Monetary Policy Operations: Theory, Evidence, and Tools for Quantitative Analysis*

Ricardo Lagos
New York University

Gastón Navarro
Federal Reserve Bank of Richmond

October 29, 2025

Abstract

We formulate a quantitative dynamic equilibrium theory of trade in the interbank overnight 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 rate target, and the heterogeneous incidence of policy actions on the shadow cost of funding across banks.

Keywords: demand for reserves, monetary policy, interbank market
JEL classification: G1, C78, D83, E44

*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 Bank of Richmond or the Federal Reserve System.

1 Introduction

Interbank overnight rates are the primary instruments through which modern central banks communicate and implement their monetary policy stance. Currently, the implementation of a rate target is guided by the stylized demand-and-supply framework illustrated in Figure 1. The goal of this paper is to develop a quantitative theory of interbank overnight rates that will be a better guide for monetary policy implementation.

Before the Great Financial Crisis of 2007–2008 (GFC), aggregate reserves in the United States were scarce (e.g., at a level such as Q_0 in the first panel of Figure 1). The interbank overnight market then operated on the steep portion of the demand curve for reserves, and target rates such as r_0^* were achieved by adjusting the supply of reserves through open-market operations. This operating framework is known as a *corridor system*, since it can implement any target rate within the corridor defined by a *ceiling rate* (e.g., the discount-window rate, ι_w) and a *floor rate* (e.g., the interest paid on reserves, ι_r).

After the onset of the GFC, the Federal Reserve undertook large-scale asset purchases that raised banks' reserve balances to very high levels (e.g., Q_1 in the top-right panel of Figure 1). When the interbank market operates on the flat portion of the demand curve for reserves, the central bank can no longer rely on conventional (small-scale) open-market operations to implement changes in the target rate. Target rates such as r_1^* are instead achieved by adjusting the administered rates offered at standing facilities: the discount-window rate (DWR), the interest on reserves (IOR), and the offering rate on overnight reverse repurchase agreements (ONRRP). This operating framework is known as a *floor system*.

To describe the ranges of reserves compatible with these two operating frameworks, policymakers often use the terms "scarce reserves" for the range where the slope of the demand is "steep," "ample reserves" for the range where it is "gentle," and "abundant reserves" for the range where it is "flat," as illustrated in the top-right panel of Figure 1.[1] The Federal Reserve has announced its intention to continue operating a floor system in which an "ample" supply of reserves ensures that the policy rate is controlled by the administered rates and in which "active management of the supply of reserves is not required."[2]

In terms of the schematic in Figure 1, a floor system may appear easy to manage: the central

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

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

bank only needs to ensure that the supply of reserves is "ample," i.e., near Q_1 in the top-right panel. In practice, however, this is challenging because identifying the range of "ample" reserves requires *global* estimates of the slope of the aggregate demand for reserves, whereas standard empirical approaches can at best deliver *local* estimates. This *local–global gap* is a significant obstacle to effectively operating a floor system such as the one adopted by the Federal Reserve.[3]

In this paper, we develop a quantitative model of the interbank overnight market, calibrate it to match a wide set of bank- and market-level statistics—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 exploit the equilibrium relationship between the aggregate supply of reserves and the interbank rate to estimate the global shape of the aggregate demand for reserves in the United States.

The theory incorporates search and bilateral bargaining to capture the well-documented over-the-counter microstructure of the overnight interbank market. It accounts for key institutional features, including the differential regulatory treatment of government-sponsored enterprises (GSEs), foreign banking organizations (FBOs), and domestic banks. The framework also incorporates the full set of policy instruments and regulations shaping participants' demand for reserves—namely, administered policy rates (DWR, IOR, ONRRP), regulatory requirements, and the aggregate quantity of reserves supplied to the system. In addition, it accommodates substantial heterogeneity among market participants along multiple dimensions: market power in bilateral negotiations; the frequency and size distribution of idiosyncratic payment shocks from non-trading motives; various measures of trading activity (including trade frequency, number of counterparties, and share of aggregate volume); and the degree of centrality in the endogenous market-making that reallocates reserves throughout the trading network.

We calibrate the model's parameters governing heterogeneity in payment and trading activity using daily, minute-by-minute bank-level data from Fedwire. The model fits the targeted cross-sectional features well—for example, as in the data, a small number of highly active banks account for most lending and intermediation. The calibration strategy also ensures that

[3]By *local* estimates we mean estimates based on instrumented variation around a relatively narrow range of the supply of reserves. These estimates often cannot be extrapolated to infer the effects of large changes in reserves (e.g., in the presence of structural or policy changes). The empirical identification challenges are illustrated in the bottom panels of Figure 1, which show situations in which the 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) shifts the demand for reserves, one may be led to believe the observations $\{(Q_i, r_i^*)\}_{i\in\{0,1\}}$ lie on a single demand curve, thereby overestimating (bottom-left panel) or underestimating (bottom-right panel) the relevant slope.

the model matches empirical estimates of the *local* slope of aggregate reserve demand at the prevailing level of total reserves. In addition, the calibrated model is broadly consistent with non-targeted empirical patterns, including the cross-sectional distribution of bilateral interbank rates, the distribution of bid-ask spreads, and the intraday flow of reserves and associated rates between bank pairs occupying different positions in the trading network.

We use the quantitative theory to guide the practice of monetary policy implementation. First, we provide global estimates of the aggregate demand for reserves, which are useful to central banks operating floor systems. Second, we devise two "navigational instruments" for monetary policy implementation. The first, which we term the *Monetary Confidence Band* (MCB), is a hybrid of theory and data: a procedure that uses the empirical distribution of daily reserve-supply shocks to construct a confidence band around the aggregate demand for reserves. For each outstanding quantity of reserves, the band contains the equilibrium overnight interbank rate with a desired degree of confidence, e.g., 99%. Using this instrument, we estimate that total reserves of about 10% of GDP would be sufficient to ensure the overnight interbank rate remains within the target range with 99% probability on any given day. The second instrument is the cross-sectional distribution of banks' shadow cost of overnight funding implied by the theory.

This paper contributes to the empirical and theoretical literature on the overnight interbank market, e.g., Poole (1968), Hamilton (1996), Furfine (1999), Carpenter and Demiralp (2006), Ashcraft and Duffie (2007), Bech and Atalay (2010), Afonso et al. (2011), Bech and Klee (2011), Afonso and Lagos (2014, 2015a,b), Ennis and Weinberg (2013), Armenter and Lester (2017), Afonso et al. (2019), Ennis (2019), Chiu et al. (2020), Beltran et al. (2021), Copeland et al. (2021), and Afonso et al. (2022). Methodologically, we draw on 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), Afonso and Lagos (2015a,b), Üslü (2019), and Hugonnier et al. (2020). In particular, we generalize the model in Afonso and Lagos (2015b) to make it a serviceable quantitative tool for monetary policy implementation.

2 Theory

There is a unit measure of *banks*, heterogeneous along several dimensions. We represent this heterogeneity with a finite set $\mathbb{N}$ of *bank types*, where $n_i \in [0,1]$ denotes 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, which can be traded during the time interval $\mathbb{T} = [0, T]$. A bank's reserve balance is represented by a real number, e.g., $a \in \mathbb{R}$. The cumulative distribution function of reserve balances across banks at time $t \in \mathbb{T}$ is $F_t(a) = \sum_{i \in \mathbb{N}} n_i F_t^i(a)$, where $F_t^i(a) : \mathbb{R} \times \mathbb{T} \to [0, 1]$ is the distribution for type i at time t. The initial distributions, $\left\{F_0^i(\cdot)\right\}_{i \in \mathbb{N}}$, are given, as is the aggregate supply of reserve balances, $Q \equiv \int a dF_0(a)$.

Banks trade reserves in a bilateral over-the-counter market. A bank of type $i \in \mathbb{N}$ contacts another bank at random times governed 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. Upon contact, the two banks 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 $j \in \mathbb{N}$, the bargaining power of the former is $\theta_{ij} = 1 - \theta_{ji} \in [0, 1]$. After the transaction, the banks part ways.

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

Banks receive payment shocks that reallocate reserve balances across pairs of banks. With Poisson rate $\lambda_i \in \mathbb{R}_+$, a bank of type i must immediately transfer reserves to a counterparty drawn uniformly from the population. The arrival of payment shocks is independent across banks and from the processes that generate bilateral trading opportunities. Conditional on a payment shock, the transfer from a type i bank to a type j bank is a random variable with cumulative distribution function $G_{ij} : \mathbb{Z} \to [0, 1]$, where $\mathbb{Z} \subseteq \mathbb{R}$ is its support, and $dG_{ij}(z) = dG_{ji}(-z)$, capturing that payments between banks net to zero in aggregate.

Let $u_i : \mathbb{R} \to \mathbb{R}$ and $U_i : \mathbb{R} \to \mathbb{R}$ denote the payoffs to a bank of type $i \in \mathbb{N}$ from holding reserve balance $a \in \mathbb{R}$ during the trading session and at the end of the trading session, respectively. A bank type $i \in \mathbb{N}$ is defined by the primitives $(n_i, \beta_i, \lambda_i, \{\theta_{ij}, G_{ij}\}_{j \in \mathbb{N}}, u_i, U_i)$: 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, payment-size distributions $\{G_{ij}\}_{j \in \mathbb{N}}$, intraday payoff function u_i, and end-of-day payoff function U_i.

2.1 Discussion

In the United States, the interbank overnight market is the epicenter of monetary policy implementation. Staff of the System Open Market Account (SOMA) closely monitor interbank

overnight rates to determine the operations needed to implement the directives of the Federal Open Market Committee (FOMC). The two main benchmarks are the *Effective Federal Funds Rate* (EFFR) and the *Secured Overnight Financing Rate* (SOFR).[4]

Market participants include domestic banks (such as commercial banks, investment banks, thrift institutions, and credit unions), Foreign Banking Organizations (FBOs, i.e., agencies and branches of foreign banks operating 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 the Federal Home Loan Banks). Trade is *over the counter*: participants must first identify a willing counterparty, and then negotiate bilaterally the loan size and interest rate.

We use a search-based model with ex post bargaining to capture the bilateral, over-the-counter nature of the interbank overnight market. The outcome of bilateral negotiations is represented by the generalized Nash bargaining solution. Search frictions in the model reflect three layers of randomness in trading activity. First, the time until a bank of type $i \in \mathbb{N}$ contacts a counterparty is exponentially distributed with mean $1/\beta_i$. Second, conditional on contact, the counterparty's type is drawn uniformly at random. Third, conditional on meeting a type-j counterparty at time t, its reserve balance is a random variable with cumulative distribution function $\{F_t^j(\cdot)\}_{j\in\mathbb{N}}$. We interpret $t = 0$ as 9:00 a.m. of a typical trading day, with the initial condition $\{F_0^i(\cdot)\}_{i\in\mathbb{N}}$ representing the distribution of reserve balances at that time.

Financial institutions participate in the interbank overnight market primarily to manage daily cash positions. They borrow and lend reserve balances to offset random payment shocks—arising from transactions initiated by clients or internal profit centers—that would otherwise leave them with reserve positions either above or below desired levels, for example relative to internal targets or regulatory requirements. In the model, the Poisson rate λ_i governs the frequency of such payment shocks for a bank of type $i \in \mathbb{N}$, and G_{ij} denotes the distribution of shock sizes for payments between banks of types i and j. The fundamental motives for holding reserves, including regulatory motives, are represented by the intraday and end-of-day payoff functions $\{u_i(\cdot), U_i(\cdot)\}_{i\in\mathbb{N}}$.

[4]The EFFR is the volume-weighted average rate of overnight, uncollateralized loans of reserve balances at Federal Reserve Banks, known as *federal funds*. The SOFR is the volume-weighted average rate of overnight *repurchase agreements*, known as *repos*. At present, the Federal Reserve uses the EFFR as its operating target, though there are ongoing proposals to modernize the framework by adopting the SOFR (see, e.g., Logan (2025)).

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 $t \in \mathbb{T}$, holds a reserve balance $a \in \mathbb{R}$ and a net credit position $c \in \mathbb{R}$. Appendix F (Lemma 1) shows 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 meet 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 amount 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 amount 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 the 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. The implied gross interest rate on this loan is

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

Appendix F (Lemma 2) shows that the value function $V_t^i(a)$ satisfies

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

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 that 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 evolves according to the law of motion

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

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 Interbank trading network: evidence

Most interbank transfers of reserves in the United States are executed through *Fedwire Funds Services* (Fedwire), an electronic, large-value, real-time gross settlement system operated by the Federal Reserve Banks. Our empirical analysis begins with the daily, minute-by-minute universe of bilateral transfers between Fedwire participants, from which we identify bilateral interbank overnight loans and interest rates by applying a version of the *Furfine algorithm*. We refer to the subset of Fedwire transfers that the algorithm identifies as overnight loans and repayments as the *Fedwire Overnight Loan* (FWOL) database, and to the corresponding daily value-weighted average interest rate as the *Overnight Fedwire Rate* (OFR).[5]

[5]In Section C.3 we provide a detailed explanation of the construction of the FWOL database. In Section C.4 we describe a procedure that matches loans on the FWOL database to loans in the FR 2420 filings, which allows us to identify certain *loan types* within the FWOL database (e.g., fed funds, Eurodollars). Empirically, the OFR for the loans that our matching procedure identifies as repos in the FWOL database is statistically very close to the SOFR, on average (see Section C.4.1).

We use the FWOL database to document two cross-sectional measures of trading activity that we will use as data targets to discipline the quantitative implementation of the theory. The joint distribution of these statistics reveals an empirical trading network characterized by substantial heterogeneity of trading activity across banks. Most banks trade relatively little and are either net buyers or net sellers, while a core of very active banks accounts for the bulk of the trade volume and act as intermediaries between third parties.[6]

The first statistic is the *participation rate*, which measures a bank's share of the market-wide volume of trade. For bank n on day d, it is defined as

$$\mathcal{P}_{nd} = \frac{\upsilon^e_{nd} + \upsilon^r_{nd}}{2\upsilon_d},$$

where υ^e_{nd} is the dollar value of loans extended by bank n on day d, υ^r_{nd} is the dollar value of loans received, and $\upsilon_d = \sum_m \upsilon^e_{md}$ is market-wide trade volume on day d. Note that $\mathcal{P}_{nd} \in [0, 1/2]$, with $\mathcal{P}_{nd} = 0$ if bank n did not trade during day d, and $\mathcal{P}_{nd} = 1/2$ if bank n was a counterparty for every dollar traded in that day.

The second statistic is the *reallocation index*, which measures the extent to which a bank is a net borrower or a net lender. For bank n on day d, it is defined as

$$\mathcal{R}_{nd} = \frac{\upsilon^e_{nd} - \upsilon^r_{nd}}{\upsilon^e_{nd} + \upsilon^r_{nd}}.$$

Note that $\mathcal{R}_{nd} \in [-1, 1]$, 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 that acted as a pure intermediary—that is, lending every dollar it borrowed. A typical bank n will have $\mathcal{R}_{nd} \in (-1, 0)$ if it is a net borrower that engaged in some intermediation, or $\mathcal{R}_{nd} \in (0, 1)$ if it is a net lender that engaged in some intermediation.

We compute the average of each bank's participation rate and reallocation index across trading days in the year, and denote these averages by $\mathcal{P}_n$ and $\mathcal{R}_n$, respectively. A few banks stand out for their high participation rates.[7] For example, each of the four most active banks individually accounts for at least 10% of the total loan volume. Together, these four banks account for 45.6% of market-wide volume in 2006 and 43.1% in 2019, implying that they acted as a counterparty in almost every dollar traded on a typical day in those years. In contrast,

[6] See Appendix A for a more comprehensive description of the facts.

[7] Figure 10 in Appendix A shows the empirical cumulative distribution function (ECDF) of participation rates, $\{\mathcal{P}_n\}$, for all banks in our sample in 2006 (circles) and 2019 (crosses).

most banks have extremely low participation rates. We regard this pronounced skewness in loan trading activity across banks as a key empirical regularity of the interbank market structure.

We use the participation rate to classify banks into three *types*, denoted S, M, and F, depending on whether a bank's participation rate is low, medium, or high, respectively. Specifically, in each year we label the four banks with the highest participation rates as F; the banks outside the top four whose participation rate is at least 0.5% as M; and all remaining banks as S. We also define a fourth type, denoted G, composed of *Government Sponsored Enterprises* (GSEs)—nonbank Fedwire participants that play an active role in the interbank market but are subject to different regulations.[8] Next, we document the joint relationship between participation rates and reallocation indices across types.

Figure 2 shows the location of each type $i \in \{F, M, S, G\}$ along the axes defined by the reallocation index, $\mathcal{R}_i$, and the participation rate, $\mathcal{P}_i$, in the years 2006, 2014, 2017, and 2019. Each panel displays what we refer to as the *interbank trading network*.[9] Each node in the network graph represents the set of banks assigned to a particular type. Node positions are determined by the participation and reallocation measures of that type. Arrows indicate loans extended from one bank type to another. Node size is proportional to the volume of trades between banks of the same type, while arrow width is proportional to the volume of trades between the connected types. The color of each node and arrow—light blue, dark blue, light red, or dark red—indicates the quartile of the volume-weighted average interest rate on loans between the two types, expressed as a spread over the OFR.[10]

Several stable trading patterns emerge from Figure 2. Banks of type F buy or sell most of the loans (i.e., $\mathcal{P}_F \approx 1/2$), and intermediate a large share of what they trade, with a slight tendency to act as net lenders (i.e., $\mathcal{R}_F$ is slightly positive). 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 $1/8$ of aggregate trade volume, act almost exclusively as lenders, and their participation increases after the GFC.

The empirical interbank trading network reveals the key trading patterns and reserve flows, and is informative about the relevant dimensions of heterogeneity. Through the lens of our

[8] Type G includes the Federal Home Loan Banks, the Federal National Mortgage Association (Fannie Mae), and the Federal Home Loan Mortgage Corporation (Freddie Mac).

[9] Appendix C (Section C.5.2) describes the procedure to compute the reallocation index for each bank type.

[10] Arrow widths and node sizes are defined relative to trades within each year and are therefore not comparable across years.

theory, all else equal, banks with high, medium, and low contact rates (i.e., β_i) would tend to be sorted into types F, M, and S, respectively, since higher contact rates imply greater participation and a comparative advantage in intermediating funds.

4 Calibration

We calibrate the model to match interbank trading activity in 2019. The primitives are: the trading session $[0, T]$; the discount rate r; the 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 rates $\{\lambda_i\}_{i\in\mathbb{N}}$; 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 rates $\{\beta_i\}_{i\in\mathbb{N}}$. For the quantitative implementation of the theory, we include proportional borrowing costs $\{\kappa_i\}_{i\in\mathbb{N}}$, which proxy for institutional and regulatory factors affecting banks' incentives to borrow in the interbank market.[11] Our calibration strategy is as follows.[12]

We interpret the trading session in the model as a trading day within a typical 14-day reserve maintenance period, and let $[0, T]$ correspond to the interval from 9:00 a.m. to 6:30 p.m. (EST) on an actual trading day.[13] In the quantitative implementation of the theory, we discretize the time interval $[0, T]$ into 800 periods, so that each period in the model corresponds approximately to a 42-second interval of the trading day.[14] With such a short model period, we abstract from pure discounting and set $r = 0$.

We classify Fedwire participants into four types, i.e., $\mathbb{N} = \{F, M, S, G\}$, based on their regulatory treatment and participation rates, as described in Section 3. 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.

A bank's beginning-of-day reserve balance in the model corresponds to its beginning-of-day balance of *unencumbered reserves* in the data—defined as the bank's beginning-of-day reserve balance net of regulatory reserve requirements and predictable Fedwire transfers (both outright payments and overnight-loan repayments).[15] Accordingly, we set the theoretical beginning-of-day distributions $\left\{F_0^i(\cdot)\right\}_{i\in\mathbb{N}}$ equal to the corresponding kernel estimates of the empirical

[11] Appendix F (Section F.2) generalizes (1), (2), and (3) to include proportional borrowing costs.

[12] In Appendix E we explore alternative calibrations and verify the robustness of the main results.

[13] As reported in Appendix A, there is little trading activity between 9:00 p.m. and 9:00 a.m. on the following trading day.

[14] See Appendix G (Section G.1) for details.

[15] Appendix C (Section C.5.1) details the construction of bank-level beginning-of-day unencumbered reserves.

distributions of beginning-of-day unencumbered reserves reported in Appendix A (Section A.3).

The payment shock rates, $\{\lambda_i\}_{i\in\mathbb{N}}$, are calibrated to match the empirical one-second frequencies of payment shocks reported in Appendix A (Section A.2).[16] The size distributions of payment shocks, $\{G_{ij}\}_{i,j\in\mathbb{N}}$, are set equal to the corresponding empirical kernel estimates reported in Appendix A (Section A.2).

We set $\theta_{ij} = \underline{\theta}$ if $i \in \{G\}$ and $j \in \mathbb{N} \setminus \{G\}$, and $\theta_{ij} = 1/2$ otherwise—that is, unless one of the parties to the trade is a GSE, we abstract from differences in relative market power driven purely by bank *type*.

We set $u_i(a) = 0$ for all $(a, i) \in \mathbb{R} \times \mathbb{N}$.[17] End-of-day payoffs are parameterized by

$$U_i(a) = \left(1 + \mathbb{I}_{\{0\leq a\}}\bar{\iota}_r + \mathbb{I}_{\{a<0\}}\bar{\iota}_w\right) a, \tag{5}$$

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 the end-of-day balance in excess of reserve requirements.[18]

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 costs and benefits from holding reserves not captured by the administered rates.[19] We use ι_w to denote the Discount Window rate (DWR) that the Federal Reserve 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.[20]

For GSEs, the end-of-day payoff is $U_G(a) = \left(1 + \mathbb{I}_{\{0\leq a\}}\bar{\iota}_o + \mathbb{I}_{\{a<0\}}\bar{\iota}_w\right) a$, 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

[16] In the discrete-time approximation that we use for computation, λ_i is the probability that a bank of type i receives a payment shock within a one-second interval. See Section Appendix G (Section G.1) for details.

[17] That is, we abstract from banks' *intraday* payoffs from holding reserves, such as the regulatory costs associated with running an intraday overdraft with the Federal Reserve.

[18] Since our calibration strategy maps beginning-of-day reserve balances in the theory to *unencumbered reserves* in the data—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 end-of-day payoff functions to be *type specific* but not *bank specific*, even though in the data two banks of the same type may have different reserve requirements. To see this, let $\mathcal{U}_i(b, \underline{b})$ denote the end-of-day payoff of a bank of type i with reserve requirement $\underline{b}$ and reserve balance b (gross of the requirement). We parameterize 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 $U_i(a)$ up to 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 (5).

[19] For example, ι_ℓ may capture the additional return associated with using reserves as a means of payment, or with lending reserves outside the interbank market (e.g., loans to corporate or retail customers).

[20] Stigma associated with Discount Window borrowing is a common explanation for why, empirically, banks sometimes borrow reserves at a premium over the DWR (see, e.g., Artuç and Demiralp (2010), Ennis and Weinberg (2013), Armantier et al. (2015), Ennis (2019), and Klee et al. (2021)).

repo facility.[21] 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 in 2019: (1) the daily average value-weighted interbank rate (OFR); (2) the daily average value-weighted interest rate (OFR) for loans with rates lower than the IOR; (3) the regression estimate of the "liquidity effect"—the slope of the empirical aggregate demand for reserves in the base year, as estimated in Appendix A (Section A.5); (4) the ratio of the average number of loans traded by type-F banks to the average number traded by all banks; (5)–(8) the reallocation indices $\{\mathcal{R}_i\}_{i\in\mathbb{N}}$; and (9)–(11) the participation rates $\{\mathcal{P}_i\}_{i\in\mathbb{N}\setminus\{F\}}$.[22]

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 institutions that were active in the FWOL database in 2019, respectively. To interpret the payment-shock rates, $\{\lambda_i\}_{i\in\mathbb{N}}$, recall that λ_i represents the probability that a bank of type i receives a payment shock within a one-second interval. For example, $\lambda_M = 0.257$ implies that a bank of type M receives a payment shock approximately every four seconds on average; $\lambda_F = 0.951$ implies roughly one payment shock per second; and $\lambda_S = 0.011$ implies a payment shock approximately every 90 seconds on average. The values of ι_w (DWR), ι_r (IOR), and ι_o (ONRRP) are set to 3.00%, 2.35%, and 2.25% per annum, respectively, which were the administered policy rates in effect from May through July 2019.

