances to be repaid by the borrower. The bargaining

outcome is determined by Nash bargaining. When a bank of type $i \in \mathbb{N}$ negotiates with a bank of type $j \in \mathbb{N}$, we assume the bargaining power of the former is $\theta_{ij} = 1 - \theta_{ji} \in [0, 1]$. After the transaction, the banks part ways.

We assume all loans are settled at time $\bar{T} > T$, and that banks value reserve balances linearly at that time. Specifically, if $c \in \mathbb{R}$ is a bank's net credit position to be settled at $\bar{T}$ that has resulted from a certain history of trades, then $e^{-r(\bar{T}-t)}c$ is the bank's payoff from this credit balance at time $t \in [0, T]$, where $r \in \mathbb{R}_+$ is the discount rate common to all banks.

Banks receive payment shocks that cause reallocations of reserve balances among pairs of banks. Specifically, with Poisson rate $\lambda_i \in \mathbb{R}_+$, a bank of type $i \in \mathbb{N}$ is forced to make an immediate transfer of reserves to a counterparty that is drawn randomly (uniformly) from the population of banks. This process for the arrival of payment shocks is independent across banks and independent of the processes that generate bilateral trading opportunities. Conditional on the arrival of a payment shock, the quantity of reserves that the bank of type i sends the bank of type j is modeled as a random variable with cumulative distribution function $G_{ij} : \mathbb{Z} \to [0, 1]$, where $\mathbb{Z} \subseteq \mathbb{R}$ is the support of G_{ij}, and $dG_{ij}(z) = dG^{ji}(-z)$, which captures the notion that these payments are *transfers* between pairs of bank types.

For each $i \in \mathbb{N}$, define the function $U_i : \mathbb{R} \to \mathbb{R}$, where $U_i(a)$ represents the payoff to a bank of type i from holding reserve balance $a \in \mathbb{R}$ at the end of the trading session. Similarly, for each $i \in \mathbb{N}$, define the function $u_i : \mathbb{R} \to \mathbb{R}$, where $u_i(a)$ represents the flow payoff to a bank of type i from holding reserve balance $a \in \mathbb{R}$ during the trading session. The type of a bank is defined by a vector of primitives, i.e., type $i \in \mathbb{N}$ is defined by $(n_i, \beta_i, \lambda_i, \{\theta_{ij}, G_{ij}\}_{j\in\mathbb{N}}, u_i, U_i)$, in the sense that each of the n_i banks of type i has trading frequency β_i, bargaining powers $\{\theta_{ij}\}_{j\in\mathbb{N}}$, payment frequency λ_i, probability distributions $\{G_{ij}\}_{j\in\mathbb{N}}$ of payment sizes, intraday payoff function u_i, and end-of-day payoff function U_i.

2.1 Discussion

The market for federal funds is a market for unsecured loans of reserve balances at the Federal Reserve Banks. These unsecured loans, commonly referred to as *federal funds* (or *fed funds*) are delivered on the same day, and their maturity is typically overnight. Most fed funds transactions and interbank payments are conducted through *Fedwire Funds Services* (Fedwire), a large-value real-time gross settlement system operated by the Federal Reserve Banks. Participants in the fed funds market are institutions that hold reserve balances in accounts at the Federal Reserve,

which include commercial banks, savings banks, thrift institutions, credit unions, agencies and branches of foreign banks in the United States, government securities dealers, government agencies such as federal or state governments, and Government Sponsored Enterprises (GSEs, e.g., Freddie Mac, Fannie Mae, and Federal Home Loan Banks). The fed funds market is *over the counter*: in order to trade, a financial institution must first find a willing counterparty and then bilaterally negotiate the size and rate of the loan. The fed funds market is the epicenter of monetary policy implementation in the sense that the *effective fed funds rate* (EFFR)—the policy rate that the Federal Reserve uses to communicate and implement monetary policy—is a daily volume-weighted average of the bilateral interest rates negotiated by fed funds participants.

We use a search-based model with ex post bargaining to represent the bilateral over-the-counter nature of the fed funds market.[9] Search captures three layers of randomness in trading activity in our model. First, the time it takes a bank of type $i \in \mathbb{N}$ to contact a counterparty is an exponentially distributed random variable with mean $1/\beta_i$. Second, conditional on having contacted a counterparty, the type of the counterparty is a uniform random draw. Third, conditional on having met a counterparty of type $j \in \mathbb{N}$ at time t, the current reserve balance of the counterparty is a random variable with cumulative distribution function $\{F_t^j(\cdot)\}_{j\in\mathbb{N}}$. We use the generalized Nash bargaining solution to represent the outcome of the bilateral negotiations between counterparties in actual fed funds trades.

The motives for trading fed funds may vary across participants and their specific circumstances on any given day, but there are two main reasons in general. First, some participants may regard fed funds as an investment vehicle—an interest-yielding asset that can be used to deposit balances overnight. Also, some institutions such as commercial banks use the fed funds market to offset the effects of random payment shocks (resulting from transactions initiated by their clients or by profit centers within the institutions themselves) that would otherwise leave them with a reserve position deemed too low relative to regulatory requirements. In the theory, the Poisson rate λ_i represents the frequency of these payment shocks for a bank of type $i \in \mathbb{N}$, and G_{ij} represents the size distribution of payment shocks between two banks of types i and j. In our model, all payoff-relevant policy and regulatory considerations are captured by the intraday and end-of-day payoff functions, $\{u_i(\cdot), U_i(\cdot)\}_{i\in\mathbb{N}}$.

Fedwire and the fed funds market operate 21.5 hours each business day, from 9:00 pm

[9] In practice, there are two ways of trading federal funds. Two participants can contact each other directly and negotiate the terms of a loan, or they can be matched by a fed funds broker. Since nonbrokered transactions represent the bulk of the volume, we abstract from brokers in our baseline model.

eastern standard time (EST) on the preceding calendar day to 6:30 pm EST. Although there is occasionally some activity between 9:00 pm and 9:00 am, the bulk of the fed funds transactions and interbank payments take place between 9:00 am and 6:30 pm. Thus, in the theory, we think of $t = 0$ as standing in for 9:00 am and use the initial condition $\left\{F_0^i(\cdot)\right\}_{i\in\mathbb{N}}$ to represent the distribution of reserve balances at this time.

2.2 Equilibrium

Let $J_t^i(a,c) : \mathbb{N}\times\mathbb{T}\times\mathbb{R}^2 \to \mathbb{R}$ be the maximum attainable payoff to a bank of type i that at time $t \in \mathbb{T}$ has reserve balance $a \in \mathbb{R}$ and net credit position $c \in \mathbb{R}$. In Appendix A (Lemma 1) we show that $J_t^i(a,c) = V_t^i(a) + e^{-r(\bar{T}-t)}c$, where $V_t^i(a) : \mathbb{N}\times\mathbb{T}\times\mathbb{R} \to \mathbb{R}$ is the maximum expected discounted payoff a bank of type $i \in \mathbb{N}$ can obtain when holding $a \in \mathbb{R}$ reserve balances at time $t \in \mathbb{T}$. Whenever two banks contact each other during the trading session, they bargain over the size of the loan and the repayment. Consider a bank of type i with reserve balance a that contacts a bank of type j with reserve balance $\tilde{a}$. The pair $(b_t^{ij}(a,\tilde{a}), R_t^{ji}(\tilde{a},a))$ denotes the bilateral terms of trade negotiated by these banks at time t, where $b_t^{ij}(a,\tilde{a})$ is the quantity of reserves that the bank of type i with balance a lends to the bank of type j with balance $\tilde{a}$, and $R_t^{ji}(\tilde{a},a)$ is the quantity of balances that the latter commits to repay the former at time $\bar{T}$. For any $(a,\tilde{a},t) \in \mathbb{R}^2\times\mathbb{T}$, $(b_t^{ij}(a,\tilde{a}), R_t^{ji}(\tilde{a},a))$ is the solution to

$$\max_{(b,R)\in\bar{\mathbb{R}}\times\mathbb{R}} \left[V_t^i(a-b) + e^{-r(\bar{T}-t)}R - V_t^i(a)\right]^{\theta_{ij}} \left[V_t^j(\tilde{a}+b) - e^{-r(\bar{T}-t)}R - V_t^j(\tilde{a})\right]^{\theta_{ji}},$$

with $\bar{\mathbb{R}} \equiv \left[-\bar{b},\bar{b}\right]$, where $\bar{b} \in \mathbb{R}_+ \cup \{\infty\}$ represents a limit on bilateral credit exposures (there is no borrowing limit if $\bar{b} = \infty$). The first-order conditions for this problem imply

$$b_t^{ij}(a,\tilde{a}) \in \arg\max_{b\in\bar{\mathbb{R}}} S_t^{ij}(a,\tilde{a},b) \tag{1}$$

$$e^{-r(\bar{T}-t)}R_t^{ji}(\tilde{a},a) = \theta_{ij}\left[V_t^j(\tilde{a}+b_t^{ij}(a,\tilde{a})) - V_t^j(\tilde{a})\right] + \theta_{ji}\left[V_t^i(a) - V_t^i(a-b_t^{ij}(a,\tilde{a}))\right], \tag{2}$$

where

$$S_t^{ij}(a,\tilde{a},b) \equiv V_t^i(a-b) + V_t^j(\tilde{a}+b) - V_t^i(a) - V_t^j(\tilde{a}).$$

Condition (1) characterizes the loan size, and (2) gives the repayment given the loan size. The implied gross interest rate on this loan is

$$1 + \rho_t^{ji}(\tilde{a},a) = \frac{R_t^{ji}(\tilde{a},a)}{b_t^{ij}(a,\tilde{a})}.$$

In Appendix A (Lemma 2) we show that the value function $V_t^i(a)$ satisfies

$$rV_t^i(a) - \dot{V}_t^i(a) = u_i(a) + \lambda_i \sum_{j\in\mathbb{N}} \pi_j \int \left[V_t^i(a-z) - V_t^i(a)\right] dG_{ij}(z) + \beta_i \sum_{j\in\mathbb{N}} \sigma_j \theta_{ij} \int \max_{b\in\mathbb{R}} S_t^{ij}(a,\tilde{a},b)\, dF_t^j(\tilde{a}), \tag{3}$$

with boundary condition $V_T^i(a) = U_i(a)$, where

$$\pi_j \equiv \frac{\lambda_j n_j}{\sum_{i\in\mathbb{N}} \lambda_i n_i}$$

is the probability the counterparty in a bilateral payment is of type j, and

$$\sigma_j \equiv \frac{\beta_j n_j}{\sum_{k\in\mathbb{N}} \beta_k n_k}$$

is the probability the counterparty in a bilateral trade is of type j.

Let $f_t^i \equiv dF_t^i$ denote the probability density function of reserve holdings among banks of type i at time t. This density follows the law of motion

$$\dot{f}_t^i(a) + (\beta_i + \lambda_i) f_t^i(a) = \lambda_i \sum_{j\in\mathbb{N}} \pi_j \int\int \mathbb{I}_{\{x-z=a\}} dG_{ij}(z)\, dF_t^i(x) + \beta_i \sum_{j\in\mathbb{N}} \sigma_j \int\int \mathbb{I}_{\{x-b_t^{ij}(x,\tilde{a})=a\}} dF_t^j(\tilde{a})\, dF_t^i(x). \tag{4}$$

Hereafter, let $\boldsymbol{U}(\cdot) = \{U_i(\cdot)\}_{i\in\mathbb{N}}$, $\boldsymbol{V}_t(\cdot) = \left\{V_t^i(\cdot)\right\}_{i\in\mathbb{N}}$, $\boldsymbol{b}_t(\cdot,\cdot) = \{b_t^{ij}(\cdot,\cdot)\}_{i,j\in\mathbb{N}}$, $\boldsymbol{R}_t(\cdot,\cdot) = \{R_t^{ij}(\cdot,\cdot)\}_{i,j\in\mathbb{N}}$, and $\boldsymbol{F}_t(\cdot) = \left\{F_t^i(\cdot)\right\}_{i\in\mathbb{N}}$.

Definition 1 *An equilibrium is a time-path* $\{\boldsymbol{b}_t(\cdot,\cdot), \boldsymbol{R}_t(\cdot,\cdot), \boldsymbol{V}_t(\cdot), \boldsymbol{F}_t(\cdot)\}_{t\in\mathbb{T}}$ *that satisfies (1), (2), (3), and (4), given the initial condition* $\boldsymbol{F}_0$ *and the terminal condition* $\boldsymbol{V}_T = \boldsymbol{U}$.

3 Data

In this section we document the fed funds market facts that will guide the quantitative implementation of the theory. Section 3.1 presents the joint distribution of two bank-level measures of fed funds trading activity: a bank's *participation rate* in marketwide trade volume, and a *reallocation index* that quantifies the degree to which a bank is a net borrower or lender of funds. Section 3.2 reports estimates of the frequency and size distribution of micro-level intraday payments between banks. Section 3.3 presents estimates of a typical beginning-of-day

cross-sectional distribution of reserve balances. Section 3.4 reports estimates of the distribution of aggregate daily reserve-draining shocks for the fed funds market since the GFC of 2007-2008. Section 3.5 presents empirical estimates of the slope of the aggregate demand for reserve balances. Finally, Section 3.6 describes an empirical interpolation procedure to map changes in the aggregate quantity of reserves into changes in the cross-sectional distributions of reserves that is consistent with available observations, and will be used in our quantitative analysis.

Since some of the regulations introduced in the wake of the GFC are likely to have affected trading incentives in the fed funds market, we report facts separately for the period before, and after these regulations had been implemented.[10] In this section we use the years 2006 and 2019 as typical pre- and post-GFC-regulation periods, respectively. However, since some of our quantitative exercises will require sample variation in the aggregate quantity of reserves while keeping regulation constant, we will also report facts for the years 2014 and 2017.[11]

We use transaction data from the Fedwire Funds Service (*Fedwire*). Our typical Fedwire participant, which we call a *bank*, corresponds to a *bank holding company*. Our sample consists of 754 Fedwire participants for the year 2006, 404 for the year 2014, 395 for the year 2017, and 412 for the year 2019.[12] We use a modified version of the *Furfine algorithm* to identify overnight loans of reserves from the universe of Fedwire transfers; and we regard the remaining transactions as payments (presumably unrelated to loan issuance or repayment).[13] We focus on transactions that occur between 9:00 am and 6:30 pm EST.

[10]Some of these regulations increased the shadow value of liquid assets (including reserves), or introduced leverage constraints that increased the shadow cost of borrowing funds (including overnight fed funds purchases). Two prominent examples of such regulations are the *Liquidity Coverage Ratio* (LCR) and the *Supplementary Leverage Ratio* (SLR) requirements. We discuss these regulations in Appendix B.

[11]By January 2015, LCR regulation had been fully phased in, and SLR disclosures had become mandatory, so we regard 2006 and 2014 as pre-GFC-regulation years, and the 2017 and 2019 as post-GFC-regulation years. In terms exploiting sample variability in the quantity of outstanding reserves in the system, the years 2006, 2014, 2017, and 2019 are natural benchmarks for the following reasons. The year 2006 is a typical pre-GFC period with excess reserves close to zero, and the year 2014 is a post-GFC but pre-GFC-regulation period with very high level of excess reserves (close to the pre-2020 historical peak). The year 2017 is a post-GFC-regulation period with very high level of excess reserves (again, close to the pre-2020 historical peak), while the year 2019 has the lowest level of excess reserves in the post-GFC-regulation era.

[12]In Appendix D (Section D.1.2) we explain our sample selection criteria, and how we assigned Fedwire transactions to bank holding companies.

[13]The algorithm, which is based on Furfine (1999), was made available to us by the Money Market Analysis Section at the Monetary Affairs Division of the Federal Reserve Board.

3.1 Fed funds trading network

Let $\mathbb{B}$ denote the collection of banks in our sample in a given year, and $\mathbb{Y}$ denote the collection of all *trading periods* in that year.[14] Let v_{nd}^e be the dollar value of all loans extended by bank $n \in \mathbb{B}$ in period $d \in \mathbb{Y}$, and use $v_d \equiv \sum_{n \in \mathbb{B}} v_{nd}^e$ to denote the dollar value of all the loans traded in period d. Also, let v_{nd}^r be the dollar value of all loans received by bank $n \in \mathbb{B}$ in period $d \in \mathbb{Y}$. For each bank n and period d, define

$$\begin{aligned} \mathcal{P}_{nd} &\equiv \frac{v_{nd}^e + v_{nd}^r}{v_d} \\ \mathcal{R}_{nd} &\equiv \frac{v_{nd}^e - v_{nd}^r}{v_{nd}^e + v_{nd}^r}. \end{aligned}$$

We refer to $\mathcal{P}_{nd}$ as bank n's *participation rate* during period d, since it measures the share of the value of all loans traded during period d in which bank n participated as a counterparty. For any given bank n in period d, $\mathcal{P}_{nd} \in [0, 1]$, with $\mathcal{P}_{nd} = 0$ corresponding to a bank that did not trade, and $\mathcal{P}_{nd} = 1$ corresponding to a bank that acted as a counterparty in every trade. In general, if a bank n participated as a counterparty in $x\%$ of the dollar value of all the loans traded in the market in period d, then $\mathcal{P}_{nd} = x/100$.[15] We refer to $\mathcal{R}_{nd}$ as bank n's *reallocation index* during period d, since it is an index of the degree to which a bank is a net borrower or lender of funds. For any given bank n in period d, $\mathcal{R}_{nd} \in [-1, 1]$, with $\mathcal{R}_{nd} = -1$ corresponding to a bank that only borrowed, $\mathcal{R}_{nd} = 1$ corresponding to a bank that only lent, and $\mathcal{R}_{nd} = 0$ corresponding to a bank whose trading activity in period d consisted of pure intermediation. A typical bank n will have either $\mathcal{R}_{nd} \in (-1, 0)$, meaning it is a net borrower that engaged in some intermediation, or $\mathcal{R}_{nd} \in (0, 1)$, meaning it is a net lender that engaged in some intermediation.[16] To provide a parsimonious description of the typical trading activity for each bank, we construct a bank-level participation rate and reallocation index averaged over

[14] In our empirical work, a *trading period* will correspond either to a *trading day*, or to a typical 14-day *(reserve) maintenance period* used to calculate a bank's reserve requirement. Our convention is to use $\mathbb{Y}$ to denote a generic set of trading periods in a year, $\mathbb{D}$ to denote the set of trading days in a year, and $\mathbb{H}$ to denote the set of maintenance periods in a year. See Section B.1 in Appendix B for institutional details on reserve requirements and maintenance periods.

[15] It is customary to define the aggregate volume of trade as $2v_d$, so $\mathcal{S}_{nd} \equiv \mathcal{P}_{nd}/2$ is bank n's share of the aggregate volume of trade in period d.

[16] Notice that $\mathcal{X}_{nd} \equiv 1 - |\mathcal{R}_{nd}|$ is a measure of the proportion of the total volume of funds traded by bank n in period d that the bank *intermediated* during that period, and $(v_{nd}^e + v_{nd}^r)\,\mathcal{X}_{nd}$ is what Afonso and Lagos (2015b) call *excess funds reallocation* (a measure of the volume of funds that an individual bank trades over and above what is required to accommodate its daily net trade).

all trading periods in a given year, i.e.,

$$\begin{aligned}\mathcal{P}_n &= \frac{1}{N_Y}\sum_{d\in\mathbb{Y}}\mathcal{P}_{nd}\\ \mathcal{R}_n &= \frac{1}{N_Y}\sum_{d\in\mathbb{Y}}\mathcal{R}_{nd},\end{aligned}$$

where $N_Y \equiv \sum_{d\in\mathbb{Y}}\mathbb{I}_{\{d\in\mathbb{Y}\}}$ is the number of trading periods in the year, and each trading period corresponds to one of bank n's (reserve) maintenance periods during the year.[17]

Figure 2 shows the empirical cumulative distribution function (ECDF) of participation rates for the banks that are in our sample in the year 2006 (the circles) and the banks that are in our sample in the year 2019 (the crosses). We use the bank-level participation rate to sort each bank into one of three *groups*, denoted S, M, and F, depending on whether the bank's participation rate is low, medium, or high, respectively.[18] Specifically, in each year we label the 4 banks with highest participation rate as F, the banks outside the top 4 that have participation rate at least as large as 1% as M, and all other banks as S. Notice that individually, each of the top four most active banks that compose group F participated as a counterparty roughly in at least 10% of the total volume of loans traded in an average reserve maintenance period. And together, these four banks accounted for a large share of the aggregate trade volume.[19] In contrast, the large majority of banks, which belong to group S, have extremely low participation rates. We regard this large skewness in loan trading activity across banks as a key empirical regularity of the fed funds market structure.

Among the institutions assigned to group S based on the ECDF in Figure 2, there is a subgroup of non-bank Fedwire participats typically referred to as *Government Sponsored Enterprises* (GSEs), which includes the Federal Home Loan Banks, the Federal National Mortgage Association (Fannie Mae), and the Federal Home Loan Mortgage Corporation (Freddie Mac). Even though on the basis of their trading activity GSEs would belong in group S, in what follows we consider them a different type of participant because their business model and regulatory treatment make their payoffs from holding reserves quite different from the rest of the participating institutions.[20] To offer a parsimonious representation of the data, we will sort

[17] See Appendix B (Section B.1) for institutional information on maintenance periods.

[18] The pneumonic is that banks of type S, M, and F, are slow, medium, and fast, at contacting counterparties.

[19] The top four most active banks participated as counterparties in 97.1% of loans in 2006, and 86.2% of loans in 2019. Or equivalently, according to the measure $\mathcal{S}_{id}$ introduced in footnote 15, they accounted for 45.6% of the aggregate trade volume in 2006, and 43.1% of the aggregate trade volume in 2019.

[20] In contrast to banks, GSEs have very predictable cashflows (so payment shocks are not relevant for their day-

institutions into four *types*, i.e., $\mathbb{N} = \{F, M, S, GSE\}$. The first two types are identical to the F and M groups defined above. Type S consists of the institutions in group S, excluding GSEs, and type GSE is composed exclusively of GSEs.[21]

Figure 3 shows the location of each bank type $i \in \{F, M, S, GSE\}$ in the coordinate axes defined by the reallocation index, $\mathcal{R}_i$, and the participation rate, $\mathcal{P}_i$, in the years 2006, 2014, 2017, and 2019. The figure shows an empirical trading network that conveys information on the distribution of trading activity across bank types, the flows of reserves implied by the fed funds lending among the four types of banks, and the average interest rates on the underlying loans. The participation, reallocation, and loans measures are all computed at the bank-type level.[22] Each node represents the set of banks assigned to a particular type, labeled accordingly as F, M, S, or GSE. The arrows from one node to another represent loans extended from banks of that type to the other. The position of each node indicates how active the corresponding bank type is in the fed funds market and whether banks of that type are, on average, net borrowers, net lenders, or intermediaries. The size of each node is proportional to the volume of trade between banks of the that type. The width of each arrow is proportional to the volume of trade between the bank types connected by the arrow. The colors of the arrows and nodes are: light blue, dark blue, light red, or dark red, if the volume-weighted average interest rate on the loans between the two types of banks, expressed as a spread over the EFFR, falls in the first, second, third, or fourth quartile, respectively.[23]

While specifics vary somewhat across years, several stable trading patterns emerge from Figure 3. Banks of type F account for about 1/2 of aggregate trade volume (i.e., $\mathcal{P}_F/2 \approx 1/2$)

to-day trading motives), and for most of our sample they did not earn interest on reserves—although nowadays they may lend reserves in the Federal Reserve's overnight reverse repo (ONRRP) facility.

[21] Our sample for 2006 consists of 4 banks of type F, 22 banks of type M, 716 banks of type S, and 12 GSEs. Our sample for 2019 consists of 4 banks of type F, 18 banks of type M, 379 banks of type S, and 11 GSEs. If we apply the same classification criteria for the years 2014 and 2017, we find that our sample for 2014 consists of 4 banks of type F, 15 banks of type M, 373 banks of type S, and 12 GSEs, while our sample for 2017 consists of 4 banks of type F, 18 banks of type M, 362 banks of type S, and 11 GSEs.

[22] The participation rate for each bank type $i \in \{F, M, S, GSE\}$ on a given year was calculated as follows. For each maintenance period, we summed the participation rates of all the banks of a given type, and then averaged across all maintenance periods in the year. The reallocation index for each bank type is calculated as follows. For each maintenance period, we sumed all the loans sent, and all the loans received, by banks of a given type, and used these aggregate measures of loans sent and received by the type to calculate the reallocation index for that bank type in that given maintenance period, and then averaged across all maintenance periods in the year. We followed the same aggregation procedure to calculate volume-weighted interest rates across groups. See Appendix D.2 (Section D.2.2) for more details.

[23] The arrow widths and node sizes are defined relative to the trades within a year, so they are not comparable across years.

and intermediate a large share of what they trade, with a tendency to act as net lenders. Banks of type M and banks of type S tend to be net borrowers; the former account for more than 1/4 of aggregate trade volume, and the latter much less (e.g., less than a quarter in 2006, and less than 1/8 in later years). GSEs account for about a 1/8 of aggregate trade volume, and participate almost exclusively as lenders.

3.2 Interbank payments

In the previous section we analyzed transfers of reserves associated with overnight borrowing and lending between banks. In this section we focus on transfers that are unrelated to loan issuance or repayment. We regard these transfers as *payments*, which may reflect transactions originated by the banks' clientele, or by sections of the bank other than the ones in charge of actively managing reserve balances.

We identify as *payments* all Fedwire transfers that are not flagged as loans or repayments by the Furfine algorithm. These payments are likely to have a predictable component, but also a random component, which we refer to as *payment shocks*. Since these components affect trading incentives differently in the theory, we must construct a measure of the predictable component, and estimate a process for the payment shocks of a typical bank of type F, M, or S.[24] As in the theory, we model payment shocks as a compound process with a parameter that determines the frequency with which a bank of type i receives a payment shock (i.e., λ_i in the theory), and a conditional probability distribution for the payment size, which is allowed to depend on the types of the banks sending and receiving the payment (i.e., G_{ij} in the theory). Next, we describe our procedure to estimate the process for high-frequency interbank payment shocks.

Let $\boldsymbol{T}$ denote the set of all one-second time intervals in a trading day $d \in \mathbb{D}$. For every pair of banks $m, n \in \mathbb{B}$, let $\mathrm{s}_{mn}(t,d) \in \mathbb{R}$ denote the dollar value of all payments from bank m to bank n in the one-second time interval $t \in \boldsymbol{T}$ during trading day $d \in \mathbb{D}$.[25] Let $\bar{\mathrm{s}}_{mn}$ denote the value of the average payment between banks m and n in a given year, and define $\tilde{\mathrm{s}}_{mn}(t,d) \equiv \mathrm{s}_{mn}(t,d) - \bar{\mathrm{s}}_{mn}$ for all $(t,d) \in \boldsymbol{T} \times \mathbb{D}$. In this way, we decompose every high-frequency payment $\mathrm{s}_{mn}(t,d)$ between a pair of banks into a *predictable component*, $\bar{\mathrm{s}}_{mn}$, and a *payment shock*, $\tilde{\mathrm{s}}_{mn}(t,d)$. For each pair of bank types $i, j \in \mathbb{N}$, we pool all payment shocks to

[24] The business model of a GSE makes its reserve balances unlikely to be subject to unexpected payment shocks of significant magnitude, so we regard all GSE payments as predictable.

[25] The bilateral payment credits bank n's account if $0 < \mathrm{s}_{mn}(t,d)$, and bank m's account if $\mathrm{s}_{mn}(t,d) < 0$.

form the data set

$$\tilde{\mathbb{S}}^{ij} = \{\tilde{\mathrm{s}}_{mn}(t,d) : m \in \mathbb{B}_i, n \in \mathbb{B}_j \text{ for all } (t,d) \in \boldsymbol{T} \times \mathbb{D}\},$$

where $\mathbb{B}_i$ is the set of banks of type $i \in \mathbb{N}$. We then use the data set $\tilde{\mathbb{S}}^{ij}$ to estimate a Gaussian kernel density that we regard as the size distribution of payment shocks between each pair of bank types i and j, i.e., the empirical counterpart of the probability density function corresponding to G_{ij} in the theory.[26] Figures 4 and 5 display the empirical histogram along with the corresponding estimated kernel of payment shocks for each pair of bank types using data from the years 2006, and 2019, respectively.