The calibration delivers a liquidity return (ι_ℓ) of 2.2 bps per annum, an additional cost associated with Discount Window borrowing (ι_s) of 75 bps per annum (about one quarter of the DWR), and $\underline{\theta} = 0.05$, which means that a GSE captures 5% of the total gains from lending to a non-GSE. The trading rate, β_i, is interpreted as the probability that a bank of type i contacts a trading partner within a 42-second interval. The calibrated values $\{\beta_i\}_{i\in\mathbb{N}}$ imply that banks of type F, M, S, and G trade 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 consistent with the range of empirical estimates for 2019 reported in Appendix A (Section A.5).[23] The borrowing costs, $\{\kappa_i\}_{i\in\mathbb{N}}$, which proxy for institutional and regulatory

[21] 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.

[22] 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 $\sum_{i\in\mathbb{N}} \mathcal{P}_i = 1$.

[23] Figure 18 shows the magnitude of the liquidity effect in the calibrated model along with the 95% confidence

factors affecting banks' incentives to borrow, are positive for banks of type F and S, and zero for banks of type M.[24]

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 five 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 distribution of bilateral borrowing and lending rates for each pair of bank types. The fifth, on the trading network. The main takeaway is that the model successfully matches a wide range of moments that were not targeted by the calibration, which lends credibility to its quantitative predictions.

5.1 Distribution of loan rates

Figure 3 shows the empirical and theoretical cumulative distribution functions of interest rates negotiated in the year 2019, along with the administered rates prevailing in the sample period (all expressed in percent per annum).[25] The model delivers a reasonable fit for the cross-sectional distribution of interest rates, which was not targeted in the calibration.[26]

bands for the regression estimates for the period 2019/05/02–2019/09/13 presented in Section A.5. In the model, the liquidity effect is computed by extracting $100 bn of reserves (approximately two standard deviations of the size distribution of reserve-supply shocks) using the procedure described in Section A.6.

[24]The value of κ_G is set sufficiently high to match the observation that GSEs essentially do not borrow in the interbank market, but its exact value is inconsequential.

[25]Data are for every trading day in the period 2019/06/06–2019/07/31, and 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.

[26]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 heterogeneity 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 overnight loans being misclassified as interbank loans in our dataset (e.g., loans between non-banks who access Fedwire through correspondent banks).

5.2 Conditional distribution of loan rates above the DWR

As in the data, banks in the model sometimes trade at rates above the DWR.[27] During the sample period 2019/06/06–2019/07/31, the DWR was set at 3%; the 10th percentile, mean, and 90th percentile of loan rates were 3%, 3.1%, and 3.3%, respectively, both in the data and in the calibrated model. The maximum loan rate observed in the sample was 3.45%, while the maximum rate a bank is willing to pay in the model is $\iota_w + \iota_\ell + \iota_s = 0.038$. Hence, the model delivers an excellent fit of the conditional distribution of loan rates above the DWR—an outcome that was not targeted in the calibration.

5.3 Bid-ask spreads

Each panel on the right side of Figure 4 shows an empirical cumulative distribution function of borrowed reserves over borrowing rates, denoted $\mathcal{H}_i^B$ (solid line), and an empirical cumulative distribution function of lent reserves over lending rates, denoted $\mathcal{H}_i^L$ (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 at interest rates lower than ι, and $\mathcal{H}_i^L(\iota)$ is the proportion of reserves lent by banks of type i at interest rates lower than ι.

Each panel on the left side of Figure 4 shows the theoretical counterpart of the adjoining right-side panel. The top-left and middle-left panels show that 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 those they earn when lending. This theoretical prediction also holds in the data, provided we focus on loans with rates not below the IOR (2.35%). In contrast, the bottom-left panel shows that the theory predicts $\mathcal{H}_S^B(\iota) \leq \mathcal{H}_S^L(\iota)$—that is, banks of type S tend to borrow at higher rates than those they earn when lending.[28] This theoretical prediction also holds in the data, and the fit is remarkably good for loans with rates not below the IOR.

5.4 Distributions of loan rates between pairs of bank types

Each panel on the right side of Figure 5 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 any interest rate ι on the horizontal axis, the height of the curve labeled "S" in the top-right panel

[27] This is possible because the calibration has $\iota_s > 0$.

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

represents the fraction of the total volume of loans extended from banks of type F to banks of type S with interest rates less than or equal to ι. Each panel on the left side of Figure 5 shows the theoretical counterpart 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—a prediction clearly validated by the data.

5.5 Interbank trading network

The model replicates several features of the fed-funds trading network, including the direction and volume of loans between and within bank types—represented by the direction of the arrows, their width, and the sizes of the nodes in the bottom panel of Figure 19 in Appendix A. However, the model predicts a significant volume of loans from GSEs to banks of type S that is not observed in the data, perhaps reflecting institutional details that lead GSEs to lend reserves only to a relatively small subset of counterparties.

6 Aggregate demand for reserves

In this section we show that our quantitative theory generates an *aggregate demand for reserves*, $\iota^* = \mathcal{D}(Q;\Pi)$—a negative relationship between the total quantity of reserves, Q, and the volume-weighted average of bilateral loan rates, ι^*, whose position and shape depend on the structural parameters $\Pi \equiv \{\beta_i, \lambda_i, \{\theta_{ij}, G_{ij}\}_{j\in\mathbb{N}}, u_i, U_i\}_{i\in\mathbb{N}}$.

Section 6.1 explains how we compute the demand for reserves in the model, and Section 6.2 illustrates how the structural parameters determine its position and shape. Section 6.3 then uses the calibrated model to produce a global estimate of the demand for reserves for the United States and compares it with the estimates implied by standard reduced-form econometric approaches. We find that seemingly reasonable econometric specifications fail to deliver robust results: they generate out-of-sample predictions inconsistent with elementary theory and estimates highly sensitive to specification details. Finally, Section 6.3.1 summarizes the general lessons for reserve-demand estimation that emerge from contrasting our quantitative-theoretic approach with atheoretical reduced-form approaches.

6.1 Reserve demand in the model and reserve supply in the data

We compute the demand mapping $\mathcal{D}$ using the following procedure. First, we calibrate the parameters Π to the year 2019 (as in Table 1). Second, we calculate the volume-weighted average

of equilibrium loan rates, ι^*, for a range of values of Q. Recall that $Q \equiv \sum_{i\in\mathbb{N}} n_i \int a\, dF_0^i(a)$, so to vary Q we must specify how to vary the beginning-of-day distribution of reserves, $\{F_0^i, n_i\}_{i\in\mathbb{N}}$. Our approach is to vary this distribution along a linear interpolation between two empirical beginning-of-day distributions estimated for two base years. Specifically, we start from two empirical distributions, $\{\bar{\boldsymbol{F}}_{\mathrm{Y}}, \bar{\boldsymbol{n}}_{\mathrm{Y}}\}_{\mathrm{Y}\in\{\mathrm{Y}_0,\mathrm{Y}_1\}}$, with $\bar{\boldsymbol{F}}_{\mathrm{Y}} \equiv \{\bar{F}_{\mathrm{Y}}^i\}_{i\in\mathbb{N}}$ and $\bar{\boldsymbol{n}}_{\mathrm{Y}} \equiv \{\bar{n}_{\mathrm{Y}}^i\}_{i\in\mathbb{N}}$, where $\bar{n}_{\mathrm{Y}}^i$ is the empirical estimate of the proportion of banks of type i in year Y, and $\bar{F}_{\mathrm{Y}}^i$ is the beginning-of-day distribution of reserve balances across banks of type i as estimated for year Y.[29]

We let $\{x_{\mathrm{Y}}^i(p_n)\}_{n=1}^N$ denote the set of N quantiles of the empirical distribution $\bar{F}_{\mathrm{Y}}^i$, where $x_{\mathrm{Y}}^i(p_n)$ is defined by $F_{\mathrm{Y}}^i\big(x_{\mathrm{Y}}^i(p_n)\big) = p_n$ for each $n \in \{1,\ldots,N\}$, and $\{p_n\}_{n=0}^{N+1}$ satisfies $p_{N+1} = 1 - p_0 = 1$ with $p_n < p_{n+1}$. We then use the two empirical distributions, $\{\bar{\boldsymbol{F}}_{\mathrm{Y}}, \bar{\boldsymbol{n}}_{\mathrm{Y}}\}_{\mathrm{Y}\in\{\mathrm{Y}_0,\mathrm{Y}_1\}}$, to generate a family of beginning-of-day distributions, $\{\bar{\boldsymbol{F}}_{\mathrm{Y}_\omega}, \bar{\boldsymbol{n}}_{\mathrm{Y}_\omega}\}_{\omega\in\mathbb{G}}$, where $\mathbb{G} \subset \mathbb{R}$ is a grid, $\bar{n}_{\mathrm{Y}_\omega}^i \equiv \omega\bar{n}_{\mathrm{Y}_1}^i + (1-\omega)\bar{n}_{\mathrm{Y}_0}^i$, and $\bar{F}_{\mathrm{Y}_\omega}^i$ is constructed by quantile interpolation, with quantiles $x_{\mathrm{Y}_\omega}^i(p_n) \equiv \omega x_{\mathrm{Y}_1}^i(p_n) + (1-\omega)x_{\mathrm{Y}_0}^i(p_n)$ for $n \in \{1,\ldots,N\}$.[30] For each element of $\{\bar{\boldsymbol{F}}_{\mathrm{Y}_\omega}, \bar{\boldsymbol{n}}_{\mathrm{Y}_\omega}\}_{\omega\in\mathbb{G}}$, we compute $Q_{\mathrm{Y}_\omega} \equiv \sum_{i\in\mathbb{N}} \bar{n}_{\mathrm{Y}_\omega}^i \int a\, d\bar{F}_{\mathrm{Y}_\omega}^i(a)$ and the corresponding equilibrium value-weighted interest rate $\iota_{\mathrm{Y}_\omega}^*$. This procedure delivers a collection of pairs $\{(Q_{\mathrm{Y}_\omega}, \iota_{\mathrm{Y}_\omega}^*)\}_{\omega\in\mathbb{G}}$ that define the mapping $\iota_{\mathrm{Y}_\omega}^* = \mathcal{D}(Q_{\mathrm{Y}_\omega}; \Pi)$—the aggregate demand for reserves implied by the theory.

The empirical measure of Q that we feed into the model to calculate the demand for reserves is a transformation of beginning-of-day reserves that we term *active excess reserves*—because it is net of predictable transfers, net of Regulation D and LCR requirements, and includes only banks that were *active* (i.e., borrowed or lent at least once) in the baseline calibration year. This is the measure of Q used on the primary horizontal axis of any figure that displays the demand for reserves implied by the theory. To make our results easier to interpret, we sometimes express the quantitative implications of the theory in terms of the better-known measure of *total reserves*, which differs from *active excess reserves* in that it is gross of reserve requirements and includes *all* institutions that hold reserve balances at the Federal Reserve Banks.[31] To this end, in Appendix C (Section C.5.5) we devise an empirical transformation to

[29] The estimation of the beginning-of-day distributions is described in Section 4 and Appendix A (Section A.3). We use $\mathrm{Y}_0 = 2017$ and $\mathrm{Y}_1 = 2019$ 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 in response to the COVID shock in 2020. As shown in Figure 20 (Appendix C), 2017 is the post-GFC-regulation year with the highest level of total reserves ($2,254.27 bn, roughly the pre-2020 historical peak), whereas 2019 has the lowest level of total reserves in the post-GFC-regulation era (roughly $1,568.26 bn).

[30] See Appendix A (Section A.6) for details on this interpolation procedure.

[31] Total reserves at weekly frequency are published in *Federal Reserve Balance Sheet: Factors Affecting Reserve Balances—H.4.1* (shown in Figure 20 in Appendix C), and are available at monthly frequency as "TOTRESNS"

translate *active excess reserves* into *total reserves*. To facilitate the translation between these units, we sometimes complement the primary horizontal axis of *active excess reserves* with a secondary horizontal axis (above the figure) of total reserves.

6.2 Reserve demand counterfactuals

In all panels of Figure 6, the curve labeled "Benchmark" represents the reserve demand $\iota^*_{\mathrm{Y}_\omega} = \mathcal{D}(Q_{\mathrm{Y}_\omega}; \Pi)$ for the model calibrated as in Table 1. We highlight two features of this demand for reserves implied by the theory. First, it exhibits the logistic sigmoid shape characteristic of the popular "Poole model."[32] Second, the demand lies within the DWR–IOR corridor. Thus, despite the presence of GSEs that earn a lower interest on reserves than banks (the ONRRP rather than the IOR), the average rate is above the IOR for all levels of reserves in the baseline calibration. This finding is consistent with the data: the OFR was consistently above the IOR during 2019, when the administered policy rates were as in our baseline calibration.[33] Our structural approach is well suited to analyze how policy or market-structure counterfactuals affect the position and shape of the reserve demand, which we turn to next.

The top-left panel of Figure 6 shows two experiments. In the first, the DWR is increased by 50 bps (so that it equals the ONRRP plus 125 bps, rather than the ONRRP plus 75 bps as in the baseline calibration). This shift moves the demand upward, with the magnitude of the shift decreasing in the quantity of reserves. Intuitively, the DWR has little effect on the equilibrium average interest rate when reserves are abundant, and a stronger effect when reserves are scarce. The second experiment increases the IOR by 15 bps (so that it equals the ONRRP plus 25 bps, rather than the ONRRP plus 10 bps as in the baseline calibration). This policy change raises the equilibrium average rate when reserves are relatively abundant, and it also implies that—if reserves are large enough—the equilibrium rate lies between the IOR and the ONRRP. This observation is consistent with the data: the OFR was consistently between the IOR and the ONRRP during most of the post-GFC period, from 2008 to 2018, when—as in this experiment—the IOR was set 25 bps above the ONRRP. The top-right panel of Figure 6 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.

at https://fred.stlouisfed.org. The average quantity of active excess reserves was about $1,150.86 bn in 2017 and $910.73 bn in 2019. The corresponding quantities of total reserves in 2017 and 2019 were $2,254.27 bn and $1,568.26 bn, respectively.

[32] See, e.g., p. 784 in Poole (1968).

[33] The same is true of the EFFR, as shown in Figure 20 in Appendix C.

The counterfactuals involving changes in administered policy rates deliver a valuable insight for empirical work: in a floor system, the demand for reserves rotates and shifts in response to changes in *any* of the administered policy rates—not only to changes in the IOR or in the DWR–IOR spread. We stress this point because, as discussed in Section 6.3 and Section 6.3.1, existing reduced-form econometric estimations of the demand for reserves in the United States have, to date, never controlled for the IOR–ONRRP spread.

The bottom-left panel of Figure 6 shows three experiments. The first multiplies the trading probabilities of all bank types by a factor of ten, making the market structure more competitive (i.e., "less OTC"), which reduces rates—and increases the slope of the demand—when the quantity of reserves is low to moderate. The second market-structure experiment sets $\beta_F = 0$, effectively excluding all banks of type F from trading and causing the demand for reserves to rotate clockwise around an intermediate quantity of aggregate reserves (about \$700 bn). This experiment raises the average rate for relatively low levels of reserves and lowers it for relatively high levels. The rotation reflects the intermediation role that type-F banks play in equilibrium: when reserves are scarce, many banks have deficient reserve balances and, absent type-F counterparties, find it more difficult to meet lenders, reducing their effective market power and leading to higher negotiated loan rates on average; conversely, when reserves are abundant, many banks hold excess reserves and, absent type-F counterparties, find it more difficult to meet borrowers, leading to lower average negotiated rates. The third experiment removes the proportional borrowing costs from the baseline calibration. This shifts the demand for reserves upward, reflecting that borrowing costs dampen banks' incentives to borrow.

The bottom-right panel of Figure 6 shows two experiments involving idiosyncratic payment risk. One experiment sets $\lambda_i = 0$ for all $i \in \mathbb{N}$, eliminating payment shocks for all banks. The other sets $\lambda_i = \lambda_F$ for all $i \in \mathbb{N}$, so that all bank types experience the same (very high) frequency of payment shocks as type F. The results of these experiments are a downward shift and an upward shift in the demand for reserves, respectively, driven by banks' precautionary motive for holding reserves.

6.3 Reserve demand estimation

As discussed in Section 1, the floor system that the Federal Reserve has adopted as its operating framework for monetary policy implementation relies on the ability to ascertain the level of reserves that is "ample enough" for active management of the supply of reserves to be

unnecessary in implementing the rate target. In other words, operating a floor system requires *global* knowledge of the aggregate demand for reserves, and in particular, reliable estimates of its slope over a wide range of reserve supplies. This poses 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 diagram in the first panel of Figure 1, the issue is identifying *exogenous* variation in the quantity of reserves that allows estimation of the slope of the demand curve. This problem is well understood and has been addressed by the empirical literature studying the "liquidity effect."[34]

The second challenge is to obtain *global* estimates of the reserve demand to identify its slope over a range of reserve supplies 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 complication is that spanning substantial variation in reserve supply requires covering a substantial period of time, during which the demand is likely to have shifted and rotated due to structural changes in regulation or market structure. Moreover, as shown in Section 6.2, even technical adjustments to the administered rates, such as the IOR–ONRRP spread, shift and rotate the demand for reserves, making it difficult to bridge the *local–global estimation gap* with empirical methods that are not guided by theory.

This global-demand identification problem has not been overcome by the empirical literature on "liquidity effects"—perhaps because of limited theoretical guidance on the structural variables that determine the shape and position of the demand for reserves. Empirical estimates of liquidity effects tend to be *local*—estimated from daily time-series variation in the quantity of reserves over short sample periods during which the average quantity of reserves remains relatively stable.[35] Our methodological innovation is to use the quantitative model to bridge

[34]We discuss these identification issues in Appendix A (Section A.5), where we also report 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).

[35]Hamilton (1997), Carpenter and Demiralp (2006), and our estimates of the "liquidity effect" in Appendix A (Section A.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 ten-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 to 2020). Their estimation, however, cannot recover the entire 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 show that the reduced-form estimates from Afonso et al. (2022) can be a useful guide once complemented with our quantitative theory, which helps identify the structural shifts in the demand for reserves.

this *local–global gap*. The idea is to exploit the structural demand relationship $\iota^* = \mathcal{D}(Q;\Pi)$ implied by the theory—with the microstructure and policy parameters, Π, calibrated to match key micro- and market-level data targets, including the econometric *local* estimate of the liquidity effect—to obtain a global estimate of the demand for reserves.[36]

Figure 7 contrasts our quantitative-theoretic identification approach (top panel) with two atheoretical reduced-form econometric specifications commonly used to estimate reserve demand (middle and bottom panels). In each panel, the vertical axis plots the OFR–IOR spread, and the main horizontal axis plots the quantity of *active excess reserves* (the secondary, top horizontal axis shows the corresponding quantity of *total reserves*). Each panel displays pairs of empirical observations of the quantity of active excess reserves and the corresponding OFR–IOR spread for every trading day in the sample period 2017/01/20–2019/09/13. Guided by the theoretical result in Section 6.2 that reserve demand shifts and rotates with changes in administered-rate spreads, the sample is divided into four subsamples defined by the size of the IOR–ONRRP spread: 10 bps (2019/05/02–2019/09/13, red dots), 15 bps (2018/12/20–2019/05/01, green dots), 20 bps (2018/06/14–2018/12/19, orange dots), and 25 bps (2017/01/20–2018/06/13, blue dots).[37]

The top panel of Figure 7 depicts four theoretical reserve-demand curves, $\iota^*_{Y_\omega} = \mathcal{D}(Q_{Y_\omega};\Pi)$. The solid (red) curve is generated by the baseline calibration (Table 1).[38] That this curve fits the (red) data points in the subsample with IOR–ONRRP = 10 bps is not surprising, since the baseline calibration targets the average OFR–IOR spread and the local slope of the empirical demand curve in this subsample. What is informative about our structural, quantitative-theoretic approach is that it allows us to trace the reserve demand over the full range of reserve levels, including regions beyond those observed in the narrow calibration subsample with IOR–ONRRP = 10 bps. The remaining three curves are counterfactual demands obtained by changing only the IOR–ONRRP spread—to 15 bps (green dashed curve), 20 bps (orange dash–dot curve), and 25 bps (blue long–short dashed curve). Despite the fact that no parameters other than the IOR–ONRRP spread were altered, these counterfactual demands fit the data points in the corresponding subsamples remarkably well—an out-of-sample validation of the quantitative predictions of the theory.

[36] 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.

[37] The DWR–ONRRP spread is constant at 75 bps throughout the sample period.

[38] For comparison with the standard econometric approaches reviewed below, the figure shows the model-implied demand expressed in terms of the spread between the equilibrium rate, $\iota^*_{Y_\omega}$, and the IOR, ι_r.

The standard econometric approach to reserve-demand estimation is to posit an ad hoc functional form, $s_t = D(Q_t)$, where s_t is an interest rate and Q_t a measure of the quantity of reserves. It is common practice to assume a generalized logistic function such as

$$D(Q_t) \equiv \underline{s} + \frac{\overline{s} - \underline{s}}{1 + e^{(Q_t - Q_0)\xi}}, \tag{6}$$

and to estimate the parameters $(\underline{s}, \overline{s}, \xi, Q_0)$ by nonlinear least squares (NLS).[39] The middle panel of Figure 7 displays the four fitted demand curves obtained from this procedure, with s_t measured by the OFR–IOR spread on day t and Q_t by the quantity of active excess reserves at the end of day t. The solid red line corresponds to the subsample with IOR–ONRRP = 10 bps; the dashed green line to 15 bps; the dash–dot orange line to 20 bps; and the long–short dashed blue line to 25 bps.

There are two takeaways from comparing the theoretical demands in the top panel of Figure 7 with the reduced-form logistic demands in the middle panel. First, the theoretical demands fit the data about as well as the logistic specifications. Second, their predictions diverge sharply at lower levels of Q. The theoretical demands predict that the OFR–IOR spread starts to increase when active excess reserves fall below about \$800 bn (or about \$1.3 tn of total reserves) and that it approaches the DWR–IOR spread as reserves approach zero. In contrast, three of the four reduced-form logistic demands predict that the OFR–IOR spread remains constant—indeed, close to zero—as the total quantity of reserves declines from \$1.3 tn to zero, which is implausible and inconsistent with elementary theory.

Another common econometric strategy is to replace (6) with the semi-log specification

$$D(Q_t) \equiv a + b \ln(Q_t), \tag{7}$$

and to estimate the demand equation $s_t = D(Q_t)$ using ordinary least squares (OLS).[40] The bottom panel of Figure 7 displays the four fitted demand curves obtained from this procedure, with s_t measured by the OFR–IOR spread on day t and Q_t by the quantity of active excess reserves at the end of day t. The solid red line corresponds to the subsample with IOR–ONRRP = 10 bps; the dashed green line to 15 bps; the dash–dot orange line to 20 bps; and the long–short dashed blue line to 25 bps.

There are two takeaways from comparing the theoretical demands in the top panel of Figure 7 with the reduced-form semi-log demands in the bottom panel. First, the theoretical demands

[39] See Afonso et al. (2022).

[40] See López-Salido and Vissing-Jorgensen (2023).

fit the data about as well as the semi-log specifications. Second, the global shape of the theoretical demands is quite different from the global shape of the semi-log demands. The theoretical demands predict that the OFR–IOR spread remains constant for levels of active excess reserves above about $800 bn (or about $1.3 tn of total reserves), that it starts to increase noticeably only when reserves fall below this level, and that it approaches the DWR–IOR spread when reserves approach zero. In contrast, the semi-log specifications predict that the OFR–IOR spread is positively sloped even when active excess reserves are in excess of $1.5 tn (or about $2.5 tn of total reserves), increases exponentially as total reserves decline, and eventually becomes implausibly large–much larger than the DWR–IOR spread—and diverging to infinity as reserves approach zero.