For each bank type $i \in \mathbb{N}$ we estimate the empirical counterpart of λ_i in our theory, as the average number of payment shocks that a typical bank of type i receives in a one-second time interval, $t \in \boldsymbol{T}$, during a trading day, d, in year $\mathbb{Y}$. Let $\mathrm{f}_m(t,d)$ denote the number of payment shocks between a bank $m \in \mathbb{B}$ and any other bank during the one-second time interval t in trading day d, i.e., $\mathrm{f}_m(t,d) = \sum_{n \in \mathbb{B}\setminus\{m\}} \mathbb{I}_{\{\mathrm{s}_{mn}(t,d)\neq 0\}}$. The corresponding average across seconds in a trading day, and trading days in the year is $\bar{\mathrm{f}}_m = \frac{1}{N_D}\sum_{d\in\mathbb{D}}\left[\frac{1}{N_{\boldsymbol{T}}}\sum_{t\in\boldsymbol{T}}\mathrm{f}_m(t,d)\right]$, where $N_{\boldsymbol{T}} \equiv \sum_{t\in\boldsymbol{T}}\mathbb{I}_{\{t\in\boldsymbol{T}\}}$ is the number of seconds in a trading day, and $N_D \equiv \sum_{d\in\mathbb{D}}\mathbb{I}_{\{d\in\mathbb{D}\}}$ is the number of trading days in a year. We use these bank-level empirical frequencies of payment shocks to estimate the probability that an average bank of type $i \in \{F, M, S\}$ receives a payment shock in a typical one-second time period, i.e., we set $\lambda_i = \frac{1}{N_i}\sum_{m\in\mathbb{B}_i}\bar{\mathrm{f}}_m$, where $N_i \equiv \sum_{m\in\mathbb{B}}\mathbb{I}_{\{m\in\mathbb{B}_i\}}$ denotes the number of banks of type i in our sample. The estimates for $\{\lambda_i\}_{i\in\{F,M,S\}}$ for the years 2019 and 2006 are reported in Table 1 and Table 3, respectively.[27]

3.3 Distribution of reserve balances

In this section we estimate beginning-of-day distributions of reserve balances (for each bank type) that are the empirical counterparts of the beginning-of-day distributions in the theory, i.e., $\{F_0^i\}$. Our calculations begin with a primitive bank-level quantity of reserves, and involve constructing a notion of *unencumbered* excess reserves by subtracting regulatory reserve requirements, and netting predictable Fedwire transfers (both, outright payments, and fed funds repayments).

For each bank in our sample, the Monetary Policy Operations and Analysis (MPOA) section at the Monetary Affairs Division at the Federal Reserve Board calculates the daily reserve

[26] See Appendix D (Section D.2.3) estimation details.

[27] We set $\lambda_{GSE} = 0$ for every year, for the reasons explained in footnote 24.

balance at 6:30 pm. We device an algorithm that uses this end-of-day balance to calculate the "basic" beginning-of-day balance at 9:00 am on the following day for each bank. Specifically, the algorithm starts with the bank's end-of-day balance for day $d-1$ provided by MPOA, adds all fed funds repayments received during day d, and subtracts all fed funds repayments sent during day d that correspond to fed funds loans originated during day $d-1$.[28] For each bank m, and each reserve maintenance period, h, that belongs to the set $\mathbb{H}$ of all maintenance periods in a given year, we calculate the average beginning-of-day balance across trading days in the maintenance period, which we denote $\mathrm{a}_m(h)$.[29] We make two additional adjustments to this average "basic" measure of beginning-of-day balance at the bank level.

The first adjustment consists of subtracting the quantity of *required reserves*, i.e., the minimum level of reserves that the bank must hold during the maintenance period in order to comply with Regulation D and the minimum *Liquidity Coverage Ratio* requirement (LCR).[30] Specifically, for each individual bank m, we compute the average beginning-of-day *excess reserves* during a maintenance period h, as $\mathrm{x}_m(h) = \mathrm{a}_m(h) - \underline{\mathrm{a}}_m^D(h) - \underline{\mathrm{a}}_m^L(h)$, where $\underline{\mathrm{a}}_m^D(h)$ and $\underline{\mathrm{a}}_m^L(h)$ denote the Regulation D and LCR reserve requirements, respectively.[31]

[28] Repayments are identified using the send-receive matching from the Furfine algorithm. The rationale for netting the *predictable transfers*, which include the *repayments of fed funds* borrowed in the previous trading day, as well as the *predictable component of payments* (discussed below), is that through the lens of our theory, the beginning-of-day balance that is relevant for a bank's incentives to trade reserves during the day ought to be net of *anticipated* transfers that the bank knows will receive or have to make during the trading day. The beginning-of-day-d balance for each GSE is constructed by taking the GSE's end-of-day balance for day $d-1$ provided by MPOA, and netting all repayments of fed funds loans traded during day $d-1$ (between the GSE and any other bank that meets the sample selection criteria described in Section D.1.2), as well as *payments* sent or received during trading day d (and that involve *any* bank, not only those that meet the sample selection criteria described in Section D.1.2). The rationale for netting all transfers that will occur during day d to obtain the GSE's balance at the beginning of day d is that a GSE's business model generates very predictable cashflows, so through the lens of our theory, we regard the GSE as being able to predict all its intraday Fedwire transfers at the beginning of the trading day.

[29] What motivates our focus on beginning-of-day balances *averaged over all trading days in a reserve maintenance period* is the fact that the reserve requirement regulations that influence banks' payoffs from holding reserves must be met not on a daily basis, but on average over all days in the maintenance period. See Section B.1 in Appendix B for details on the reserve requirements stipulated by Regulation D.

[30] Appendix B gives an overview of the relevant regulation. Our motivation for estimating reserves net of regulatory requirements is that this notion of *excess reserves* will play an important role in our quantitative theoretical exercises, e.g., it will be a key input to determine whether the central bank is implementing a monetary policy framework with "ample reserves" or a "corridor system". For this reason, in the quantitative implementation of the theory we specify banks' end-of-day payoffs in terms of excess reserves.

[31] The bank-level data for Regulation D requirements are provided by MPOA. The LCR regulation requires a bank to maintain (typically on a daily basis) a quantity of *High Quality Liquid Assets* (HQLA) at least as large as a measure of total net cash outflows in a 30-day standardized stress scenario. Specifically, if we let $\mathrm{H}_m(d)$ denote the quantity of qualifying HQLA held by bank m in a trading period d, and $\mathrm{L}_m(d)$ denote the corresponding measure of outflows in the stress scenario, the LCR regulation requires $\mathrm{L}_m(d) \leq \mathrm{H}_m(d)$. Both these quantities

The second adjustment to the average basic measure of a bank's beginning-of-day balance consists of subtracting the *predictable component of payments*. Specifically, for each bank m we compute $\hat{s}_m = \sum_{n \in \mathbb{B}} \hat{s}_{mn}$, where $\hat{s}_{mn} \equiv \sum_{d \in \mathbb{D}} \frac{1}{N_D} \sum_{t \in \boldsymbol{T}} s_{mn}(t,d)$ is the average (over the set $\mathbb{D}$ of N_D trading days in the year) net daily payment from bank m to bank n. Then, for each bank m and reserve maintenance period h, we construct $\mathrm{q}_m(h) = \mathrm{x}_m(h) - \hat{s}_m$, which is a bank's average (across days in the maintenance period) beginning-of-day measure of *unencumbered reserves*.[32]

For each bank type $i \in \mathbb{N}$, define

$$\mathbb{Q}^i = \{\mathrm{q}_m(h) : m \in \mathbb{B}_i \text{ for all } h \in \mathbb{H}\}.$$

We pool the data in the set $\mathbb{Q}^i$ and use it to estimate a Gaussian kernel density that we regard as the empirical counterpart of the beginning-of-day distribution of reserves, F_0^i, in the theory.[33]

Figures 6-9 show the kernel density estimates of the distributions of reserves for each bank type $i \in \mathbb{N}$ for the years 2006, 2014, 2017, and 2019, respectively. In every year, the distribution of unencumbered reserves across banks of type S is fairly concentrated around zero. In 2006 (a typical year before the GFC), about 60% of bank-period observations for type S have beginning-of-day reserves close to zero, with dispersion in both directions. In 2014, 2017, and 2019 (the post-GFC period with very high level of total reserves), the pattern for banks of type S is similar: about 60% of bank-period observations have beginning-of-day reserves close to zero, with some bank-period observations with positive reserves, and almost no bank-period observations with negative reserves. The distributions of beginning-of-day reserves for banks of type F and M, on the other hand, exhibit significant dispersion. For type M there are virtually no bank-period

are publicly available for each bank at a quarterly frequency (see Section D.1.3 in Appendix D for details). The set of qualifying HQLA includes reserves in excess of Regulation D, as well as securities issued or guaranteed by the U.S. Treasury (and also other securities, but subject to caps and haircuts). The fact that the LCR regulation allows banks to meet the requirement with assets other than reserves presents a challenge when trying to identify the quantity of reserves that bank m treats as "required" to satisfy the LCR constraint in period d, i.e., $\underline{\mathrm{a}}_m^L(d)$. Our strategy to tackle this identification problem is to set $\underline{\mathrm{a}}_m^L(d) = \max(0, \mathrm{L}_m(d) - \mathrm{A}_m(d))$, where $\mathrm{A}_m(d) \equiv \mathrm{H}_m(d) - \max\left(0, \mathrm{a}_m(d) - \underline{\mathrm{a}}_m^D(d)\right)$ is the quantity of qualifying HQLA in excess of (i.e., other than) reserves net of the Regulation D requirement. Notice that the resulting measure of excess reserves, $\mathrm{x}_m(d)$, selects the largest level of excess reserves net of the Regulation D requirement that is consistent with the LCR constraint. (Section B.2.1 in Appendix B discusses our strategy to identify the quantity of required reserves induced by the LCR regulation.) For banks that are not subject to LCR regulation, we set $\underline{\mathrm{a}}_m^L(d) = 0$. Since GSEs are not subject to Regulation D or LCR regulation, we set $\underline{\mathrm{a}}_m^D(d) = \underline{\mathrm{a}}_m^L(d) = 0$ for $m \in \mathbb{B}_{GSE}$.

[32] Unless otherwise specified, whenever we refer to "beginning-of-day reserves", we will be alluding to *unencumbered reserves*, i.e., the quantity of reserves in excess of Regulation D and LCR requirements, and net of predictable interbank Fedwire transfers.

[33] See Appendix D (Section D.2.3) estimation details.

observations with negative reserves for the years 2014, 2017, and 2019, and the dispersion over positive holdings is sizeable. For type F there is significant dispersion of reserves around zero in the years 2017 and 2019, largely due to the predictable component of payments.

3.4 Reserve-draining shocks

The aggregate demand for reserves is determined by the decisions of individual banks, who demand reserve balances as payment instruments, as safe short-term investment vehicles, and to meet regulatory requirements. The aggregate supply of reserves, on the other hand, is largely determined by the central bank's actions. But the central bank does not have complete control over the supply of reserves: The supply of reserves available to private banks also depends on transactions for which the Federal Reserve is not a counterparty, such as those that involve private-sector bank accounts and the account that the U.S. Treasury holds at the Federal Reserve. We will term the changes in the aggregate quantity of reserves resulting from the actions of entities other than the Federal Reserve, *exogenous supply shocks*. For example, whenever corporations or households pay taxes or purchase issuances of treasury securities, reserves are transferred from private banks to the Treasury's account at the Federal Reserve, which from the perspective of domestic banks, amounts to an aggregate contractionary (reserve-draining) supply shock. Conversely, expansionary (reserve-augmenting) supply shocks take place whenever the Treasury makes payments to the private sector (e.g., when redeeming outstanding debt instruments).[34] In this section we use daily data for the 2011-2019 sample period to estimate the size distribution of exogenous shocks to the supply of reserves.

Reserves were relatively scarce before 2007, and the Open Market Trading Desk ("the Desk") at the Federal Reserve Bank of New York (FRB-NY) routinely conducted open-market operations to offset the effects of exogenous supply shocks on the fed funds rate. These systematic policy responses make it challenging to identify exogenous shifts in the supply of reserves in the pre-2007 period. The sharp increase in excess reserves and the very low fed funds rate target that followed the GFC made it unnecessary for the Desk to actively respond to daily market conditions in order to implement the target. In fact, post-2008, the Federal Reserve interventions that affected the stock of reserves were driven by longer-term objectives (e.g.,

[34]Three other common sources of reserve-draining or reserve-augmenting shocks are: foreign official reverse repurchase agreements, changes in the quantity of currency in circulation (which imply swaps of currency for reserves or viceversa), and Federal Reserve "float" that is caused by the mismatch in timing between the debiting of reserves from a paying bank and the crediting of reserves to a receiving bank.

implementation of quantitative easing policies) rather than by day-to-day managing of the fed funds rate in response to high-frequency exogenous supply shocks to the quantity of reserves. Thus, in the post-GFC era we can identify exogenous supply shocks using high-frequency (e.g., daily) changes in the aggregate quantity of reserves held by financial institutions. The middle panel of Figure 10 shows that the variation in total reserves has been much larger since 2008, which validates our identifying assumption that the Desk did not react to exogenous supply shocks to the stock of reserves in the post-GFC period.

We estimate the distribution of reserve-draining shocks as follows. For each trading day d in the set $\mathbb{D}$ of all trading days in a given year, let A_d denote the aggregate quantity of reserves held by all banks at the end of day d, and define the corresponding 40-day (two-sided) moving average, $\overline{\mathrm{A}}_d \equiv \frac{1}{41}\sum_{k=-20}^{20} \mathrm{A}_{d+k}$.[35] The top panel of Figure 10 shows the time series $\{\mathrm{A}_d, \overline{\mathrm{A}}_d\}$ between the years 2001 and 2019. The middle panel of Figure 10 shows the deviations between total reserves and its own moving average, i.e., $\{\mathrm{z}_d\}$, with $\mathrm{z}_d \equiv \mathrm{A}_d - \overline{\mathrm{A}}_d$.[36] In time periods when the Federal Reserve does not react systematically to exogenous shocks to the suppply of reserves, $\{\mathrm{z}_d\}$ can be interpreted as a measure of the supply shocks themselves. Define the set $\mathbb{Z} = \{\mathrm{z}_d : d \in \mathbb{D}\}$, where $\mathbb{D}$ denotes the collection of trading days during the sample period January 2011-July 2019. We use the pooled data in the set $\mathbb{Z}$ to estimate a Gaussian kernel density for the distribution of shocks to the aggregate quantity of reserves.[37] The bottom panel of Figure 10 displays the empirical histogram based on the daily observations in $\mathbb{Z}$, along with its kernel estimate. The figure also depicts the intervals that contain the daily realization of the "aggregate supply shock" with 99% or 95% probabilities, i.e., [−\$279 bn, \$130 bn], and [−\$115 bn, \$99 bn], respectively.

To assess the plausibility of our estimates, consider Anbil et al. (2020), who in the context of the market events of September 16-17, 2019, estimate a reserve-draining shock of \$120 bn, and remark "it is not uncommon for reserves to fall about \$100 bn over a day or two" (p. 5). Our estimates imply that the probability of a reserve-draining shock of \$110 bn or larger is about 2.5%.

[35] For the purpose of these calculations we include *all* banks, not only those that meet the sample selection criteria based on fed funds trading activity described in Section D.1.2.

[36] Since the daily time series cannot be made public, the top and middle panels of Figure 10 show the *weekly* versions. But we use the daily time series for the purposes of the kernel estimation discussed below.

[37] See Appendix D (Section D.2.3) estimation details.

3.5 Liquidity effect

In this section we present empirical estimates of the change in the fed funds rate in response to an exogenous change in the aggregate quantity of reserves—the so-called *liquidity effect*.[38] It is customary to think of the fed funds rate as resulting from the intersection of a vertical supply and a downward sloping demand for reserves (e.g., as in Poole (1968)). Framed in this way, the slope of the demand for reserves is the key determinant of the liquidity effect, and the main challenge for estimation is to identify *exogenous* shifts in the supply of reserves.[39]

In an influential paper, Hamilton (1997) proposed a proxy for exogenous shifts in the aggregate quantity of reserves, and Carpenter and Demiralp (2006) subsequently proposed another.[40] The range of estimates obtained by Hamilton (1997) (for the period 1989/04/06–1991/11/27) and Carpenter and Demiralp (2006) (for the period 1989/05/19–2003/06/27) is similar: the estimated increase in the fed funds rate in response to an unexpected, temporary (one-day) $1 bn aggregate reserve-draining shock, ranges between 1 and 2 basis points (and can be as high as 3 basis points on "settlement Wednesdays").[41]

To estimate the liquidity effect for the post-GFC sample period with large excess reserves that were not actively managed by the central bank, we run the following regression:

$$s_t - s_{t-1} = \gamma_0 + \gamma(Q_t - Q_{t-1}) + \varepsilon_t, \tag{5}$$

[38] See Carpenter and Demiralp (2006) for a review, and Afonso et al. (2022) for more recent references.

[39] The estimation challenges are different for the pre-GFC regime in which the Desk was actively conducting open-market operations reacting to market conditions in order to manage the fed funds rate, than for the post-GFC era (until mid September 2019), when reserves were not actively managed by the central bank. And even within the post-GFC sample when the central bank began managing the fed funds rate by setting administered interest rates rather than the quantity of reserves, our theory prescribes controlling for the spreads between the administered rates (see Section 6.1).

[40] Hamilton (1997) proposed the deviations between the actual end-of-day balance of the Treasury's Fed account and an empirical forecast of the end-of-day balance of the Treasury's Fed account as a proxy for unexpected changes in the quantity of reserves. Carpenter and Demiralp (2006) build on the work of Hamilton (1997) by replacing his measure of unexpected changes in the Treasury's Fed account with a more accurate and comprehensive measure: the difference between the realized quantity of reserves on a given day, and the forecast for the quantity of reserves for that day that is used by the Desk (or the FRB) to perform its daily accommodative open-market operations. Relative to Hamilton's, the Carpenter-Demiralp measure of unexpected changes in reserves is more comprehensive because it contemplates all possible sources of variation in the supply of rerserves (not only fluctuations in the Treasury's Fed account), and it is more accurate because, by definition, these daily "forecast misses" are changes in the quantity of reserves that the Desk did not accommodate.

[41] Since this range of estimates was obtained from time series during a period in which reserves in excess of Regulation D were very close to zero, and post-GFC regulation had not been introduced, we will use it in our historical calibration exercise (Appendix E) to discipline the parameters that determine the magnitude of the liquidity effect in our quantitative theory *locally*, i.e., around the equilibrium point that results when excess reserves are close to zero.

where s_t denotes the spread between the effective fed funds rate (EFFR, published by the FRB-NY) and the administered interest rate on reserves (IOR) on day t, Q_t denotes the aggregate quantity of reserves at the end of day t (provided by the Monetary Affairs Division at the Federal Reserve Board), ε_t is an error term, and γ is the coefficient of interest.

We estimate regression (5) at daily frequency for the sample period 2019/05/02–2019/09/13. We base our estimation on the year 2019 because it is the baseline year we will use to calibrate our theory in Section 4. Our identifying assumption is that the daily changes in the aggregate quantity of reserves can be regarded as exogenous because, as discussed in Section 3.4, the Federal Reserve was not actively managing the quantity of reserves in response to developments in the fed funds market during the post-GFC sample periods that we consider for this regression.[42] The estimate is $\gamma = -0.0119$ (significant at the 1% level), with 95% confidence interval $[-0.0187, -0.0052]$. Since the independent variable is measured in billions of dollars and the dependent variable in basis points, these estimates mean that a \$1 bn increase in the quantity of reserves decreases the EFFR-IOR spread by 0.01 basis points (i.e., about one hundred times smaller than the estimates obtained by Hamilton (1997) and Carpenter and Demiralp (2006) for the pre-GFC corridor system with scarce reserves).

The sample period that we use in our estimation is chosen so that the spread between the (primary credit) Discount-Window rate (DWR) and the overnight reverse repo rate (ONRRP), and the spread between the IOR and the ONRRP, are constant (and in particular, equal to 75 and 10 bps per annum, respectively, as in our baseline calibration of Section 4).[43] This is important because, as we show in Section 6.1, our theory predicts that changes in these spreads shift the aggregate demand for reserves. To illustrate the perils of not controlling for these spreads, we run regression (5) at daily frequency for an extended sample period: 2019/01/01–2019/09/13. This sample period consists of two subperiods with different spreads between

[42] The sample goes up to mid-September 2019, when the overnight money market rates exhibited unusual spikes and exhibited significant volatility. This sample includes 2019/09/13 (Friday) and deliberately stops there because on 2019/09/16 (Monday), in response to the fed funds rate printing at the upper limit the target range, the Desk announced an overnight repo operation to be conducted at 9:30 AM on 2019/09/17 (Tuesday), offering up to \$75 billion against Treasury, agency, and agency MBS collateral. This operation, which injected \$53 billion in additional reserves and led to an immediate decline in rates, was the first time since the GFC that the Desk conducted an open-market operation to manage the fed funds rate. The sample we use to estimate γ ought to end before this policy response since it would clearly violate our identifying assumption. See Afonso et al. (2022) for a more comprehensive estimation exercise under different identifying assumptions. See Anbil et al. (2020) for a detailed narrative of the money-market rate spikes of mid-September 2019, and Section 8 below for a quantitative theoretical analysis of this episode.

[43] The time series of the three administered rates (DWR, IOR, ONRRP) are displayed in Figure 16.

administered rates: a first subperiod (from 2019/01/01 to 2019/05/01) with IOR-ONRRP spread equal to 15 bps, and a second subperiod (starting on 2019/05/02) with IOR-ONRRP spread equal to 10 bps. (The DWR-ONRRP spread is equal to 75 bps throughout.) The resulting estimate is $\gamma = -0.0062$ (significant at the 1% level), with 95% confidence interval $[-0.00975, -0.00264]$. Since the independent variable is measured in billions of dollars and the dependent variable in basis points, this estimate means that a \$1 bn increase in the quantity of reserves decreases the EFFR-IOR spread by about 0.006 basis points.[44]

To validate our estimates, we can compare them with those from Afonso et al. (2022), who provide time-varying estimates for the period 2009-2021 of the slope of the aggregate demand for reserves using an instrumental variable approach combined with a time-varying vector autoregressive model of the joint dynamics of reserves and federal fund rates. The slope of the aggregate demand for reserves for the year 2019 estimated by Afonso et al. (2022) implies that a 1 percentage point increase in the ratio of total reserves to total assets held by commercial banks leads to a 1 basis point reduction in the EFFR-IOR spread (see the entry in panel (a), row 1 of the column labeled "2019" in their Table 1). Since the value of total assets held by commercial banks was about \$17,000 bn in 2019, a 1 percentage point daily increase in the ratio of total reserves to total assets held by commercial banks corresponds roughly to a \$170 bn increase in total reserves. Thus, the estimate for 2019 that Afonso et al. (2022) report in Table 1 means that a \$1 bn increase in the quantity of reserves decreases the EFFR-IOR spread by about 0.00588 basis points, which is essentially the same as the estimate we obtain from regression (5) when we do not control for variation in the IOR-ONRRP spread.

Estimates of the liquidity-effect coefficient (e.g., γ in our regression equation (5), or the analogous estimates from Hamilton (1997), Carpenter and Demiralp (2006), and Afonso et al. (2022)) are to be interpreted as *local* estimates of the slope of the aggregate demand for reserves, since they can be thought of as the empirical counterparts of the slope of the demand for reserves in the Poole (1968) model—calculated using a relatively narrow range of variation in the aggregate supply of reserves. Unlike the Poole (1968) model, our theory does not have a primitive "demand for reserves". But as we change the exogenous quantity of reserves, the model traces out a series of equilibrium interest rates, which together with the respective

[44] Since in the quantitative implementation of the theory we focus on a subset of fed funds participants (see Section D.1.2 in Appendix D for our sample selection criteria), we have also run a version of (5) where the loan rate used to compute the spread s_t is the volume weighted average of loans in our sample, and the quantity of reserves Q_t is the aggregate level of reserves held by all banks in our sample. The estimate is $\gamma = -0.0057$, which is within the 95% confidence interval of the estimate reported above.

quantities of reserves, can be regarded as a model-generated "demand for reserves".

3.6 An interpolation procedure for counterfactual experiments

Several of the counterfactual and policy experiments that we conduct below involve changes in the aggregate quantity of reserves. Since our theory features ex ante heterogenity in reserve balances, changing the aggregate supply of reserves requires us to specify the underlying change in the distributions of reserve balances across banks. For example, in order to implement a \$1 bn decrease in the aggregate quantity of reserves in the model, we must specify the associated changes in the beginning-of-day distributions of reserve balances of the four bank types. How is the \$1 bn being drained exactly? Only from fast banks? Only from slow banks? Uniformly from all banks? We tackle this issue with a simple interpolation procedure that allows us to map changes in the aggregate quantity of reserves into changes in the cross-sectional distributions of reserves that is consistent with available observations.[45] The procedure is as follows.

Let $\bar{n}^i_{\mathrm{Y}}$ denote the proportion of banks of type i in our sample for the year Y, and let $\bar{F}^i_{\mathrm{Y}}$ denote the empirical beginning-of-day distribution of reserve balances across banks of type i, estimated from all trading days in year Y (as described in Section 3.3). Let Y_0 and Y_1 denote two sample years for which we have estimates of $\left\{\bar{F}^i_{\mathrm{Y}_0}, \bar{F}^i_{\mathrm{Y}_1}\right\}_{i\in\mathbb{N}}$. For each $i \in \mathbb{N}$, and each $\mathrm{Y} \in \{\mathrm{Y}_0, \mathrm{Y}_1\}$, discretize the continuous cumulative distribution function $\bar{F}^i_{\mathrm{Y}}$ with N quantiles, denoted $\left\{x^i_{\mathrm{Y}}(p_n)\right\}_{n=1}^{N}$, where $\{p_n\}_{n=0}^{N+1}$ is a sequence that satisfies $p_{N+1} = 1 - p_0 = 1$, with $p_n < p_{n+1}$ for all $n \in \{0, ..., N\}$, and $x^i_{\mathrm{Y}}(p_n)$ is the number that satisfies $F^i_{\mathrm{Y}}\left(x^i_{\mathrm{Y}}(p_n)\right) = p_n$ for each $n \in \{1, ..., N\}$.[46] For each $i \in \mathbb{N}$, $\mathrm{Y} \in \{\mathrm{Y}_0, \mathrm{Y}_1\}$, $n \in \{1, ..., N\}$, and $\omega \in \mathbb{R}$, use the pair of quantiles $\left\{x^i_{\mathrm{Y}_0}(p_n), x^i_{\mathrm{Y}_1}(p_n)\right\}$ to define the *synthetic quantile*,

$$x^i_{\mathrm{Y}_\omega}(p_n) \equiv \omega x^i_{\mathrm{Y}_1}(p_n) + (1-\omega)\, x^i_{\mathrm{Y}_0}(p_n). \tag{6}$$

[45] Empirical studies (e.g., the ones that estimate the liquidity effect discussed in Section 3.5) typically abstract from how reserve-draining or reserve-augmenting shocks are distributed in the cross section of banks. The theoretical challenge of having to specify a path for the distribution of reserve balances associated with a certain path for the aggregate quantity of reserves (which is the variable we usually regard as being under direct control of the central bank) is common to all existing micro-based models of the fed funds market that allow for heterogeneity in initial reserve holdings across banks. Afonso and Lagos (2015b), for example, parametrize the beginning-of-day distribution of reserves with a Gaussian mixture with two components, and implement changes in the aggregate quantity of reserves by draining reserves from the two components in a way that their variances and the ratio of their means remain constant (see footnote 26, and Section C.2 in the Supplemental Material of Afonso and Lagos (2015b) for details). Afonso et al. (2019), whose main quantitative experiment involves draining a large quantity of aggregate reserves, assume a two-stage draining scheme: Reserves are drained exclusively from the banks with the largest initial holdings until their reserves become low enough; and are drained proportionately from all banks thereafter.

[46] See Appendix C (Section C.1) for more details on the grids that we use in our quantitative implementation.

We then use $\omega \in \mathbb{R}$ to define a family of economies indexed by the following distribution of banks across types and distributions of reserves for each bank type $i \in \mathbb{N}$:

$$\bar{n}^i_{\mathrm{Y}_\omega} \equiv \omega \bar{n}^i_{\mathrm{Y}_1} + (1-\omega)\, \bar{n}^i_{\mathrm{Y}_0} \tag{7}$$

$$\bar{F}^i_{\mathrm{Y}_\omega}(a) \equiv \sum_{n \in \{\{1,\ldots,N\}: x_{\mathrm{Y}_\omega}(p_n) \leq a\}} (p_n - p_{n-1}), \tag{8}$$

so the corresponding aggregate quantity of reserves is

$$Q_{\mathrm{Y}_\omega} \equiv \sum_{i \in \mathbb{N}} \bar{n}^i_{\mathrm{Y}_\omega} \int a d\bar{F}^i_{\mathrm{Y}_\omega}(a). \tag{9}$$

Notice that for $\omega = 1$ the distribution of banks across types and the distributions of reserves for each bank type are as in the base year Y_1, and for $\omega = 0$ they are as in the base year Y_0. Thus, by varying ω on $[0,1]$ we can use (9) to span any aggregate level of reserves between Q_{Y_1} (the aggregate supply of reserves held by all banks in our sample in base year Y_1) and Q_{Y_0} (the aggregate supply of reserves held by all banks in our sample in base year Y_0). Conversely, for any aggregate quantity of reserves, Q, between Q_{Y_1} and Q_{Y_0}, there is an $\omega \in [0,1]$ implied by (9), denoted $\omega(Q)$, that decomposes Q into a particular distribution of banks across types and distributions of reserves for each bank type, namely $\left\{\bar{n}^i_{\mathrm{Y}_{\omega(Q)}}, \bar{F}^i_{\mathrm{Y}_{\omega(Q)}}\right\}_{i \in \mathbb{N}}$ implied by (7) and (8). For any $\omega \in [0,1]$ our procedure produces a distribution of banks across types and a set of distributions of reserves for each bank type that are linear interpolations of the corresponding distributions for the base years. We will use this procedure to conduct counterfactual and policy experiments in our quantitative model.[47]

[47] The procedure also allows for linear extrapolations, e.g., corresponding to parametrizations with $\omega < 0$ or $\omega > 1$. An alternative to our empirical interpolation/extrapolation procedure would be to integrate a fully specified capital-structure theory of the bank into our dynamic stochastic heterogeneous-bank fed-funds trading model in order to establish a theoretical link between market conditions (e.g., policy choices of administered rates and aggregate supply of reserves) and the cross section of the composition of banks' assets, and in particular, their choices of reserve balances. The main challenge would then be to ensure that the endogenous portfolio choices implied by the theory are quantitatively consistent with the empirical paths for the cross-sectional distributions of reserves that have accompanied the observed long- and medium-term changes the aggregate supply of reserves. An attractive feature of our empirical interpolation procedure is that, by construction, it ensures that this is the case (at least for moderate deviations in the aggregate supply of reserves from those prevailing the base years). We think that integrating fed funds microstructure theory with a macroeconomic theory of the capital structure of the banking sector is a promising avenue of research (see Bianchi and Bigio (2022) for work along these lines).

4 Calibration

In this section we calibrate the model to match the key statistics that describe fed funds trading activity in the year 2019.[48] The model primitives are: trading session, $[0, T]$, discount rate, r, set of bank types, $\mathbb{N}$, population shares of bank types, $\{n_i\}_{i\in\mathbb{N}}$, beginning-of-day distributions of reserve balances, $\left\{F_0^i(a)\right\}_{i\in\mathbb{N}}$, payment shock frequencies, $\{\lambda_i\}_{i\in\mathbb{N}}$, conditional size distributions of payment shocks, $\{G_{ij}\}_{i,j\in\mathbb{N}}$, bargaining powers, $\{\theta_{ij}\}_{i,j\in\mathbb{N}}$, intraday payoffs, $\{u_i\}_{i\in\mathbb{N}}$, end-of-day payoffs, $\{U_i\}_{i\in\mathbb{N}}$, and trading frequencies, $\{\beta_i\}_{i\in\mathbb{N}}$. In the quantitative implementation it is useful to augment the model to include proportional borrowing costs, $\{\kappa_i\}_{i\in\mathbb{N}}$, that proxy for institutional and regulatory considerations that affect banks' incentives to borrow in the fed funds market. In Appendix A (Section A.2) we derive a generalization of (1), (2), and (3) to the case of proportional borrowing costs.

Our calibration strategy is as follows. We regard the trading session in the model as an average trading day in a typical 14-day reserve maintenance period. As discussed in Section 3, there is little trading activity between 9:00 pm on day $h-1$ and 9:00 am on day h, so we think of $[0, T]$ as corresponding to the time interval that starts at 9:00 am and ends at 6:30 pm EST on an actual trading day. In the quantitative implementation of the theory we discretize the time interval $[0, T]$ into 800 periods, so each period in the model corresponds approximately to a 42-second interval of the trading day.[49] With a model period this short, we abstract from pure

[48] We use 2019 as the baseline year for our calibration because it has the lowest level of total reserves of the current post-GFC-regulation policy regime (see Figure 16). As we explain in Section 6, this allows us to test the quantitative predictions of the theory by varying the level of excess reserves from their lowest level in the post-GFC-regulation era, to the level they reached in the year 2017 (a post-GFC-regulation year with a level of excess reserves that is close to the pre-2020 historical peak).

[49] A model period corresponding to 42 seconds is short enough to approximate the empirical frequency of loans even for the most active banks. Payment shocks, however, are much more frequent than loans: In Section 3.2 we had to use a period length of 1 second in order to get a good approximation to the empirical frequency of payment shocks (especially for fast banks, which typically experience several payment shocks per minute, and sometimes even more than one payment shock per second). In order to allow for such high frequency of payment shocks, we could simply discretize $[0, T]$ into 34,200 periods, each corresponding to 1 second. With so many periods, however, the computational burden would increase significantly, so we took a different approach. Payment shocks, although very frequent, are computationally cheap since they involve no optimization (they are just "forced" transfers between banks). Loans on the other hand, are computationally more expensive (they involve maximization of the joint surplus), but are also significantly less frequent than payment shocks in the data. In the quantitative implementation of the model, we balance these considerations as follows. We regard each model period as being composed of 42 *subperiods*, each corresponding to 1 second in the actual trading day. We then treat the first 41 subperiods as "payment-shock rounds" (in each of these rounds, there are only bilateral payment shocks among banks), and treat the 42^{nd} subperiod as a "loan round" in which banks get bilateral opportunities to negotiate loans. In sum, this allows us to have payment shocks that are as frequent as 1 per second, and loans that are as frequent as one every 42 seconds, while economizing on computation time.

discounting, and set $r = 0$. As described in Section 3.1, we sort institutions into four types, i.e., $\mathbb{N} = \{F, M, S, GSE\}$, based on their participation rates in the volume of fed funds trade, business model, and regulatory treatment. We set $n_i = N_i / \sum_{j \in \mathbb{N}} N_j$, where N_i denotes the number of banks of type $i \in \mathbb{N}$ in the base year. We interpret reserve balances in the theory as a bank's *unencumbered reserves* in the data, and therefore set the theoretical beginning-of-day distributions, $\left\{F_0^i(\cdot)\right\}_{i \in \mathbb{N}}$, equal to the corresponding empirical kernel estimates reported in Section 3.3. The frequencies of payment shocks, $\{\lambda_i\}_{i \in \mathbb{N}}$, are calibrated to match the empirical one-second frequencies of payment shocks reported in Section 3.2.[50] The size distributions of payment shocks, $\{G_{ij}\}_{i,j \in \mathbb{N}}$, are set equal to the corresponding empirical kernel estimates reported in Section 3.2. We set $\theta_{ij} = \underline{\theta}$ if $i \in \{GSE\}$ and $j \in \mathbb{N} \setminus \{GSE\}$, and $\theta_{ij} = 1/2$ otherwise (i.e., unless one of the parties in the trade is a GSE, we abstract from differences in relative market power purely driven by a bank's *type*). We set $u_i(a) = 0$ for all $(a, i) \in \mathbb{R} \times \mathbb{N}$ (i.e., we abstract from banks' *intraday* payoffs from holding reserves, such as the regulatory costs associated with running an intraday overdraft with the Federal Reserve).

End-of-day payoffs are parametrized by

$$U_i(a) = \left(1 + \mathbb{I}_{\{0 \leq a\}} \bar{\iota}_r + \mathbb{I}_{\{a<0\}} \bar{\iota}_w\right) a, \tag{10}$$

for any $(a, i) \in \mathbb{R} \times \{F, M, S\}$, where $\mathbb{I}_{\{a<0\}}$ is an indicator function that equals 1 if $a < 0$ and 0 otherwise, $\bar{\iota}_r \equiv \iota_r + \iota_\ell$, $\bar{\iota}_w \equiv \iota_w + \iota_\ell + \iota_s$, and a denotes end-of-day balance in excess of reserve requirements.[51] We use ι_r to denote the interest rate that a bank earns from the Federal Reserve per dollar of end-of-day reserves (IOR), and ι_ℓ to represent a *liquidity return* that proxies for a bank's benefits from holding reserves that are not captured by the administered rates.[52] We use ι_w to denote the (primary credit) Discount-Window rate (DWR) that the Federal Reserve

See Appendix C for a more detailed discussion of computational issues.

[50] In the discrete-time approximation that we compute, λ_i corresponds to the probability that a bank of type i receives a payment shock in a one-second time interval (see footnote 49).

[51] Since our calibration strategy maps beginning-of-day reserve balances in the theory to *unemcumbered reserves* in the data, which are reserves in excess of reserve requirements (and net of predictable payments), we specify a bank's end-of-day payoff as a function of its *excess reserves*. This allows us to have end-of-day payoff functions that are *type specific* but not *bank specific*, despite the fact that in the data, two banks will typically have different reserve requirements even if they are of the same type, $i \in \mathbb{N}$. To see this, let $\mathcal{U}_i(b, \underline{b})$ be the end-of-day payoff of a bank of type i with reserve requirement $\underline{b}$, and reserve balance b (gross of the reserve requirement). We would parametrize this function as $\mathcal{U}_i(b; \underline{b}) = b + \bar{\iota}_r \underline{b} + \left(\mathbb{I}_{\{0 \leq b - \underline{b}\}} \bar{\iota}_r + \mathbb{I}_{\{b - \underline{b} < 0\}} \bar{\iota}_w\right)(b - \underline{b})$, which is equivalent to $U_i(a)$ in the sense that they only differ by a constant, i.e., $\mathcal{U}_i(b; \underline{b}) = U_i(a) + (1 + \bar{\iota}_r)\underline{b}$, where $a \equiv b - \underline{b}$ denotes excess reserves, as in (10).

[52] For example, ι_ℓ may stand in for the additional return associated with the use of reserves as means of payment, or for the additional return resulting from lending reserves outside the fed funds market (e.g., in repo markets, or as loans to corporate or retail bank customers).

charges a bank that needs to borrow to make up an end-of-day shortfall of reserves relative to the required level, and ι_s to represent the additional costs associated with borrowing from the Discount Window.[53] For GSEs, the end-of-day payoff is $U_{GSE}(a) = \left(1 + \mathbb{I}_{\{0 \leq a\}}\bar{\iota}_o + \mathbb{I}_{\{a<0\}}\bar{\iota}_w\right) a$, with $\bar{\iota}_o \equiv \iota_o + \iota_\ell$, where ι_o denotes the interest rate that the Federal Reserve offers on the overnight reverse repo facility.[54] The administered rates, i.e., ι_r, ι_w, and ι_o, are set equal to their empirical counterparts in the base year.

The remaining eleven parameters, $\underline{\theta}$, ι_ℓ, ι_s, and $\{\beta_i, \kappa_i\}_{i\in\mathbb{N}}$, are calibrated so that the equilibrium of the model matches the following eleven empirical moments: (1) average value-weighted fed funds rate; (2) average value-weighted fed funds rate for loans with rates lower than the IOR; (3) regression estimates of the "liquidity effect" (at the average level of aggregate reserves outstanding in the base year, as reported in Section 3.5); (4) ratio of the average number of loans traded by banks of type F relative to the average number of loans traded by all banks; (5)-(8) reallocation indices $\{\mathcal{R}_i\}_{i\in\mathbb{N}}$ (as defined in Section 3.1); (9)-(11) participation rates $\{\mathcal{P}_i\}_{i\in\mathbb{N}\setminus\{F\}}$ (as defined in Section 3.1).[55]

Table 1 reports the parameter values, the targeted moments, and the corresponding theoretical moments for the 2019 calibration. Banks of type F, M, and S, accounted for about 1%, 4%, and 92%, of all the institutions that were active in the fed funds market in 2019, respectively. To interpret the frequencies of payment shocks, $\{\lambda_i\}_{i\in\mathbb{N}}$, recall that λ_i represents the probability that a bank of type i receives a payment shock in a one-second time interval, so for example, $\lambda_M = 0.257$ implies a bank of type M receives a payment shock approximately every 4 seconds, on average. Similarly, $\lambda_F = 0.920$ implies a bank of type F receives approximately a payment shock per second, and $\lambda_S = 0.011$ implies a bank of type S receives a payment shock approximately every 90 seconds, on average. The values of ι_w (DWR), ι_r (IOR), and ι_o (ON-RRP) are set to 3.00%, 2.35%, and 2.25% per annum, respectively, which were the administered

[53] It is a well documented phenomenom that banks often borrow from other banks at a premium over the DWR. The most common explanation is stigma associated with Discount-Window borrowing (see e.g., Artuç and Demiralp (2010), Ennis and Weinberg (2013), Armantier et al. (2015), Ennis (2019), and Klee et al. (2021)). Another reason why fed funds may trade above the DWR is that Discount-Window loans must be collateralized, and Reserve Banks require a perfected security interest in all collateral pledged to secure these loans, which entails costs for the borrower. Assets accepted as collateral are assigned a lendable value deemed appropriate by the Reserve Bank that issues the loan (a market value or an internally-modelled fair market value estimate multiplied by a margin, possibly adjusted as a function of the financial condition of the borrowing institution). For details, see https://www.frbdiscountwindow.org/Pages/General-Information/The-Discount-Window.

[54] We use ι_o rather than ι_r in the payoff for GSEs because regulation prevents them from earning interest on reserves. We use ι_r rather than ι_o in the payoffs of other bank types because $\iota_o \leq \iota_r$ throughout our sample.

[55] The participation rate of type F banks is not an explicit calibration target because it is implied by the participation rates of the other three bank types, since $\sum_{i\in\mathbb{N}} \mathcal{P}_i = 2$.

policy rates in effect from May through July of 2019.

The calibration strategy delivers a liquidity return (ι_ℓ) of 4.9 bps per annum, an additional cost associated to Discount-Window borrowing (ι_s) of 75.8 bps per annum (i.e., about one quarter of the DWR), and $\underline{\theta} = 1/20$, which means that a GSE reaps 5% of the total gains from lending to a non-GSE. The frequency of trade, β_i, is interpreted as the probability that a bank of type i contacts a trading partner during a 42-second time interval. Thus, the calibrated values $\{\beta_i\}_{i\in\mathbb{N}}$ imply that banks of type F, M, S, and GSE, trade fed funds approximately every 23 minutes, 4.86 hours, 16.7 hours, and 3.24 hours, respectively. The calibration also ensures that the magnitude of the "liquidity effect" in the theory is in line with the range of empirical estimates for the year 2019 reported in Section 3.5.[56] The borrowing costs needed to match the calibration targets, $\{\kappa_i\}_{i\in\mathbb{N}}$, which proxy for institutional and regulatory considerations that affect banks' incentives to buy fed funds, are positive for banks of type F and S, and zero for banks of type M.[57]

5 Validation

In this section we report the model fit of empirical price and quantity observations not targeted in the calibration. We organize the material in four sections. The first focuses on the cross-sectional distribution of loan rates for all transactions. The second, on the distribution of loan rates for transactions with rates higher than the DWR. The third, on the distribution of borrowing and lending rates for each bank type. The fourth, on the trading network.

5.1 Distribution of loan rates

Figure 12 shows the empirical and theoretical cumulative distribution functions of bilateral negotiated fed funds rates in the year 2019, along with the administered rates prevailing in the sample period (all expressed in percent per annum).[58] The model delivers a reasonable fit for

[56] Figure 11 shows the magnitude of the liquidity effect in the calibrated model along with the confidence bands for the regression estimates for the sample period 2019/05/02–2019/09/13 presented in Section 3.5. In the model, the liquidity effect is computed by extracting $100 bn reserves (approximately 2 standard deviations of the size distribution of reserve-draining shocks) using the procedure described in Section 3.6. The figure shows that the model-generated liquidity effect is within the 95% confidence bands of the empirical estimate.

[57] The value of κ_{GSE} is set large enough to match the observation that GSEs essentially do not borrow in the fed funds market, but its exact value is inconsequential.

[58] Data are for every trading day in the period 2019/06/06–2019/07/31, which covers eight reserve maintenance periods during which the policy rate remained constant and the administered rates (DWR, IOR, ONRRP) were as in our baseline calibration. To obtain the equilibrium rates for 2019, the model is calibrated as in Table 1.

the distribution of bilateral fed funds rate, which was not targeted in the calibration.[59]

5.2 Conditional distribution of loan rates in excess of DWR

In the model, as in the data, banks sometimes trade at rates higher than the DWR. In the model this is possible because ι_s is calibrated to a positive value. In this section we compare the theoretical and empirical distributions of traded rates conditional on the rate being higher than the DWR, which were not targeted in our calibration. During the sample period 2019/06/06–2019/07/31 the DWR was set at 3%; the 10th percentile, mean, and 90th percentile were 3%, 3.1%, and 3.3%, respectively, both in the data and in the calibrated model. The maximum loan rate observed in our data sample was 3.45%, and the maximum possible rate a bank is willing to pay in the equilibrium of the model is $\iota_w + \iota_\ell + \iota_s = 0.038$.

5.3 Bid-ask spreads

Each of the panels on the right side of Figure 13 shows an empirical cumulative distribution function of borrowed reserves over borrowing rates, denoted $\mathcal{H}_i^B$ (represented by a solid line), and an empirical cumulative distribution function of lent reserves over lending rates, denoted $\mathcal{H}_i^L$ (represented by a dashed line), for $i \in \{F, M, S\}$. In words, $\mathcal{H}_i^B(\iota)$ is the proportion of reserves borrowed by banks of type i that bear interest rates lower than ι, and $\mathcal{H}_i^L(\iota)$ is the proportion of reserves lent by banks of type i that bear interest rates lower than ι.

Each of the panels on the left side of Figure 13 shows the theoretical counterpart of the adjoining right-side panel. The top-left and middle-left panels show the theory predicts $\mathcal{H}_i^L(\iota) \leq \mathcal{H}_i^B(\iota)$ for $i \in \{F, M\}$. That is, banks of type F and M tend to borrow at lower rates than they earn when they lend. This theoretical prediction also holds in the data, as long as we focus on loans with rates that are not lower than the IOR (2.35%).[60] In contrast, according to the bottom-left panel, the theory predicts $\mathcal{H}_S^B(\iota) \leq \mathcal{H}_S^L(\iota)$, i.e., banks of type S tend to borrow

[59] The model, however, does not generate enough dispersion of rates relative to the data. This is the case for loans that trade above the IOR, but also for loans that trade below the IOR. One way to match the larger empirical dispersion of loans with above-IOR rates would be to allow for heterogenity in bargaining powers across banks of types, i.e., to let θ_{ij} differ in trades between two non-GSEs. Notice that a significant part of the large dispersion for below-IOR trades in the data comes from trades with rates lower than the ONRRP. This observation is difficult to rationalize through the lens of the theory, and may be indicative of some repo loans being misclassified as fed funds in our dataset (e.g., as suggested by Armantier and Copeland (2015)).

[60] As mentioned in footnote 59, rates below the IOR are likely to correspond to repo loans that are misclassified as fed funds by the Furfine algorithm.

at higher rates than they earn when they lend.[61] This theoretical prediction also holds in the data, and the fit is remarkably good for loans with rates that are not lower than the IOR.

5.4 Distributions of loan rates between pairs of bank types

Each of the panels on the right side of Figure 14 shows an empirical cumulative distribution of rates for loans extended from bank type $i \in \{F, M, S\}$ to bank type $j \in \{F, M, S\}$. For example, for each interest rate ι on the horizontal axis, the height of the curve labeled "S" in the top right panel represents the fraction of the total volume of loans extended from banks of type "F" to banks of type "S" with interest rate less than or equal to ι. Each of the panels on the left side of Figure 14 shows the theoretical counterparts of the adjoining right-side panel. The theory predicts that, regardless of lender type, banks of type "S" tend to borrow at higher rates than banks of other types, and this is true in the data.[62]

5.5 Fed funds trading network

Figure 15 shows the empirical fed funds trading network for the year 2019 (bottom panel) and the corresponding trading network generated by the model (top panel). As explained in Section 3.1, these network plots show the location of the four bank types in the coordinate axes defined by the reallocation index, $\mathcal{R}_i$, and the participation rate, $\mathcal{P}_i$, and convey information on the sizes of the flows of reserves associated with fed funds lending across and within bank types, as well as on the average interest rates on the underlying loans.[63]

The theoretical network matches several characteristics of the empirical one. For example, the model replicates quite well the direction and volume of the loans between and within bank types (represented by the direction of the arrows, their width, and the sizes of the nodes). In this regard, one difference is that the model predicts a significant volume of loans from GSEs to banks of type S that is not present in the data.[64] The model predicts that GSEs tend to

[61] The model counterparts of $\mathcal{H}_S^B(\iota)$ and $\mathcal{H}_S^L(\iota)$ are constructed excluding loans between a GSE and a bank of type S. The rationale is that our model abstracts from the institutional details that make these trades very rare in the data. For example, there was only one loan of this kind in our sample period.

[62] The theory also predicts that banks of type "M" tend to borrow at higher rates than banks of type "F", but this is not as evident in the data.

[63] In comparing the top and bottom panels of Figure 15, notice that while the positions of the four nodes in $\mathcal{R}_i$-$\mathcal{P}_i$ space have been used as calibration targets, the remaining collection of statistics that shape these network representations were not targeted. This includes the size and color of each node, and the direction, color, and width of each arrow.

[64] This discrepancy is likely due to the fact that our theory abstracts from the real-world institutional details that cause GSEs to lend reserves only to a relatively small subset of counterparties, which tend to be big banks

lend at lower rates than banks, and in particular, lower than the rates that banks of type F tend to charge banks of type M. However, the oposite is true in the data.[65]

6 Aggregate demand for reserves

It is customary to think of the fed funds rate in the context of a static, perfectly competitive loan market, i.e., as being determined by the intersection of a vertical supply of reserves controlled by the central bank, and an aggregate downward-sloping demand for reserves implied by the solution to a reserve-management problem faced by individual banks.[66] Our over-the-counter theory does not involve a bank-level reserve demand.[67] However, as we vary the aggregate quantity of reserves, Q, e.g., by changing the beginning-of-day distributions of reserves, our theory can generate a negative relationship between Q and a volume-weighted average of all the equilibrium bilateral loan rates, ι^*, which can be interpreted as the *aggregate demand for reserves* implied by the theory. We can write this relationship as $\iota^* = \mathcal{D}(Q;\Pi)$, where $Q = \sum_{i\in\mathbb{N}} n_i \int a dF_0^i(a)$ is the supply of reserves, and $\Pi \equiv \{\beta_i, \lambda_i, \{\theta_{ij}, G_{ij}\}_{j\in\mathbb{N}}, u_i, U_i\}_{i\in\mathbb{N}}$ is the full set of model primitives, which makes clear that the mapping depends on all the structural parameters of the model.[68]

Consider the model calibrated to the year 2019, as described in Table 1. Then, using the notation introduced in Section 3.6, let $\mathrm{Y}_0 = 2017$ and $\mathrm{Y}_1 = 2019$, i.e., Y_0 and Y_1 represent the years 2017, and 2019, respectively, with $\bar{n}^i_{2017}$ and $\bar{F}^i_{2017}$ given by the estimates reported in Section 3.3. Construct a grid, $\mathbb{G} \subset \mathbb{R}$ for ω, and for each $\omega \in \mathbb{G}$, use the interpolation procedure described by (7) and (8) to generate the sample $\left\{(\bar{n}^i_{\mathrm{Y}_\omega}, \bar{F}^i_{\mathrm{Y}_\omega})\right\}_{(i,\omega)\in\mathbb{N}\times\mathbb{G}}$. For each pair $(\bar{n}^i_{\mathrm{Y}_\omega}, \bar{F}^i_{\mathrm{Y}_\omega})$, compute the equilibrium value-weighted fed funds rate, which we denote $\iota^*_{\mathrm{Y}_\omega}$, and

that are very active in fed funds trading.

[65] This discrepancy may be due to the fact that the Furfine algorithm, used to identify overnight uncollateralized loans from the universe of Fedwire transfers, may pick up some overnight *collateralized* loans (i.e., repos) that trade at lower rates. Armantier and Copeland (2015) provide some evidence consistent with this interpretation.

[66] See footnote 1.

[67] This is because in our dynamic over-the-counter theory, the end-of-day reserve holding of an individual bank is a random variable that depends on the bank's beginning-of-day balance, the number of counterparties it encounters throughout the trading session, and the bilateral bargaining outcomes that in turn depend on the counterparties' individual characteristics (such as their reserve holdings at the time of the trade, bargaining powers, and abilities to contact other counterparties). In other words, the fact that our theory is distinctively non-Walrasian implies there is no natural or useful counterpart to the notion of an optimal quantity of reserves chosen by an individual bank who can borrow and lend frictionlessly at a given market interest rate.

[68] For example, equation (16) in Afonso and Lagos (2015b) gives an explicit formula for this mapping for a special case of our model that allows an analytical solution (identical banks and heterogeneity in reserve balances restricted to the set $\{0, 1, 2\}$).

let $Q_{\Upsilon_\omega} \equiv \sum_{i\in\mathbb{N}} \bar{n}^i_{\Upsilon_\omega} \int a d\bar{F}^i_{\Upsilon_\omega}(a)$. This procedure delivers a collection of pairs, $\left\{(Q_{\Upsilon_\omega}, \iota^*_{\Upsilon_\omega})\right\}_{\omega\in\mathbb{G}}$, that define the mapping $\iota^*_{\Upsilon_\omega} = \mathcal{D}(Q_{\Upsilon_\omega}; \Pi)$. This mapping—the aggregate demand for reserves generated by the theory—is the curve labeled "Benchmark" in all panels of Figure 17.[69]

We have calibrated the model using empirical beginning-of-day distributions of reserves that are net of predictable transfers, net of Regulation D and LCR requirements, and that only include banks that had at least one fed funds transaction in the baseline year. Hence, the measure "Q" of aggregate reserves reported in the primary horizontal axis of any figure that displays the aggregate demand for reserves generated by the theory (e.g., Figure 17, Figure 19, Figure 21, or the right panels of Figure 18) is the sum of *excess* reserves (net of Regulation D and LCR requirements) across *active* banks (in the sense of having had at least one fed funds transaction in the baseline year). We term this measure of aggregate reserves *active excess reserves*, to distinguish it from *total reserves*, which is gross of reserve requirements, and includes *all* institutions that hold reserve balances at the Federal Reserve Banks.[70]

The notion of active excess reserves arises naturally in our theory, since reserve requirements determine incentives to hold reserves, and reserve balances at banks that are inactive in the fed funds market are inconsequential. However, we want to establish a mapping between our notion of active excess reserves and the notion of total reserves for two reasons. First, doing so will make our results easier to interpret, since the latter is a well known and readily available measure of aggregate reserves.[71] Second, for some of our quantitative exercises (e.g., the demand estimation illustrated in the top-right panel of Figure 18) we will want to overlay empirical observations for total reserves, which we may denote $\{Q^D_t\}$, on the theoretical demand for reserves, which is computed as a function of active excess reserves, which we may denote Q^M_t. For these two reasons, in Appendix D (Section D.2.5) we show how to "translate" the value of Q^D_t into a value of Q^M_t using a linear mapping that preserves the variation in the relevant sample $\{Q^D_t\}$ and that is consistent with the observed relationship between the sample mean of $\{Q^D_t\}$ and the

[69] We use 2017 and 2019 as endpoints for our interpolation procedure because this choice maximizes the sample variation in total reserves during the post-GFC-regulation era (prior to the large reserve injection that took place in response to the COVID shock in the year 2020). Specifically, as illustrated in Figure 16, 2017 is the post-GFC-regulation year with highest level of total reserves ($2,254.27 bn, which is roughly the pre-2020 historical peak), while the year 2019 has the lowest level of total reserves in the post-GFC-regulation era (roughly $1,568.26 bn).

[70] Fed funds transfers approximately sum to zero in our sample of active banks, so the fact that the beginning-of-day distributions that we feed the theory are net of predictable transfers does not contribute much to the difference between *total reserves* and *active excess reserves*.