6.3.1 Main lessons

Three lessons emerge from comparing the results of our quantitative-theoretic reserve-demand estimation with the reduced-form econometric specifications (6) and (7). Appendix D (Section D.3) shows that these conclusions generalize to other reduced-form estimation strategies, including variants of the logistic NLS estimation proposed by Afonso et al. (2022) (Section D.3.1) as well as versions of the semi-log OLS estimation with the controls proposed by López-Salido and Vissing-Jorgensen (2023) (Section D.2).

First, our theory identifies a set of structural variables that shift and rotate the demand for reserves, and thereby help with the identification problems that pervade reduced-form econometric estimations. The theoretical counterfactuals in Section 6.2 show that this set includes: the spreads between administered policy rates, the parameters governing the trading frequencies and bargaining powers of the different types of market participants, the bank-level idiosyncratic payment-shock processes, and balance-sheet borrowing costs induced by regulation.

Second, while our quantitative-theoretic approach delivers estimates of reserve demands that fit available data as well as the reduced-form approaches, the two approaches generate vastly different out-of-sample predictions. For low levels of reserves, the most common reduced-form econometric specifications produce predictions that are implausible and inconsistent with elementary theory.

Third, the out-of-sample shape of the reserve demand implied by reduced-form econometric specifications is highly sensitive to their ad hoc functional-form assumptions. As shown in Fig-

ure 7, seemingly reasonable specifications yield markedly different out-of-sample predictions.[41] In this regard, the advantage of our quantitative-theoretic framework is that the global (out-of-sample) shape of the reserve demand is governed by "deep" microstructure parameters that can be disciplined with micro data.

These three lessons underscore the advantage of our quantitative-theoretic approach over reduced-form econometric ones, especially when global estimation involves large extrapolations.

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 interbank market.

7.1 Confidence bands for monetary policy implementation

The demand for reserves is determined by private banks' decisions. The supply of reserves is largely controlled by the Federal Reserve but is also affected by transactions in which the Fed is not a counterparty, such as those involving private-sector bank accounts and the account that the U.S. Treasury holds at the Fed. Whenever corporations or households pay taxes or purchase Treasury securities, reserves are transferred from private banks to the Treasury's account, which, from the perspective of domestic banks, constitutes an aggregate reserve-supply "shock." Conversely, reserve-augmenting shocks occur whenever the Treasury makes payments to the private sector. Appendix A (Section A.4) reports estimates of the daily distribution of supply shocks for the period January 2011–July 2019.

A central bank that wishes to operate a floor system must address an elementary question: What is the smallest quantity of reserves needed to ensure that the policy rate remains within its target range under plausible shocks to reserve supply? In this section, we use our quantitative-theoretic framework to answer this question in terms of a new policy-evaluation instrument: the *Monetary Confidence Band* (MCB).

Let $\iota = \mathcal{D}(Q)$ denote the aggregate-demand relationship between the equilibrium rate ι and the aggregate supply of reserves Q.[42] Let Z_p denote the p^{th} percentile of the empirical distribution of reserve-supply shocks estimated in Appendix A (Section A.4). We define the

[41] See Appendix D for a more detailed discussion of these issues.

[42] In this section, we omit the parameters Π as arguments of the demand function $\mathcal{D}$ to simplify the exposition.

"p% MCB" as a pair of functions $(\underline{\iota}(Q), \bar{\iota}(Q))$, where $\underline{\iota}(Q) \equiv \mathcal{D}\Big(Q + Z_{\frac{100+p}{2}}\Big)$ and $\bar{\iota}(Q) \equiv \mathcal{D}\Big(Q + Z_{\frac{100-p}{2}}\Big)$. The idea is that reserve-augmenting and reserve-draining shocks introduce randomness in the supply of reserves, which in turn induces randomness in the interbank rate. For a given beginning-of-day supply of reserves Q, the equilibrium rate lies within the 95% MCB, $(\mathcal{D}(Q + Z_{97.5}), \mathcal{D}(Q + Z_{2.5}))$, with probability 95%.

Figure 8 displays the 95% and 99% MCBs around the demand for reserves under the baseline calibration. There are two ways to use the MCBs. First, for a given beginning-of-day supply of reserves, they can be used to estimate the probability that the interbank rate will lie within a certain range. For example, during the sample period targeted by the baseline calibration, the typical beginning-of-day quantity of active excess reserves, Q, was about \$900 bn and the IOR was 235 bps. Under these conditions, the MCB implies that the central bank could implement any target rate in the range [IOR, IOR + 25 bps] with probability very close to one. Second, for a given target range for the interbank rate, the MCBs yield the minimum quantity of reserves needed to meet the target with a desired confidence level. For instance, to keep the interbank rate within the [IOR, IOR + 25 bps] range with 95% confidence, the central bank would have to supply at least about \$670 bn in beginning-of-day active excess reserves. A 99% confidence level would instead require about \$850 bn. In terms of total reserves, a 99% confidence level corresponds to about \$1.3 tn, or roughly 6.2% of GDP in the calibration year 2019.

7.2 Distribution of shadow price of reserves

When analyzing commercial banks' decisions to lend to households and non-financial corporations, an average interbank rate is typically viewed as a measure of the opportunity cost of loanable funds. The logic is that banks with excess reserves can lend in the interbank market rather than to a client, and any bank can borrow in the interbank market and lend elsewhere. Thus, under a competitive market structure, the cost of funds for *all* banks is summarized by a single interest rate. In contrast, in an over-the-counter market structure—where loans are negotiated bilaterally and sequentially over time, as in the actual interbank market—each bank faces different borrowing and lending rates depending on its own characteristics and those of its counterparties, such as their reserve balances at the time of trade, degree of market power (e.g., θ_{ij}), abilities to find counterparties (e.g., β_i), and regulatory treatment (e.g., the administered rates it earns on reserves or pays for overdrafts).

In a dynamic over-the-counter market structure, each bank of type $i \in \mathbb{N}$ with reserve bal-

ance $a \in \mathbb{R}$ at time t has its own opportunity cost—or "shadow price"—of reserves, summarized by $\frac{\partial V_t^i(a)}{\partial a}$. At any given time, the opportunity cost of loanable funds is described by a cross-bank distribution rather than by a single number, which may or may not be representative of most banks. As we show below, under the baseline calibration, neither the average interbank rate nor the distribution of traded rates is representative of the distribution of shadow prices of reserves for the majority of participants.

While outside the scope of our model, one can envision that banks make lending decisions to non-financial clients in a first stage, knowing that they will later participate in an interbank trading stage like the one modeled above. In this setup, the relevant opportunity cost of loanable funds in the first stage for a bank of type i is given by the shadow price $\mu_i(a) \equiv \frac{\partial V_0^i(a)}{\partial a} - 1$, where a denotes the bank's residual balance after making loans to non-financial clients in the first stage. We summarize this heterogeneity by the cumulative distribution function $\mathcal{M}_i(\iota) \equiv \int \mathbb{I}_{\{a:\mu_i(a)\leq\iota\}}\, dF_0^i(a)$; that is, $\mathcal{M}_i(\iota)$ represents the proportion of banks of type $i \in \mathbb{N}$ whose shadow price of reserves at the beginning of the trading day is below $\iota \in \mathbb{R}$.

The top-left panel of Figure 9 plots

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

together with the cumulative distribution function of *all* loan rates negotiated throughout the day, denoted $\mathcal{H}$ (under the baseline calibration). Intuitively, $\mathcal{H}(\iota)$ represents the proportion of reserves traded at rates below ι. The dashed vertical line marks the volume-weighted average rate on all trades implied by the theory, while the IOR and DWR are shown as solid vertical lines. Notice that $\mathcal{H}$ is highly concentrated around the average rate: about 60% of funds are traded at that rate. Thus, although there is heterogeneity in negotiated loan rates, the average is fairly representative of the overall distribution of traded rates. By contrast, neither the distribution of traded rates nor the average rate is representative of the distribution of shadow prices of reserves across all banks, represented by $\mathcal{M}$. For instance, about 80% of banks have a shadow price of reserves above the average rate, yet only about 10% of reserves are traded at rates above that level. This discrepancy arises because banks of type S, which account for more than 90% of the population, represent only a small share of total trades and are therefore underrepresented in transaction-based statistics such as the volume-weighted average rate and the distribution $\mathcal{H}$.

The remaining three panels of Figure 9 display, for each bank type i, the beginning-of-day

cumulative distribution of shadow prices, $\mathcal{M}_i$, and the cumulative distribution of all loan rates paid or received by that type, $\mathcal{H}_i$. These panels show that the average rate and the distribution of traded rates, $\mathcal{H}_i$, are fairly representative of the distribution of shadow prices of reserves, $\mathcal{M}_i$, only for types $i \in \{F, M\}$, but not for type S. This implies that, for roughly 90% of the banks participating in the interbank market, an average rate—such as the conventional EFFR or our OFR—does not adequately capture the shadow cost of procuring funding and is therefore not the relevant cost of lending in retail or corporate loan markets.

8 Conclusion

We have taken several steps toward developing structural over-the-counter models of the interbank market into serviceable tools for guiding monetary policy implementation. Our framework incorporates the key microstructure features of interbank markets, accounts for salient institutional details, and embeds the policy instruments and regulations that shape participants' demand for reserves. The model also captures the high degree of heterogeneity among participants across multiple dimensions, including market power in bilateral loans, the frequency and size distribution of payment shocks, and the degree of centrality in market making.

We documented a comprehensive set of novel market-wide and micro-level observations that characterize the dynamics of the interbank market and showed that our quantitative model is sufficiently flexible to replicate them. We then used the model to obtain structural estimates of the aggregate demand for reserves and developed two diagnostic tools to evaluate the central bank's ability to track its policy rate target and to measure the heterogeneous incidence of policy actions on banks' shadow cost of funding.

While we think we have made significant progress, we have also touched upon several issues that merit further exploration. We allowed for heterogeneity in contact rates across bank types to capture the core–periphery structure of the interbank market, but we treated these contact rates as exogenous. While this may be a reasonable assumption during periods when regulation and deeper market-structure parameters are relatively stable, it is easy to envision settings in which it would be desirable to endogenize search intensity. A similar observation applies to the beginning-of-day distributions of reserves, which in many applications would be best derived from an explicit portfolio problem faced by banks *prior* to the interbank trading stage that we have focused on.

Finally, a monetary-policy operating framework consists of two parts: an *operating target*

(e.g., an average interbank rate such as the EFFR or the SOFR) and a set of *policy instruments* (e.g., standing facilities or open-market operations). Monetary models in the macroeconomic tradition focus on the effects of choosing different values—or rules—for the operating target, leaving implementation issues outside the scope of analysis. In this paper, we have instead concentrated on the operational side of the monetary policymaking process, leaving macroeconomic considerations for future work. We think the macroeconomic implications of the microstructure of interbank lending and payments are a promising avenue for research.[43]

[43] See Bianchi and Bigio (2022) for a recent example of work along these lines.

References

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., 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 (2020): "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 (2014): "An Empirical Study of Trade Dynamics in the Fed Funds Market," Working Paper 708, Federal Reserve Bank of Minneapolis.

——— (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.

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.

Armantier, O., J. McAndrews, and J. Arnold (2008): "Changes in the Timing Distribution of Fedwire Funds Transfers," *Economic Policy Review*, 14.

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.

Bartolini, L., S. Hilton, and J. J. McAndrews (2010): "Settlement Delays in the Money Market," *Journal of Banking & Finance*, 34, 934–945.

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.

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.

Demiralp, S., B. Preslopsky, and W. Whitesell (2006): "Overnight Interbank Loan Markets," *Journal of Economics and Business*, 58, 67–83.

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

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. (2002): "The Interbank Market During a Crisis," *European Economic Review*, 46, 809–820.

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

——— (2001): "Banks as Monitors of Other Banks: Evidence from the Overnight Federal Funds Market," *The Journal of Business*, 74, 33–57.

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.

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.

Kovner, A. and D. Skeie (2013): "Evaluating the Quality of Fed Funds Lending Estimates Produced from Fedwire Payments Data," Staff Report 629, Federal Reserve Bank of New York.

Kuo, D., D. Skeie, J. Vickery, and T. Youle (2013): "Identifying Term Interbank Loans from Fedwire Payments Data," Staff Report 603, Federal Reserve Bank of New York.

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.

Logan, L. (2025): "The Case for Modernizing the FOMC's Operating Target Rate," Speech at the Federal Reserve Bank of Richmond CORE Week Workshop: *The Fed's Balance Sheet*.

López-Salido, D. and A. Vissing-Jorgensen (2023): "Reserve Demand, Interest Rate Control, and Quantitative Tightening," Manuscript, Federal Reserve Board.

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.

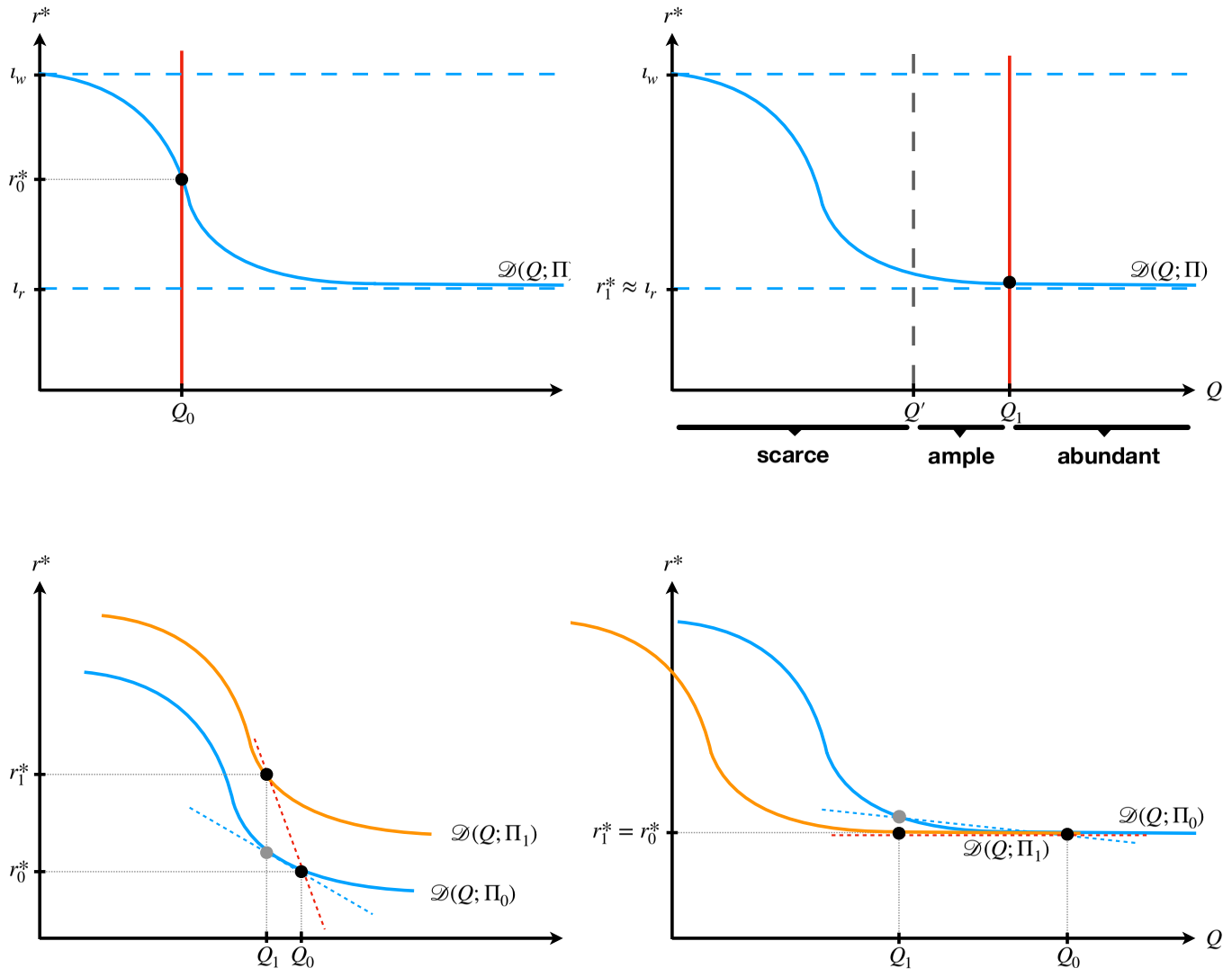

Figure 1: Stylized model of the determination of the interbank rate.

Notes: In every panel, Q denotes the aggregate quantity of reserves, a vertical line represents the actual quantity of reserves supplied by the central bank, r^* denotes the interbank rate, and Π is 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.

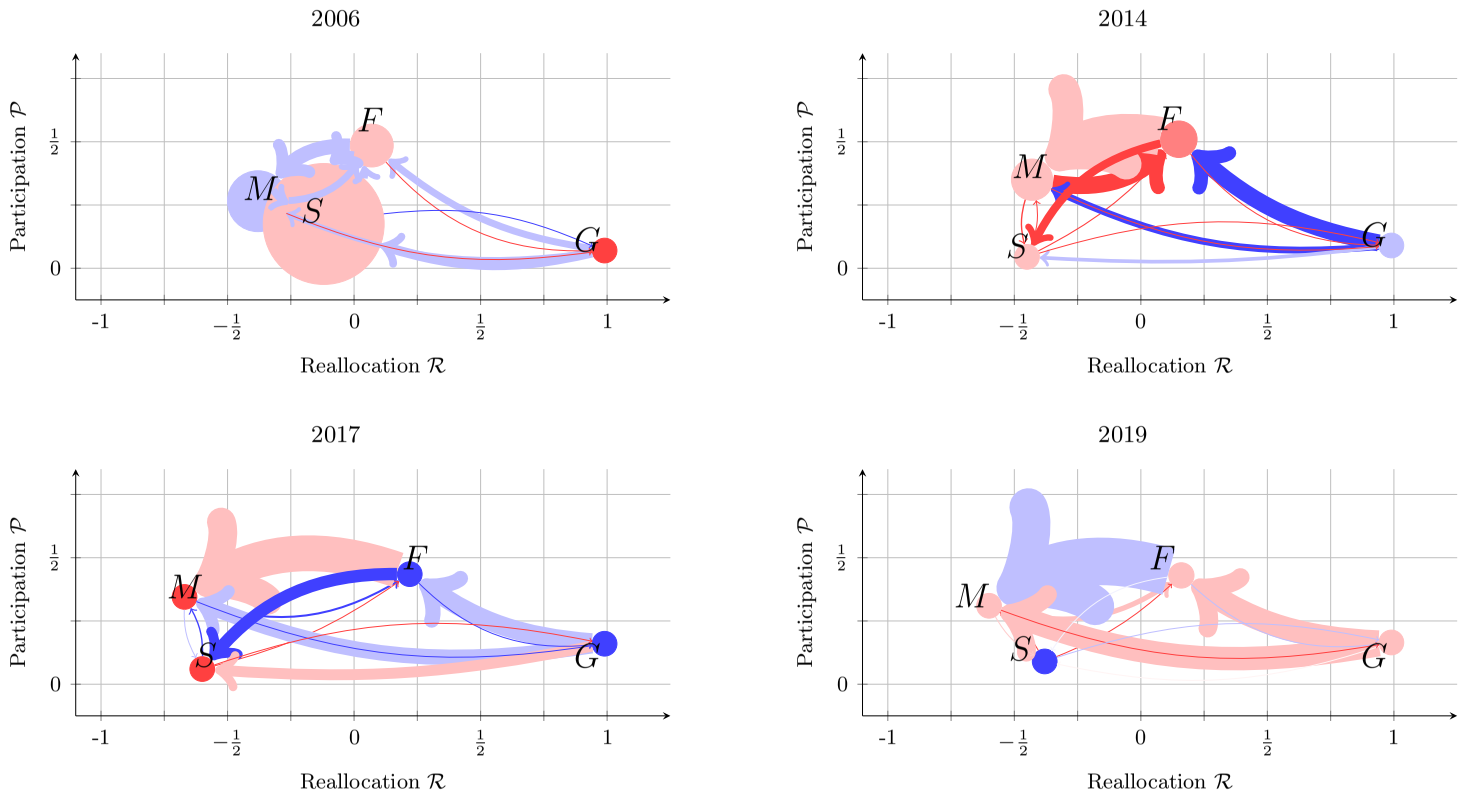

Figure 2: Interbank trading networks.

Notes: Each node corresponds to one of the four bank types, and labeled accordingly as F, M, S, or G. 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 *Overnight Fedwire Rate* (OFR), falls in the first, second, third, or fourth quartile, respectively.

Parameter	Target	Moment	
		Data	Model
$n_F = 0.010$	proportion of financial institutions of type F	0.010	0.010
$n_M = 0.044$	proportion of financial institutions of type M	0.044	0.044
$n_S = 0.920$	proportion of financial institutions of type S	0.920	0.920
$n_G = 0.026$	proportion of financial institutions of type G	0.026	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
$\lambda_G = 0$	bank-level share of unexpected payments per second for type G	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
$\iota_\ell = 0.00022/360$	average value-weighted rate	0.0237/360	0.0237/360
$\iota_s = 0.00752/360$	estimated liquidity effect for 2019 (bps per \$1 bn decrease in reserves)	$\in [-0.0145, -0.0035]$	-0.0073
$\underline{\theta} = 0.05$	conditional (below the IOR) average value-weighted overnight rate	0.0229/360	0.0228/360
$\beta_F = 0.0300$	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., $\mathcal{P}_M$)	0.31	0.27
$\beta_S = 0.0007$	participation rate of financial institutions of type S (i.e., $\mathcal{P}_S$)	0.09	0.08
$\beta_G = 0.0036$	participation rate of financial institutions of type G (i.e., $\mathcal{P}_G$)	0.17	0.14
$\kappa_F = 0.039e\text{-}3$	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
$\kappa_G = 1.25e\text{-}3$	reallocation index of financial institutions of type G (i.e., $\mathcal{R}_G$)	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, $r = 0$, $\mathbb{N} = \{F, M, S, G\}$ (as discussed in Section A.1), $\theta_{i,j} = 1/2$ for all $i, j \in \mathbb{N} \setminus \{G\}$, $\theta_{ij} = \underline{\theta}$ if $i \in \{G\}$ and $j \in \mathbb{N} \setminus \{G\}$, $\{G_{ij}\}_{i,j\in\mathbb{N}}$ are estimated as described in Section A.2, $\{F_0^i\}_{i\in\mathbb{N}}$ are estimated as described in Section A.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 A.6.