[71] E.g., Total reserves at weekly frequency is published in *Federal Reserve Balance Sheet: Factors Affecting Reserve Balances - H.4.1* (shown in Figure 16), and available at monthly frequency as "TOTRESNS" at https://fred.stlouisfed.org.

sample mean of $\{Q_t^M\}$ in the two base years that we use to derive the theoretical demand for reserves (i.e., 2017 and 2019). Whenever a figure shows active excess reserves on the primary horizontal axis, we often include a secondary horizontal axis (above the figure) that shows the corresponding values for total reserves, to facilitate the translation between these units.[72]

In Section 6.1 we study how changes in structural parameters affect the position and shape of the theoretical aggregate demand for reserves. In Section 6.2 we use these insights and our quantitatve theory to tackle the well-known empirical challenges involved in obtaining global estimates of the slope of the aggregate demand for reserves.

6.1 Reserve demand counterfactuals

In this section we study how the theoretical aggregate demand responds to changes in the administered rates and key marketstructure parameters. The results are reported in Figure 17.

In all panels of Figure 17, the curve labeled "Benchmark" is the theoretical aggregate demand $\iota^*_{\mathrm{Y}_\omega} = \mathcal{D}(Q_{\mathrm{Y}_\omega}; \Pi)$ for the model calibrated as in Table 1, and with $\iota^*_{\mathrm{Y}_\omega}$ and Q_{Y_ω} computed with the interpolation procedure described in Section 3.6, for $\mathrm{Y}_0 = 2017$ and $\mathrm{Y}_1 = 2019$. We wish to make two observations about this demand for reserves generated by our theory. First, it exhibits the kind of logistic sigmoid shape that is characteristic of the popular "Poole model" (see, e.g., p. 784 in Poole (1968)). Second, for the baseline calibration, the demand lies within the DWR-IOR corridor. This means that despite there being GSEs that earn a lower interest on reserves than banks (i.e., the ONRRP rather than the IOR), the average equilibrium fed funds rate is above the IOR for all levels of reserves in the baseline calibration.[73]

The top-left panel of Figure 17 shows two experiments. In the first, the DWR is increased by 50 bps (so that it is equal to the ONRRP plus 125 bps, rather than equal to the ONRRP plus 75 bps as in the baseline calibration). This shifts the demand up, with the size of the shift being decreasing in the quantity of reserves. Intuitively, the DWR has little effect on the equilibrium average interest rate when reserves are abundant, but a stronger effect when reserves are scarce. The second experiment consists of increasing the IOR by 15 bps (so that

[72] See, e.g., Figure 19, Figure 21, and the right panels of Figure 18. The average quantity of active excess reserves was about \$1,150.86 bn in 2017, and \$910.73 bn in 2019. Thus, by varying ω on $[0, 1]$ and using (9), our choice of endpoints ($\mathrm{Y}_0 = 2017$ and $\mathrm{Y}_1 = 2019$) allow us to interpolate any level of active excess reserves between \$1,150.86 bn and \$910.73 bn. The corresponding quantities of total reserves for 2017 and 2019 are \$2,254.27 bn and \$1,568.26 bn, respectively.

[73] This observation is in line with the data, since the EFFR was consistently above the IOR during the 2019 sample period that we used as baseline for our calibration (see Figure 16).

it is equal to the ONRRP plus 25 bps, rather than equal to the ONRRP plus 10 bps as in the baseline calibration). This policy change increases the equilibrium average rate when the quantity of reserves is relatively large, and it also implies that—if reserves are abundant enough—the equilibrium average fed funds rate lies between the IOR and the ONRRP.[74] The top-right panel of Figure 17 shows that increasing all administered rates (DWR, IOR, and ONRRP) by 75 bps simply causes a parallel upward shift in the aggregate demand for reserves.

The bottom-left panel of Figure 17 shows three experiments. The first, is to multiply the trading probabilities of all bank types by a factor of 10, which makes the marketstructure more competitive (i.e., "less OTC"), reducing rates (and increasing the slope of the demand for reserves) when the quantity of reserves is low to moderate. The second marketstructure experiment is to set $\beta_F = 0$, which effectively excludes all banks of type F from fed funds trading, and causes the aggregate demand for reserves to rotate clockwise around an intermediate quantity of aggregate reserves (about \$700 bn). This experiment causes the average fed funds rate to rise for relatively low levels of reserves, and to fall for relatively high levels of reserves. This rotation reflects the intermediation role that banks of type F play in the equilibrium: When reserves are scarce there are many banks with deficient reserve balances who, absent type-F counterparties, find it more difficult to meet a counterparty eager to lend, which reduces their effective market power thus leading to higher negotiated loan rates on average. When reserves are abundant there are many banks with excess reserves who, absent type-F counterparties, find it more difficult to meet a counterparty eager to borrow, thus leading to lower average negotiated rates. The third experiment is to eliminate the proportional borrowing costs in the baseline calibration. This shifts up the aggregate demand for reserves, reflecting that the borrowing costs stifle individual banks' incentives to borrow.

The bottom-right panel of Figure 17 shows two experiments involving payment risk: One where we eliminate payment shocks for all banks, i.e., we set $\lambda_i = 0$ for all $i \in \mathbb{N}$, and another where we set $\lambda_i = \lambda_F$ for all $i \in \mathbb{N}$, i.e., we assume all bank types experience the same—very high—frequency of payment shocks as banks of type F. In both cases the result is an upward shift in the demand for reserves. In the second experiment the demand shifts up due to a heightened precautionary motive for holding reserves. In the first experiment the upward shift occurs because of a compositional effect: In an equilibrium with payment shocks, there

[74] This observation is in line with the data, since the EFFR was consistently between the IOR and the ONRRP during most of the post-GFC period ranging from 2008 until 2018 when, as in this experiment, the IOR was set 25 bps above the ONRRP.

are banks that borrow because their balances are deficient, and banks with moderate positive reserves that borrow to self-insure against payment shocks. The former values reserves more, and thus is willing to pay higher rates than the latter. The precautionary motive for borrowing disappears when $\lambda_i = 0$ for all $i \in \mathbb{N}$, and therefore the average negotiated rate increases.[75]

6.2 Quantitative-theoretic reserve demand estimation

As discussed in Section 1, the floor system that the Federal Reserve has chosen as operating framework for monetary policy implementation relies on the ability to ascertain what level of reserves is "ample enough" so that active management of the supply of reserves is not required to instrument the fed funds rate target. In other words, operating a floor system requires *global* estimates of the aggregate demand for reserves, and in particular, reliable estimates of its slope for wide ranges of the aggregate supply of reserves. This presents two empirical challenges.

The first challenge is the potential endogeneity of the supply of reserves, which complicates the estimation of the demand equation. In terms of the simple demand-and-supply picture in the first panel of Figure 1, the issue is to identify *exogenous* variation in the quantity of reserves that allow to estimate the slope of the demand. This problem is well-understood, and has been addressed by the empirical literature that studies the *liquidity effect*.[76]

The second challenge is to obtain *global* estimates for the slope of the demand; i.e., to identify the slope of the demand for a range of values of the supply of reserves that is wide enough to span the "abundant", "ample", and "scarce" segments of the demand curve, as illustrated in the top-right panel of Figure 1. The issue is that, empirically, spanning substantial variation in the supply of reserves usually entails spanning a substantial period of time during which the demand for reserves itself is likely to have shifted due to structural changes, e.g., in the marketstructure of the fed funds market, or in banks' incentives to hold reserves (due to changes in policy, regulation, or portfolio allocation frameworks within banks).[77]

[75] The fact that the size of the upward shift that results from setting $\lambda_i = 0$ for all $i \in \mathbb{N}$ is decreasing in the quantity of reserves is consistent with this intuition.

[76] We discussed these identification issues in Section 3.5, where we also reported estimates of the slope of the reserve demand for different sample periods based on the identification strategies of Hamilton (1997), Carpenter and Demiralp (2006), and Afonso et al. (2022).

[77] These are the kinds of shifts in the demand for reserves that we studied in Section 6.1. The bottom panels of Figure 1 show situations in which structural parameters are Π_i at the time the quantity-price pair (Q_i, r_i^*) is observed, for $i \in \{0, 1\}$. The problem is that, without controlling for the structural change from Π_0 to Π_1, other considerations may lead one to assume the observations $\{(Q_i, r_i^*)\}_{i \in \{0,1\}}$ lie on a single demand curve, and therefore overestimate (in the example in the bottom-left panel) or underestimate (in the example in the bottom-right panel) the (absolute value of the) slope.

This low-frequency demand-shift identification problem has not been overcome by the empirical literature on *liquidity effects*—possibly due to limited theoretical guidance on the key structural variables that determine the shape and position of the aggregate demand for reserves. Available empirical estimates of liquidity effects tend to be *local*, i.e., estimated from daily time-series variation in the quantity of reserves over relatively short sample periods during which the average quantity of reserves remains relatively stable.[78] We will use our quantitative model to bridge this *local-global gap*. The idea is to use the structure imposed by the theory, i.e., the equilibrium *aggregate demand relationship*, $\iota^* = \mathcal{D}(Q;\Pi)$, with the microstructure and policy parameters, Π, calibrated to match the key micro-level and market-level moments that describe the fed funds market—and in particular the available *local* estimates of the liquidity effect in the base year—to estimate the *global* shape of the aggregate demand for reserves.[79]

To motivate and illustrate our quantitative-theoretic identification approach, consider Figure 18. The top-left panel displays pairs of empirical observations of the total quantity of reserves, and the corresponding EFFR-IOR spread for every trading day in the sample period 2017/01/20–2019/09/13. Through the lens of standard theory (e.g., Poole (1968)), each of these observations depicts the intersection point of the supply and demand for reserves on a given day. To inform monetary policy operations, one needs to estimate the liquidity effect for each level of reserves over a wide range of reserves, which can be done by estimating an aggregate demand for reserves. A natural approach is to posit a flexible reduced-form model of the demand for reserves, e.g., $s_t = D(Q_t)$, where s_t denotes the EFFR-IOR spread on day t and Q_t denotes the aggregate quantity of reserves at the end of day t, with

$$D(Q_t) \equiv \underline{s} + \frac{\overline{s} - \underline{s}}{1 + e^{(Q_t - Q_0)\xi}}, \tag{11}$$

and estimate the parameters $(\underline{s}, \overline{s}, \xi, Q_0)$. The top-left panel of Figure 18 displays the fitted

[78] Hamilton (1997), Carpenter and Demiralp (2006), and our estimate of the coefficient γ in (5) are examples of this standard methodology. Afonso et al. (2022) follow an alternative methodology that involves estimating a time-varying vector autoregressive model at daily frequency (with an instrumental variable approach to address endogeneity of the supply of reserves) to obtain a 10-year time series of daily estimates of the elasticity of the fed funds rate to (instrumented) variation in the aggregate quantity of reserves (from 2010 until 2020). Their estimation, however, cannot recover the whole demand function. The reason is that without information on whether structural factors have shifted the demand schedule during the sample period, it is not possible to infer the global shape of the reserve demand from a sequence of (local, linear, daily) estimates of the sensitivity of the fed funds rate to (instrumented) changes in aggregate reserves. Having said this, below we will find that the reduced-form estimates from Afonso et al. (2022) can be a useful guide once complemented with our quantitative theory, which can help identify the structural shifts in the demand for reserves.

[79] Alvarez and Argente (2023) use a similar strategy to extrapolate a demand for cash-paid Uber rides in Mexico using relatively narrow empirical variation in prices.

demand curve that results from estimating (11) on the full sample (2017/01/20–2019/09/13) by nonlinear least-squares.[80] This estimation presumes all observations in the sample lie on a single demand curve.[81] The estimated slope evaluated at the mean quantity of total reserves for the full sample (about $1,974.69 bn) is −0.016, which means a $1 bn decrease in total reserves increases the EFFR by 0.016 bps when the supply of total reserves is around $2 tn. This local estimate (at about $2 tn) is similar to the linear estimates reported in Section 3.5.

In Section 6.1 we showed that keeping the DWR-ONRRP spread constant, changes in the IOR-ONRRP spread shift the demand for reserves. This minimal theoretical insight implies that not all the data points from the sample period 2017/01/20–2019/09/13 plotted in the top-left panel of Figure 18 lie on the same demand curve (contrary to what we implicitly assumed when running (11) on the full sample). The bottom-left panel of Figure 18 displays the same data points as the top-left panel, but partitioned into four subsamples, each determined by the size of the IOR-ONRRP spread: 10 bps (2019/05/02–2019/09/13), 15 bps (2018/12/20–2019/05/01), 20 bps (2018/06/14–2018/12/19), or 25 bps (2017/01/20–2018/06/13).[82] The bottom-left panel also displays the four fitted demand curves that result from estimating (11) on each subsample by nonlinear least-squares.

To illustrate the perils associated with the atheoretical demand estimation in the top-left panel of Figure 18, we focus on the demand estimation for the policy regime with IOR-ONRRP equal to 10 bps in the bottom-left panel, and highlight two discrepancies. First, the liquidity effect at about $1,974.69 bn (the mean of *total reserves* for the full sample) is −0.0001 bps; but it was estimated to be −0.016 bps in the full sample—much bigger in absolute value.[83] Second, suppose we want to use the estimated demand to identify the quantity of reserves that determines the end of the "ample" and the beginning of the "abundant" range for reserves, i.e., we want to estimate a quantity such as the Q_1 illustrated in the top-right panel of Figure 1. For practical purposes we adopt the convention that a supply of reserves, Q, is considered

[80] See Appendix D (Section D.2.4) for details.

[81] Afonso et al. (2022, Sec. 6), for example, justify this particular identifying assumption by splitting their sample period (2010–2021/03/29) according to the different low-frequency cycles of expansion and contraction of the Federal Reserve balance sheet. Specifically, they split it into three periods: the initial post-GFC expansionary period (2010–2014), the subsequent post-GFC and pre-COVID contractionary period (2015–2020/3/13), and the most recent post-COVID expansionary period (2020/03/16–2021/03/29). Thus, all the data points displayed in our Figure 18 belong to their pre-COVID contractionary period, which Afonso et al. (2022) fit with a single reduced-form demand curve (the gray curve in their Figure 9, p. 30), like we do in the top panel of Figure 18.

[82] The DWR-ONRRP spread was constant (equal to 75 bps) throughtout the full sample (see Figure 16).

[83] The slope of the demand estimated for the subsample with IOR-ONRRP equal to 10 bps evaluated at the mean for the *subsample* (about $1,521.48 bn of total reserves) is −0.0186 bps.

"abundant" if reducing Q by \$1 bn increases the EFFR by no more than on hundredth of a basis point. Given this definition of "abundant", the demand estimated for the subsample with IOR-ONRRP equal to 10 bps implies $Q_1 =$ \$1,300 bn, while the demand estimated on the full sample implies $Q_1 =$ \$2,943 bn. Discrepancies this large make one wary of relying on these kinds of estimations to guide monetary policy operations.

A shortcoming of the atheoretical reduced-form approach to estimating a global aggregate demand curve for reserves is that the extrapolations the empirical model makes for ranges of Q for which there are not many observations (e.g., very low values of Q) can be very sensitive to our ability to identify the structural parameters that shift the aggregate demand being estimated. It seems sensible to try to control for these "policy regimes", and the sample split in the bottom-left panel of Figure 18 is an attempt to do so; but is this the right way to split the sample? Can variation in other policy or microstructure parameters shift or rotate the aggregate demand for reserves? To tackle these questions, we propose a quantitative theory-based approach.

The top-right panel of Figure 18 depicts several theoretical demands, $\iota^*_{\Upsilon_\omega} = \mathcal{D}(Q_{\Upsilon_\omega}; \Pi)$. The curve labeled "IOR-ONRRP = 10 bps" is the demand generated by the baseline calibration.[84] The curves labeled "IOR-ONRRP = 15 bps", "IOR-ONRRP = 20 bps", and "IOR-ONRRP = 25 bps" are the theoretical demands corresponding to $\iota_r - \iota_o = 0.0015/360$, $\iota_r - \iota_o = 0.0020/360$, and $\iota_r - \iota_o = 0.0025/360$, respectively, with all other parameters as in the baseline calibration. The top-right panel of Figure 18 also displays pairs of empirical observations of the quantity of active excess reserves, and the corresponding EFFR-IOR spread for every trading day in the sample period 2017/01/20–2019/09/13. As before, the sample is partitioned into four subsamples, each determined by the size of the IOR-ONRRP spread: 10 bps (2019/05/02–2019/09/13), 15 bps (2018/12/20–2019/05/01), 20 bps (2018/06/14–2018/12/19), or 25 bps (2017/01/20–2018/06/13). The bottom-right panel of Figure 18 displays the same data points as the top-right panel, along with the four fitted demand curves that result from estimating (11) on each subsample by nonlinear least-squares.

There are two main takeaways from the top-right and bottom-right panels of Figure 18. First, the theoretical demands fit the data reasonably well.[85] Second, *locally*, i.e., for range

[84] That is, with the parameter values reported in Table 1, and $\iota^*_{\Upsilon_\omega}$ and Q_{Υ_ω} computed with the interpolation procedure described in Section 3.6, for $\Upsilon_0 = 2017$ and $\Upsilon_1 = 2019$.

[85] The height and slope of the demand curve labeled "IOR-ONRRP = 10 bps" were calibrated to match the average EFFR-IOR spread and the local liquidity effect for the corresponding subsample, but the other subsamples were not targeted. The theoretical demand labeled "IOR-ONRRP = 25 bps" predicts an EFFR-IOR spread that is somewhat high, but it only takes a 2 bp reduction in the liquidity return parameter, ι_ℓ, to bring

of Q for which there is available data, the theoretical demand in the top-right panel and the reduced-form demands in the bottom-right panel fit about as well.[86] However, as is evident from the figure, their predictions for lower levels of Q are drastically different. To illustrate this point, focus on the subsample with IOR-ONRRP spread equal to 10 bps. The reduced-form model in the bottom-right panel estimates the steepest point on the corresponding demand at about \$934 bn (or about \$1,637 bn of total reserves), and predicts that reducing the supply of reserves below \$800 (or about \$1,255 bn of total reserves) would have essentially no effect on the equilibrium EFFR-IOR spread.[87] In contrast, our theory estimates the steepest point on the demand at about \$500 bn (or about \$400 bn of total reserves), and predicts that reductions in the supply of reserves start to cause significant increases in the EFFR-IOR spread for levels of reserves roughly below \$700 (below \$970 bn of total reserves). For the reduced-form approach, the extrapolation to levels of Q for which there are no empirical observations is essentially driven by the functional form assumed. On the other hand, the theoretical extrapolation is based on the explicit equilibrium borrowing-and-lending activity that underlie the equilibrium *aggregate demand relationship*, $\iota^* = \mathcal{D}(Q; \Pi)$, with the microstructure and policy parameters, Π, calibrated to match the key micro-level and market-level moments that describe the fed funds market (documented in Section 3).

7 Navigational instruments for central banks

In this section we propose two diagnostic tools, or "navigational instruments" to aid monetary policy operations: (i) the *Monetary Confidence Band* (MCB), and (ii) the theory-based cross-sectional distribution of banks' shadow cost of procuring funding in the fed funds market.

the theoretical EFFR-IOR spread in line with the data.

[86] In terms of local fit, the reduced-form specification is, as expected, no worse than the theory since it is more flexible. E.g., it allows us to choose *four* parameters, i.e., $(\underline{s}, \overline{s}, \xi, Q_0)$, to match the data corresponding to each subsample, while the theoretical demands corresponding to each subsample are generated by changing only one parameter, i.e., the policy spread $\iota_r - \iota_o$.

[87] The reduced-form demand curves estimated using *active excess reserves* (reported in the bottom-right panel of Figure 18) are essentially identical to the ones estimated using *total reserves* (reported in the bottom-left panel). For example, the steepest point on the reduced-form demand estimated on the subsample with IOR-ONRRP spread equal to 10 bps in the bottom-left panel is at \$1,637 bn of total reserves; the slope at that point is -0.0002 bps, which is the same slope that the reduced-form demand estimated using active excess reserves achieves at \$934 bn).

7.1 Confidence bands for monetary policy implementation

Monetary policy implementation changed drastically during the last decade as the supply of reserve balances increased to unprecedented levels, effectively turning the Fed's operating framework into a *floor system*. There are still unanswered operational questions about this system. The most elementary is: what is the smallest quantity of outstanding aggregate reserves needed to ensure that plausible market shocks do not cause significant deviations of the fed funds rate from its policy target? In this section we use the estimated quantitative theory to frame our answer in terms of a new policy-evaluation instrument: the *Monetary Confidence Band* (MCB).

Let $\iota = \mathcal{D}(Q)$ denote an aggregate-demand relationship between the equilibrium fed funds rate, ι, and the aggregate supply of reserves, Q. The mapping $\mathcal{D}(\cdot)$ could be obtained from the equilibrium of the theory, as in Section 6.1, or from another procedure (e.g., by estimating something like (11)). Let Z_p denote the p^{th} percentile of the empirical distribution of reserve-draining shocks estimated in Section 3.4. We define the "p% MCB" as a pair of functions, $(\underline{\iota}(Q), \bar{\iota}(Q))$ with $\underline{\iota}(Q) \equiv \mathcal{D}\left(Q + Z_{\frac{100+p}{2}}\right)$ and $\bar{\iota}(Q) \equiv \mathcal{D}\left(Q + Z_{\frac{100-p}{2}}\right)$. The idea is that the reserve-augmenting or reserve-draining shocks induce randonmness in the supply of reserves, which in turn induces randomness in the fed funds rate. For example, for a given beginning-of-day supply of reserves, Q, the equilibrium fed funds rate lies inside the 95% MCB, $(\mathcal{D}(Q + Z_{97.5}), \mathcal{D}(Q + Z_{2.5}))$, with 95% probability. Figure 19 presents several examples of MCBs where $\mathcal{D}(\cdot)$ is the aggregate demand for reserves derived from our theory.

The top-left panel in Figure 19 displays the 95% and 99% MCBs around the aggregate demand corresponding to the baseline calibration, which is labeled "Mean (volume weighted) rate". There are two ways to use the MCB. First, for a given beginning-of-day supply of reserves, we can use the MCB to estimate the probability that the fed funds rate will be in a certain range. For example, for a typical day in the sample period targeted by this baseline calibration, the beginning-of-day quantity of active excess reserves, Q, was about \$900 bn, and the IOR was 235 bps. Under these conditions, the MCB indicates that the Desk should be able to implement any target rate in the range IOR-IOR+25bps with certainty. Second, for any target range for the fed funds rate, the MCB yields the minimum quantity of reserves needed to meet the target with a desired degree of confidence. For example, if the Desk wanted the fed funds rate to be within the IOR-IOR+25bps range with 95% confidence, it would have to supply the market at least about \$670 bn in active excess beginning-of-day reserves. A 99% degree of confidence would instead require at least about \$850 bn.

The other panels in Figure 19 report the MCBs for calibrations that differ in one parameter from the baseline calibration. The top-right panel sets $\beta_F = 0$ (the baseline has $\beta_F = 0.03$). This could be interpreted as a day in which all banks of type F withdraw from the fed funds market. Under these conditions, the Desk would have to supply the market at least about \$700 bn beginning-of-day active excess reserves to keep the fed funds rate within the IOR-IOR+25bps with 95% confidence (about \$30 bn more than in the baseline). The bottom-left panel increases the IOR by 15 bps (from ONRRP + 10 bps in the baseline, to ONRRP + 25 bps). Notice that in this case the Desk would have to make beginning-of-day active excess reserves very scarce—less than \$500 bn—to ensure the FFR is higher than the IOR with 95% confidence. This is in contrast with the baseline calibration, which guarantees the FFR will be higher than the IOR with certainty for any level of reserves.

The bottom-right panel assumes $u_i(a) = \iota_d a \mathbb{I}_{\{a<0\}}$ for all i, with $\iota_d = \frac{0.2}{800}\iota_w$ (the baseline has $u_i(a) = 0$ for all $(a, i) \in \mathbb{R} \times \mathbb{N}$). The parameter ι_d captures the regulatory, reputational, or other costs associated with running an *intraday overdraft* (defined as a negative intraday excess reserve balance), which gained notoriety after the spikes in money-market rates of September 2019.[88] The main takeaway from this exercise is that even modest costs of not meeting the LCR and Regulation-D thresholds *on an intraday basis* can cause a significant upward shift in the demand for reserves. For example, from the bottom-right panel of Figure 19 we see that, with a level of beginning-of-day active excess reserves of about \$800 bn, the Desk cannot keep the FFR within the IOR-IOR+25bps range with 99% confidence (but the Desk can do so if $\iota_d = 0$, as in the baseline of the top-left panel).

7.2 Distribution of shadow price of reserves

When analyzing commercial banks' decisions to lend to households, corporations, or money-market participants, the fed funds rate is usually regarded as a measure of the (opportunity) cost of the loanable funds. The logic is that a bank that is long in reserves could lend in the fed funds market rather than to a client, and a bank that is short may borrow in the fed funds market and lend elsewhere. Thus, in a competitive marketstructure the (opportunity) cost of

[88] See, e.g., Copeland et al. (2021, Section 4). With no available evidence on the value of ι_d, for illustrative purposes, here we have chosen it so that a bank that incurs intraday overdraft for a whole trading day (composed of 800 model periods) suffers a per-dollar cost equal to 20% of the DWR. In Section 8 we use our theory augmented with $0 < \iota_d$ to rationalize the spikes in the EFFR of September 16 and 17, 2019. In Appendix D (Section D.3) we give a more detailed account of the events that took place during September 13–20, 2019 (reserve-draining shocks, associated rate spikes, and ensuing policy interventions).

funds for *all* banks is summarized by a single statistic—the equilibrium fed funds rate. But in an over-the-counter marketstructure where loans are negotiated bilaterally and sequentially over time (as in the actual fed funds market), each bank faces different borrowing and lending rates depending on their own and their counterparties' characteristics, such as their reserve balance at the time of the trade, degree of market power (e.g., θ_i), ability to find counterparties (e.g., β_i), and regulatory treatment (e.g., the administered rates they earn for holding reserves or pay for overdrafts).

In a dynamic OTC marketstructure like the fed funds market, each participating bank of type $i \in \mathbb{N}$ with reserve balance $a \in \mathbb{R}$ at time t has its own opportunity cost, or "shadow price" of reserves, which is summarized by $\frac{\partial V_t^i(a)}{\partial a}$. In this context, at any point in time the opportunity cost of loanable funds is characterized by a whole cross-bank distribution rather than by a single number, which may be more or less representative of the majority of banks. Below, we show that according to our baseline calibration, neither the EFFR nor the distribution of traded rates are representative of the distribution of shadow prices of reserves of the majority of the banks that participate in the fed funds market.

While outside the scope of our model, one could envision a more general model in which banks make lending decisions to outside clients at a first stage knowing they will later participate of a fed funds trading stage like the one we have modeled above.[89] In this setup, the relevant opportunity cost of loanable funds in the first stage for a bank of type i is given by $\mu_i(a) \equiv \frac{\partial V_0^i(a)}{\partial a} - 1$, where a is the bank's residual balance after having made loans to outside clients in the first stage. We summarize this heterogenity with a cumulative distribution function $\mathcal{M}_i(\iota) \equiv \int \mathbb{I}_{\{a:\mu_i(a)\leq\iota\}} dF_0^i(a)$, i.e., $\mathcal{M}_i(\iota)$ is the proportion of banks of type $i \in \mathbb{N}$ whose shadow price of reserves at the beginning of the fed funds market trading day is lower than $\iota \in \mathbb{R}$.

The top-left panel of Figure 20 shows

$$\mathcal{M}(\iota) \equiv \sum_{i\in\{F,M,S\}} \frac{n_i \mathcal{M}_i(\iota)}{\sum_{i\in\{F,M,S\}} n_i},$$

along with the cumulative distribution function of *all* bilateral loan rates negotiated throughout the day, denoted $\mathcal{H}$ (both calculated for the baseline calibration). Intuitively, $\mathcal{H}(\iota)$ is the

[89] This setup would be a natural way to incorporate a repo market into the theory, since the majority of repo transactions are executed early in the business day. Copeland et al. (2021), for example, report that a large fraction of interdealer repo trades are conducted between 7:00 am and 7:20 am, EST, and use this fact to argue that when intermediating the Treasury repo market, the marginal value to a dealer bank of holding balances at the Fed is sensitive to anticipated intraday payment stresses on these balances.

proportion of reserves traded at rates below ι. The dashed vertical line labeled "EFFR" denotes the volume-weighted average fed funds rate on *all* trades implied by the theory. The IOR and DWR are denoted by solid vertical lines. Notice that $\mathcal{H}$ is very concentrated around the EFFR (about 60% of the funds are traded at the EFFR), so although there is heterogeneity in negotiated loan rates, the EFFR is quite representative of the overall distribution of traded rates. On the other hand, neither the distribution of traded rates nor the EFFR are representative of the distribution of shadow prices of reserves across all banks, represented by $\mathcal{M}$. For example, 80% of banks have a shadow price of reserves higher than the EFFR, but only about 10% of reserves are traded at rates higher the EFFR. The reason is that banks of type S, which constitute more than 90% of the population of banks, account for a small share of trades, and are therefore underreprecented in the satistics computed on actual trades, such as the EFFR and the distribution $\mathcal{H}$.

The remaining three panels of Figure 20 display the beginning-of-day cumulative distribution function of shadow prices for banks of type i, denoted $\mathcal{M}_i$, and the cumulative distribution function of all loan rates paid or received by banks of type i, denoted $\mathcal{H}_i$. These panels show that the EFFR and the distribution of traded rates, $\mathcal{H}_i$, are fairly representative of the distribution of shadow prices of reserves across banks, $\mathcal{M}_i$, only for types $i \in \{F, M\}$, but not for type S. This means that for about 90% of banks that participate in the fed funds market, the EFFR does not adequately capture the shadow cost of procuring funding, and is therefore not the relevant cost of lending in the retail and corportate loan markets.

8 Tuesday, September 17, 2019

On Tuesday September 17, the EFFR printed at 230 bps, exceeding the upper limit of the FOMC's target range by 5 bps.[90] This event garnered the attention of market analysts and policymakers for two reasons. First, it was the first upward deviation from target in the 11 years since the FOMC began announcing a target range for the EFFR in December 2008. Second, it seemed inconsistent with the widespread view that the $1.3 tn of reserves in excess of Regulation D outstanding at the time ought to be "ample enough" to run a floor system in which the Federal Reserve can implement its EFFR target without having to micromanage the

[90]The 99^{th} percentile of the distribution of fed funds rates reached about 400 bps on September 17. Repo markets also experienced rate spikes, e.g., the secured overnight financing rate (SOFR) printed at 243 bps on Monday September 16 (13 bps higher than the previous business day), and exceeded 500 bps on September 17. See Afonso et al. (2020a) and Anbil et al. (2020) for detailed accounts of these money-market events.

supply of reserves.

To frame the discussion, consider the top panel of Figure 21, which displays a data scatterplot with the EFFR-IOR spread on the vertical axis (in percent per annum), and the quantity of reserves on the horizontal axis (in billions of dollars).[91] The data points labeled "IOR-ONRRP = 10 bps" are all trading days in the sample 2019/05/02–2019/09/13 (the period we used to estimate the liquidity effect in Section 3.5). The six darkest data points labeled "Sept 13-20 2019" are September 13, 16, 17, 18, 19, and 20. The dashed lines labeled "Target Upper Limit" and "Target Lower Limit" are the top and bottom of the fed funds target range minus the IOR for the period 2019/05/02–2019/09/18. On the scatterplot we have overlayed the MCB implied by the baseline calibration of the model (this is the same MCB displayed in the top-left panel of Figure 19, but this time with the EFFR-IOR spread on the vertical axis).

Friday, September 13 is the dark dot that sits on the demand for reserves generated by the theory—well within the EFFR target range. Monday, September 16 is the rightmost dark dot that sits on the upper limit of the target range for the EFFR-IOR spread, and September 17 is the uppermost dark dot, with an EFFR-IOR spread of 20 bps (5 bps higher than the spread between the upper limit of the EFFR target range and the IOR). Wednesday, September 18 is the leftmost dark dot that sits on the upper limit of the target range for the EFFR-IOR spread.[92] The most cited culprits for the rate spikes of September 16 and 17 are two anticipated reserve-draning shocks that reduced the supply of reserves by about $120 bn over two business days.[93] From the top panel of Figure 21, we see that the EFFR-IOR spreads for September 16–18 lie outside the 99% MCB. This means that (under our baseline calibration) our quantitative model cannot rationalize these observations as resulting from a "typical" daily reserve-draining shock—even if we define a "typical" shock as one with probability larger than 1%.

These events raised several questions: In a context with $1.3 tn of excess reserves in the banking system, how could an anticipated $120 bn reserve-draining shock cause such large spikes in money-market rates? Why didn't banks lend some of their excess reserves to exploit the high overnight rates? In response to these questions during an earnings call on October

[91] As in Figure 19, the primary horizontal axis represents *active excess reserves* (as defined in Section 6), and the secondary horizontal axis translates them into *total reserves* (as explained in Section D.2.5 of Appendix D).

[92] September 19 and 20 are the dark dots with an EFFR-IOR spread of 10 bps.

[93] The first was a quarterly corporate tax payment transferred from corporations' bank and money market mutual fund accounts to the Treasury's account. The second, a $54 bn settlement of Treasury debt paid by primary dealers into the Treasury's account on September 16. In Section D.3 we give a more detailed account of the reserve-draining shocks, associated rate spikes, and ensuing policy interventions that took place during September 13–20, 2019. Table 2 summarizes the main facts.

15, 2019, Jamie Dimon (Chairman and CEO of JPMorgan Chase) famously alluded to internal reserve management practices to ensure compliance with liquidity regulations:

> *As I said, we have $120 bn in our checking account at the Fed, and it goes down to $60 bn and then back to $120 bn during the average day. But we believe the requirement under CLAR (Comprehensive Liquidity Analysis and Review) and resolution and recovery is that we need enough in that account, so if there's extreme stress during the course of the day, it doesn't go below zero. If you go back to before the crisis, you'd go below zero all the time during the day. So the question is, how hard is that as a red line? Was the intent of regulators between CLAR and resolution to lock up that much of reserves in the account with Fed? And that'll be up to regulators to decide. But right now, we have to meet those rules and we don't want to violate anything we've told them we're going to do.*[94]

To explore this hypothesis, the middle and bottom panels of Figure 21 overlay, on the same data scatterplot of the top panel, the MCB implied by the baseline calibration of the model, but with $u_i(a) = \iota_d a \mathbb{I}_{\{a<0\}}$ for all i, with $\iota_d = \frac{x}{800}\iota_w$. The middle panel has $x = 0.1$, and the bottom panel, $x = 0.2$.[95] The parameter ι_d stands in for a bank's perceived penalty from going below Dimon's "red line" (e.g., associated to the possible loss of reputation with regulatory supervisors for failing to maintain prudent liquidity buffers, as suggested by Copeland et al. (2021)). The middle panel of Figure 21 then shows that a shadow cost of intraday overdraft equal to 10% of the DWR, e.g., caused by precautionary reserve internal management practices designed to ensure compliance with liquidity regulations, is enough for the model to rationalize the September 16-18 EFFR-IOR observations as resulting from "typical" daily reserve-draining shocks, in the sense of being within the 99% MCB. The bottom panel of Figure 21 shows that a shadow cost of intraday overdraft equal to 20% is enough to put these observations within the 95% MCB (marginally within, in the case of September 16).

[94] For a full transcript of the call, see: https://tinyurl.com/29scwszt. There is other evidence that the introduction of post-GFC liquidity regulations and associated supervisory programs have changed banks' liquidity risk management practices. Afonso et al. (2020a), for example, point to a recent survey conducted by the Federal Reserve in which the majority of bank respondents identified "meeting routine intraday payments flows and satisfying internal liquidity stress metrics as the main drivers of their demand for reserves". See, e.g., the August 2019 Senior Financial Officer Survey, https://www.federalreserve.gov/data/sfos/sfos.htm.

[95] The baseline calibration in the top panel corresponds to the special case with $x = 0$. The case with $x = 0.2$ was one of the counterfactual exercises considered in Section 7.1. Recall that $x = 0.1$, for example, implies a bank that incurs intraday overdraft for a whole trading day suffers a per-dollar cost equal to 10% of the DWR.

In another segment of the JPM earnings call of October 15, 2019, Dimon also alluded to the LCR requirement as a relevant determinant of banks' demand for reserves:

> *We have a checking account at the Fed with a certain amount of cash in it. That cash, we believe, is required under resolution and recovery and liquidity stress testing. And therefore, we could not redeploy it into repo market, which we would've been happy to do. [...] You're also going to hit a red line in LCR, like HQLA, which cannot be redeployed either.*

Throughout the paper we have defined a bank's "excess reserves" as its reserve balance net of the Regulation D reserve requirement, and net of the minimum quantity of reserves necessary to meet the LCR requirement given the bank's holdings of other qualifying HQLA.[96] In other words, our baseline calculation of excess reserves corresponds to a world in which banks have a preference for satisfying the LCR requirement with non-reserve assets to the extent possible.[97] In reality, however, there are anecdotal accounts that banks appear to have a preference for meeting LCR requirements with reserves rather than with other HQLA. By incorporating a very modest preference for complying with the LCR requirement with reserves, e.g., something that reduces our measure of aggregate beginning-of-day excess reserves by as little as $50 bn, the baseline calibration is able to rationalize the events of September 16–17, 2019 (in the sense that the compliance preference would shift all the dark dots in the top panel of Figure 21 leftwise by $50 bn, and into the 99% MCB).

9 Conclusion

In this paper we have taken several steps toward developing models of the fed funds market with explicit over-the-counter microstructures into useful tools to guide monetary policy operations. Our framework incorporates the main microstructure ingredients of the fed funds market, accounts for the most salient institutional features, and includes the collection of policy instruments and regulations that shape participants' demands for reserves. The model also incorporates the large degree of heterogeneity among participants across several dimensions,

[96] See footnote 31 and the more comprehensive discussion in Section B.2.1.

[97] We have adopted this identifying assumption for our baseline because we regard it the most conservative option in the sense that—even to this day—the common definition of "excess reserves" considers the pre-GFC Regulation D requirement but not the post-GFC LCR requirement.

such as: market power in bilateral loans, frequency and size distribution of payment shocks, and degree of centrality in market-making.

We documented a comprehensive set of novel marketwide and micro-level observations that describe the market dynamics, and showed that the quantitative model is flexible enough to match these observations. We then used the quantitative theory to deliver structural estimates of the aggregate demand for reserves, and developed two policy instruments to assess the cross-bank inequality in the shadow cost of procuring funding, and the central bank's ability to implement a given fed funds target.

While we think we have made significant progress, we also realize we have touched upon some questions and ideas that would be worth studying further in future work. First, we have allowed for heterogenity in contact rates across bank types to capture the core-periphery structure of the fed funds market, but we have treated these contact rates as parameters. While the exogeneity of contact rates may be a reasonable assumption during periods when regulation and the deeper marketstructure parameters are relatively constant, it is not difficult to imagine settings or questions where it would be desirable to endogenize search intensity, (e.g., perhaps along the lines of Farboodi et al. (2023)). A similar point can be made about the beginning-of-day distributions of reserves, which for many applications would be best derived from an explicit portfolio problem of banks that takes place *prior* to the fed-funds trading stage that we have focused on.

Finally, a monetary-policy operating framework consists of two parts: an *operating target* (e.g., the fed funds rate), and *policy instruments* (e.g., standing facilities, open-market operations). Monetary models in the macro tradition focus on the macroeconomic effects of choosing different values (or rules) for the operating target, and leave operational implementation considerations outside the scope of their analysis. Here we have instead focused on the operational side of the monetary policymaking process, and have left macro considerations outside the scope of our analysis. We think that exploring the macroeconomic implications of the microstructure of interbank lending and payments is a promising avenue of research (examples of work along these lines include Arce et al. (2020), Bianchi and Bigio (2022), De Fiore et al. (2018), Li and Li (2021), and Piazzesi and Schneider (2018)).

References

Åberg, P., M. Corsi, V. Grossmann-Wirth, T. Hudepohl, Y. Mudde, T. Rosolin, and F. Schobert (2021): "Demand for Central Bank Reserves and Monetary Policy Implementation Frameworks: The Case of the Eurosystem," *European Central Bank Occasional Paper*.

Afonso, G., R. Armenter, and B. Lester (2019): "A Model of the Federal Funds Market: Yesterday, Today, and Tomorrow," *Review of Economic Dynamics*, 33, 177–204.

Afonso, G., M. Cipriani, A. M. Copeland, A. Kovner, G. La Spada, and A. Martin (2020a): "The Market Events of Mid-September 2019," *FRB of New York Staff Report*.

Afonso, G., D. Giannone, G. La Spada, and J. C. Williams (2022): "Scarce, Abundant, or Ample? A Time-Varying Model of the Reserve Demand Curve," Working Paper 1019, Federal Reserve Bank of New York.

Afonso, G., K. Kim, A. Martin, E. Nosal, S. Potter, and S. Schulhofer-Wohl (2020b): "Monetary Policy Implementation with an Ample Supply of Reserves," *Finance and Economics Discussion Series 2020-020. Washington: Board of Governors of the Federal Reserve System*.

Afonso, G., A. Kovner, and A. Schoar (2011): "Stressed, Not Frozen: The Federal Funds Market in the Financial Crisis," *Journal of Finance*, 66, 1109–1139.

Afonso, G. and R. Lagos (2012): "An Empirical Study of Trade Dynamics in the Fed Funds Market," *FRB of New York Staff Report*.

——— (2015a): "The Over-the-Counter Theory of the Fed Funds Market: A Primer," *Journal of Money, Credit and Banking*, 47, 127–154.

——— (2015b): "Trade Dynamics in the Market for Federal Funds," *Econometrica*, 83, 263–313.

Alvarez, F. and D. Argente (2023): "Consumer Surplus of Alternative Payment Methods: Paying Uber with Cash," Manuscript, University of Chicago.

Anbil, S., A. G. Anderson, and Z. Senyuz (2020): "What Happened in Money Markets in September 2019?" *FEDS Notes*, 27.

Arce, O., G. Nuno, D. Thaler, and C. Thomas (2020): "A Large Central Bank Balance Sheet? Floor vs Corridor Systems in a New Keynesian Environment," *Journal of Monetary Economics*, 114, 350–367.

Armantier, O. and A. M. Copeland (2015): "Challenges in Identifying Interbank Loans," *Federal Reserve Bank of New York, Economic Policy Review*, 1–17.

Armantier, O., E. Ghysels, A. Sarkar, and J. Shrader (2015): "Discount Window Stigma During the 2007–2008 Financial Crisis," *Journal of Financial Economics*, 118, 317–335.

Armenter, R. and B. Lester (2017): "Excess Reserves and Monetary Policy Implementation," *Review of Economic Dynamics*, 23, 212–235.

Artuç, E. and S. Demiralp (2010): "Discount Window Borrowing After 2003: The Explicit Reduction in Implicit Costs," *Journal of Banking & Finance*, 34, 825–833.

Ashcraft, A. B. and D. Duffie (2007): "Systemic Illiquidity in the Federal Funds Market," *American Economic Review*, 97, 221–225.

Basel Committee on Banking Supervision (2010): *Basel III: International Framework for Liquidity Risk Measurement, Standards and Monitoring*, Bank for International Settlements.

Bech, M. L. and E. Atalay (2010): "The Topology of the Federal Funds Market," *Physica A: Statistical Mechanics and its Applications*, 389, 5223–5246.

Bech, M. L. and E. Klee (2011): "The Mechanics of a Graceful Exit: Interest on Reserves and Segmentation in the Federal Funds Market," *Journal of Monetary Economics*, 58, 415–431.

Beltran, D. O., V. Bolotnyy, and E. Klee (2021): "The Federal Funds Network and Monetary Policy Transmission: Evidence from the 2007–2009 Financial Crisis," *Journal of Monetary Economics*, 117, 187–202.

Bernanke, B. S. and D. Kohn (2016): "The Fed's Interest Payments to Banks," *Brookings Institution blog*.

Bianchi, J. and S. Bigio (2022): "Banks, Liquidity Management, and Monetary Policy," *Econometrica*, 90, 391–454.

Botev, Z. I., J. F. Grotowski, D. P. Kroese, et al. (2010): "Kernel Density Estimation via Diffusion," *The Annals of Statistics*, 38, 2916–2957.

Carpenter, S. and S. Demiralp (2006): "The Liquidity Effect in the Federal Funds Market: Evidence from Daily Open Market Operations," *Journal of Money, Credit and Banking*, 38, 901–920.

Chiu, J., J. Eisenschmidt, and C. Monnet (2020): "Relationships in the Interbank Market," *Review of Economic Dynamics*, 35, 170–191.

Copeland, A., D. Duffie, and Y. Yang (2021): "Reserves Were Not So Ample After All," Working Paper 29090, National Bureau of Economic Research.

De Fiore, F., M. Hoerova, and H. Uhlig (2018): "Money Markets, Collateral and Monetary Policy," Working Paper 25319, National Bureau of Economic Research.

Duffie, D., N. Gârleanu, and L. H. Pedersen (2005): "Over-the-Counter Markets," *Econometrica*, 73, 1815–1847.

Ennis, H. M. (2019): "Interventions in Markets with Adverse Selection: Implications for Discount Window Stigma," *Journal of Money, Credit and Banking*, 51, 1737–1764.

Ennis, H. M. and T. Keister (2008): "Understanding Monetary Policy Implementation," *FRB Richmond Economic Quarterly*, 94, 235–263.

Ennis, H. M. and J. A. Weinberg (2013): "Over-the-Counter Loans, Adverse Selection, and Stigma in the Interbank Market," *Review of Economic Dynamics*, 16, 601–616.

Farboodi, M., G. Jarosch, and R. Shimer (2023): "The Emergence of Market Structure," *Review of Economic Studies*, 90, 261–292.

Federal Reserve Board (2019a): *Reserve Maintenance Manual*, Board of Governors of the Federal Reserve System.

——— (2019b): "Statement Regarding Monetary Policy Implementation," *Press Release, Washington, DC, October 11, 11:00 a.m. EDT*.

——— (2019c): "Statement Regarding Monetary Policy Implementation and Balance Sheet Normalization," *Press Release, Washington, DC, January 30, 2:00 p.m. EST*.

Feinman, J. N. (1993): "Reserve Requirements: History, Current Practice, and Potential Reform," *Fed. Res. Bull.*, 79, 569.

Furfine, C. H. (1999): "The Microstructure of the Federal Funds Market," *Financial Markets, Institutions & Instruments*, 8, 24–44.

Hamilton, J. D. (1996): "The Daily Market for Federal Funds," *Journal of Political Economy*, 104, 26–56.

——— (1997): "Measuring the Liquidity Effect," *American Economic Review*, 87, 80–97.

Hugonnier, J., B. Lester, and P.-O. Weill (2020): "Frictional Intermediation in Over-the-Counter Markets," *Review of Economic Studies*, 87, 1432–1469.

Ireland, P. (2018): "Fed Should Stop Paying Interest on Reserves," *Economics 21 blog*.

Keister, T. (2012): "Corridors and Floors in Monetary Policy," *Liberty Street Economics*.

Keister, T., A. Martin, and J. McAndrews (2008): "Divorcing Money from Monetary Policy," *Economic Policy Review*, 14.

Klee, E. et al. (2021): "The First Line of Defense: The Discount Window during the Early Stages of the Financial Crisis," *International Journal of Central Banking*, 17, 143–190.

Lagos, R. and G. Rocheteau (2007): "Search in Asset Markets: Market Structure, Liquidity, and Welfare," *American Economic Review*, 97, 198–202.

——— (2009): "Liquidity in Asset Markets With Search Frictions," *Econometrica*, 77, 403–426.

Lagos, R., G. Rocheteau, and P.-O. Weill (2011): "Crises and Liquidity in Over-the-Counter Markets," *Journal of Economic Theory*, 146, 2169–2205.

Li, Y. and Y. Li (2021): "Payment Risk and Bank Lending," Working Paper 2021-03, Ohio State University Fisher College of Business.

Piazzesi, M. and M. Schneider (2018): "Payments, Credit and Asset Prices," Manuscript, Stanford University.

Poole, W. (1968): "Commercial Bank Reserve Management in a Stochastic Model: Implications for Monetary Policy," *The Journal of Finance*, 23, 769–791.

Üslü, S. (2019): "Pricing and Liquidity in Decentralized Asset Markets," *Econometrica*, 87, 2079–2140.

Weill, P.-O. (2007): "Leaning Against the Wind," *Review of Economic Studies*, 74, 1329–1354.

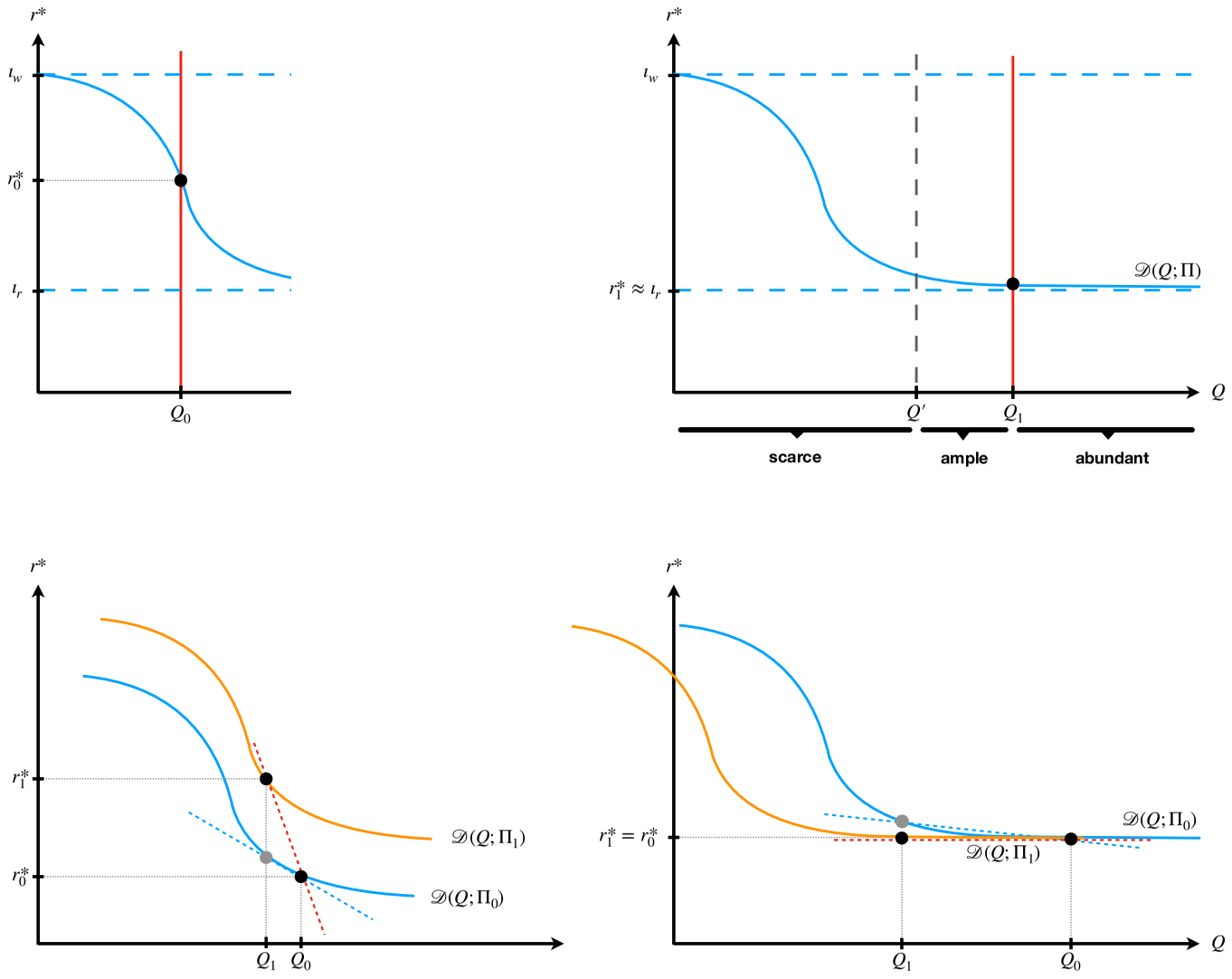

Figure 1: Stylized model of the determination of the fed funds rate.

Notes: In this figure, Q denotes the aggregate quantity of reserves, r^* the fed fund rate, and Π a set of parameters that determine the position of the aggregate demand for reserves, $\mathcal{D}$. The administered rates in the lending and deposit facility are denoted ι_w and ι_r, respectively.

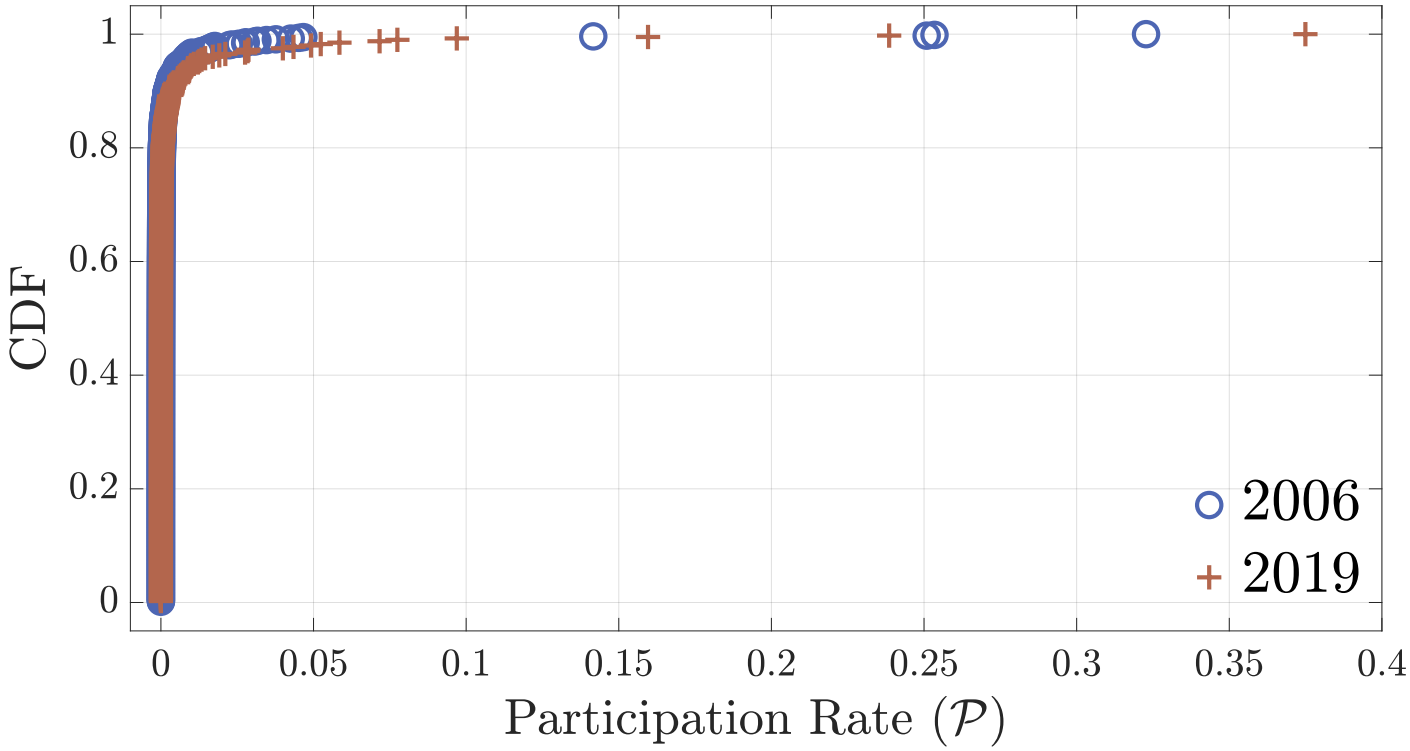

Figure 2: Empirical cumulative distribution function of bank-level participation rates for 2006 and 2019.