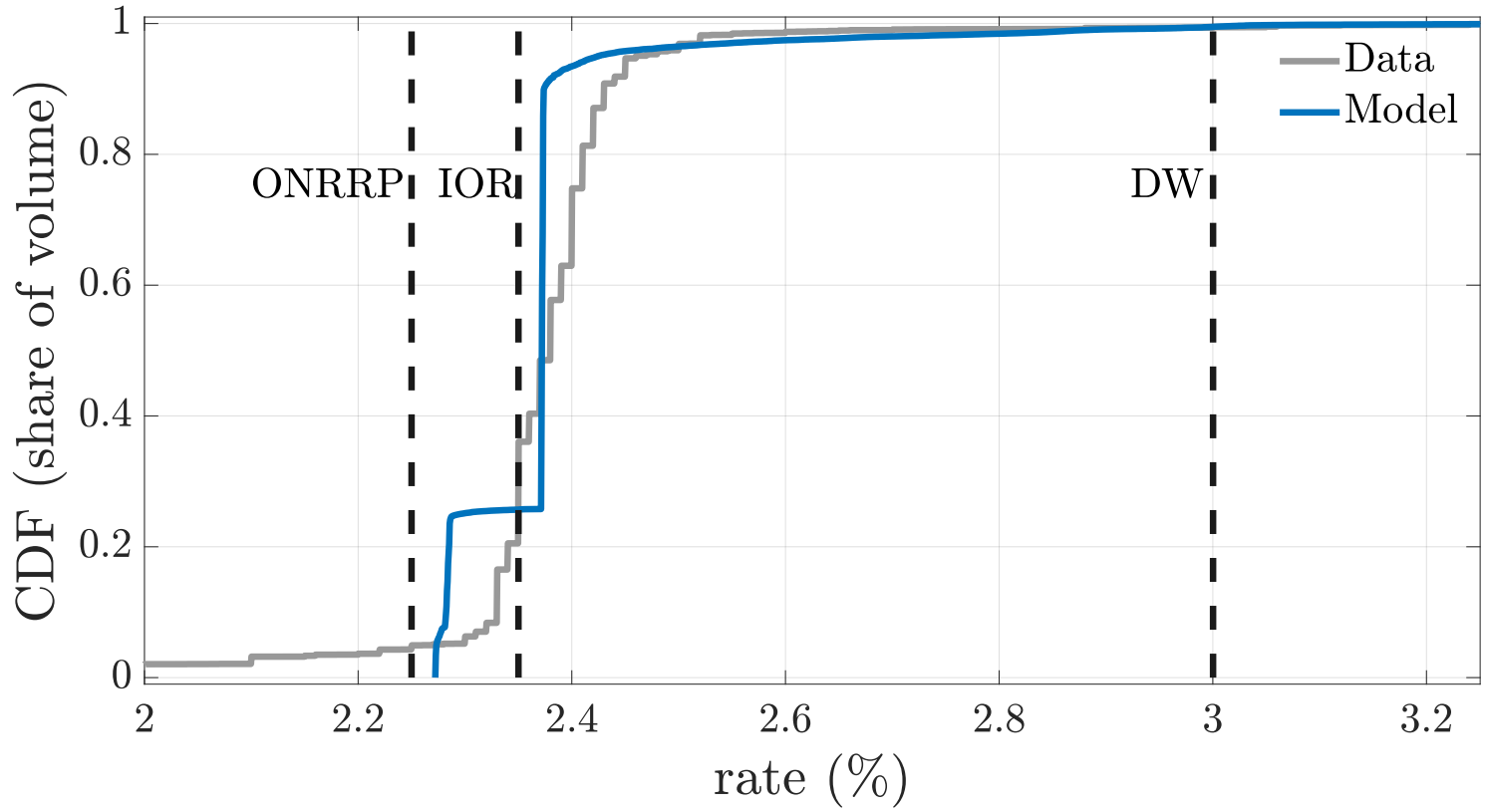

Figure 3: Empirical and theoretical cumulative distribution functions of bilateral rates for the year 2019.

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.

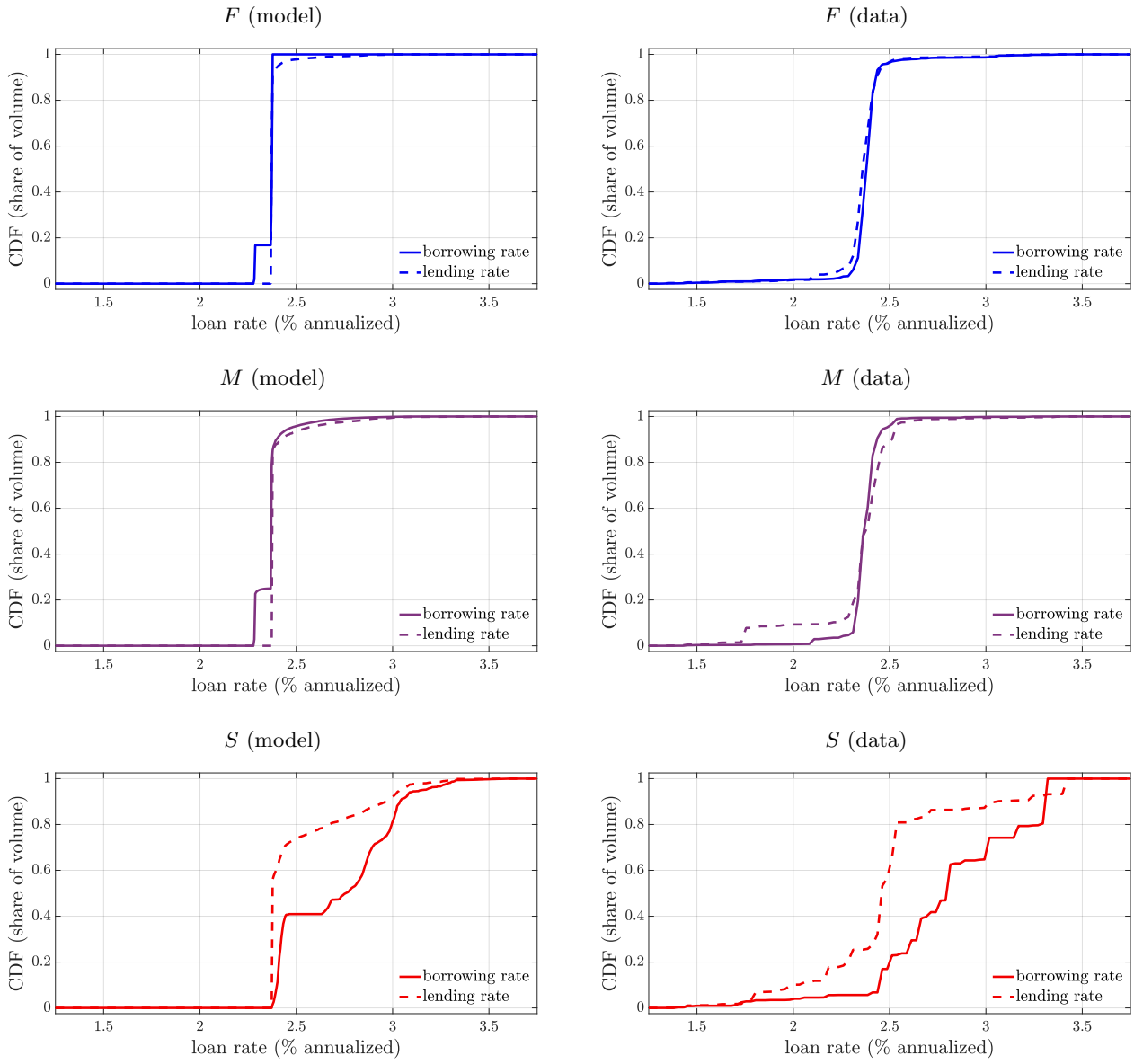

Figure 4: Cumulative distributions of borrowing and lending rates by bank type.

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.

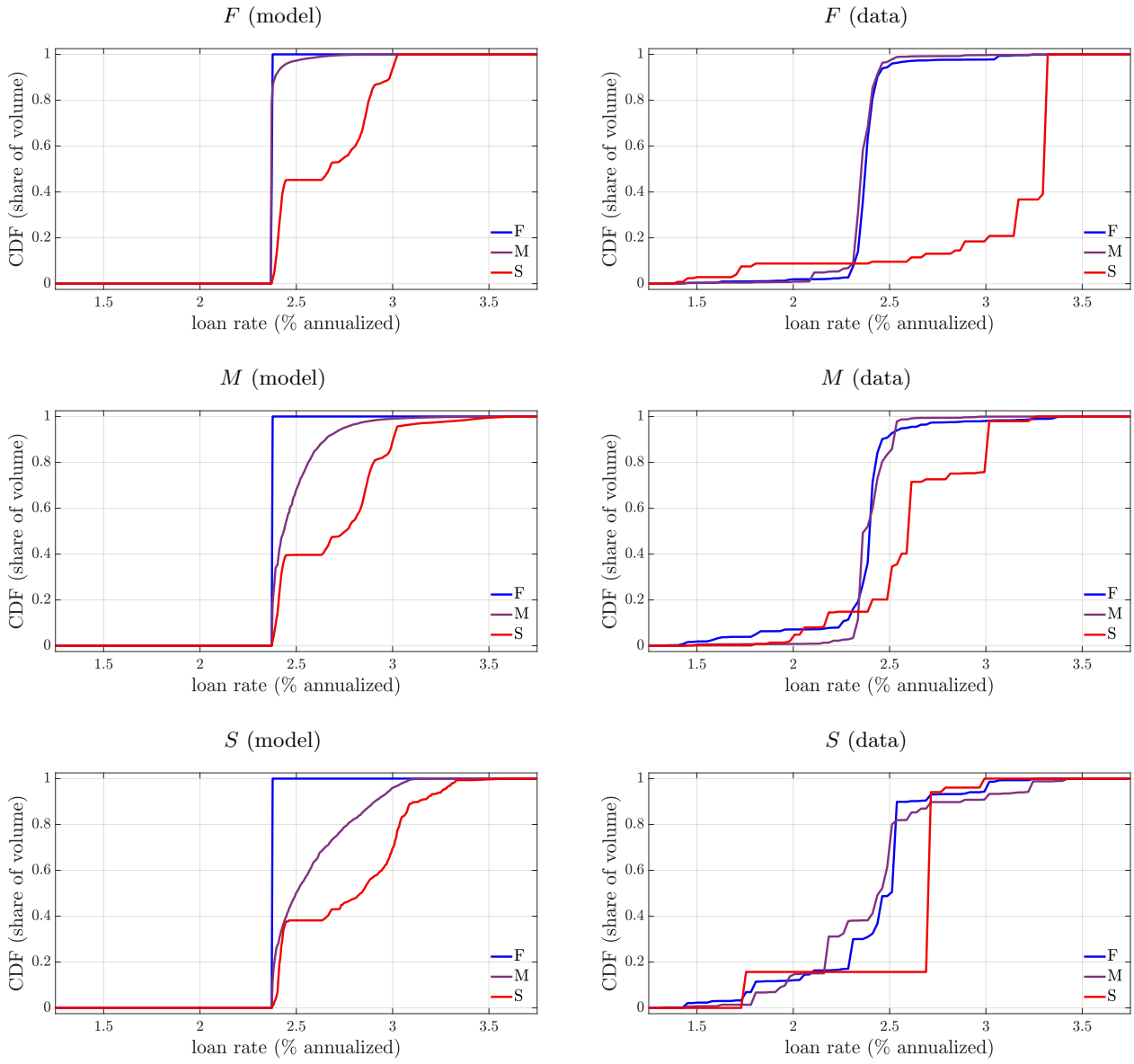

Figure 5: Cumulative distributions of loan rates between pairs of bank types.

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.

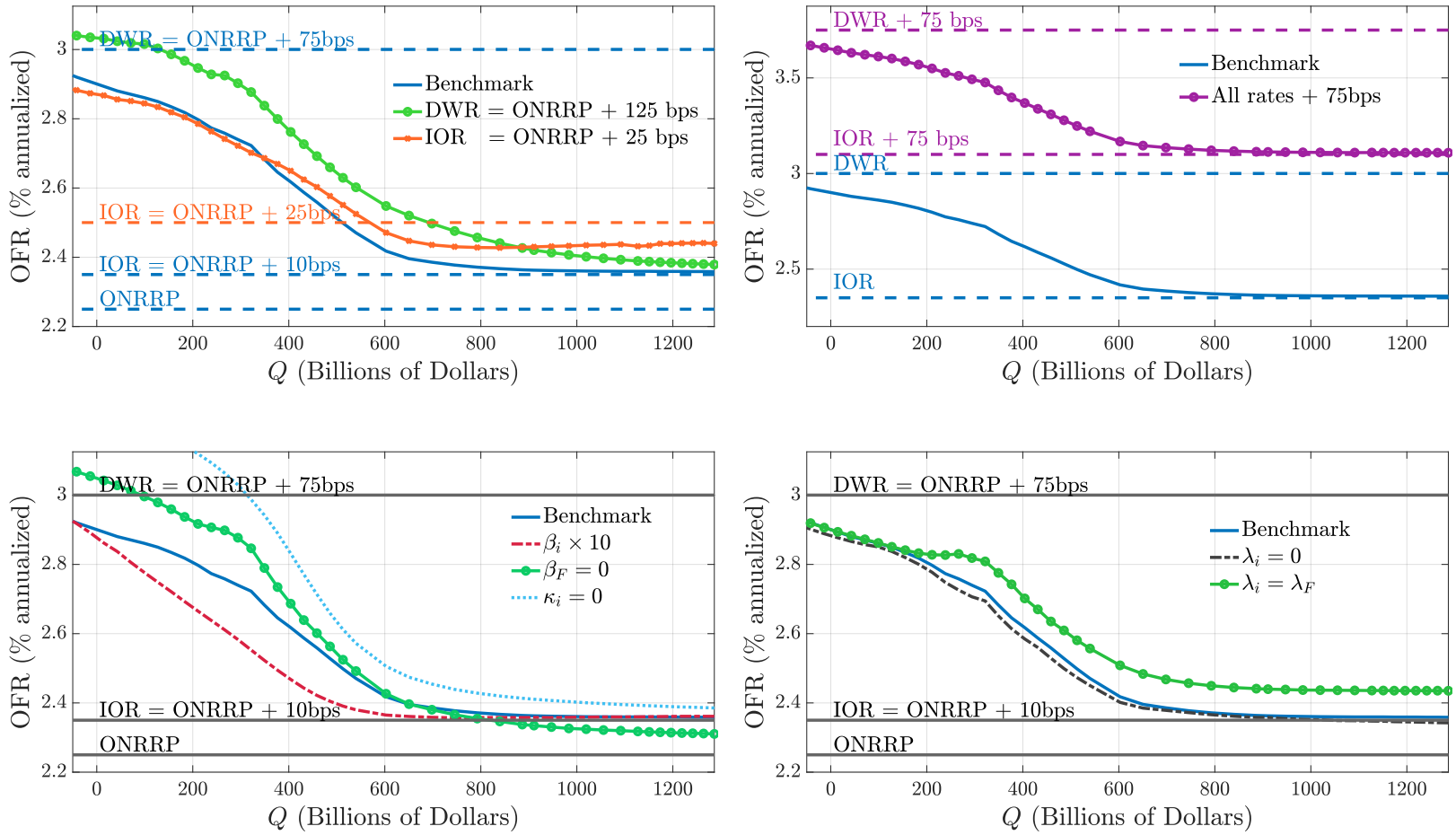

Figure 6: Theoretical aggregate demand for reserves: shifts and rotations.

Notes: In all panels, 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 Appendix A (Section A.6), for $\mathrm{Y}_0 = 2017$ and $\mathrm{Y}_1 = 2019$. 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 $i \in \mathbb{N}$.

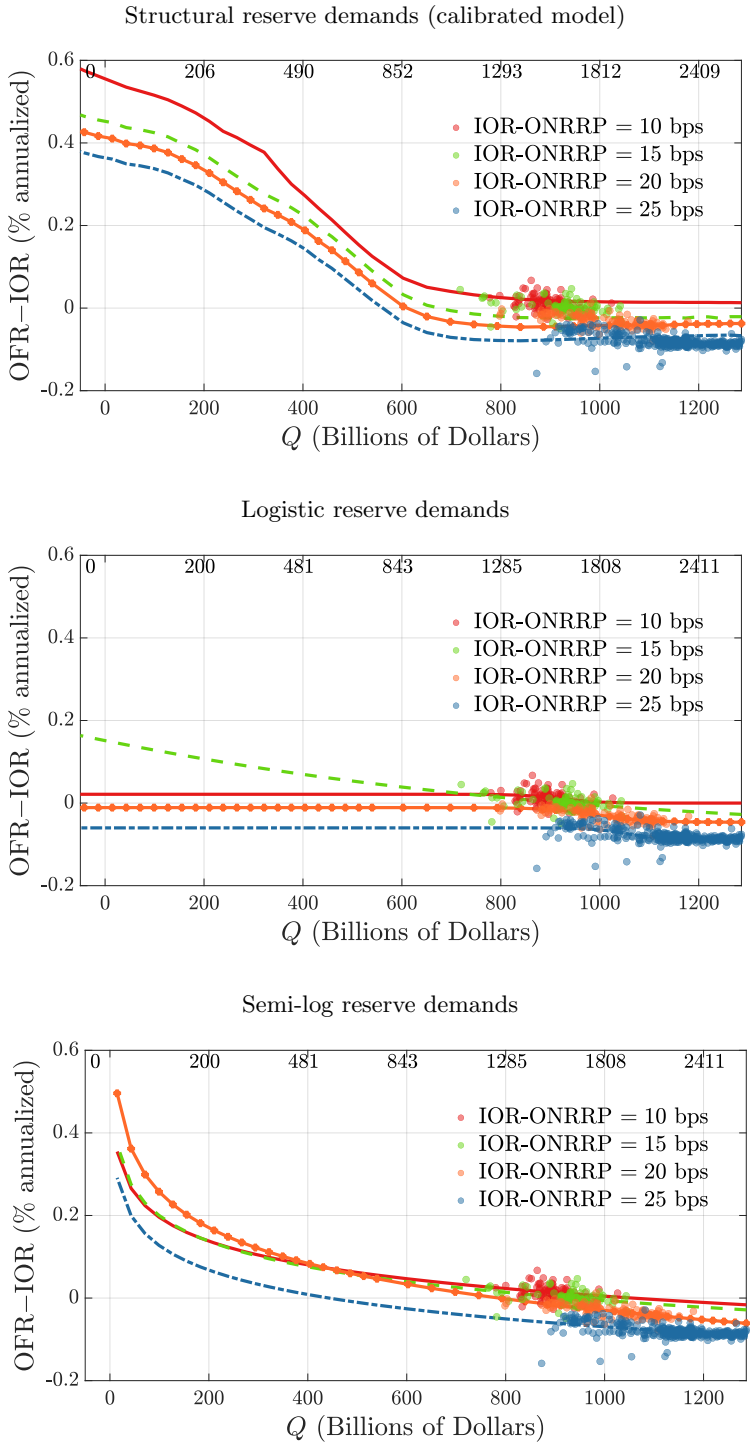

Figure 7: Reserve-demand estimates: calibrated model, logistic, and semi-log.

Notes: In each panel: vertical axis shows the OFR–IOR spread; horizontal axis shows *active excess reserves*, as defined in Section 6; the secondary (top) horizontal axis shows *total reserves*. Sample includes all trading days from 2017/01/20 to 2019/09/13. Top panel: reserve demands from the calibrated model for each subsample defined by the IOR–ONRRP spread. Middle panel: reserve demands obtained from NLS estimation of (6) for each subsample defined by the IOR–ONRRP spread. Bottom panel: reserve demands obtained from OLS estimation of (7) for each subsample defined by the IOR–ONRRP spread.

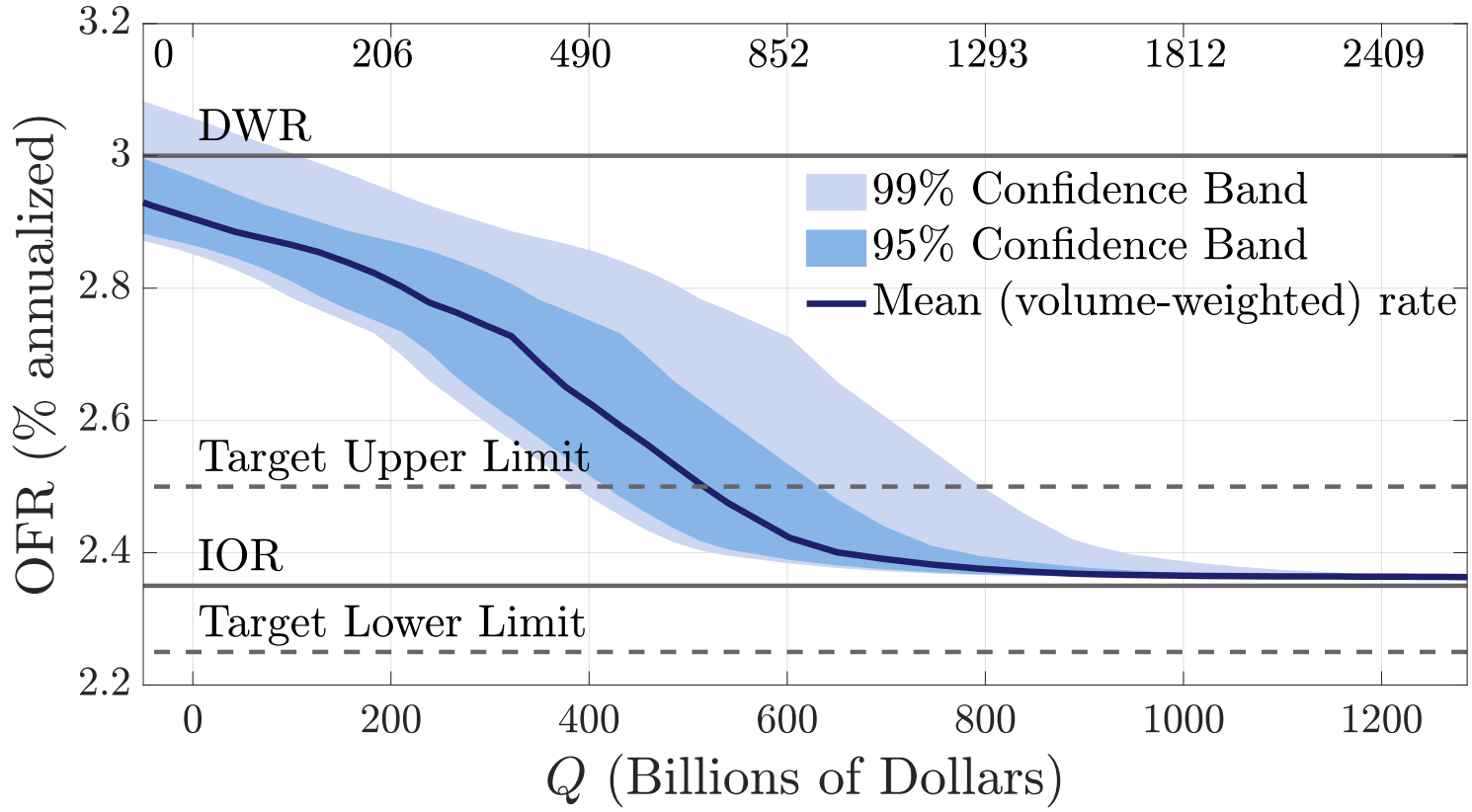

Figure 8: Monetary confidence bands.

Notes: The curve labeled "Mean (volume-weighted) rate" is the theoretical money demand for the baseline calibration, $\mathcal{D}(Q)$. 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-supply 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.

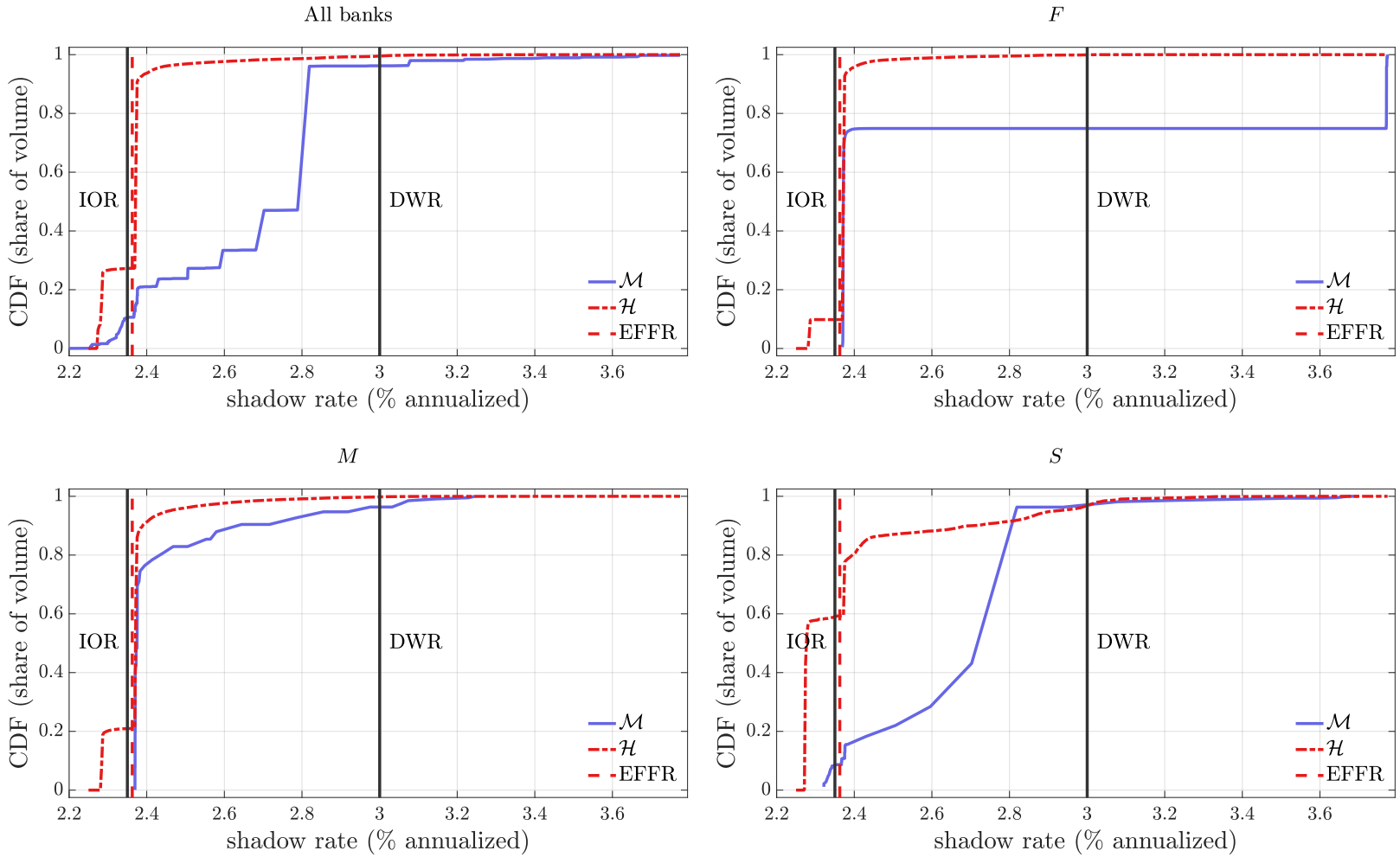

Figure 9: Cross-bank distributions of the shadow price of reserves.

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 denotes the volume-weighted average 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.

Monetary Policy Operations: Theory, Evidence, and Tools for Quantitative Analysis

Supplementary Appendix

Ricardo Lagos
New York University

Gastón Navarro
Federal Reserve Board

A Facts

In this section we document the facts that will guide the quantitative implementation of the theory. Section A.1 presents the joint distribution of two bank-level measures of trading activity: a bank's *participation rate* in market-wide trade volume, and a *reallocation index* that quantifies the degree to which a bank is a net borrower or lender of funds. Section A.2 reports estimates of the frequency and size distribution of micro-level intraday payments between banks. Section A.3 presents estimates of a typical beginning-of-day cross-sectional distribution of reserve balances. Section A.4 reports estimates of the distribution of aggregate daily reserve-supply shocks since the GFC of 2007-2008. Section A.5 presents empirical estimates of the slope of the aggregate demand for reserve balances. Finally, Section A.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 interbank market, we report facts separately for the period before, and after these regulations had been implemented.[44] 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.[45]

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,

[44]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 interbank borrowing). 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.

[45] The LCR was phased in between January 2015 and January 2017. Non-foreign bank organizations began reporting SLR to U.S. regulators in July of 2013, SLR disclosures become mandatory in January of 2015, and SLR compliance became mandatory in January of 2018. We regard 2006 and 2014 as pre-GFC-regulation years, and the 2017 and 2019 as post-GFC-regulation years. In terms of 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.

and 412 for the year 2019.[46] 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).[47] We focus on transactions that occur between 9:00 am and 6:30 pm EST.

A.1 Interbank 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.[48] Let υ^e_{nd} be the dollar value of all loans extended by bank $n \in \mathbb{B}$ in period $d \in \mathbb{Y}$, and use $\upsilon_d \equiv \sum_{n\in\mathbb{B}} \upsilon^e_{nd}$ to denote the dollar value of all the loans traded in period d. Also, let υ^r_{nd} 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

$$\begin{aligned} \mathcal{P}_{nd} &\equiv \frac{\upsilon^e_{nd} + \upsilon^r_{nd}}{2\upsilon_d} \\ \mathcal{R}_{nd} &\equiv \frac{\upsilon^e_{nd} - \upsilon^r_{nd}}{\upsilon^e_{nd} + \upsilon^r_{nd}}. \end{aligned}$$

We refer to $\mathcal{P}_{nd}$ as bank n's *participation rate* during period d, since it measures the share of the total period-d trade volume that is accounted for by bank n's trading activity. For any given bank n in period d, $\mathcal{P}_{nd} \in [0, 1/2]$, with $\mathcal{P}_{nd} = 0$ corresponding to a bank that did not trade, and $\mathcal{P}_{nd} = 1/2$ 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 period d, then $2\mathcal{P}_{nd} = x/100$. 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,

[46] In Appendix C (Section C.3) we explain our sample selection criteria, and how we assigned Fedwire transactions to bank holding companies.

[47] 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.

[48] 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.

or $\mathcal{R}_{nd} \in (0,1)$, meaning it is a net lender that engaged in some intermediation.[49] 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 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.[50]

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.[51] Figure 10 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). 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 0.5%, as M; and all other banks, as S. 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: 45.6% in 2006, and 43.1% in 2019. 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 interbank market structure.

Among the institutions assigned to group S based on the ECDF there is a subgroup of non-bank Fedwire participants 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

[49] 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 $(\upsilon^e_{nd} + \upsilon^r_{nd})\,\mathcal{X}_{nd}$ 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).

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

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

their payoffs from holding reserves different from the rest of the participating institutions.[52] To offer a parsimonious representation of the data, we will sort institutions into four *types*, i.e., $\mathbb{N} = \{F, M, S, G\}$. Types F, M, and S correspond to the F, M, and S groups defined above, but excluding GSEs, and type G is composed exclusively of GSEs.[53]

Figure 2 shows the location of each bank type $i \in \{F, M, S, G\}$ 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 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.[54] Each node represents the set of banks assigned to a particular type, labeled accordingly as F, M, S, or G. 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 interbank 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 OFR, falls in the first, second, third, or fourth quartile, respectively.[55]

While specifics vary somewhat across years, several stable trading patterns emerge from

[52] In contrast to banks, GSEs have very predictable cash flows (so payment shocks are not relevant for their day-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.

[53] 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.

[54] The participation rate for each bank type $i \in \{F, M, S, G\}$ 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 summed 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 C (Section C.5.2) for details.

[55] Arrow widths and node sizes are defined relative to trades within a year; thus not comparable across years.

Figure 2. Banks of type F account for about 1/2 of aggregate trade volume (i.e., $\mathcal{P}_F \approx 1/2$) 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.

A.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 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.[56] 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 $d \in \mathbb{D}$.[57] 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

[56] 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.

[57] 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 dataset

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

where $\mathbb{B}_i$ is the set of banks of type $i \in \mathbb{N}$. We then use the dataset $\tilde{\mathbb{S}}^{ij}$ 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.[58] Figures 11 and 12 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.[59]

A.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 loan 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

[58] See Appendix C (Section C.5.3) estimation details.

[59] We set $\lambda_G = 0$ for every year, for the reasons explained in footnote 56.

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 loan repayments received during day d, and subtracts all loan repayments sent during day d that correspond to loans originated during day $d-1$.[60] 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 $\mathrm{a}_m(h)$.[61] 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).[62] 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.[63]

[60] Repayments are identified using the send-receive matching from the Furfine algorithm. The rationale for netting the *predictable transfers*, which include the *repayments of loans* 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 loans traded during day $d-1$ (between the GSE and any other bank that meets the sample selection criteria described in Section C.3 of Appendix C), 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 C.3 of Appendix C). 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 cash flows, 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.

[61] 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.

[62] 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.

[63] 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

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*.[64]

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.[65]

Figures 13-16 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

these quantities are publicly available for each bank at a quarterly frequency (see Section C.2 in Appendix C 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 $m \in \mathbb{B}_G$.

[64] Unless otherwise specified, whenever we refer to "beginning-of-day reserves" we are alluding to *unencumbered reserves*, i.e., the reserves in excess of Regulation D and LCR requirements, and net of predictable Fedwire transfers, computed as described in Appendix C (Section C.5.1).

[65] See Appendix C (Section C.5.3) estimation details.

no bank-period 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.

A.4 Reserve-supply 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).[66] 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 interbank 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 rate 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., implementa-

[66] 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 vice versa), 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.

tion of quantitative easing policies) rather than by day-to-day managing of the interbank 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 17 shows that the variation in total reserves has been much larger since 2008, which is in line with 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-supply 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, $\overline{\mathrm{A}}_d \equiv \frac{1}{41}\sum_{k=-20}^{20} \mathrm{A}_{d+k}$.[67] The top panel of Figure 17 shows the time series $\{\mathrm{A}_d, \overline{\mathrm{A}}_d\}$ between the years 2001 and 2019. The middle panel of Figure 17 shows the deviations between total reserves and its own moving average, i.e., $\{\mathrm{z}_d\}$, with $\mathrm{z}_d \equiv \mathrm{A}_d - \overline{\mathrm{A}}_d$.[68] In time periods when the Federal Reserve does not react systematically to exogenous shocks to the supply 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.[69] The bottom panel of Figure 17 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., $[-\$279 \text{ bn}, \$130 \text{ bn}]$, and $[-\$115 \text{ bn}, \$99 \text{ bn}]$, 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%.

[67] For the purpose of these calculations we include *all* banks, not only those that meet the sample selection criteria based on trading activity described in Appendix C.3.

[68] Since the daily time series cannot be made public, the top and middle panels of Figure 17 show the *weekly* versions. But we use the daily time series for the purposes of the kernel estimation discussed below.

[69] See Appendix C (Section C.5.3) estimation details.

A.5 Liquidity effect

In this section we present empirical estimates of the change in the interbank overnight rate in response to an exogenous change in the aggregate quantity of reserves—the so-called *liquidity effect*.[70] It is customary to think of the interbank 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. Traditionally the main challenge for estimation has been to identify *exogenous* shifts in the supply of reserves.[71]

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.[72] 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").[73]

To estimate the liquidity effect for the post-GFC sample period with large excess reserves

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

[71] This was the main estimation challenge for the pre-GFC regime in which the Desk was actively conducting open-market operations reacting to market conditions in order to manage the interbank rate. The challenges are different for the post-GFC era (e.g., until mid September 2019), when reserves were not actively managed by the Desk. For example, within the post-GFC period when the Federal Reserve began managing the interbank rate by setting administered rates rather than the quantity of reserves, our theory prescribes controlling for the spreads between the administered rates (see Section 6.2).

[72] 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 reserves (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.

[73] 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.3) 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.

that were not actively managed by the central bank, we run the following regression:

$$s_t - s_{t-1} = \gamma_0 + \gamma(Q_t - Q_{t-1}) + \varepsilon_t, \tag{8}$$

where s_t denotes the spread between an interbank overnight rate 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 consider two measures of the overnight rate: the EFFR, and the OFR that we compute from our FWOL database. The former allows us to compare our results with existing estimates. The latter produces the estimate of the liquidity effect that we use to calibrate the model.

We estimate regression (8) 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 A.4, the Federal Reserve was not actively managing the quantity of reserves in response to developments in overnight money markets during the post-GFC sample periods that we consider for this regression.[74]

When we use the OFR as dependent variable the estimate is $\gamma = -0.009$ (significant at the 1% level), with 95% confidence interval $[-0.0145, -0.0035]$. 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 OFR–IOR spread by 0.009 basis points.

When we use the EFFR as dependent variable 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

[74]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 EFFR 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 EFFR. 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.

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).[75] This is important because, as we show in Section 6.2, 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 (8) 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 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.[76]

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

[75] The time series of the administered rates (DWR, IOR, ONRRP) are displayed in Figure 20 (Appendix C).

[76] Since in the quantitative implementation of the theory we focus on a subset of participants (see Section C.3 in Appendix C for our sample selection criteria), we have also run a version of (8) 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.

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

A.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 heterogeneity 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.[77] The procedure is as follows.

[77] Empirical studies (e.g., those that estimate the liquidity effect discussed in Appendix A.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 interbank market that allow for heterogeneity in 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.

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 Appendix A.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 $n \in \{1, ..., N\}$.[78] 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*,

$$x^i_{\mathrm{Y}_\omega}(p_n) \equiv \omega x^i_{\mathrm{Y}_1}(p_n) + (1-\omega)\, x^i_{\mathrm{Y}_0}(p_n). \tag{9}$$

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 $i \in \mathbb{N}$:

$$\bar{n}^i_{\mathrm{Y}_\omega} \equiv \omega \bar{n}^i_{\mathrm{Y}_1} + (1-\omega)\, \bar{n}^i_{\mathrm{Y}_0} \tag{10}$$

$$\bar{F}^i_{\mathrm{Y}_\omega}(a) \equiv \sum_{n\in\{\{1,...,N\}:x_{\mathrm{Y}_\omega}(p_n)\le a\}} (p_n - p_{n-1}), \tag{11}$$

so the corresponding aggregate quantity of reserves is

$$Q_{\mathrm{Y}_\omega} \equiv \sum_{i\in\mathbb{N}} \bar{n}^i_{\mathrm{Y}_\omega} \int a d\bar{F}^i_{\mathrm{Y}_\omega}(a). \tag{12}$$

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 (12) 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 (12), 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 (10) and (11). 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.[79]

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

[79] 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 interbank 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).

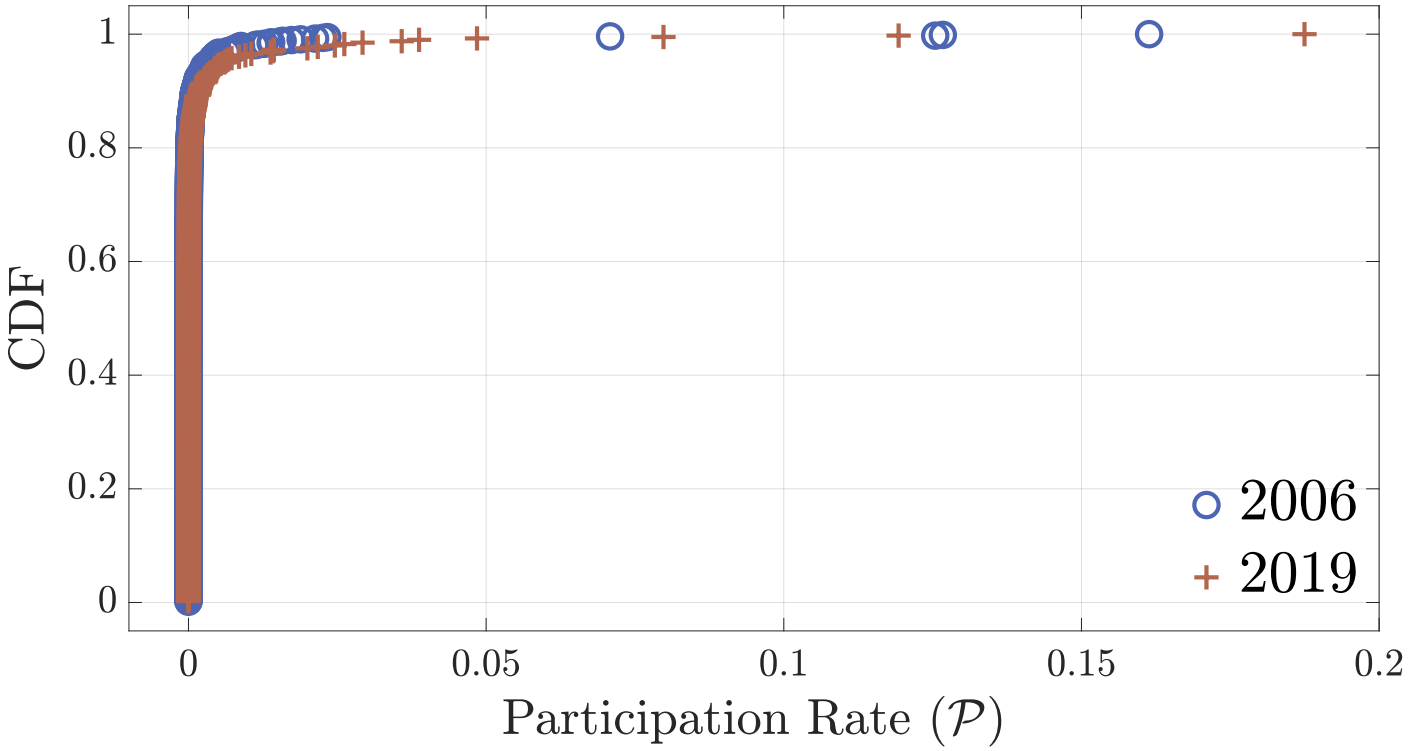

Figure 10: Empirical cumulative distribution function of bank-level participation rates for 2006 and 2019.

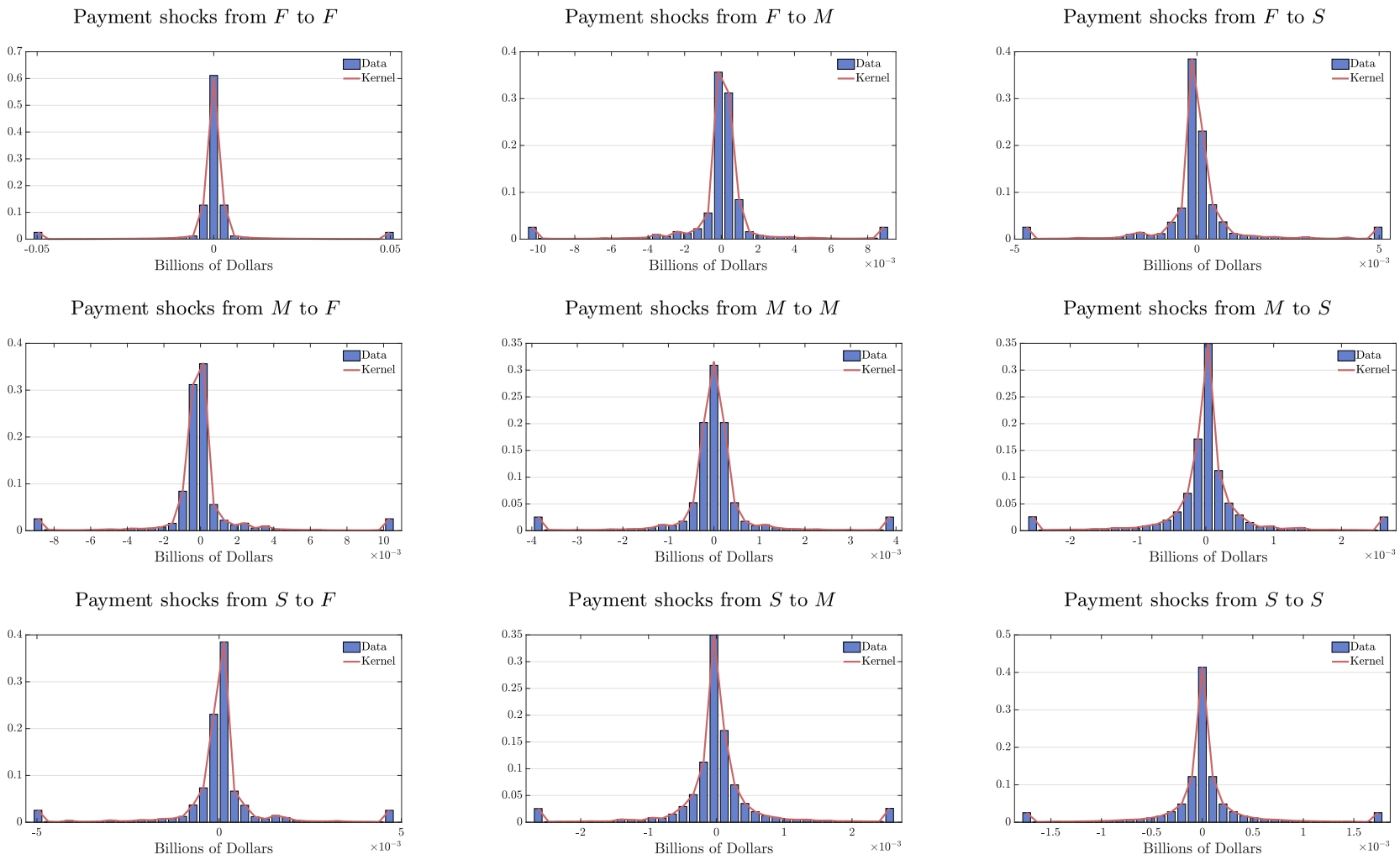

Figure 11: High-frequency payment shocks between pairs of bank types in 2006.

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).

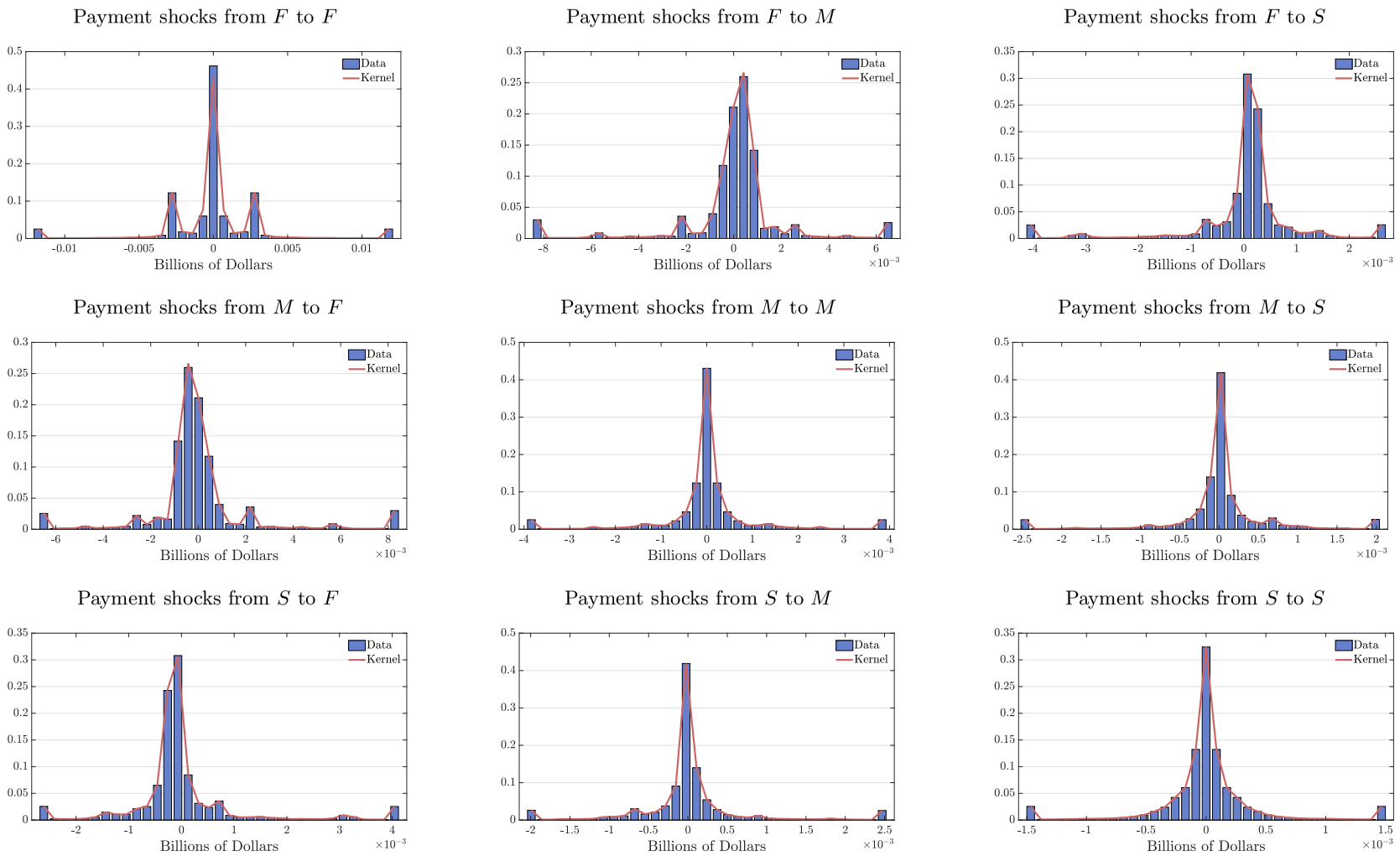

Figure 12: High-frequency payment shocks between pairs of bank types in 2019.

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).