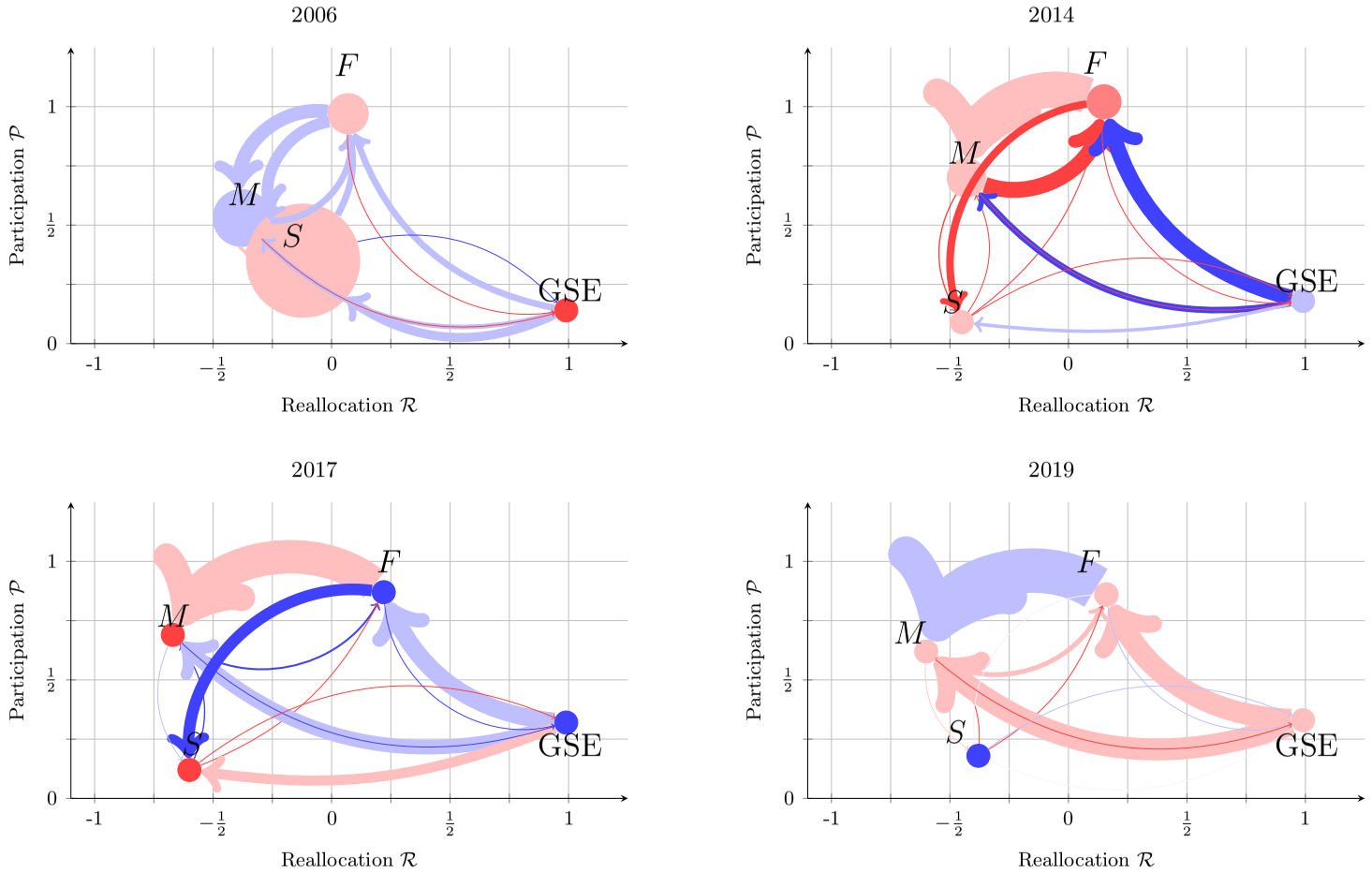

Figure 3: Fed funds trading networks.

Notes: Each node corresponds to one of the four bank types, and labeled accordingly as F, M, S, or GSE. An arrow from a type to another represents loans extended from banks in the former to banks in the latter, with the size of the arrow proportional to the volume of loans. The size of each node is proportional to the volume of loans extended by banks of that type to other banks of that same type. The colors of the arrows and nodes are: light blue, dark blue, light red, or dark red, if the volume-weighted average interest rate on the loans between the two types of banks, expressed as a spread over the *effective fed funds rate* (EFFR), falls in the first, second, third, or fourth quartile, respectively.

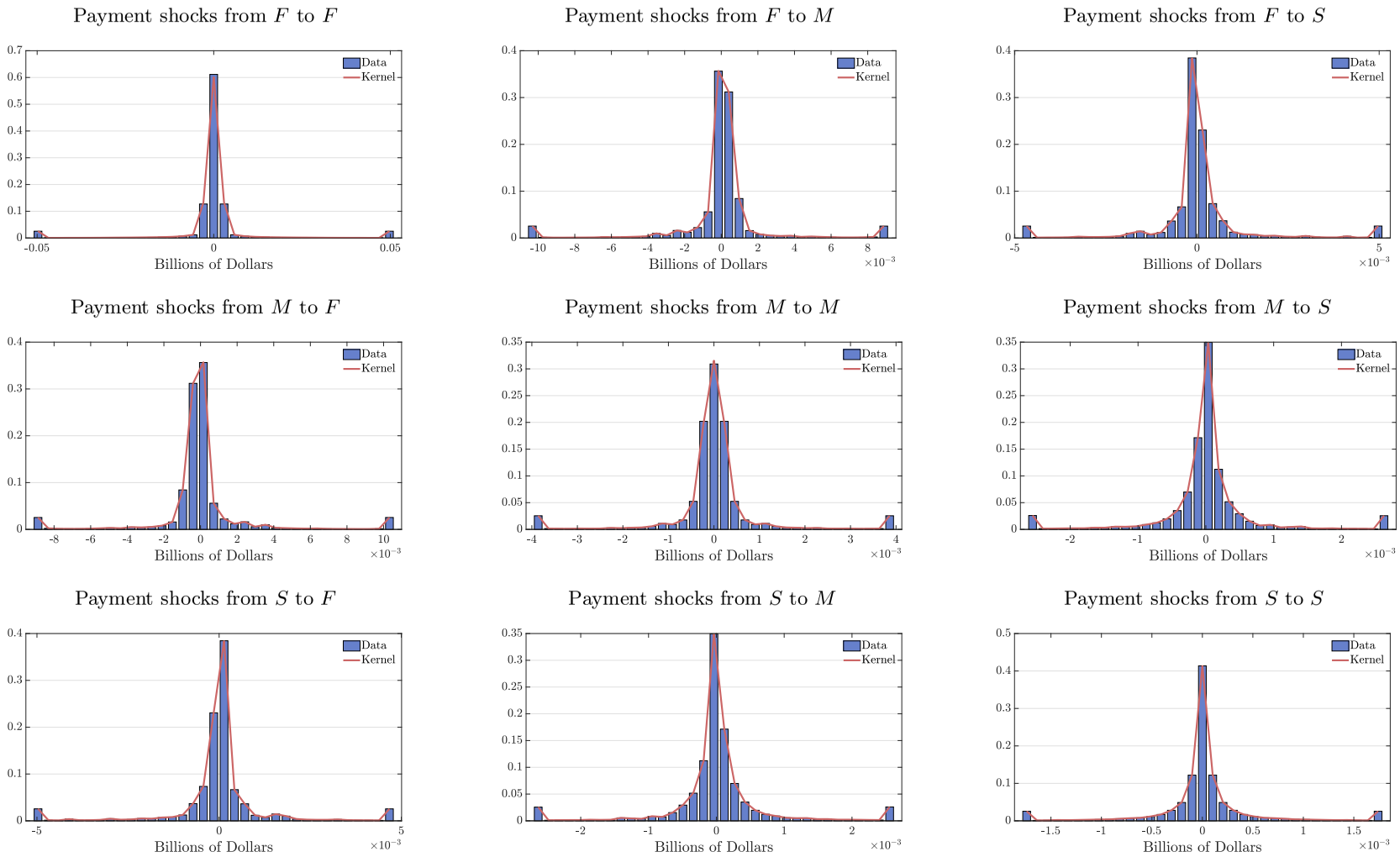

Figure 4: High-frequency payment shocks between pairs of bank types in 2006.

Notes: Empirical size distributions of high-frequency payment shocks between banks of each type (blue histograms) and their corresponding Gaussian kernel density estimates (red curves).

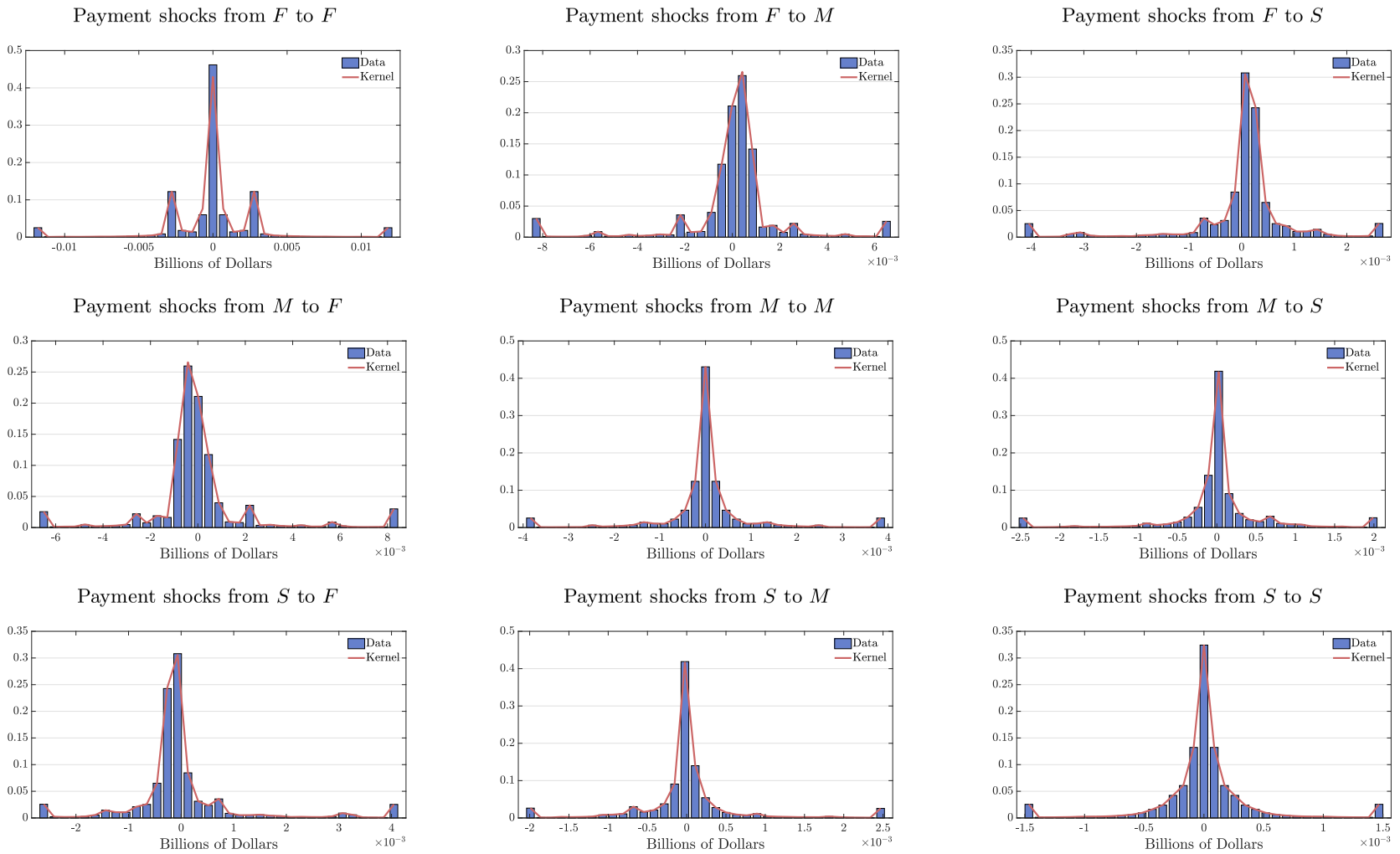

Figure 5: High-frequency payment shocks between pairs of bank types in 2019.

Notes: Empirical size distributions of high-frequency payment shocks between banks of each type (blue histograms) and their corresponding Gaussian kernel density estimates (red curves).

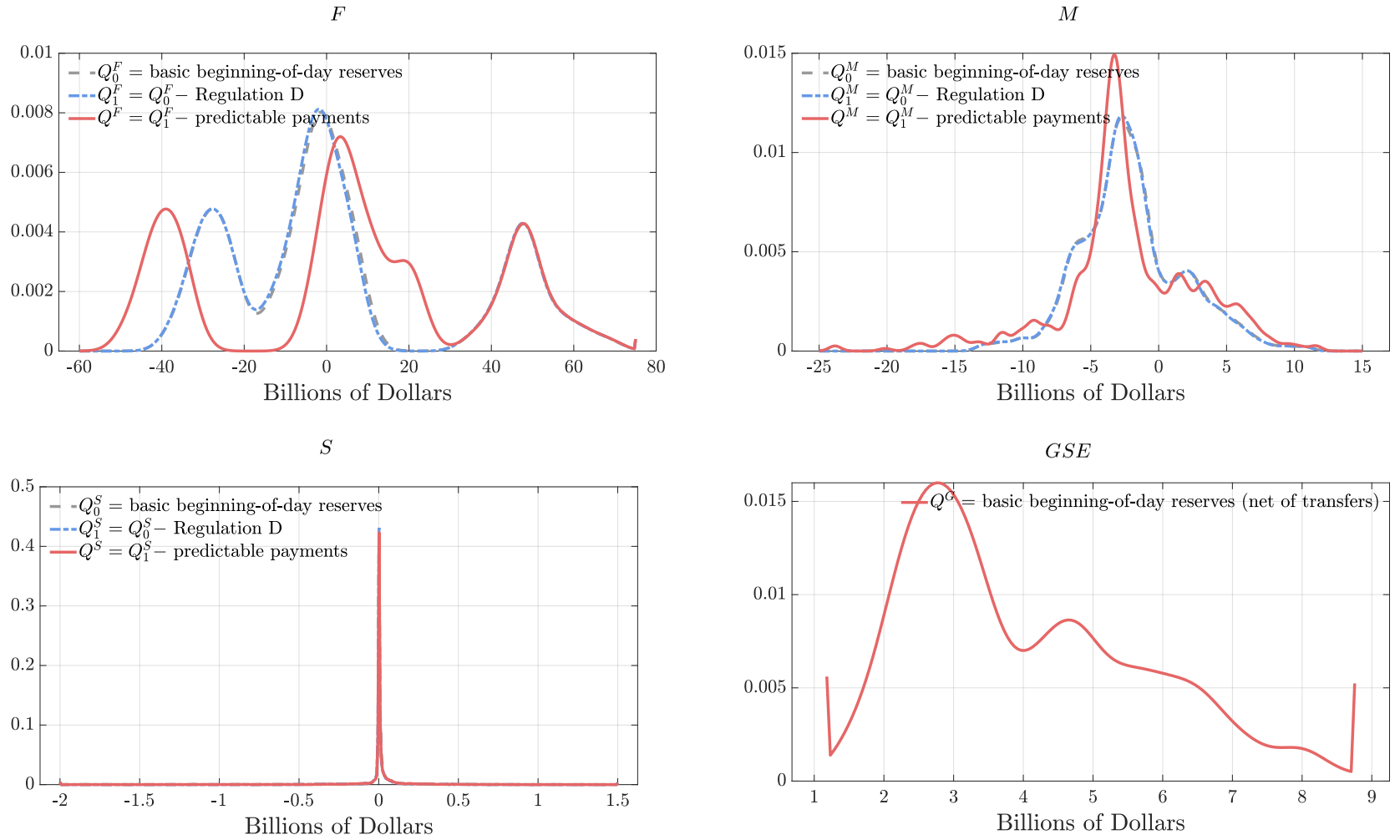

Figure 6: Estimated beginning-of-day distributions of reserves by bank type for the year 2006.

Notes: For every bank type $i \in \mathbb{N}$, the curve labeled Q^i is the (Gaussian kernel density) estimate of the empirical distribution of beginning-of-day excess reserves net of predictable transfers. For bank type $i \in \{F, M, S\}$, the curve labeled Q_0^i is the estimate of the distribution of "basic" beginning-of-day reserves (net of fed funds repayments), and the curve labeled Q_1^i is the estimate of the distribution of "basic" beginning-of-day reserves net of the Regulation D reserve requirement.

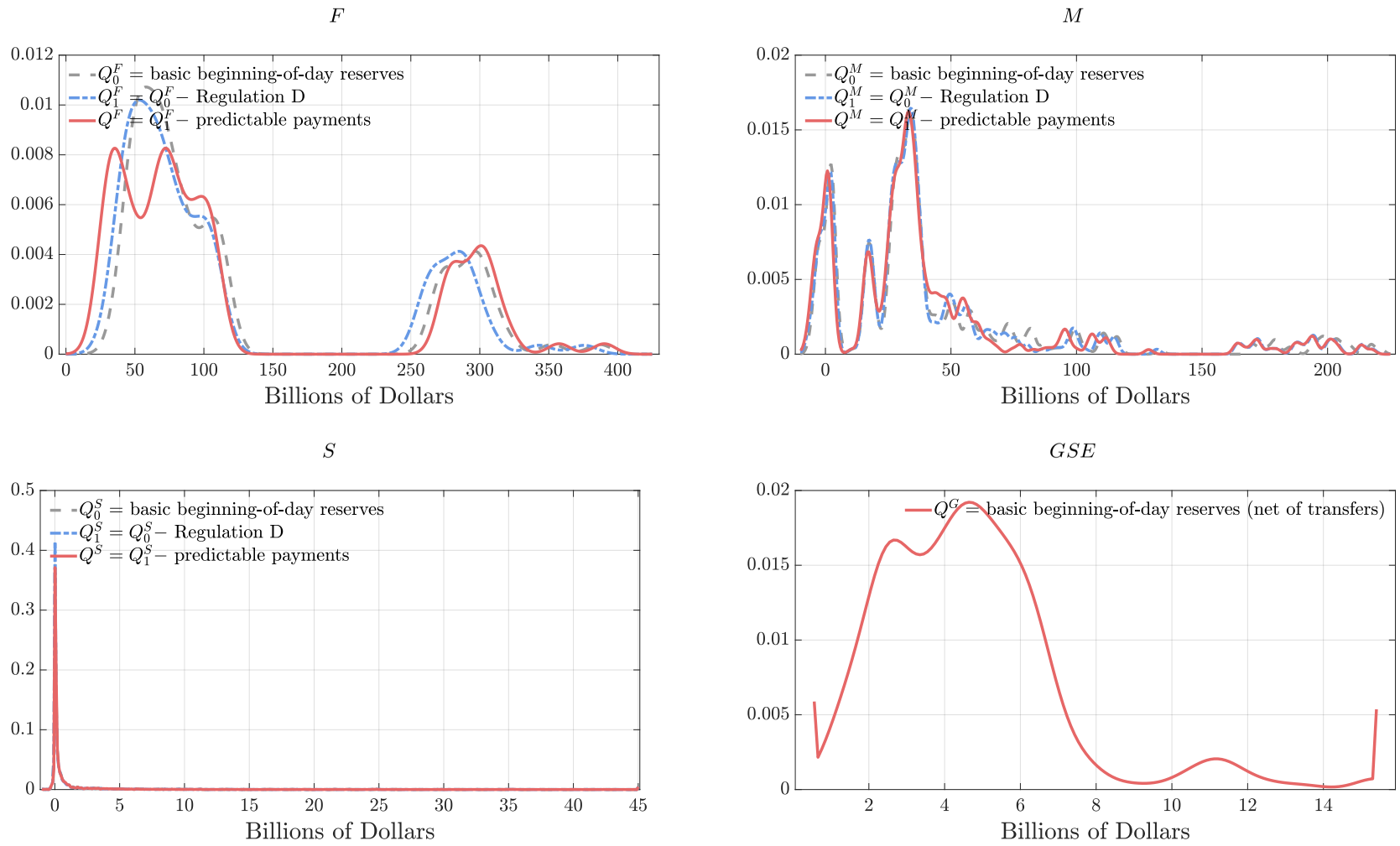

Figure 7: Estimated beginning-of-day distributions of reserves by bank type for the year 2014.

Notes: For every bank type $i \in \mathbb{N}$, the curve labeled Q^i is the (Gaussian kernel density) estimate of the empirical distribution of beginning-of-day excess reserves net of predictable transfers. For bank type $i \in \{F, M, S\}$, the curve labeled Q_0^i is the estimate of the distribution of "basic" beginning-of-day reserves (net of fed funds repayments), and the curve labeled Q_1^i is the estimate of the distribution of "basic" beginning-of-day reserves net of the Regulation D reserve requirement.

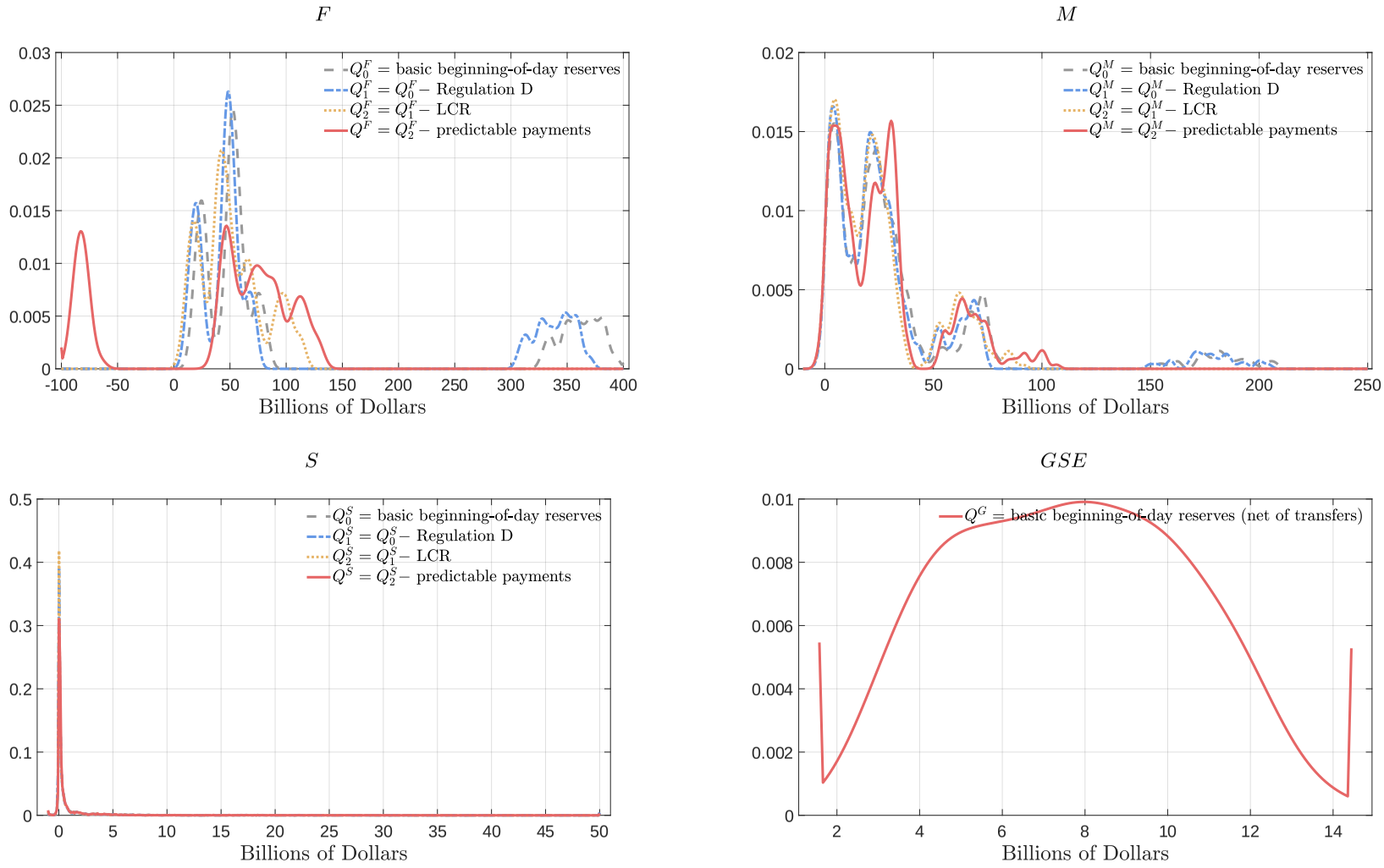

Figure 8: Estimated beginning-of-day distributions of reserves by bank type for the year 2017.

Notes: For every bank type $i \in \mathbb{N}$, the curve labeled Q^i is the (Gaussian kernel density) estimate of the empirical distribution of beginning-of-day excess reserves net of predictable transfers. For bank type $i \in \{F, M, S\}$, the curve labeled Q_0^i is the estimate of the distribution of "basic" beginning-of-day reserves (i.e., net of fed funds repayments), the curve labeled Q_1^i is the estimate of the distribution of "basic" beginning-of-day reserves net of the Regulation D reserve requirement, and the curve labeled Q_2^i is the estimate of the distribution of "basic" beginning-of-day reserves net of the Regulation D and LCR requirements.

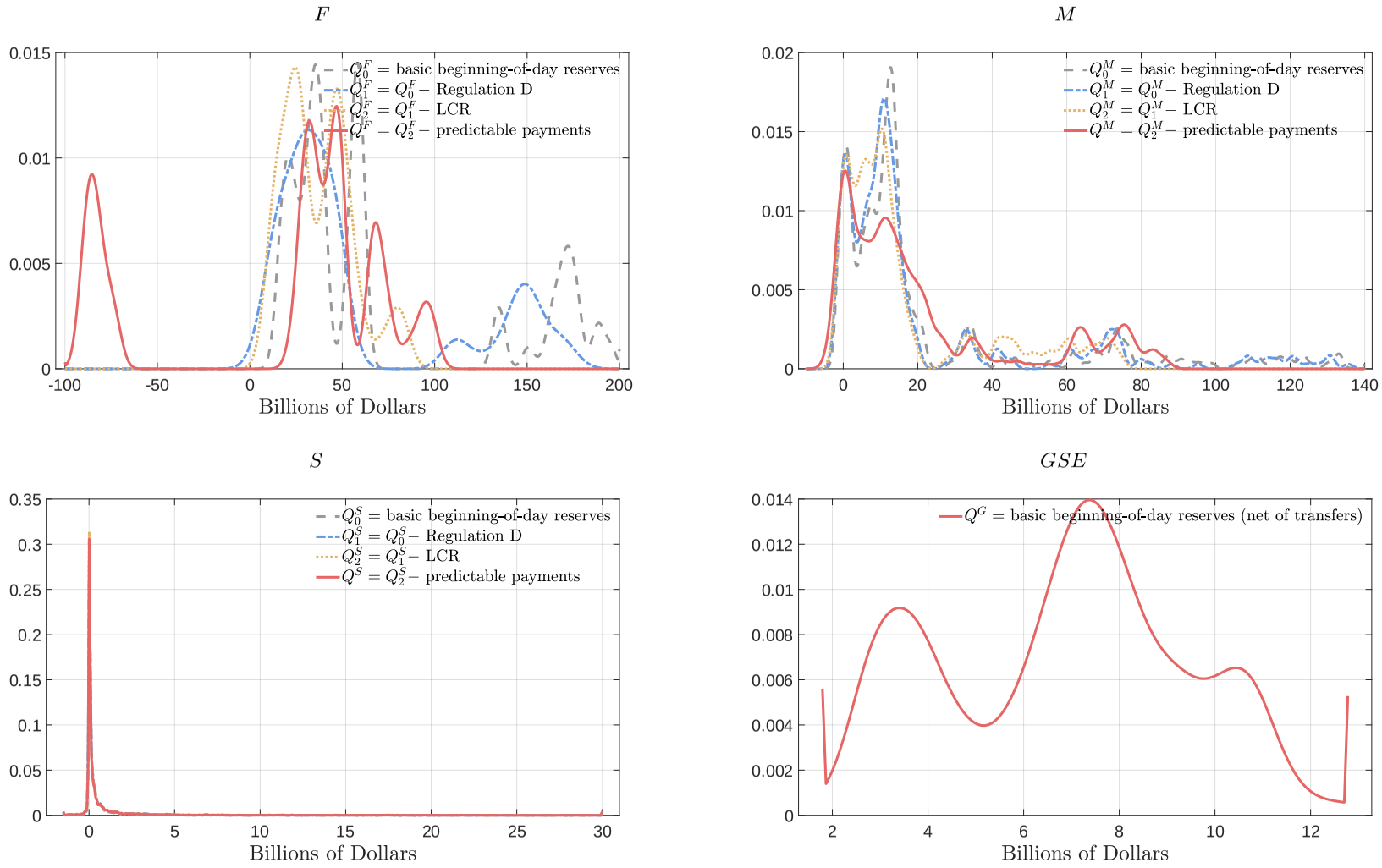

Figure 9: Estimated beginning-of-day distributions of reserves by bank type for the year 2019.

Notes: For every bank type $i \in \mathbb{N}$, the curve labeled Q^i is the (Gaussian kernel density) estimate of the empirical distribution of beginning-of-day excess reserves net of predictable transfers. For bank type $i \in \{F, M, S\}$, the curve labeled Q_0^i is the estimate of the distribution of "basic" beginning-of-day reserves (i.e., net of fed funds repayments), the curve labeled Q_1^i is the estimate of the distribution of "basic" beginning-of-day reserves net of the Regulation D reserve requirement, and the curve labeled Q_2^i is the estimate of the distribution of "basic" beginning-of-day reserves net of the Regulation D and LCR requirements.