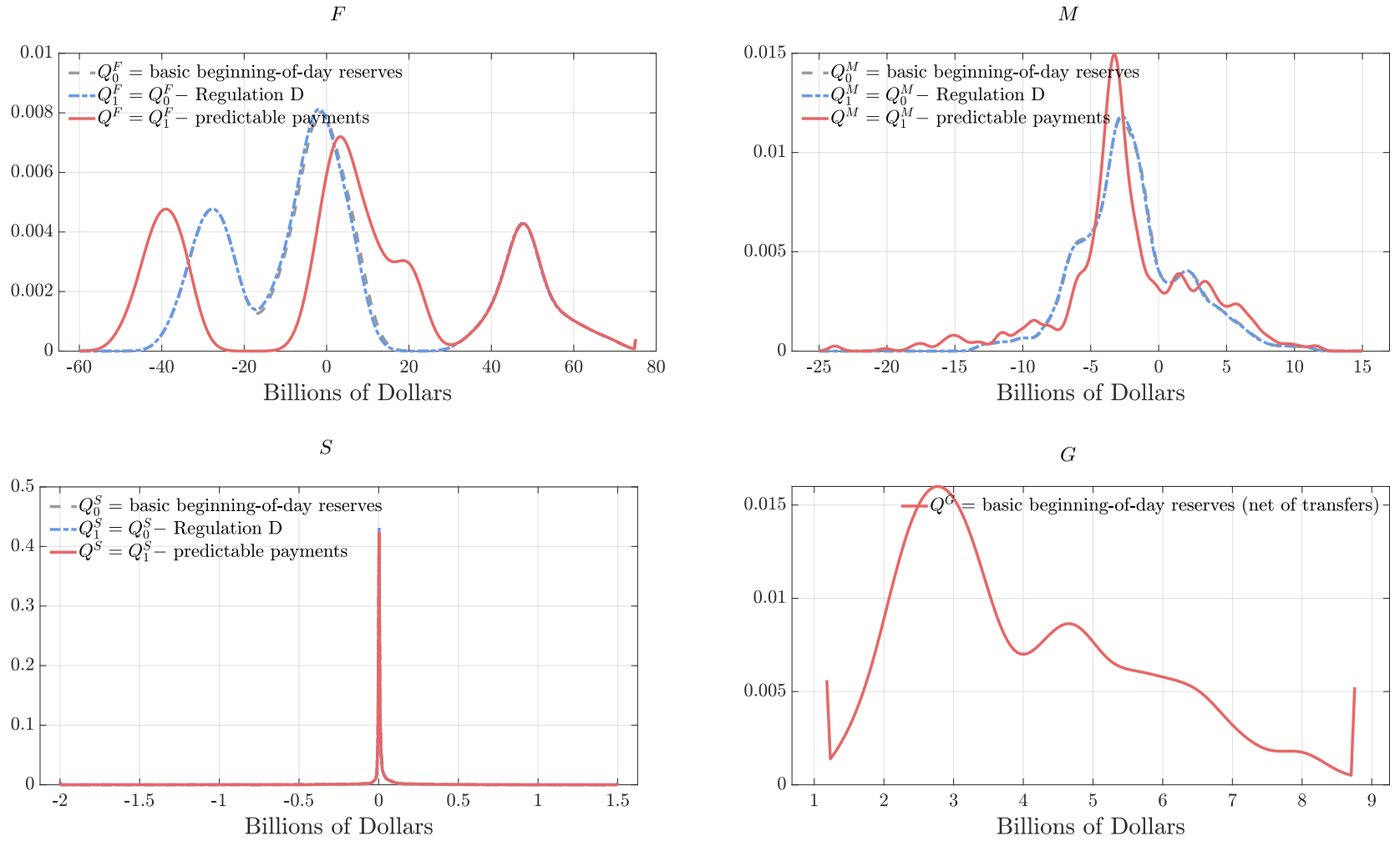

Figure 13: Estimated beginning-of-day distributions of reserves by bank type for the year 2006.

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 overnight-loan 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.

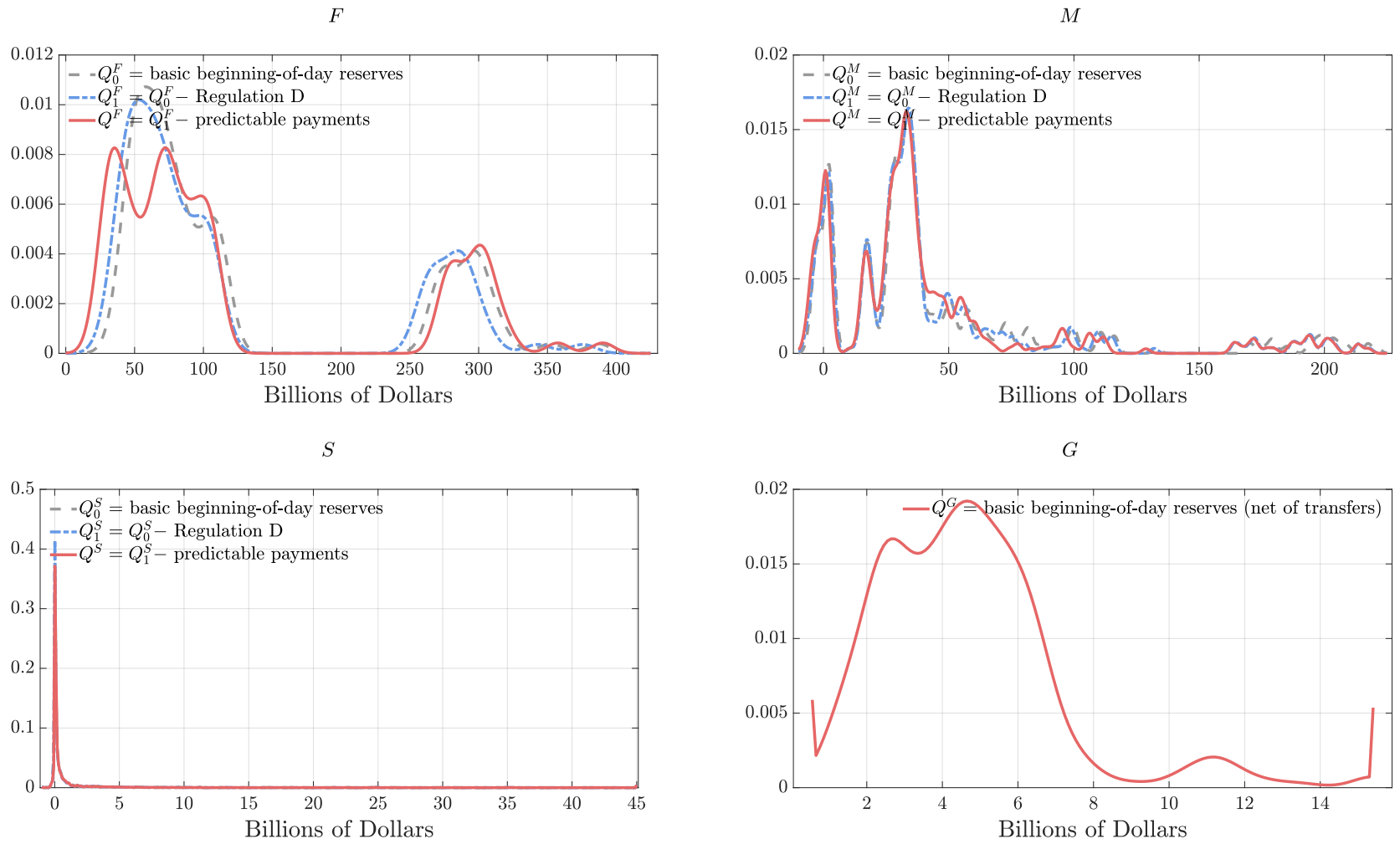

Figure 14: Estimated beginning-of-day distributions of reserves by bank type for the year 2014.

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 overnight-loan 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.

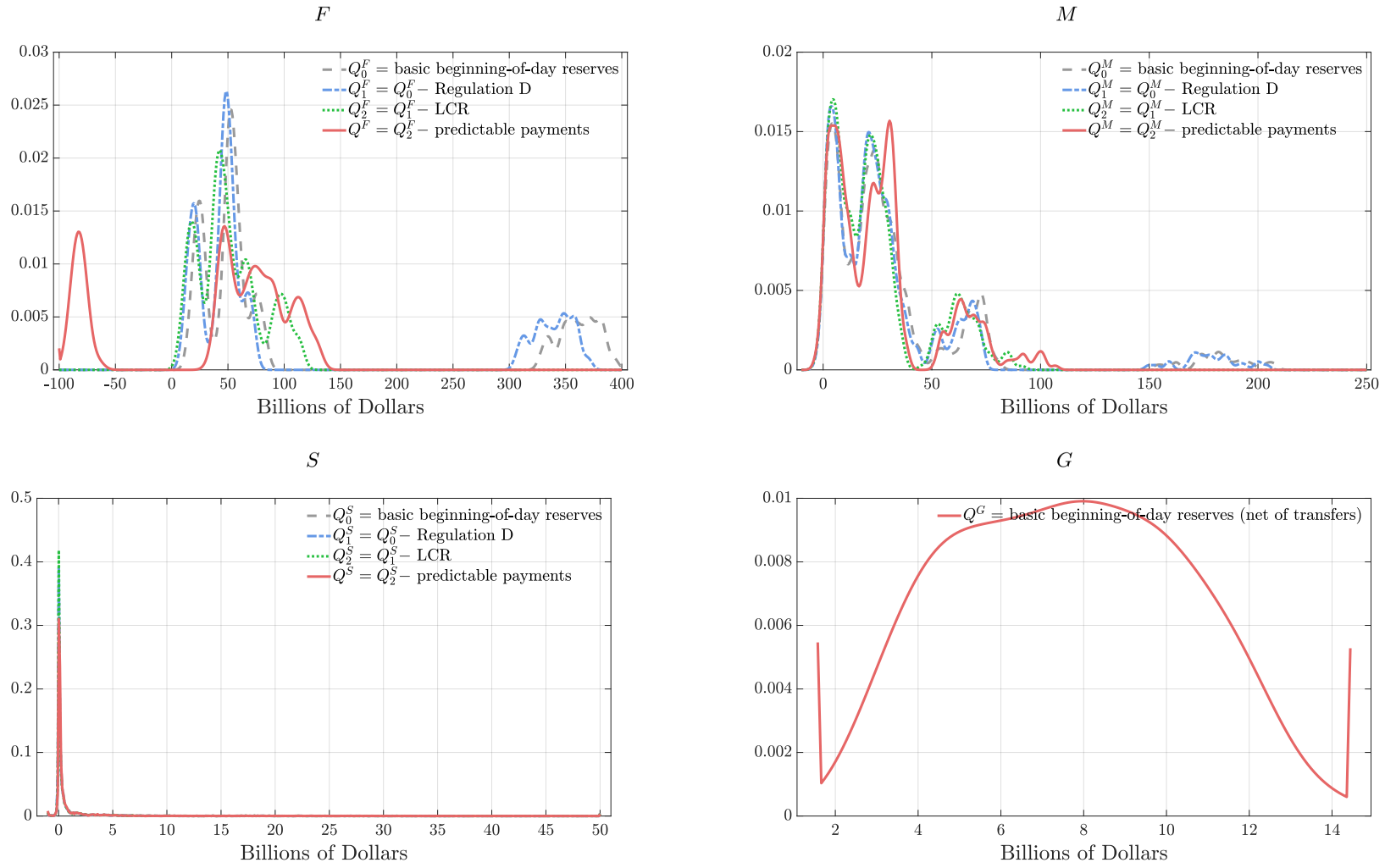

Figure 15: Estimated beginning-of-day distributions of reserves by bank type for the year 2017.

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 overnight-loan 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.

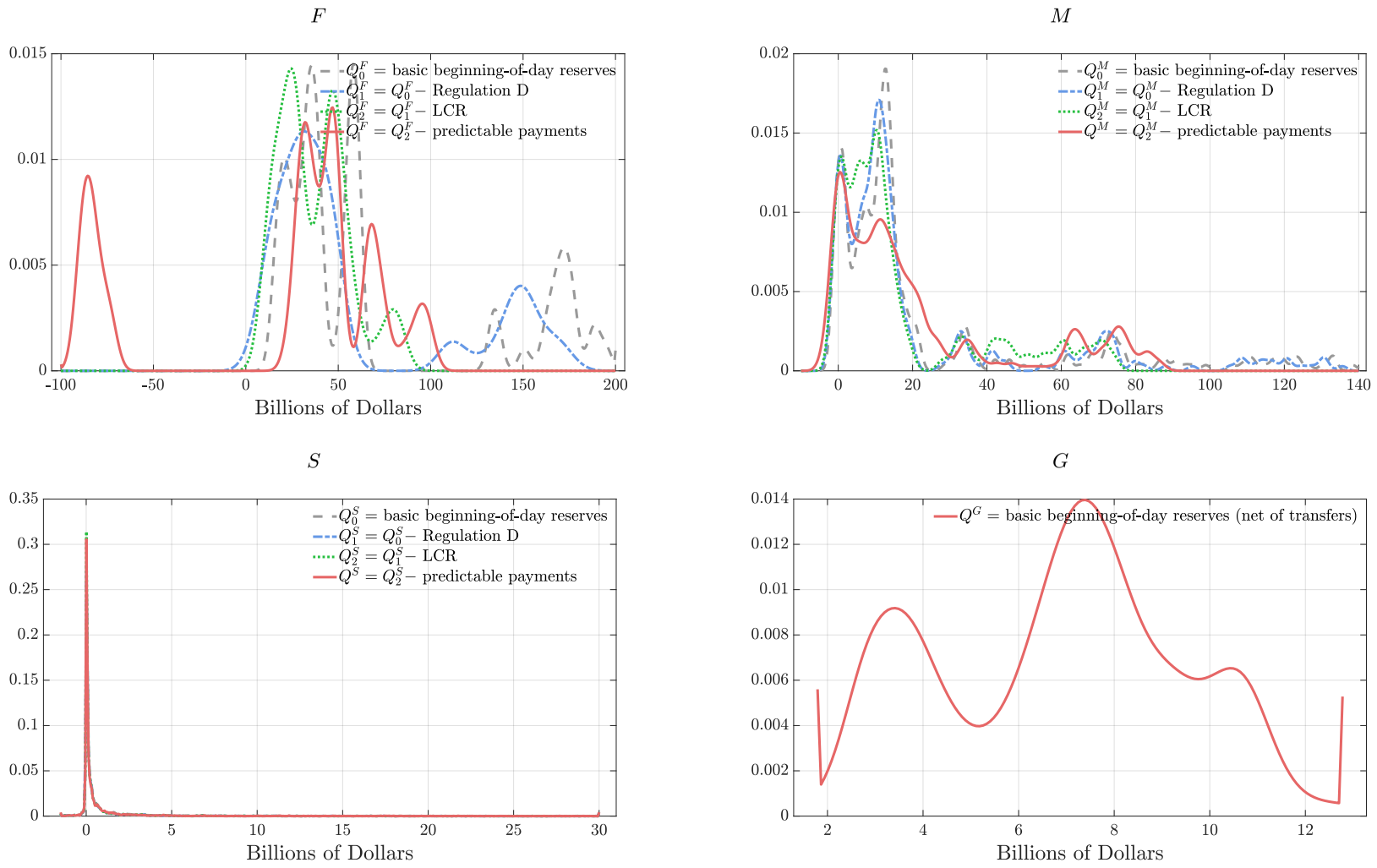

Figure 16: Estimated beginning-of-day distributions of reserves by bank type for the year 2019.

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 overnight-loan 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.

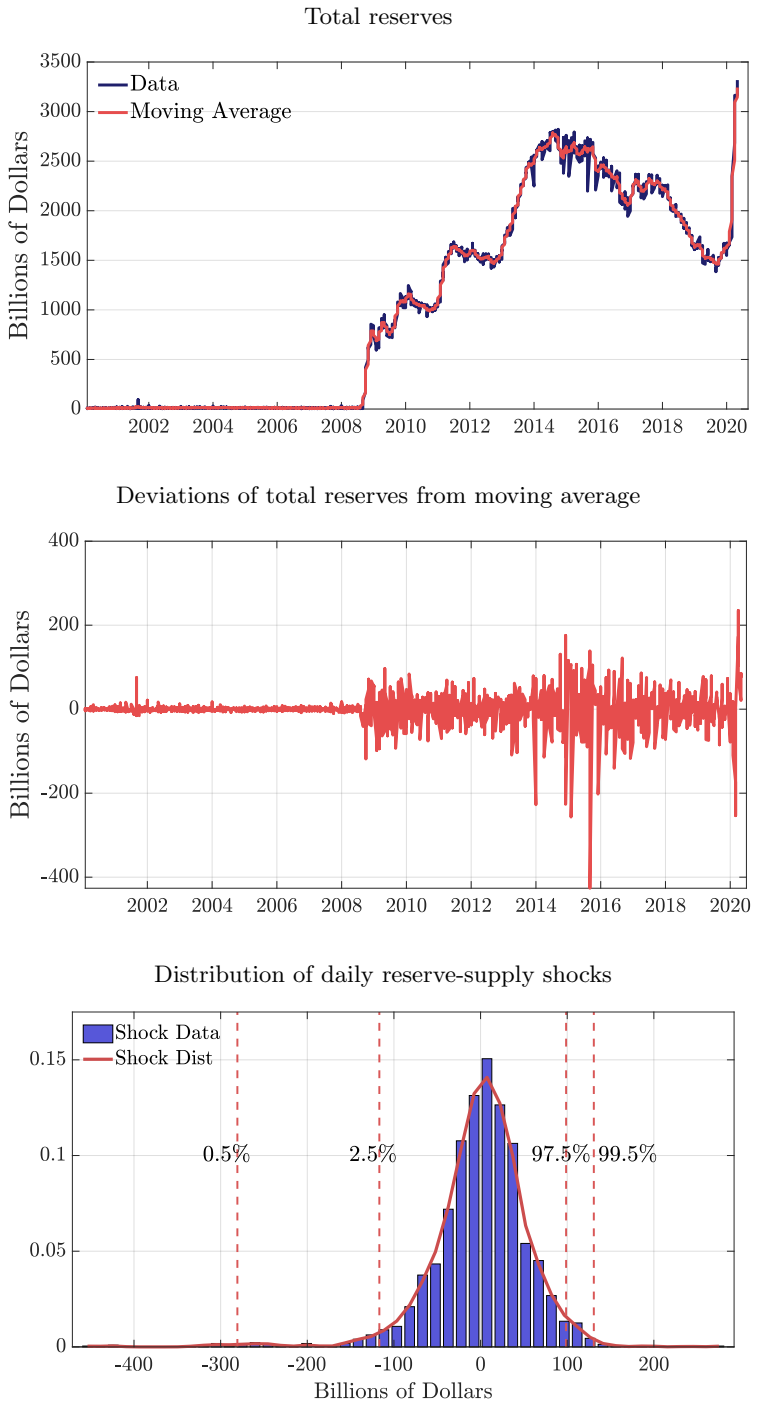

Figure 17: Aggregate supply of reserves and reserve-supply shocks.

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.

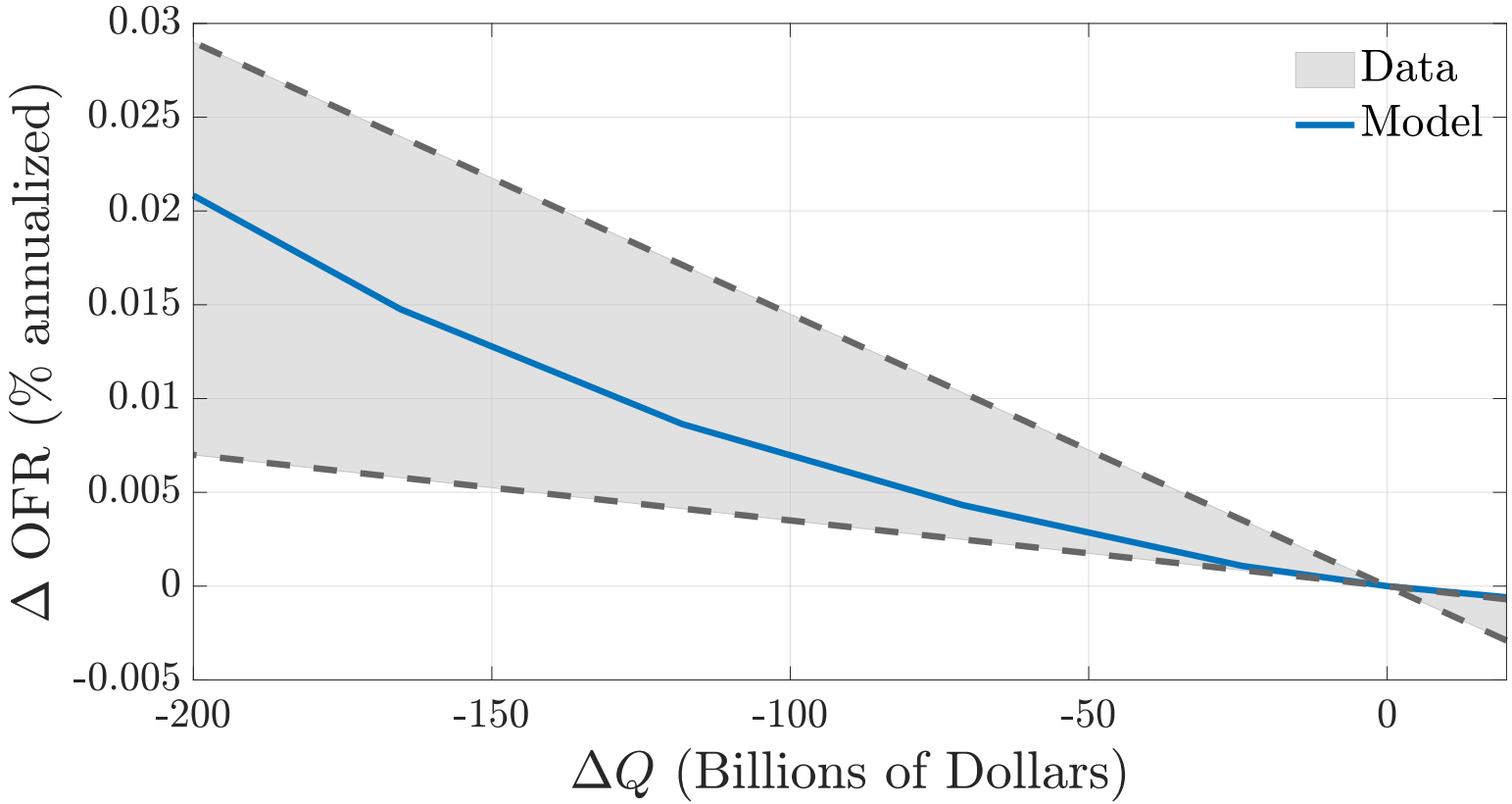

Figure 18: Liquidity effect: model and empirical estimates for the year 2019.

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 (8). The solid line is the change in the equilibrium value-weighted average interest 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 A.6.)