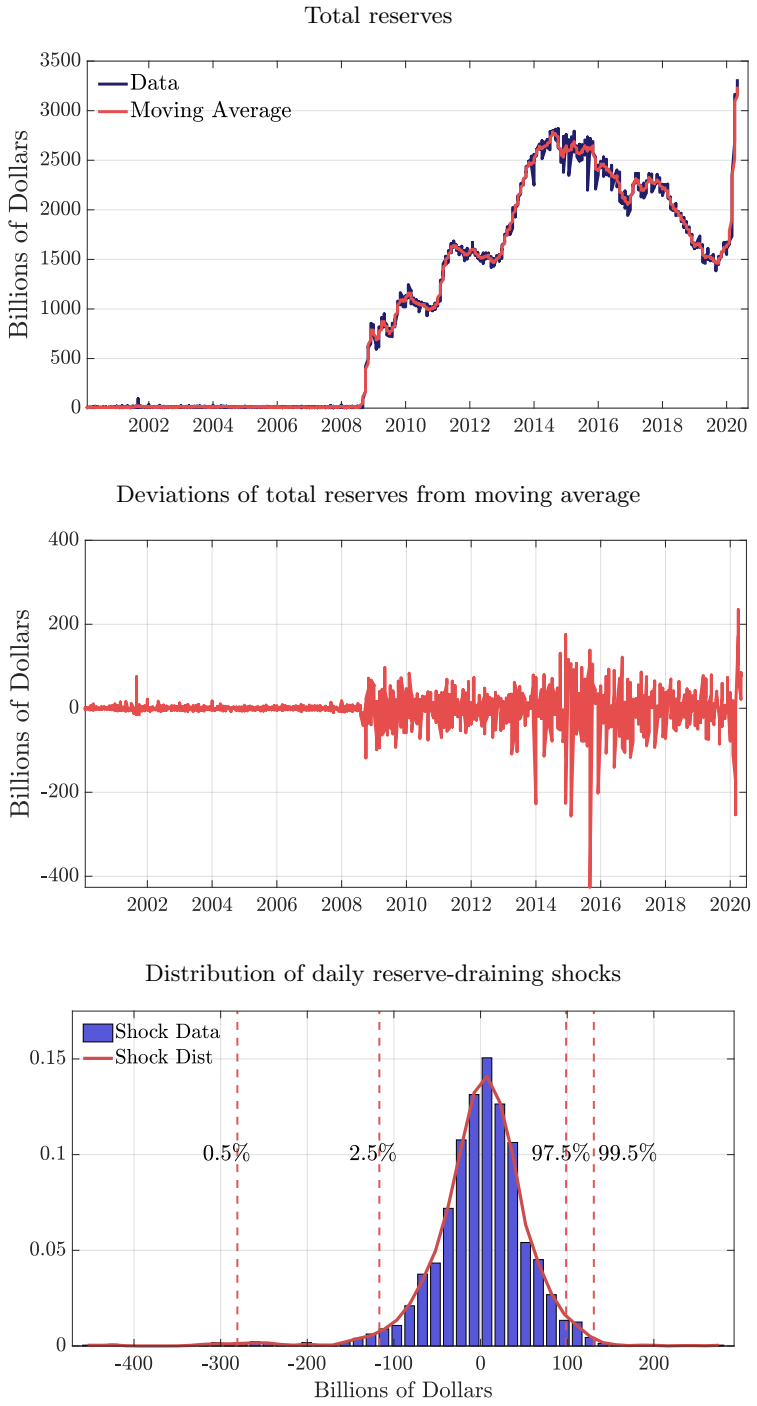

Figure 10: Aggregate supply of reserves and reserve-draining shocks.

Notes: Top panel: weekly time series of aggregate quantity of reserves and corresponding 40-day two-sided moving average. Middle panel: difference between the two time series in the top panel. Bottom panel: empirical histogram of daily deviations of the aggregate quantity of reserves from its 40-day two-sided moving average (January 2011-July 2019), and the corresponding Gaussian kernel estimate.

Parameter	Target	Moment	
		Data	Model
$n_F = 0.010$	proportion of financial institutions of type F	4/412	0.010
$n_M = 0.044$	proportion of financial institutions of type M	18/412	0.044
$n_S = 0.920$	proportion of financial institutions of type S	379/412	0.920
$n_{GSE} = 0.026$	proportion of financial institutions of type GSE	11/412	0.026
$\lambda_F = 0.951$	bank-level share of unexpected payments per second for type F	0.951	0.951
$\lambda_M = 0.257$	bank-level share of unexpected payments per second for type M	0.257	0.257
$\lambda_S = 0.011$	bank-level share of unexpected payments per second for type S	0.011	0.011
$\lambda_{GSE} = 0$	bank-level share of unexpected payments per second for type GSE	0	0
$\iota_w = 0.0300/360$	DWR (3.00% per annum, primary credit)	0.0300/360	0.0300/360
$\iota_r = 0.0235/360$	IOR (2.35% per annum)	0.0235/360	0.0235/360
$\iota_o = 0.0225/360$	ONRRP (2.25% per annum)	0.0225/360	0.0225/360
$\iota_\ell = 0.00049/360$	average value-weighted fed funds rate	0.0239/360	0.0239/360
$\iota_s = 0.00758/360$	estimated liquidity effect for 2019 (bps per 1 $bn decrease in reserves)	$\in [-0.019, -0.006]$	-0.0073
$\underline{\theta} = 1/20$	conditional (below the IOR) average value-weighted fed funds rate	0.0229/360	0.0231/360
$\beta_F = 0.0300$	number of loans of financial institutions of type F relative to average	24	25
$\beta_M = 0.0024$	participation rate of financial institutions of type M (i.e., $\mathcal{P}_M$)	0.62	0.54
$\beta_S = 0.0007$	participation rate of financial institutions of type S (i.e., $\mathcal{P}_S$)	0.18	0.15
$\beta_{GSE} = 0.0036$	participation rate of financial institutions of type GSE (i.e., $\mathcal{P}_{GSE}$)	0.33	0.27
$\kappa_F = 0.039e\text{-}3$	reallocation index of financial institutions of type F (i.e., $\mathcal{R}_F$)	0.16	0.13
$\kappa_M = 0$	reallocation index of financial institutions of type M (i.e., $\mathcal{R}_M$)	-0.61	-0.64
$\kappa_S = 0.003e\text{-}3$	reallocation index of financial institutions of type S (i.e., $\mathcal{R}_S$)	-0.38	-0.37
$\kappa_{GSE} = 1.25e\text{-}3$	reallocation index of financial institutions of type GSE (i.e., $\mathcal{P}_{GSE}$)	1	1

Table 1: Calibration for the year 2019.

Notes: Each non-shaded parameter is calibrated externally (i.e., to match a corresponding target moment, independently of the model and other parameters). Shaded parameters are calibrated internally (i.e., jointly, to match the set of shaded target moments, using the equilibrium conditions of the model, and given the values of the parameters calibrated externally). The calibration assumes a model period corresponding to approximately to 42 seconds in a trading day, $r = 0$, $\mathbb{N} = \{F, M, S, GSE\}$ (as discussed in Section 3.1), $\theta_{i,j} = 1/2$ for all $i, j \in \mathbb{N} \setminus \{GSE\}$, $\theta_{ij} = \underline{\theta}$ if $i \in \{GSE\}$ and $j \in \mathbb{N} \setminus \{GSE\}$, $\{G_{ij}\}_{i,j \in \mathbb{N}}$ are estimated as described in Section 3.2, $\{F_0^i\}_{i \in \mathbb{N}}$ are estimated as described in Section 3.3, $u_i = 0$ for all $i \in \mathbb{N}$, and $\{U_i\}_{i \in \mathbb{N}}$ are as in Section 4. The liquidity effect in the model is computed by extracting \$100 bn reserves using the procedure described in Section 3.6.

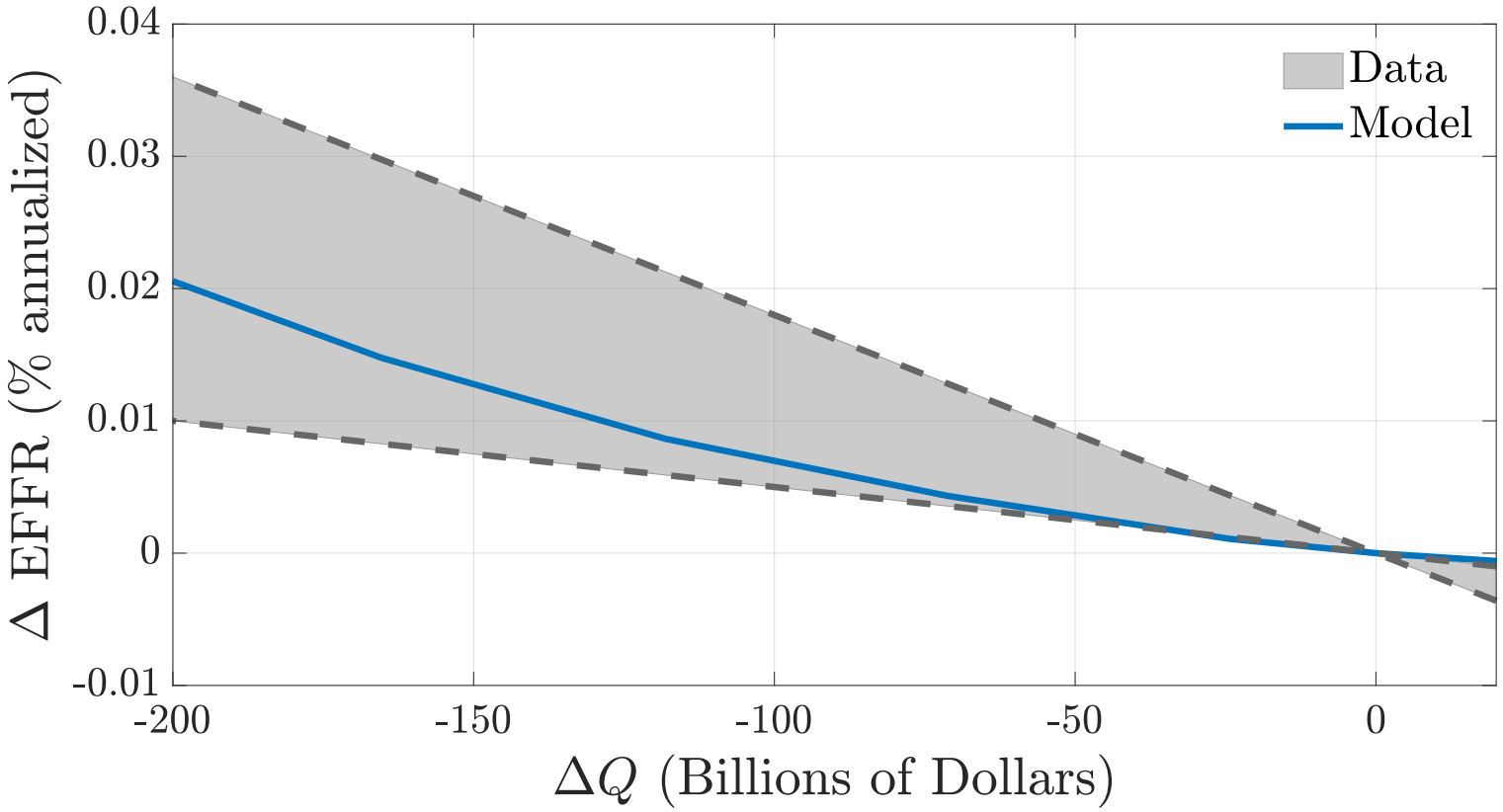

Figure 11: Liquidity effect: model and empirical estimates for the year 2019.

Notes: Rates in the vertical axis are in percent per annum. The shaded area represents the 95% confidence interval for the point estimates of the liquidity effect from specification (5). The solid line is the change in the equilibrium fed funds rate implied by the theory in response to changes in the total quantity of reserves (starting from the quantity of reserves corresponding to the 2019 calibration, and extracting reserves using the procedure described in Section 3.6.)

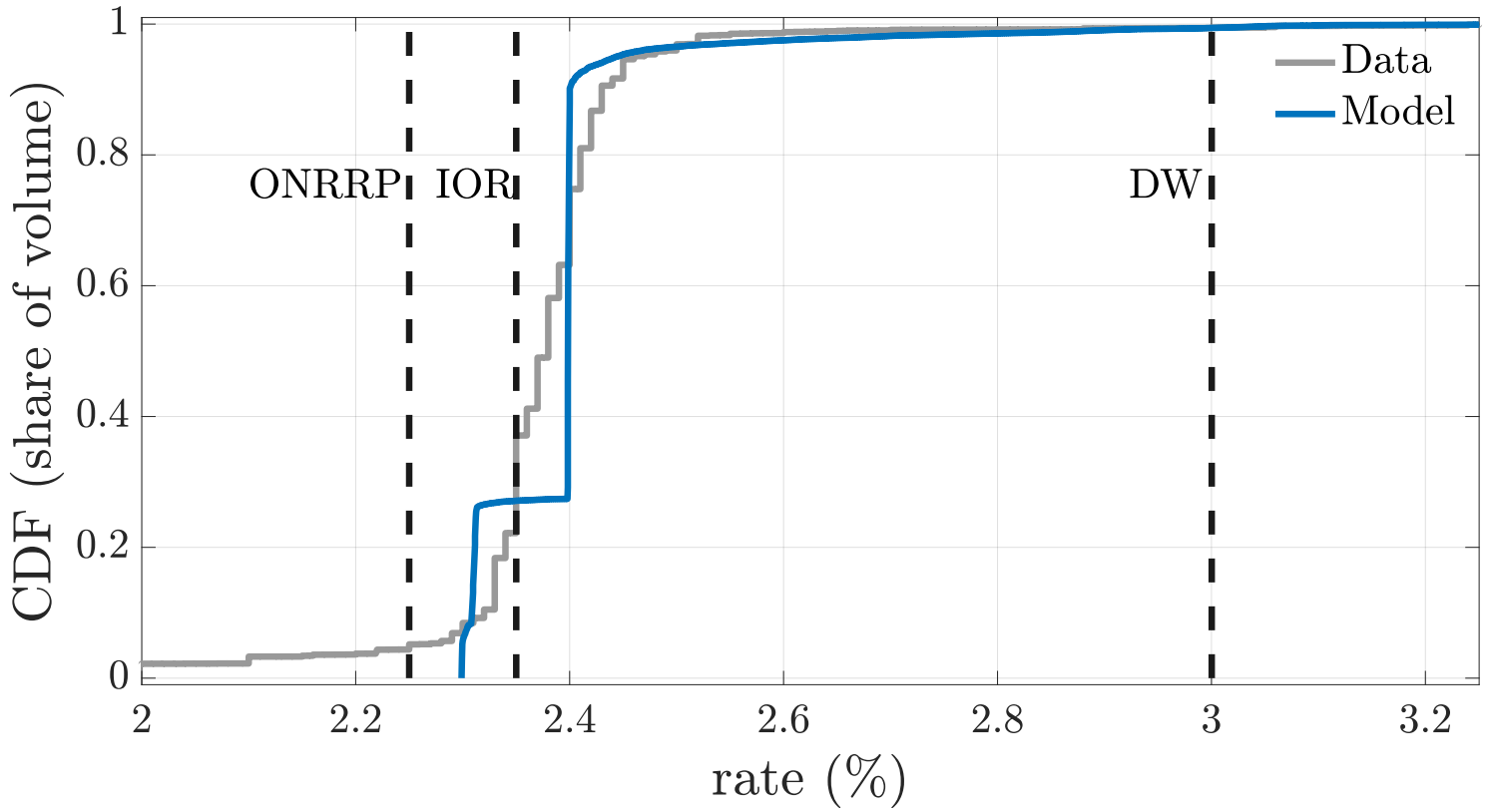

Figure 12: Empirical and theoretical cumulative distribution functions of bilateral fed funds rates for the year 2019.

Notes: For each loan rate ι, the curve labeled "Data" ("Model") gives the fraction of total loan volume traded at rates lower than ι in the data (model). Data are for every trading day in the period 2019/06/06–2019/07/31. The model calibrated as in Table 1. Rates are in percent per annum.

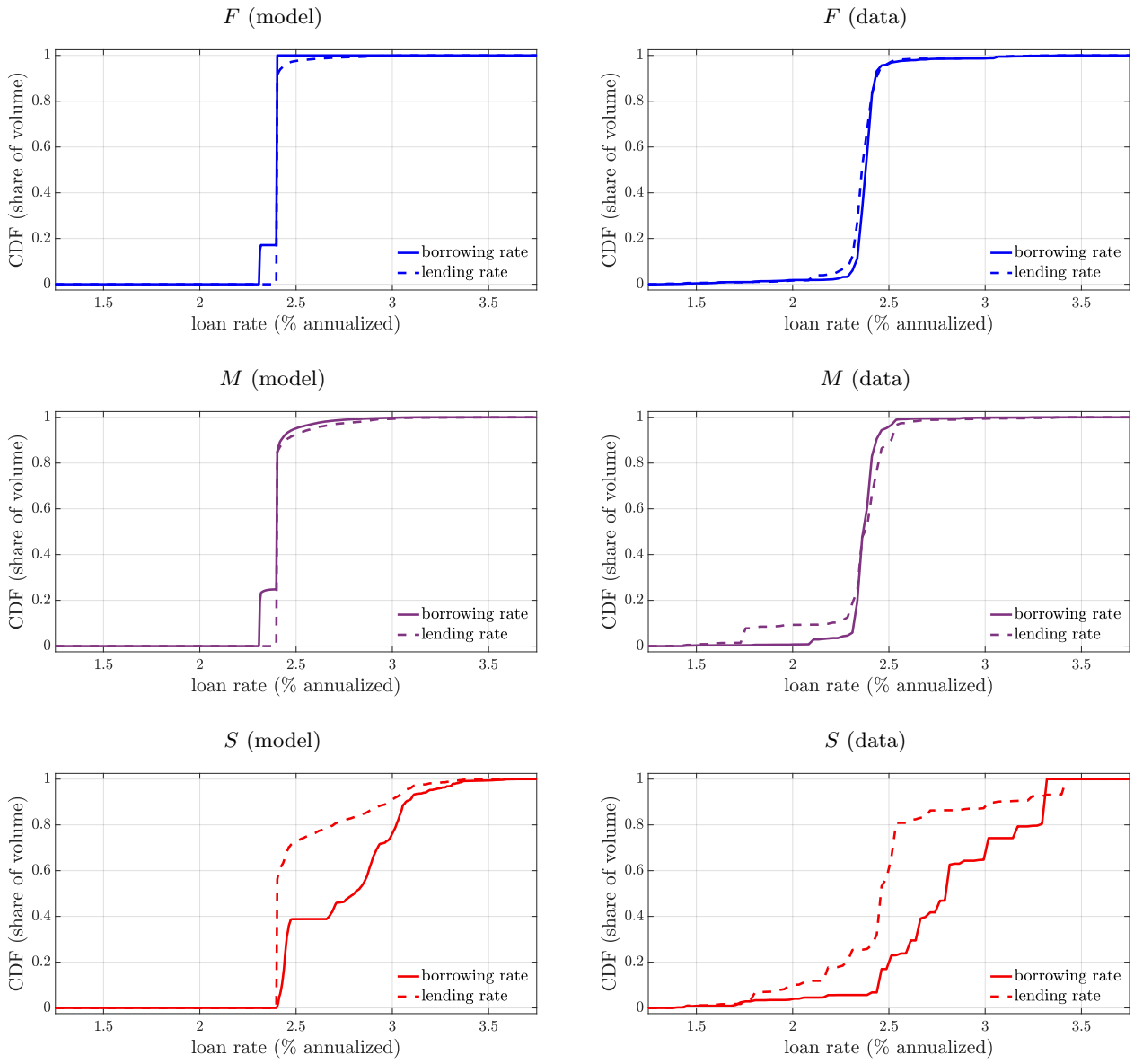

Figure 13: Cumulative distributions of borrowing and lending rates by bank type.

Notes: For each loan rate, the curve labeled "borrowing rate" ("lending rate") gives the fraction of total reserves borrowed (lent) by banks of the type indicated in the panel heading, at rates lower than that rate. The panels on the left are for the model calibrated as in Table 1. The panels on the right are from data, for every trading day in the period 2019/06/06–2019/07/31. Rates are in percent per annum.

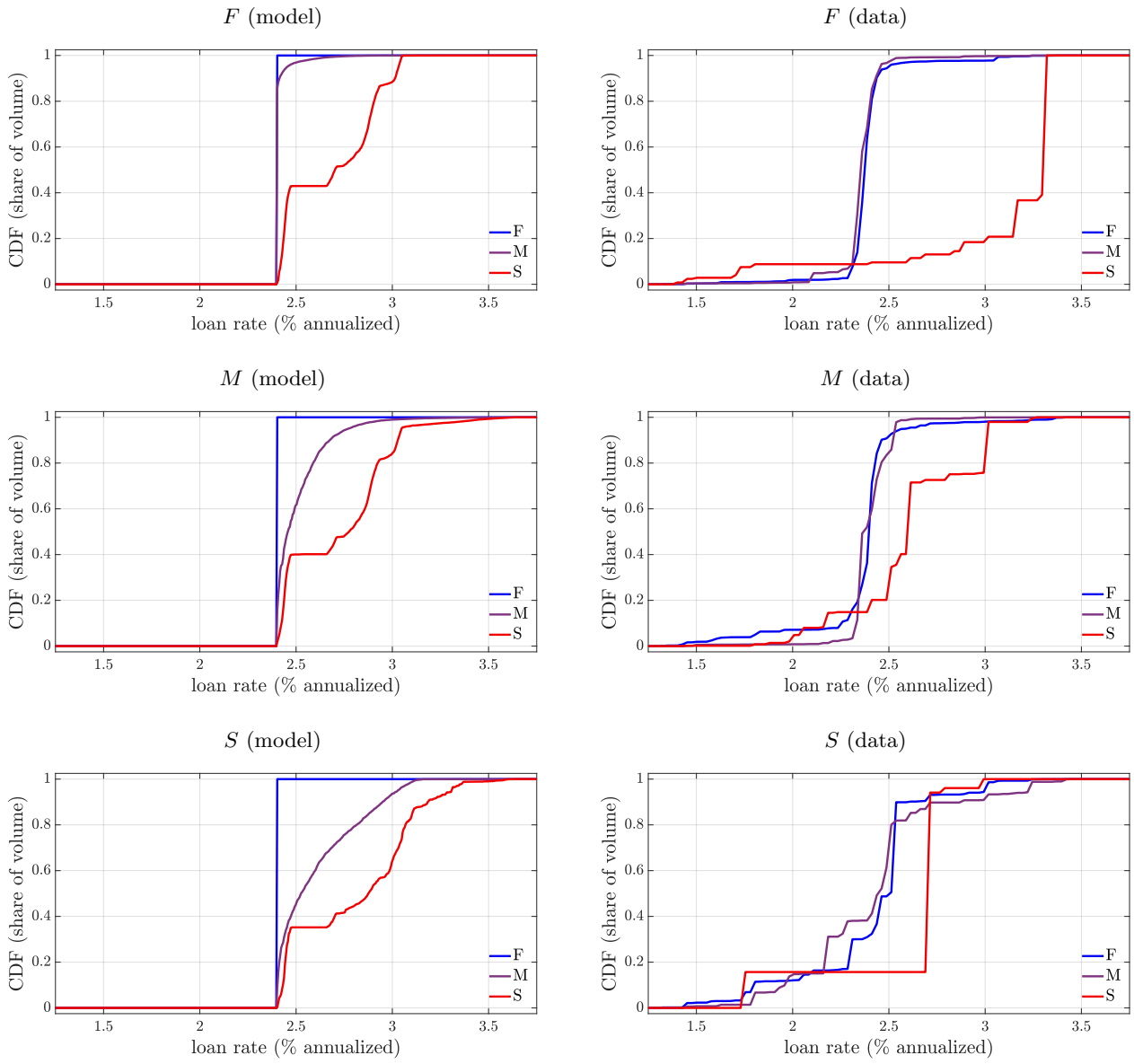

Figure 14: Cumulative distributions of loan rates between pairs of bank types.

Notes: For each loan rate, the curve labeled "i" (for $i \in \{F, M, S\}$) gives the fraction of total reserves borrowed by banks of type i from the bank types indicated in the panel heading, at rates lower than that rate. The panels on the left are for the model calibrated as in Table 1. The panels on the right are from data, for every trading day in the period 2019/06/06–2019/07/31. Rates are in percent per annum.

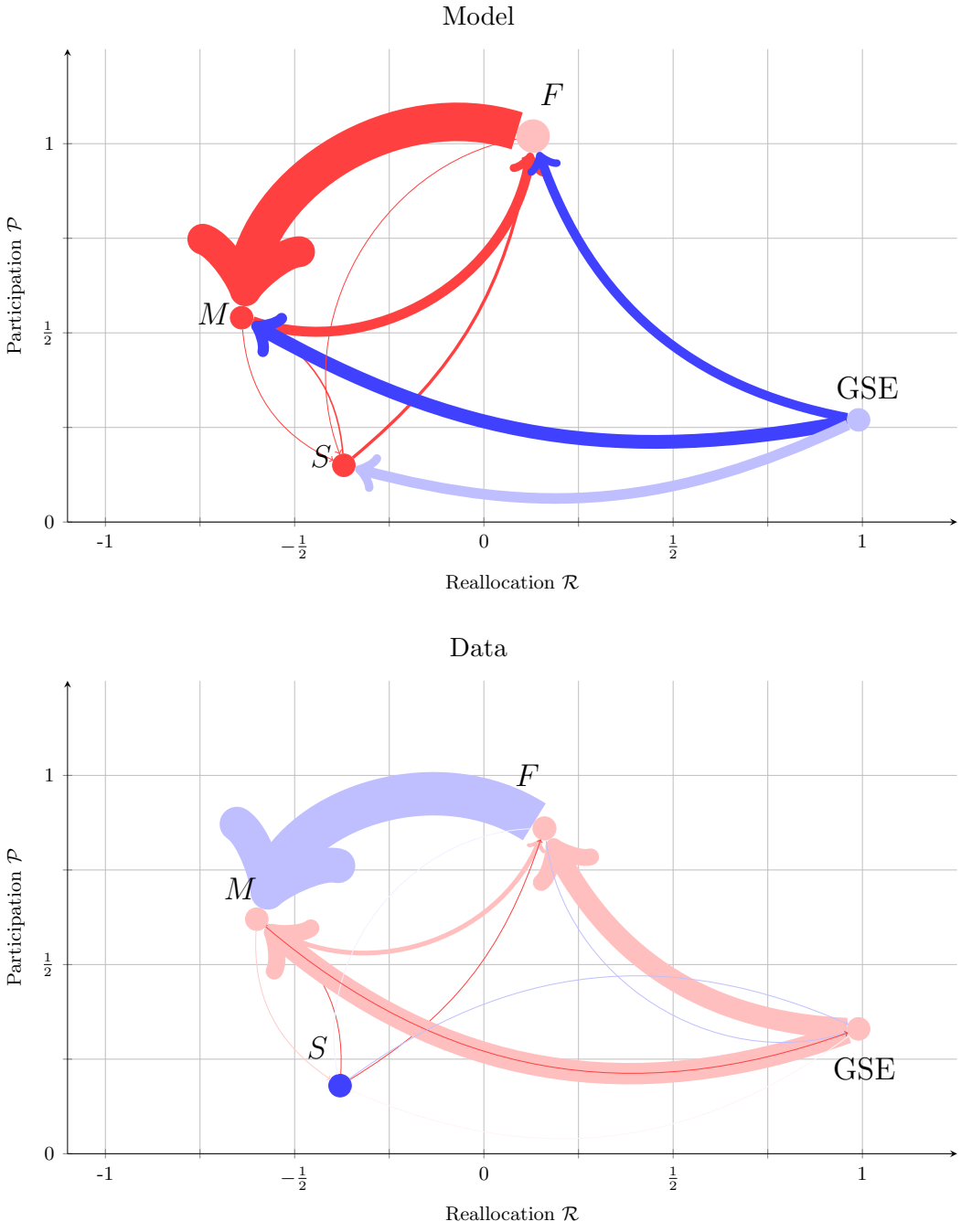

Figure 15: Theoretical and empirical fed funds trading networks for 2019.