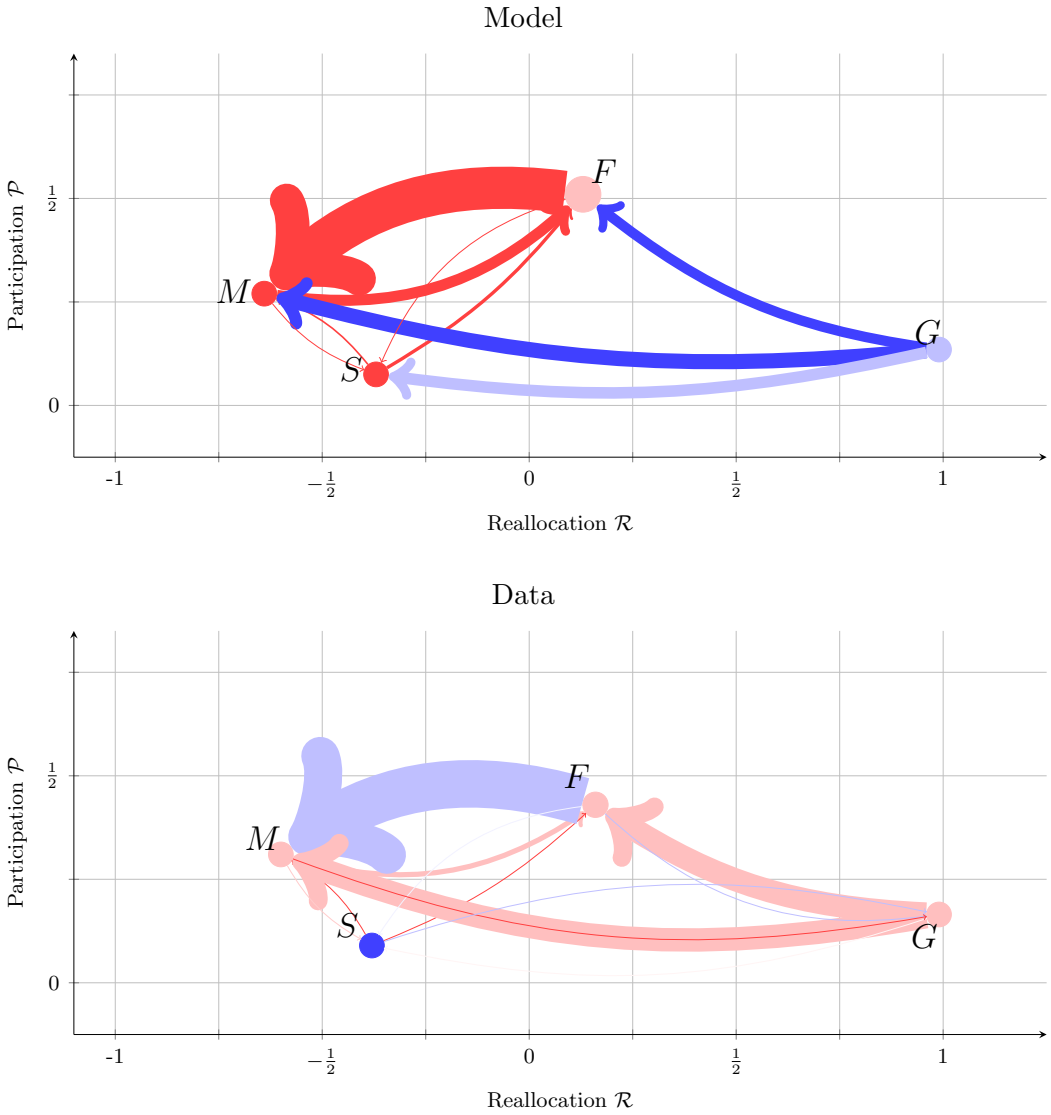

Figure 19: Theoretical and empirical interbank trading networks for 2019.

B Institutional background and regulation

In this section we review three financial regulations that affect banks' incentives to borrow and lend in the interbank market. Two of them directly increase a bank's shadow value of holding reserves by imposing regulatory balance-sheet constraints that can be satisfied with reserve balances (traditional reserve requirements, discussed in Section B.1, and the *Liquidity Coverage Ratio*, discussed in Section B.2.1). The third, is a leverage constraint that increases a bank's shadow cost of all borrowing, including interbank overnight borrowing (the *Supplementary Leverage Ratio*, discussed in Section B.2.2).

B.1 Traditional reserve requirements (Regulation D)

Reserve requirements have been a part of the financial landscape in the United States since before the Federal Reserve Act of 1913 that created the system of Reserve Banks.[80] Regulation D ("Reserve Requirements for Depository Institutions") is the Federal Reserve regulation that stipulates reserve requirements for depository institutions (i.e., commercial banks, savings banks, thrift institutions, credit unions, and agencies and branches of foreign banks located in the United States).

Until March 2020, Regulation D required depository institutions to keep a minimum amount of reserves against their transaction accounts (such as demand deposits).[81] This reserve requirement was 0%, 3%, or 10% of transaction account deposits depending on the size of the bank's reservable liabilities.[82] Institutions had to satisfy reserve requirements by holding cash in their vaults or as a balance in the institution's account at the Federal Reserve Bank in the Federal Reserve District in which the institution is located (either an account of the institution or an account of the institution's Federal Reserve pass-through correspondent).

Reserve requirements were calculated based on a bank's deposit accounts during *computation periods* that depended on the frequency (either weekly or quarterly) with which an

[80] Reserve requirements at the national level were first established with the passage of the National Bank Act in 1863. In the original Federal Reserve Act of 1913, for example, banks were required to hold in reserve different percentages of their demand deposits, depending on whether they were classified as *central reserve city banks* (18 percent), *reserve city banks* (15 percent), or *country banks* (12 percent). See Feinman (1993) for more background and references on the history of reserve requirements in the United States.

[81] There was an explicit exemption from Regulation D for bank obligations in nondeposit form to another bank, which included "federal funds purchased".

[82] The Federal Reserve Board reduced all reserve requirement ratios to 0% effective March 26, 2020.

institution files an FR 2900 report.[83] Each reserve computation period was used to calculate the reserve requirement that a bank had to satisfy on a lagged basis, i.e., during a 14-day *(reserve) maintenance period* in the future.

For institutions that file the FR 2900 report weekly, a *(FR 2900) reporting period* is one week long, covering the seven consecutive calendar days beginning on a Tuesday and ending on the following Monday. The *computation period* for weekly reporters consisted of two *reporting periods*, i.e., 14 consecutive days beginning on a Tuesday and ending on the second Monday thereafter. A *maintenance period* consisted of 14 consecutive days beginning on a Thursday and ending on the second Wednesday thereafter. Each reserve computation period was used to calculate the reserve requirement that a bank had to satisfy on a lagged basis: The reserve balance requirement that had to be satisfied during a maintenance period was based on the average level of net transaction accounts and vault cash held during the computation period that had ended 17 days earlier.[84]

Federal Reserve Banks were authorized to assess charges for deficiencies at a rate of 1 percentage point per year above the primary credit rate in effect for borrowings from the Federal Reserve Bank on the first day of the calendar month in which the deficiencies occurred. Charges were assessed on the basis of daily average deficiencies during each maintenance period.

B.2 Post-GFC regulation

In the years following the Great Financial Crisis (GFC), the Federal Reserve Board (FRB), the Federal Deposit Insurance Corporation (FCC), and the Office of the Comptroller of the Currency (OCC) implemented versions of two regulations agreed to by the Basel Committee on Banking Supervision (BCBS), and consistent with the Dodd-Frank Wall Street Reform and Consumer Protection Act: The *Liquidity Coverage Ratio* (LCR), a prudential liquidty standard, and the *Supplementary Leverage Ratio* (SLR), a prudential leverage standard. Both affect banks' payoffs from trading in the interbank market. We discuss each in turn.

B.2.1 Liquidity Coverage Ratio (LCR)

The first objective of the Basel III accord agreed upon by the members of the Basel Committee on Banking Supervision (BCBS) is to promote the short-term resilience of the liquidity risk

[83] This report collects information on select deposits and vault cash from depository institutions.
[84] See Federal Reserve Board (2019a) for details.

profile of banks. The BCBS developed the LCR to achieve this objective.[85] Specifically, the LCR is designed to ensure that a bank maintains an adequate level of unencumbered, *High Quality Liquid Assets* (HQLA) that can be converted into cash to meet its liquidity needs for a 30-calendar-day time horizon under a liquidity stress scenario specified by supervisors.

The LCR is defined as

$$LCR \equiv \frac{H}{L}, \tag{13}$$

where H denotes HQLA, and L is a measure of total net cash outflows in a 30-day standardized stress scenario. The HQLA consist of Level 1 assets and Level 2 assets. Level 1 assets, which are not subject to haircuts or quantitative caps, include reserves in excess of Regulation D held at a Federal Reserve Bank, as well as securities issued or guaranteed by the U.S. Treasury. Level 2 assets are subject to prescribed haircuts and are capped at no more than 40% of a banking organization's total HQLA.[86] For our purposes, we can think of H as consisting of two components: (i) reserves, denoted Q_0, minus Regulation D required reserves, denoted R_D; and (ii) the value (net of haircut) of all other assets that qualify as HQLA, denoted A, i.e.,

$$H = A + M,$$

where

$$M \equiv \max(Q_1, 0) \tag{14}$$

and $Q_1 \equiv Q_0 - R_D$ denotes the quantity of reserves in excess of the Regulation D requirement. The "max" in (14) reflects that only reserves in excess of Regulation D qualify as HQLA.

Banks report H and L, and these reports are publicly available at a quarterly frequency.[87] Given H, since we have independent information on Q_0 and R_D (and therefore M), we can infer A. The LCR regulation requires

$$1 \leq LCR \tag{15}$$

[85] See Basel Committee on Banking Supervision (2010) for more details on the rationale for the regulation.

[86] Level 2 assets are further divided into Level 2A and Level 2B assets. Level 2A assets, which are subject to a 15% haircut, include claims on or guaranteed by a U.S. government-sponsored enterprise (GSE) such as Fannie Mae and Freddie Mac. Level 2B assets, which are subject to a 50% haircut and are capped at no more than 15% of a banking organization's total HQLA, include certain corporate debt securities issued by non-financial companies, and certain publicly traded common equities issued by non-financial companies that are included in the Russell 1000 Index or a foreign equivalent index for shares held in foreign jurisdictions.

[87] E.g., from the S&P Global Capital IQ database. See Appendix C (Section C.2) for details.

daily, or monthly, depending on the size and other characteristics of the bank.[88] For our purposes, the key implication of the policy constraint (15) is that it may cause a bank to treat certain holdings of HQLA as required to comply with the LCR regulation. By this we mean that the LCR constraint may cause the bank to inpute an additional shadow cost of reducing its holdings of HQLA on a typical day—including reserve balances. In the specific case of reserve balances, the bank may impute an additional shadow cost of lending, since this may drive the bank's reserves (net of the Regulation D requirement) below the level of reserves that the bank routinely allocates to comply with the LCR regulation. Thus, in practice, banks may regard some of the reserves in excess of the Regulation D requirement as being "required" to satisfy the LCR constraint. The fact that the LCR regulation allows for substitutability among the HQLA in the numerator of the left side of (15) presents us with an identification challenge when trying to estimate the share of a bank's reserve balances in excess of the Regulation D requirement that the bank treats as "required" to satisfy the LCR constraint. Next, we formalize this identification problem, and describe how we address it.

For each bank, we observe H, M, and A. We want to express M as the sum of a component, $\hat{M}^R$, that represents the quantity of reserves (in excess of the Regulation D requirement) that the bank relies on to comply with the LCR regulation, and a component, $\hat{M}^E$, that represents reserves in excess of the Regulation D *and* the LCR requirements. Similarly, a bank may hold HQLA (other than reserves) in excess of what would be necessary to meet the LCR requirement for reasons other than having to comply with the LCR regulation, so we can also decompose A into two (unobserved) components: $\hat{A}^R$, which represents the value of HQLA (other than reserves in excess of the Regulation D requirement) that the bank regards as being necessary to comply with the LCR regulation, and $\hat{A}^E$, which represents the value of HQLA (other than reserves in excess of the Regulation D requirement) that the bank regards as being in excess of

[88] Relatively large institutions regulated by the FRB must calculate and maintain a liquidity coverage ratio that is equal to or greater than 1 on each business day (or, in the case of a smaller FRB-regulated institutions, on the last business day of the applicable month). The LCR rule is codified at 12 CFR part 50 (OCC), 12 CFR part 249 (FRB), and 12 CFR part 329 (FDIC).

what is required to meet the LCR regulation. In summary, $\{\hat{M}^j, \hat{A}^j\}_{j\in\{R,E\}}$ satisfy:

$$M = \hat{M}^R + \hat{M}^E \tag{16}$$

$$A = \hat{A}^R + \hat{A}^E \tag{17}$$

$$\hat{A}^R + \hat{M}^R \leq L, \text{ with "=" if } L \leq A + M \tag{18}$$

$$\hat{A}^E + \hat{M}^E = 0, \text{ if } A + M < L \tag{19}$$

$$\hat{M}^j, \hat{A}^j \in \mathbb{R}_+ \text{ for } j \in \{R, E\}. \tag{20}$$

We are interested in using the policy constraint (15) along with data on M, A, and L, and (16)-(20), to estimate bank-level bounds for $\hat{M}^R$.

There are three special cases in which the constraint (15) together with knowledge of M, A, and L, and the definitions (16)-(20) are sufficient to identify $\hat{M}^R$ and $\hat{A}^R$. First, if a bank has $LCR \leq 1$ (i.e., if it is not complying with the LCR regulation in a given sample period), then the bank is clearly holding no excess HQLA of any type, so $\hat{M}^R = M$, $\hat{A}^R = A$, and $\hat{M}^E = \hat{A}^E = 0$, as implied by (16), (17), (19), and (20). Second, if $LCR \geq 1$ and $Q_1 \leq 0$, then $M = 0$, so the LCR requirement, L, is being satisfied exclusively with HQLA other than reserves, i.e., $\hat{A}^R = L$ and $\hat{A}^E = A - L$, with $\hat{M}^R = \hat{M}^E = 0$, as implied by (16), (17), (18), and (20). Third, if $LCR \geq 1$ and $A = 0$, then the LCR requirement, L, is being satisfied exclusively with reserves, M, i.e., $\hat{M}^R = L$ and $\hat{M}^E = M - L$, with $\hat{A}^R = \hat{A}^E = 0$, as implied by (16), (17), (18), and (20).

In practice, most banks satisfy the LCR constraint (15) with $\min(M, A) \geq 0$, and for such banks it is not obvious how to decompose the level of *required* HQLA, i.e., L, into the two unobserved components, $\hat{M}^R$ and $\hat{A}^R$. However, notice that conditions (16)-(20) imply $\hat{M}^R$ must satisfy the following bounds:

$$\hat{M}^R \begin{cases} = M & \text{if } A + M < L \\ \in [\max(0, L - A), \min(L, M)] & \text{if } L \leq A + M. \end{cases} \tag{21}$$

We can write (21) as

$$\hat{M}^R = \begin{cases} M & \text{if } A + M < L \\ \rho \min(L, M) + (1 - \rho) \max(0, L - A) & \text{if } L \leq A + M, \end{cases} \tag{22}$$

for some $\rho \in [0, 1]$. For a given ρ, (16)-(20), and (22) imply

$$\hat{A}^R = \begin{cases} A & \text{if } A + M < L \\ (1 - \rho) \min(L, A) + \rho \max(0, L - M) & \text{if } L \leq A + M, \end{cases}$$

and given $\hat{M}^R$ and $\hat{A}^R$, $\hat{M}^E$ and $\hat{A}^E$ are implied by (16) and (17).

The parameter $\rho \in [0,1]$ represents the bank's (unobserved) preference for satisfying the LCR requirement, L, with reserves (rather than with other HQLA). For example, if $\rho = 1$, the bank has a strong preference for satisfying the LCR with reserves, and this will reduce the bank's willingness to lend reserves in the interbank market. If $\rho = 0$, the bank has a strong preference for satisfying the LCR with HQLA *other* than reserves, and will be less constrained by its reserve balance when trading in the interbank market.

According to elementary theory, the quantity of reserves in excess of regulatory reserve requirements is a key determinant of a bank's "fundamental" incentive to borrow and lend in the interbank market. For example, a bank whose reserve balance is lower than the minimum regulatory requirement, has a fundamental incentive to borrow (at a rate no larger than the shadow cost of violating the regulatory requirement). Conversely, a bank whose reserve balance is higher than the regulatory requirement, would have, all else equal, an incentive to lend (e.g., to banks with negative excess reserves, at a rate between the lender's and the borrower's respective shadow prices of reserves). For this reason, it is important to impute an accurate notion of "excess reserves" in any empirical implementation of a theory of interbank loans.

The traditional definition of "excess reserves", which only subtracts the Regulation D requirement from the bank's reserve balance is not an adequate notion of excess reserves for institutions that must comply with the LCR regulation.[89] In our empirical and quantitative work we use a more comprehensive notion of "required reserves" that includes not only the level of reserves that a bank is required to hold to comply with Regulation D, but also the level of reserves that the bank holds toward meeting the LCR requirement. Specifically, our benchmark definition of "excess reserves" for any bank that is subject to, and satisfies the LCR constraint (15), is $Q_2 \equiv Q_1 - R_L$, where $R_L \equiv \max(0, L - A)$. In other words, to construct our preferred notion of excess reserves, we start from the traditional notion of reserves in excess of the Regulation D requirement, Q_1, and subtract the minimum level of reserves needed to comply with the LCR requirement, i.e., R_L.[90] Notice that our measure of excess reserves coincides with the traditional measure for a bank that has enough HQLA other than reserves to meet the LCR

[89] The LCR regulation applies to bank holding companies (BHCs) and savings and loans holding (SLHCs) with at least $50 bn in total consolidated assets.

[90] From (22), we see that R_L is the same as $\hat{M}^R$ when $\rho = 0$ (in the empirically relevant case with $L \leq A + M$). In this sense, our preferred notion of excess reserves selects the largest level of excess reserves that is consistent with the LCR constraint, (15).

requirement, i.e., $Q_1 - Q_2 = R_L = 0$ if $L \leq A$. But our measure of excess reserves is lower than the traditional measure for a bank whose holdings of HQLA other than reserves are insufficient to meet the LCR requirement, i.e., if $A < L$, then $0 < Q_1 - Q_2 = R_L = L - A$.

B.2.2 Supplementary Leverage Ratio (SLR)

The SLR is the U.S. banking agencies' implementation of the "Basel III Tier 1 Leverage Ratio", which is defined as

$$SLR \equiv \frac{\text{Tier 1 Capital}}{\text{Total Leverage Exposure}}. \tag{23}$$

The numerator (defined in U.S. Basel III) includes common stock and retained earnings. The denominator is a comprehensive measure of assets, composed of four elements: (1) on-balance sheet assets, (2) derivative exposures, (3) repo-style transaction exposures, and (4) other off-balance sheet exposures. The SLR regulation requires a bank to maintain an SLR above a threshold; specifically, either $SLR \geq 0.03$, or $SLR \geq 0.05$.[91]

B.2.3 Resolution Planning

In the aftermath of the GFC, regulatory authorities started requiring large "systemically important" financial institutions (e.g., BHCs with total consolidated assets of $50 bn or more) to periodically submit a resolution plan (also known as "living will") to the Federal Reserve and the Federal Deposit Insurance Corporation. A resolution plan describes in some detail the company's strategy for rapid and orderly resolution in the event of material financial distress or failure of the company.

B.2.4 Effects of LCR and SLR regulation on interbank trading incentives

In this section we discuss the effects of the LCR and SLR reguation on banks' incentives to borrow and lend in the interbank market.

First, we consider the effect of LCR regulation on banks' incentives to borrow and lend in the interbank market. Reserves appear (with weight =1) in the numerator of the LCR in (13), and overnight fed fund liabilities appear in the denominator (also with weight =1). Consider

[91] The threshold equals 3% for *advanced approaches firms*, which include state banks, savings associations, bank holding companies (BHCs), and saving and loan holding companies (SLHCs) with more than $250 bn in total consolidated assets, or more than $10 bn of on-balance sheet foreign exposures. The threshold equals 5% for the 8 US bank-holding companies that have been identified by the Financial Stability Board as global systemically important banks (and their U.S. insured depository institution subsidiaries).

a bank that borrows ℓ in the interbank market. The LCR before the trade is $\frac{H}{L}$ and after the trade it is $\frac{H+\ell}{L+\ell}$. Since

$$\frac{\partial}{\partial \ell}\left(\frac{H+\ell}{L+\ell}\right) = \frac{L-H}{(L+\ell)^2},$$

it follows that the trade does not affect the LCR if the bank is satisfying it exactly pre-trade (i.e., if $LCR \equiv \frac{H}{L} = 1$), increases the LCR if the borrowing bank is below the LCR target pre trade (i.e., if $LCR \equiv \frac{H}{L} < 1$), and decreases the LCR if the borrowing bank is above the LCR target pre trade (i.e., if $LCR \equiv \frac{H}{L} > 1$). For a bank that lends ℓ in the interbank market, the LCR before the trade is $\frac{H}{L}$ and after the trade it is $\frac{H-\ell}{L}$.[92] Hence, lending in the interbank market unambiguously reduces the LCR. To summarize, LCR regulation increases the shadow cost of lending overnight (because lending reserves tightens the LCR constraints of lenders), and increases the shadow cost of borrowing for banks whose LCR constraints are slack at the time of the trade (because borrowing reserves tightens the LCR constraints of such banks).

Second, we consider the effect of SLR regulation on banks' incentives to borrow and lend in the interbank market. Let $\mathcal{A}$ denote assets, $\mathcal{L}$ denote liabilities, and $\mathcal{C} \equiv \mathcal{A} - \mathcal{L}$ denote capital. Then, we can write (23) as

$$SLR \equiv \frac{\mathcal{C}}{\mathcal{A}} = \frac{\mathcal{A}-\mathcal{L}}{\mathcal{A}}. \tag{24}$$

Notice that *lending* in the interbank market does not change the SLR because the bank that acts as a lender is just exchanging reserves for an overnight credit of reserves, which leaves both $\mathcal{L}$ and $\mathcal{A}$ unchanged. However, *borrowing* in the interbank market reduces the SLR, since borrowing ℓ dollars worth of reserves increases liabilities from $\mathcal{L}$ to $\mathcal{L}+\ell$, and increases assets from $\mathcal{A}$ to $\mathcal{A}+\ell$, and therefore the SLR is *reduced* from $\frac{\mathcal{A}-\mathcal{L}}{\mathcal{A}}$ to $\frac{\mathcal{A}-\mathcal{L}}{\mathcal{A}+\ell}$. To summarize, SLR regulation has no effect on the shadow cost of lending (because lending reserves does not alter the SLR constraint), but increases the shadow cost of borrowing (because borrowing reserves tightens the SLR constraint of solvent banks).

[92] The quantity of reserves sold, ℓ, is subtracted from the HQLA of the lender, but the corresponding credit is not added to the total of HQLA of the lender because interbank loans not qualify as a HQLA.

C Data: sources, construction, and empirical estimations

This appendix describes the databases used in the paper, and how we merged them to produce bank-level datasets containing: (i) daily reserve balances and regulatory reserve requirements under Regulation D and the LCR (Section C.1 and Section C.2); and (ii) high-frequency interbank reserve transfers—comprising payments and overnight loans (Section C.3 and Section C.4). We also describe the computation of the statistics used in the empirical and quantitative analyses (Section C.5).

C.1 Reserve balances and Regulation D

The Monetary Policy Operations and Analysis (MPOA) section of the Monetary Affairs Division at the Federal Reserve Board of Governors provided the database of bank-level end-of-day reserve balances at daily frequency. MPOA also supplied bank-level data on Regulation D reserve requirements for each two-week maintenance period. Both reserve balances and Regulation D requirements are reported at the level of the bank holding company, which we adopt as the relevant unit of observation throughout. MPOA reports end-of-day balances as of 6:30 p.m. EST. We impute beginning-of-day balances for the following day (as of 9:00 a.m. EST) using the procedure described in Section C.5.1.

C.2 Liquidity Coverage Ratio

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. 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 $\mathrm{L}_m(d) \leq \mathrm{H}_m(d)$.[93]

Both, $\mathrm{L}_m(d)$ and $\mathrm{H}_m(d)$ are made public by each bank at a quarterly frequency. We obtain data on the ratio of these quantities from S&P Global Capital IQ database.[94] We used SNL Classic Data and run a Companies (Classic) screener to search for our data. We extracted quarterly LCR (LIQUIDITY_COV_RATIO) data from 1990Q1 to 2021Q2. For some banks, LCR data were missing in some quarters. For these cases, we obtained the LCR data directly

[93] Appendix B (Section B.2.1) describes the LCR regulation in greater detail.
[94] The S&P database can be accessed at: https://www.spglobal.com/marketintelligence/en/.

from the bank's website.[95]

We merged our balances data from MPOA (described in Section C.1) with the S&P LCR data using the Replication Server System Database (RSSD) ID. (The balances data from MPOA contains the RSSD of each bank holding company.) We then created a manual crosswalk to match RSSDs to parent company names in the S&P database, using the National Information Center repository from the Federal Financial Institutions Examination Council (https://www.ffiec.gov/NPW). We always matched RSSDs to the parent bank holding company to which the LCR regulation applies. In general, this procedure implies matching the RSSDs to the highest level parent company in the corporate structure, except for cases in which the parent company is a sovereign government and the LCR constraint applies to the second highest parent company level.