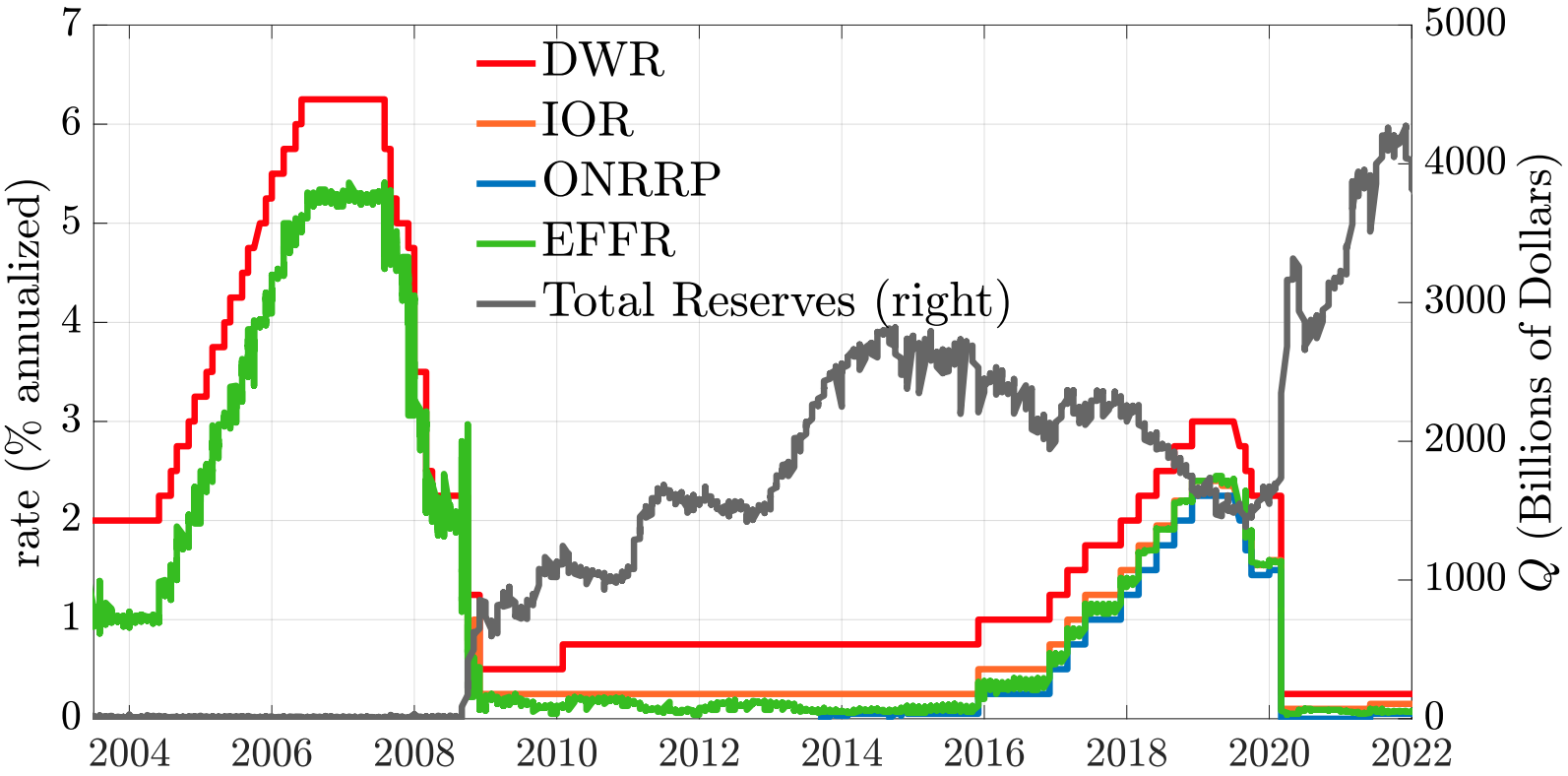

Figure 16: Time series of Total Reserves, and administered rates: Discount-Window rate (DWR), interest on reserves (IOR), and overnight reverse repo rate (ONRRP). Reserves are in billions of dollars. Rates are in percent per annum.

Notes: Total Reserves is "Reserve balances with Federal Reserve Banks: Wednesday level" from *Federal Reserve Balance Sheet: Factors Affecting Reserve Balances - H.4.1.* Administered rates are from https://fred.stlouisfed.org. DWR is "DPCREDIT"; IOR is "IOER" (until 2021/07) and "IORB" since (2021/08); ONRRP is "RRPONTSYAWARD"; EFFR is "DFF".

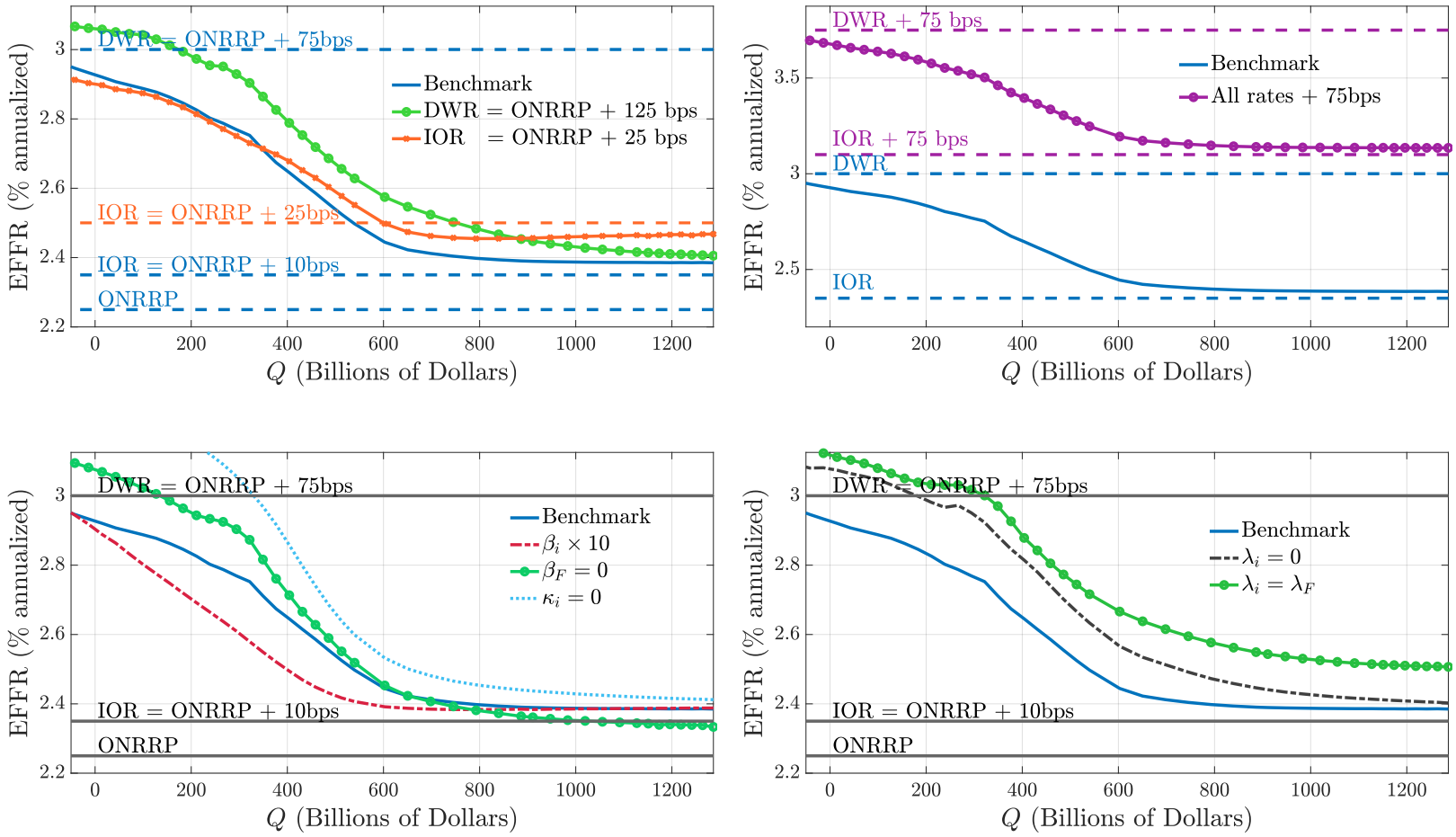

Figure 17: Theoretical aggregate demand for reserves: shifts and rotations.

Notes: In all panels, the curve labeled "Benchmark" is the theoretical aggregate demand $\iota^*_{\Upsilon_\omega} = \mathcal{D}(Q_{\Upsilon_\omega}; \Pi)$ for the model calibrated as in Table 1, and with $\iota^*_{\Upsilon_\omega}$ and Q_{Υ_ω} computed with the interpolation procedure described in Section 3.6, for $\Upsilon_0 = 2017$ and $\Upsilon_1 = 2019$. Top-left panel: benchmark aggregate demand, and aggregate demands resulting from two experiments: (i) increase DWR by 50 bps; (ii) increase IOR by 15 bps. Top-right panel: benchmark aggregate demand, and aggregate demand resulting from increasing all administered rates (i.e., DWR, IOR, and ONRRP) by 75 bps. Bottom-left panel: benchmark aggregate demand, and aggregate demands resulting from three experiments: (i) multiply $\{\beta_i\}_{i\in\mathbb{N}}$ by 10; (ii) set $\beta_F = 0$; (iii) set $\kappa_F = \kappa_S = 0$. Bottom-right panel: benchmark aggregate demand, and aggregate demands resulting from two experiments: (i) set $\lambda_i = 0$ for all $i \in \mathbb{N}$; (ii) set $\lambda_i = \lambda_F$ for all $i \in \mathbb{N}$.

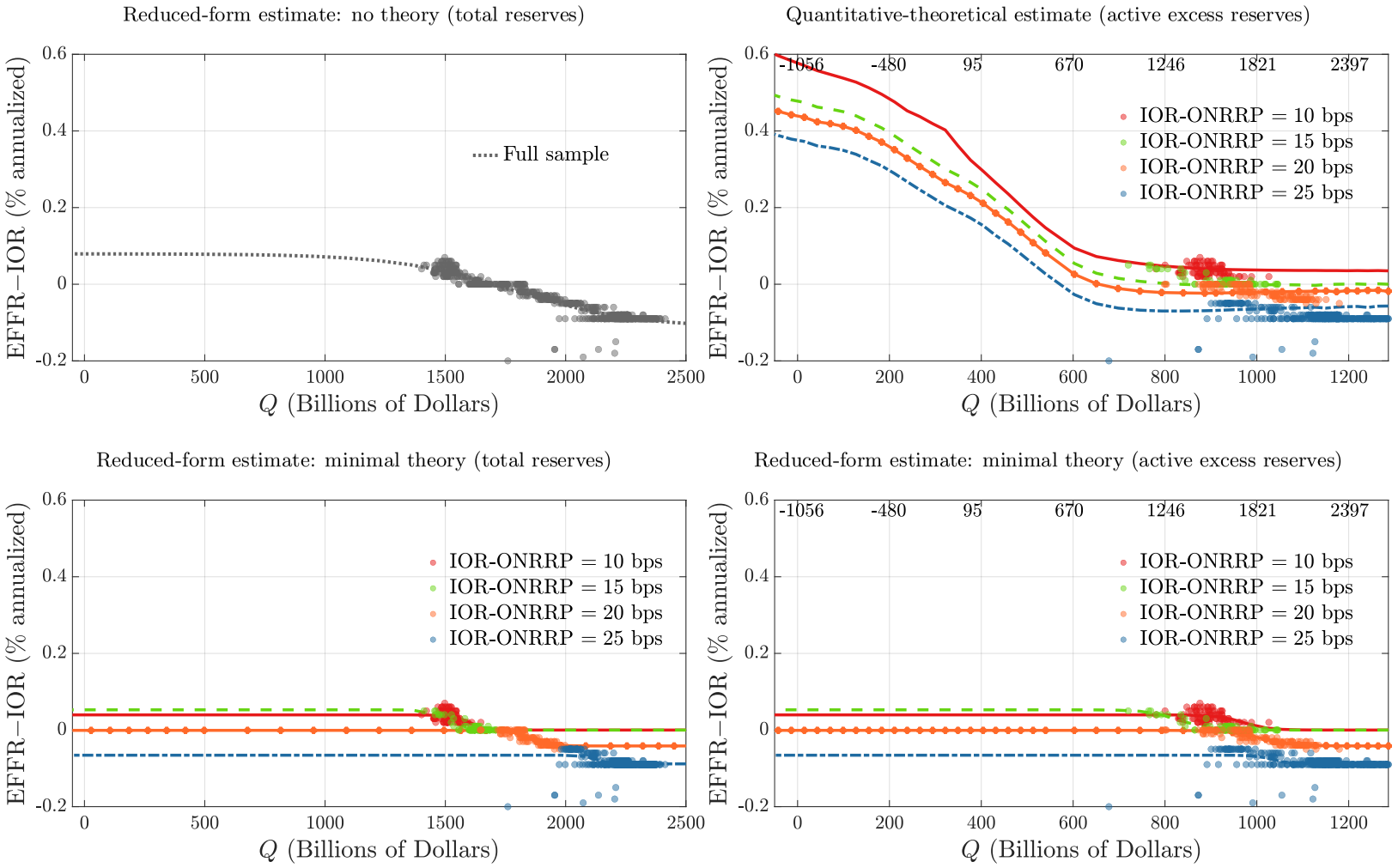

Figure 18: Aggregate demand for reserves: estimation.

Notes: In each panel: vertical axis is EFFR-IOR spread; horizontal axis is *total reserves* or *active excess reserves* as defined in Section 6. Full sample: all trading days in 2017/01/20–2019/09/13. Top-left panel: total reserves and nonlinear least square fit of (11) on full sample. Bottom-left panel: total reserves and nonlinear least square fits of (11) on each subsample (defined by IOR-ONRRP spread). Top-right panel: active excess reserves and aggregate demands implied by the theory for each IOR-ONRRP spread (all other parameters as in the baseline calibration). Bottom-right panel: active excess reserves and nonlinear least square fits of (11) on each subsample. Secondary horizontal axis (above top-left and bottom-left panels) translates active excess reserves into total reserves.

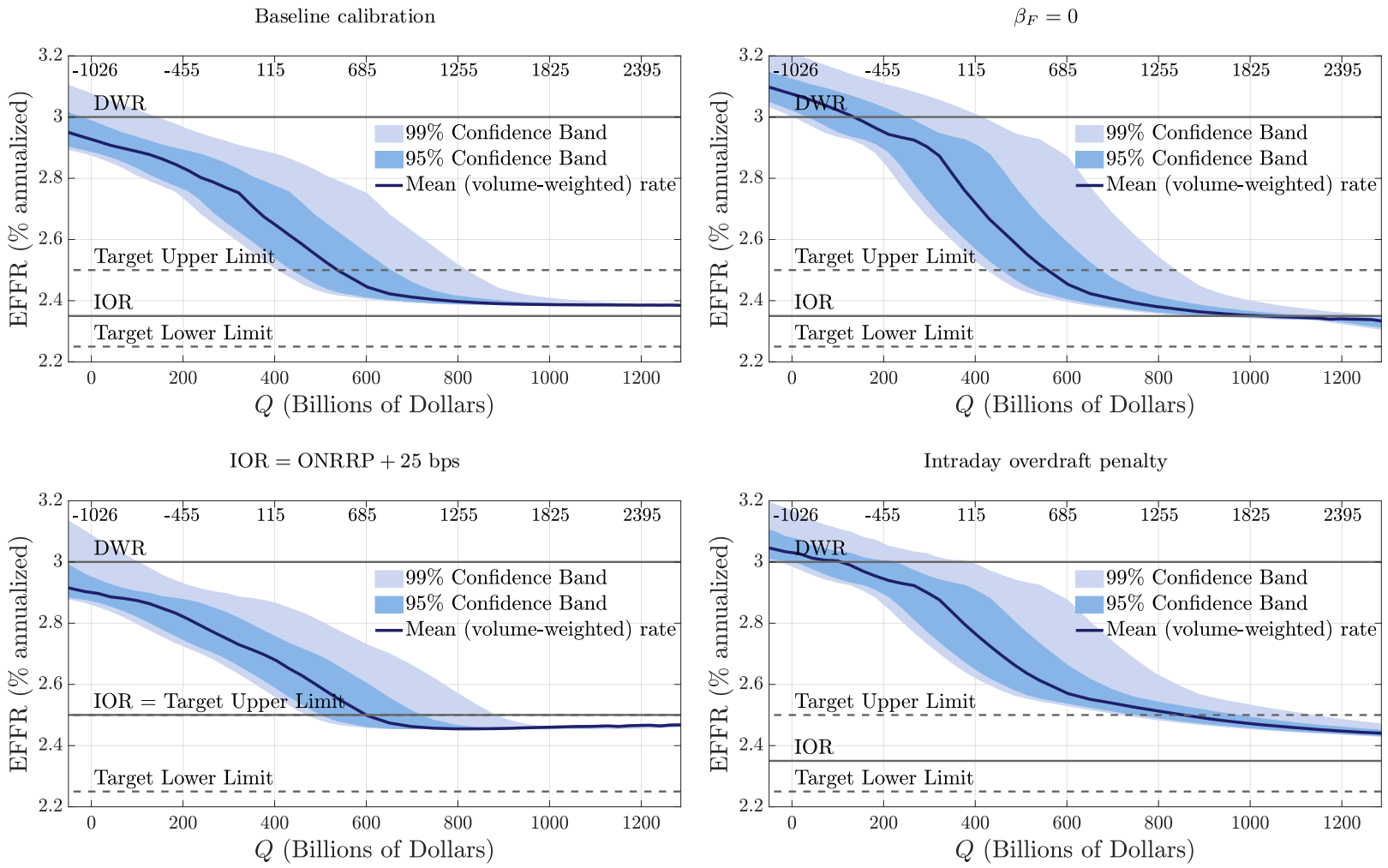

Figure 19: Monetary confidence bands.

Notes: In each panel, the curve labeled "Mean (volume-weighted) rate" is the theoretical money demand, $\mathcal{D}(Q)$. The lower and upper boundaries of the shaded area labeled "99% Confidence Band" are $\mathcal{D}(Q + Z_{99.5})$ and $\mathcal{D}(Q + Z_{0.5})$, respectively, where Z_p is the p^{th} percentile of the empirical distribution of reserve-draining shocks. The lower and upper boundaries of the shaded area labeled "95% Confidence Band" are $\mathcal{D}(Q + Z_{97.5})$ and $\mathcal{D}(Q + Z_{2.5})$, respectively. The top-left panel corresponds to the baseline calibration. The other panels are for calibrations that differ in one parameter from the baseline calibration. The top-right panel sets $\beta_F = 0$ (the baseline has $\beta_F = 0.03$). The bottom-left panel increases the IOR by 15 bps (from ONRRP + 10 bps in the baseline, to ONRRP + 25 bps). The bottom-right panel assumes $u_i(a) = \iota_d a \mathbb{I}_{\{a<0\}}$ for all i, with $\iota_d = \frac{0.2}{800}\iota_w$ (the baseline has $\iota_d = 0$).

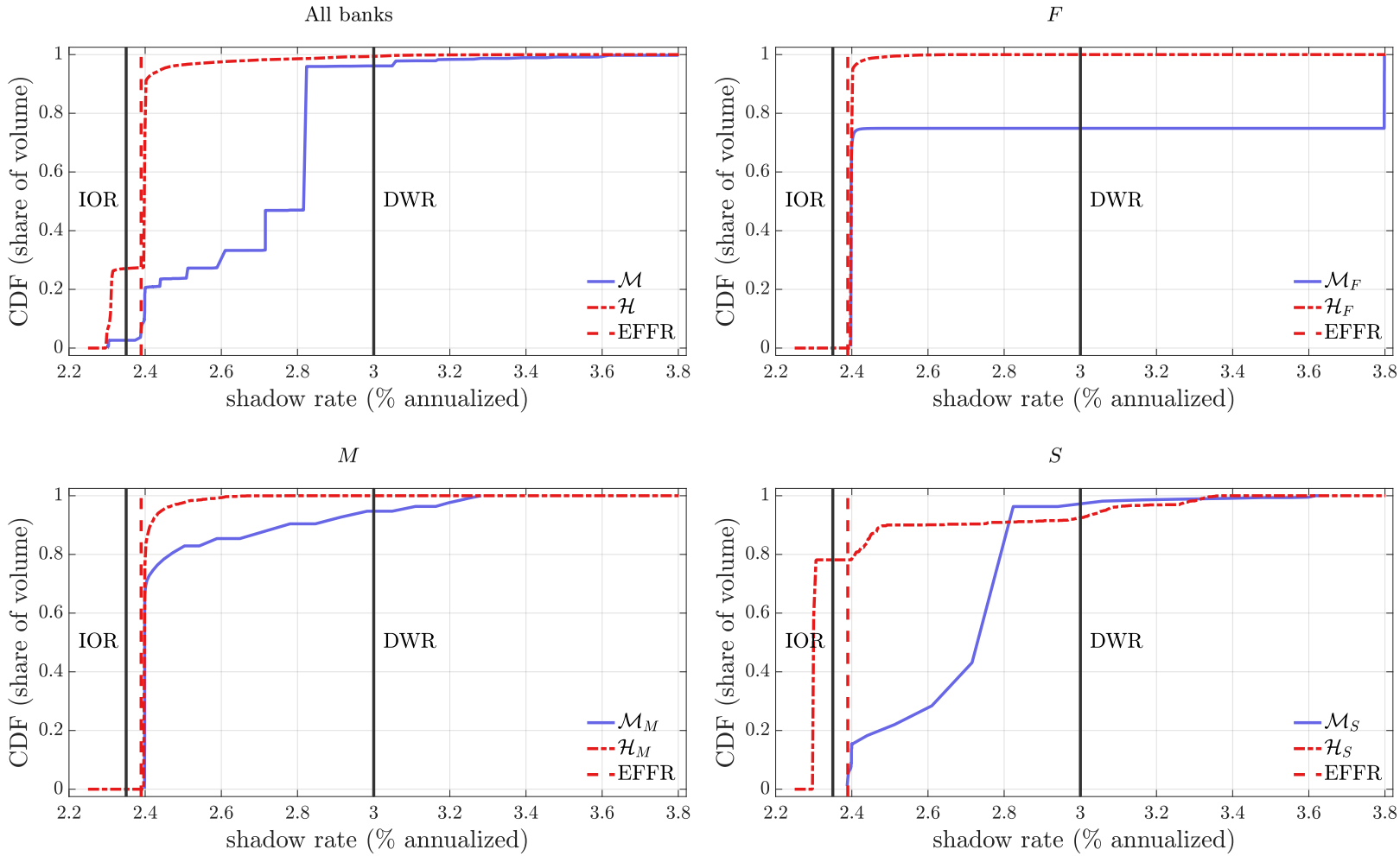

Figure 20: Cross-bank distributions of the shadow price of reserves.

Notes: All panels are constructed using data generated from the model under the baseline calibration. The beginning-of-day cumulative distribution function of shadow prices is denoted $\mathcal{M}_i$ for banks of type i, and $\mathcal{M}$ for all banks. The cumulative distribution function of all the bilateral loan rates negotiated throughout the day is denoted $\mathcal{H}$. The cumulative distribution function of all loan rates paid or received by banks of type i is denoted $\mathcal{H}_i$. The dashed vertical line labeled "EFFR" denotes the volume-weighted average fed funds rate on *all trades* implied by the theory. The IOR and DWR are denoted by solid vertical lines. All rates are in percent per annum.

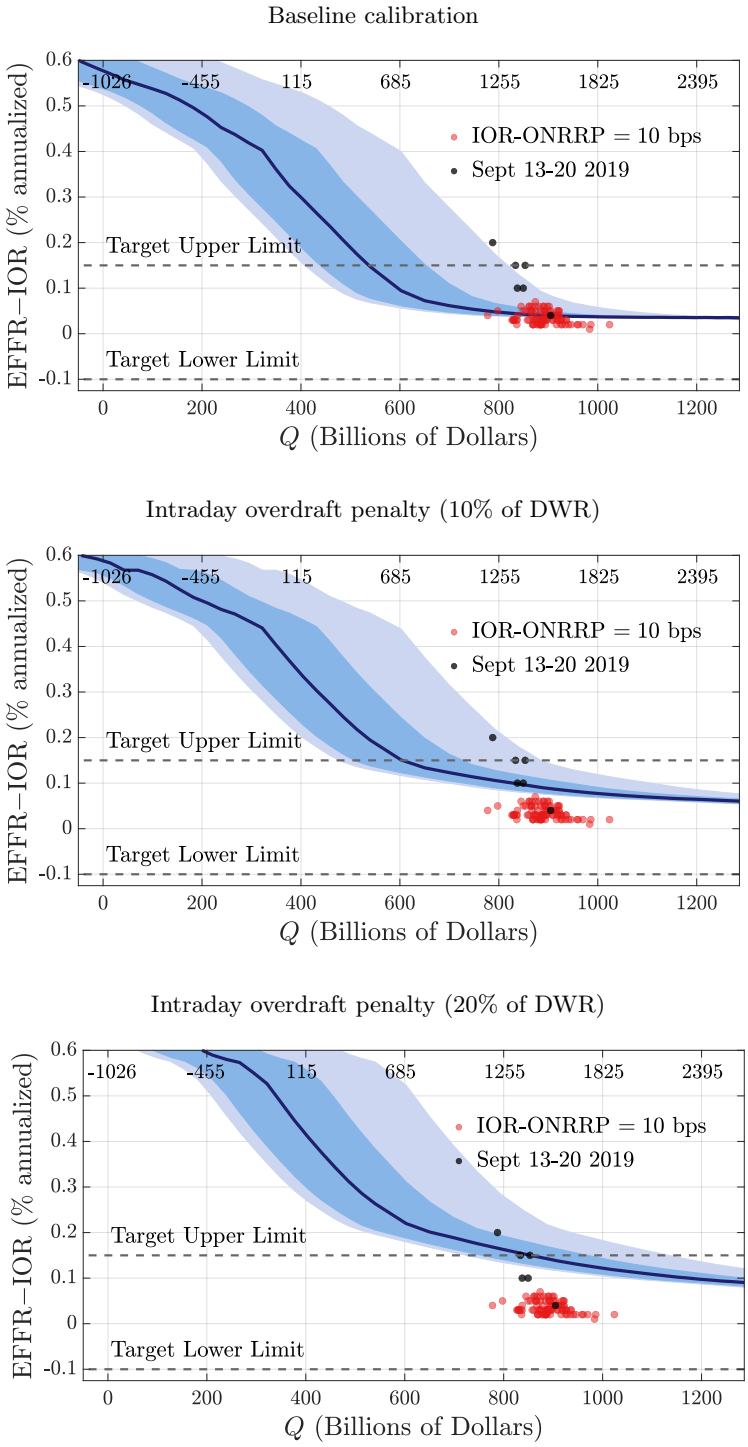

Figure 21: The events of September 13–20, 2019.

Notes: Each panel shows an MCB with the EFFR-IOR spread on the vertical axis (in percent per annum). The MCBs assume $u_i(a) = \iota_d a \mathbb{I}_{\{a<0\}}$ for all i, with $\iota_d = \frac{x}{800}\iota_w$; the top panel has $x = 0$ (the baseline calibration), the middle panel $x = 0.1$, and the bottom panel $x = 0.2$. The data points labeled "IOR-ONRRP=10 bps" are for the period 2019/05/02–2019/09/13. The dashed lines labeled "Target Upper Limit" and "Target Lower Limit" are the top and bottom of the fed funds target range minus the IOR for the period 2019/05/02–2019/09/18.

Day	Administered Rates					EFFR	EFFR-IOR	Fed Repo
	ONRRP	IOR	DWR	TRL	TRU			
September 13 (Friday)	2.00	2.10	2.75	2.00	2.25	2.14	0.04	0
September 16 (Monday)	2.00	2.10	2.75	2.00	2.25	2.25	0.15	0
September 17 (Tuesday)	2.00	2.10	2.75	2.00	2.25	2.30	0.20	53
September 18 (Wednesday)	2.00	2.10	2.75	2.00	2.25	2.25	0.15	75
September 19 (Thursday)	1.70	1.80	2.50	1.75	2.00	1.90	0.10	75
September 20 (Friday)	1.70	1.80	2.50	1.75	2.00	1.90	0.10	75

Table 2: The events of September 13–20, 2019.

Notes: ONRRP, IOR, and DWR, denote the overnight reverse repo rate, the interest rate paid on reserves, and the discount-window rate, respectively. The lower limit of the fed funds target range is denoted TRL, and the upper limit is denoted TRU. The effective fed funds rate is denoted EFFR, and EFFR-IOR denotes the EFFR-IOR spread. The column labeled "Fed Repo" reports the quantity of reserves injected by the Federal Reserve during day t through overnight repo operations. All rates are in percent per annum. All quantities in billions of dollars.