C.3 Reserve tranfers: payments and overnight loans from Fedwire

Our analysis of interbank reserve transfers begins with the daily, minute-by-minute universe of bank-level transactions settled through Fedwire Funds Service (*Fedwire*). Fedwire is an electronic, large-value, real-time gross settlement system operated by the Federal Reserve Banks that operates for 21.5 hours each business day, from 9:00 p.m. Eastern Standard Time (EST) on the preceding calendar day to 6:30 p.m. EST.[96] Participants include domestic financial institutions (such as commercial banks, investment banks, thrift institutions, and credit unions), Foreign Banking Organizations (FBOs—agencies and branches of foreign banks operating in the U.S.), government securities dealers, government agencies (including federal and state entities), and Government-Sponsored Enterprises (GSEs), such as Freddie Mac, Fannie Mae, and the Federal Home Loan Banks.

A Fedwire participant must have a *master account* at a Federal Reserve Bank in which it holds its reserves, and it uses its *Fedwire accounts* to transfer reserves from its master account to the master accounts of other Fedwire participants. Transfers may be made on a bank's own account or on behalf of its non-participant clients, and can serve two primary purposes: to lend

[95] This was the case for the following three banks:

- Credit Agricole Group (https://www.credit-agricole.com/en/pdfPreview/186985)
- DNB ASA (https://ir.dnb.no/capital-framework)
- State Street Corporation (https://investors.statestreet.com/filings-and-reports/u-s-liquidity-coverage-ratio-disclosures/default.aspx).

[96] While some transfers occur between 9:00 p.m. and 9:00 a.m., the bulk are between 9:00 a.m. and 6:30 p.m.

or repay reserves, and to make outright payments.

An institution that participates in Fedwire may hold one or more Fedwire accounts, each identified by a unique *Fedwire account number*. We followed the guidelines from the Reserve Bank Operations and Payment Systems Division at the Federal Reserve Board to link each institution's Fedwire account number(s) to a unique *bank ID* corresponding to that institution. When institutions with different bank IDs belonged to the same bank holding company, we aggregated them into a single entity, since regulatory requirements—such as reserve requirements, the liquidity coverage ratio (LCR), the supplementary leverage ratio (SLR), and interest-on-reserves calculations—typically apply at the bank holding company level.

Having mapped Fedwire account numbers to bank holding companies, we assigned each Fedwire transfer to the corresponding bank holding company. In the few cases where a bank ID could not be matched to a bank holding company, the Fedwire account numbers linked to that bank ID were excluded from the sample.

To initiate a Fedwire transfer, a participant completes several fields in an electronic form, including the Fedwire account numbers of the sending and receiving parties and the amount to be sent. The participant then submits the transfer order to the Federal Reserve, which, upon receipt, immediately debits the sender's Fedwire account and credits the receiver's Fedwire account. Fedwire transfers are irrevocable and final. The form includes an optional field where the sender can specify a *payment type code* indicating the purpose of the transfer. However, this field is often left blank or filled with nonstandard coded identifiers.[97] For this reason, it is not possible to determine the purpose of a Fedwire transfer from the Fedwire dataset alone.

Given that the purpose of a transfer is not directly observable from the Fedwire dataset, we employed a classification procedure to infer which transactions represent overnight lending. Specifically, we applied a modified version of the *Furfine algorithm* (based on Furfine (1999)), provided to us by the Money Market Analysis Section of the Monetary Affairs Division at the Federal Reserve Board. We treated the transactions identified by the algorithm as *actual* overnight loans and regarded the remaining transfers as *payments*——presumably unrelated to overnight lending or repayment activity. All individual payments between a pair of banks during a trading day with a value below $10,000 were consolidated into a single payment.

We excluded from our sample all bank IDs whose linked Fedwire accounts did not borrow

[97]The codes that senders use are not standardized; for example, a payment type code of "1" may indicate a fed funds sale for one sender, but a payment to settle a securities purchase for another.

or lend during a given year (according to the Furfine algorithm). Our final sample consists of 754 Fedwire participants in 2006, 404 in 2014, 395 in 2017, and 412 in 2019.

In this paper, whenever we refer to overnight interbank loans, we are actually referring to the subset of Fedwire transactions that were identified as overnight interbank loans by our version of the Furfine algorithm.

Despite its widespread use and intuitive appeal, the Furfine algorithm has well-known limitations. It may identify transactions that are not overnight interbank loans, and fail to identify transactions that are. It can also misidentify counterparties. Even when a loan is correctly identified, the algorithm does not distinguish among loan types—such as fed funds, Eurodollar, tri-party repos, and correspondent lending. We address these limitations in the next subsection.

C.4 Loan identification in Fedwire

We apply a modified version of the Furfine algorithm to identify overnight lending transactions in the Fedwire database. This approach is widely used in the literature, as exemplified by Furfine (1999, 2001, 2002), Demiralp et al. (2006), Ashcraft and Duffie (2007), Armantier et al. (2008), Bartolini et al. (2010), Bech and Atalay (2010), Afonso et al. (2011), Kuo et al. (2013), and Afonso and Lagos (2014, 2015b), among others. We begin by reviewing the structure of the original algorithm, then describe its limitations and how we address them.[98]

First, potential overnight loans are identified as transfers between counterparties i and j involving a payment from i to j of at least \$1 million, ending in five zeros, matched to a return payment from j to i on the following business day that exceeds the original amount by a margin consistent with a plausible overnight interest rate. The difference between the two payments is interpreted as interest on the loan.[99] Second, the set of potential overnight loans is refined by excluding transactions with implied rates that are too high or too low relative to a benchmark overnight rate, typically the effective fed funds rate.

Practitioners sometimes impose additional filters depending on the intended application. For instance, some versions attempt to identify and exclude Eurodollar loans, while others screen out transactions involving settlement institutions such as the Depository Trust Company (a

[98] Variants of the algorithm are discussed in Furfine (1999), Afonso et al. (2011), and Afonso and Lagos (2014).

[99] Consider a \$100 million payment sent from bank i to bank j on day t, matched to a \$100.0139 million payment from j to i on day $t+1$. Since $100.0139/100 = 1 + 0.05/360$, this transaction is interpreted as a \$100 million overnight loan from i to j at a 5 percent annualized interest rate. If multiple candidate repayments match an outgoing transfer, the implied median interest rate is imputed.

securities settlement system) or the Clearing House Interbank Payments System (CHIPS—a private large-value U.S. dollar payment system).

The limitations of the Furfine algorithm are well understood——many were originally noted by Furfine (1999). Given its widespread use by researchers within the Federal Reserve System, there has been ongoing interest in evaluating the algorithm's accuracy in specific applications.

Armantier and Copeland (2015), for example, examined the accuracy of the version of the Furfine algorithm used by the Federal Reserve Bank of New York at the time of their writing, evaluating how well its output corresponded to fed funds as defined under Regulation D. Their case study relied on private reports from two undisclosed banks, which allowed the authors to identify the send legs of fed funds received by those banks between 2007 and 2011. Based on this back-office information, Armantier and Copeland (2015) reported that, on average, the version of the Furfine algorithm they tested failed to identify 23 percent of the banks' actual fed funds transactions, and that 81 percent of the pairs of payments identified as overnight loans did not correspond to fed funds as defined under Regulation D.

Kovner and Skeie (2013) compared the output of the same algorithm with FR Y-9C regulatory filings of bank holding companies. They concluded that, although the identified loans may not correspond precisely to fed funds as defined under Regulation D, they can be more conservatively regarded as estimates of broader overnight interbank activity.

It is useful to group the limitations of the Furfine algorithm into three categories: (1) *loan identification*, (2) *counterparty identification*, and (3) *loan-type identification*.

(1) **Loan identification.**
May fail to identify actual interbank overnight loans, and may misidentify as an interbank overnight loan a pair of transfers that do not correspond to an interbank overnight loan.

   (a) Fails to identify interbank overnight loans that:
      i. Are not settled through Fedwire.
      ii. Are rolled over to the next business day.
      iii. Have an implied rate outside the specified selection window.
      iv. Have a send leg of a dollar amount that does not satisfy the selection criteria.
      v. Have a send or repayment leg involving multiple Fedwire transfers.
   (b) May misidentify as an interbank overnight loan:

i. A random pair of payments that meet the selection criteria.

ii. A loan that banks intermediate on behalf of *nonbanks*—wealth management funds, hedge funds, or nonfinancial corporations.

iii. The send and repayment legs of two distinct, overlapping interbank *term loans*.

(2) **Counterparty identification.**
May misidentify the borrower and lender—conditionally on correctly identifying an interbank overnight loan.

(3) **Loan-type identification.**
Cannot identify the loan type—fed funds, Eurodollar, repo, tri-party repo—conditionally on correctly identifying an interbank overnight loan.

In this paper, we make progress on all three fronts. We develop a procedure to improve loan identification and loan-type classification for fed funds and Eurodollars by matching the output of the Furfine algorithm with bank-level overnight deposit data from FR 2420 filings. To identify repos in the Furfine output, we examine the relationship between the reference rate for overnight repo transactions and the interest rates on loans that are identified by the algorithm but not matched to fed funds or Eurodollars in FR 2420.[100]

We also improve counterparty identification in two ways. First, we augment the Furfine–FR 2420 matching procedure for fed funds and Eurodollars with additional information on correspondent banking relationships for FBOs. Second, we implement a new adjustment to the Furfine algorithm to ensure that our main results are robust to its ability to correctly identify the counterparties to tri-party repo loans.

Section C.4.1 describes the cross-referencing procedures between the FR 2420 filings and the Furfine output. Section C.4.2 describes our tri-party repo adjustment to the Furfine algorithm.

C.4.1 Cross-referencing Fedwire loans with FR 2420 and repo rates

The FR 2420 report (titled *Report of Selected Money Market Rates*) is a regulatory filing required by the Federal Reserve from certain financial institutions, including domestic banks and FBOs. It collects daily, transaction-level data on money market *borrowing*. The form

[100] Loan-type identification is not a major concern for our analysis, as our application focuses on a broad class of interbank overnight loans—not a narrow subset such as fed funds under Regulation D, which was the focus of most previous studies.

consists of four parts: fed funds (Part A), Eurodollar deposits (Part B), time deposits and certificates of deposit (Part C), and other selected deposits (Part D). Loans are classified by type and maturity bucket (overnight or term), allowing us to identify overnight fed funds and Eurodollar deposits. For each loan, FR 2420 reports the *date*, *size*, *rate*, *receiver ID*, and a broad *counterparty category* for the sender—domestic bank, FBO, GSE, or other financial institution.

We use the following procedure to cross-reference the FR 2420 database for 2019 with the database of overnight interbank loans produced by the Furfine algorithm, which we refer to as the *Fedwire Overnight Loan* (FWOL) database. First, we search the FWOL database for loans that match the *date*, *size*, *rate*, and *receiver ID* of each overnight fed funds and Eurodollar deposit reported in the FR 2420 database. We find matches for 48 percent of the volume of overnight fed funds and 39 percent of the volume of overnight Eurodollar deposits. Since fed funds account for 60 percent of the combined volume of these two instruments in FR 2420, this matching procedure identifies corresponding loans in the FWOL database for 44 percent of their combined volume. Based on this step alone, the FWOL database appears to miss a significant share of fed funds and Eurodollar activity. To address this, we take a second step.

As a second step, we identify 23 receiver bank IDs in the FR 2420 filings that do not appear in the FWOL database; 22 of these correspond to FBOs that operate on Fedwire through correspondent banks. We consult the *Bankers Almanac* to construct a set of candidate correspondent banks for these 22 FBOs and identify those banks' Fedwire IDs in the FWOL database.[101] We then run a new search on the FWOL database——similar to our initial one——but replacing the receiver IDs of the 22 FBOs with the IDs of their respective correspondent banks. This second step yields matches in the FWOL database for 98 percent of the volume of overnight fed funds and 64 percent of the volume of overnight Eurodollar deposits reported in FR 2420. The unmatched Eurodollar loans in FR 2420 may reflect rollovers to the next business day, transactions missed by the Furfine algorithm due to selection criteria, or——most likely——loans not settled through Fedwire.

Our cross-referencing procedure between the FR 2420 and FWOL databases mitigates several well-known concerns about the Furfine algorithm.

First, contrary to what Armantier and Copeland (2015) report in their two-bank case study

[101] The *Bankers Almanac* is a comprehensive directory of banks and financial institutions that provides detailed information, including bank identifiers, ownership and hierarchy (parent–subsidiary structure), and correspondent relationships. We used the list of correspondent banks published for the year 2024, which was the only one available to us.

for 2007–2011, our finding that the FWOL database captures 98 percent of overnight fed funds reported in FR 2420 dispels the concern—listed as (1)(a) above—that the algorithm may miss a significant fraction of fed funds activity.

Second, the matching procedure allows us to identify the counterparties to fed funds and Eurodollar deposits, partially addressing the counterparty identification problem (2) for these instruments.[102]

Third, the matching procedure enables loan-type identification for fed funds and Eurodollar deposits within the FWOL database, thereby addressing the loan-type identification problem (3) for these instruments.

The two-step procedure described above allows us to find matches in the FWOL database for 85 percent of the combined volume of overnight fed funds and Eurodollars reported in the FR 2420 database. The remaining 15 percent most likely corresponds to loans extended or settled outside of Fedwire. However, the volume of overnight fed funds and Eurodollars in the FWOL database that we are able to match to FR 2420 loans accounts for only 36 percent of the total volume of overnight loans in the FWOL database in 2019.

This leaves open the question of whether the remaining 64 percent represent true interbank overnight loans or "false positives"—that is, pairs of transfers that do not correspond to an actual overnight loan. And if they are indeed overnight loans, then what kind are they? A natural hypothesis is that they are repurchase agreements (repos), a common form of overnight lending between banks.

To assess this hypothesis, we study the relationship between the volume-weighted average daily interest rate for overnight loans in the FWOL database that are not identified as fed funds or Eurodollars in the FR 2420 filings, which we refer to as the *Overnight Fedwire Repo Rate* (OFRR), and the *Secured Overnight Financing Rate* (SOFR), a widely used reference rate for overnight repo transactions.[103] Using daily data for the period 2019/01/01–2019/08/31 (168 observations), we estimate a linear regression of the SOFR on the OFRR, including dummies

[102] A caveat is that FR 2420 reports only the *receiver* bank ID for each loan. This prevents us from identifying *both* counterparties for loans involving banks already present in the FWOL database. For the 22 FBOs that appear in FR 2420 but not in the FWOL database, we can assign to them the fed funds or Eurodollars they receive, but we cannot observe the loans they send, as this information is not included in the FR 2420 filings. Any interbank loans made by these FBOs, if they exist, would still be attributed to their correspondent banks by the Furfine algorithm. As a result, if FBOs lend through correspondent banks over Fedwire, our two-step cross-referencing procedure would understate their reallocation measures and overstate the lending activity of the banks providing correspondent services.

[103] See https://www.newyorkfed.org/markets/reference-rates/sofr.

for the last trading day of the month, the first trading day of the year, and the first trading day after July 4. The estimated slope is 0.99, with an R^2 of 0.90. This close correspondence between SOFR and the OFRR supports the hypothesis that the bulk of overnight loans in the FWOL database not identified in FR 2420 are indeed repos.

C.4.2 Tri-party repo adjustment to the Furfine algorithm

Banks that act as custodians, correspondents, or tri-party repo agents pose a counterparty-identification challenge for the Furfine algorithm.

As a first example, consider a financial institution—such as a money market fund—that lends reserves to a Fedwire participant P through a custodian bank C. The Furfine algorithm will interpret a payment from C to P that is matched by a return payment from P to C on the following day as an overnight loan from C to P, when in fact the loan originated from the financial institution, with C acting solely as custodian. Because we are not aware of any database that records custodian relationships of this kind, it is not possible to adjust the Furfine algorithm output to correct this type of counterparty misidentification.

As a second example, consider a loan from a Fedwire participant L to a foreign banking organization (FBO) F that does not hold a Fedwire account but operates on Fedwire indirectly through a correspondent bank C. The Furfine algorithm will interpret a payment from L to C that is matched by a return payment from C to L on the following day as an overnight loan from L to C, when in fact the loan was from L to F, with C acting as correspondent. Step 2 in Section C.4.1 corrects this type of counterparty misidentification for transactions involving FBOs that borrow fed funds or Eurodollars through correspondent banks.

As a third example, consider a tri-party repo loan from a borrower B to a lender L, in which a tri-party agent A (typically a custodian or clearing bank) clears funds between the two parties. In this case, the Furfine algorithm misclassifies the transaction as two separate loans: one from L to A, and another from A to B. This misidentification inflates A's participation rate and biases its reallocation rate toward zero.

We address this problem by implementing a tri-party repo adjustment to the Furfine algorithm. Using industry reports, we identify the main tri-party agents in the United States—namely, BNY Mellon (BONY) and JPMorgan Chase (JPM). We then search the FWOL database for pairs of overnight loans in which a tri-party agent A (either BONY or JPM) appears as an intermediary—first borrowing from a lender L, then lending to a borrower B on

the same day. We treat these two transfers as a single loan from L to B, effectively removing A from the trade. We then use the tri-party-adjusted FWOL database to recompute the participation and reallocation rates, and find $\mathcal{P}_M = 0.33$, $\mathcal{P}_S = 0.09$, $\mathcal{P}_G = 0.17$, $\mathcal{R}_F = 0.20$, $\mathcal{R}_M = -0.61$, $\mathcal{R}_S = -0.37$, and $\mathcal{R}_G = 1$. These target moments are very similar to those obtained without applying the tri-party repo filter (see Table 1), which we used to calibrate the baseline model (Section 4). This implies that our baseline results are robust to this source of counterparty misidentification.[104]

C.5 Empirical computations

C.5.1 Balances: beginning-of-day imputation

This section provides further details about the construction of beginning-of-day (BOD) balances that we discussed in Appendix A.3. The BOD balances used in the paper were obtained from the following three-step procedure for each bank:

- Step 1. We started with the end-of-day (EOD) balance for trading day $d-1$ obtained from MPOA, and calculated a "basic" measure of the BOD balance for trading day d, by adding (subtracting) the repayments received (sent) corresponding to loans extended (received) during trading day d.

- Step 2. From the "basic" measure of BOD balance calculated in Step 1, we calculated an "adjusted" measure of BOD balance by 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 LCR requirement.

- Step 3. From the "adjusted" measure of BOD balance calculated in Step 2, we calculated a measure of "unencumbered" BOD reserve balance for trading day d, by the netting predictable payments that take place during trading day d.

Next, we discuss each step in more detail.

[104] We also implement a more conservative adjustment of the Furfine algorithm in which we expand the set of potentially spurious intermediaries to include the top four custodian banks (ranked by assets under custody)—namely BONY, State Street, JPM, and Citigroup. This custodian-bank adjustment also yields target moments that are very similar to those of our baseline calibration.

Step 1: Netting repayments of previous-day loans. For each bank holding company m in our sample, we obtained the EOD balance as of 6:30 pm EST of day d from MPOA (see Section C.1), which we denote $\mathrm{a}_m^{\mathrm{eod}}(d)$. For each bank m, we used the output of the Furfine algorithm to compute the repayments to be sent and received on day d corresponding to loans originated during day $d-1$. Let $\mathrm{receive}_m(d)$ and $\mathrm{send}_m(d)$ denote the amounts of reserves that bank m will receive or send, respectively, on day d, and define the net repayment corresponding to loans originated during day $d-1$, as $\mathrm{net}_m(d) \equiv \mathrm{receive}_m(d) - \mathrm{send}_m(d)$. We then computed $\mathrm{a}_m(d) = \mathrm{a}_m^{\mathrm{eod}}(d-1) + \mathrm{net}_m(d)$, which is our "basic" measure of BOD balance for bank m on day d. Finally, we computed the BOD "basic" balance for the maintenance period h as the average of $\mathrm{a}_m(d)$ for days $d \in h$: $\mathrm{a}_m(h) = \frac{1}{N_h} \sum_{d \in h} \mathrm{a}_m(d)$, where N_h is the number of trading days in a maintenance period h.

As mentioned in Appendix A.2 (footnote 56), for the purpose of calculating the "basic" BOD balance, we treated GSEs differently than banks. In the case of a GSE, we did not only net out the repayments corresponding to loans issued on day $d-1$ (i.e., $\mathrm{net}_m(d)$), but *all transfers* sent or received during trading day d—involving *any* counterparty, not only those that meet the sample selection criteria described Section C.3. The rationale for netting all transfers that will occur during day d to obtain the GSE's "basic" BOD balance for day d is that a GSE's business model generates very predictable cash flows, 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.

Step 2: Subtracting reserve requirements. For each bank m in maintenance period h, we computed "adjusted" (excess) reserves as $\mathrm{x}_m(h) \equiv \mathrm{a}_m(h) - \underline{\mathrm{a}}_m^D(h) - \underline{\mathrm{a}}_m^L(h)$, where $\underline{\mathrm{a}}_m^D(h)$ and $\underline{\mathrm{a}}_m^L(h)$ denote the Regulation D and LCR reserve requirements, respectively. The bank-level Regulation-D requirement, $\underline{\mathrm{a}}_m^D(h)$, was provided by MPOA. The reserve requirement implied by the LCR regulation is less straightforward, as we discuss next.

As explained in Appendix B (Section B.2.1), the LCR regulation requires a bank to maintain (on a daily or monthly 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 (a day or a month, depending on the type of institution, see footnote 88) and $\mathrm{L}_m(d)$ denote the corresponding measure of outflows in the stress scenario, the LCR regulation requires $\mathrm{L}_m(d) \leq$

$\mathrm{H}_m(d)$. 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.[105] Our proposed measure of excess reserves, $\mathrm{x}_m(h) \equiv \mathrm{a}_m(h) - \underline{\mathrm{a}}_m^D(h) - \underline{\mathrm{a}}_m^L(h)$, selects the largest level of excess reserves net of the Regulation D requirement that is consistent with the LCR constraint.

For banks that are not subject to LCR regulation (such as banks with assets below \$50 bn in our sample period), we set $\underline{\mathrm{a}}_m^L(h) = 0$. Since GSEs are not subject to Regulation D or LCR regulation, we set $\underline{\mathrm{a}}_m^D(h) = \underline{\mathrm{a}}_m^L(h) = 0$ for $m \in \mathbb{B}_G$. Finally, since we only have quarterly LCR observations (see Section C.2), we imputed the same LCR-induced reserve requirement for all maintenance periods within the quarter.

Step 3: Netting predictable payments. To go from the bank-level "adjusted" measure of BOD balance calculated in Step 2, to the bank-level measure of "unencumbered" BOD reserve balance for period h, we netted (at the individual bank level) all *predictable payments* that take place during period h, as explained in Section A.3.

C.5.2 Network statistics

In this section we describe the calculations of the network statistics reported in Figure 2.[106] We begin by introducing some notation. Let v_{md}^e denote the dollar value of all loans extended, and v_{md}^r denote the dollar value of loans received, by bank m on day d. Let v_{mh}^e and v_{mh}^r denote the dollar values of loans extended and received, respectively, during the maintenance period h, i.e., $v_{mh}^e = \sum_{d \in h} v_{md}^e$ and $v_{mh}^r = \sum_{d \in h} v_{md}^r$. Finally, let $v_h = \sum_m v_{mh}^e$ denote the total dollar value of loans extended in maintenance period h. We compute the participation and reallocation values by *bank type*, $i \in \{F, M, S, G\}$, as follows.

[105] See Section B.2.1 in Appendix B for a more detailed explanation of our strategy to identify the quantity of required reserves induced by the LCR regulation.

[106] The theoretical counterparts of these computations are discussed in Appendix G (Section G.2).

Participation rate by bank type. We computed the participation rate for bank type $i \in \{F, M, S, G\}$ during maintenance period h as $\mathcal{P}_{ih} = \sum_{m \in i} \frac{v^e_{mh} + v^r_{mh}}{2v_h}$. We then computed the participation rate of bank type $i \in \{F, M, S, G\}$ in a given year as $\mathcal{P}_i = \frac{1}{N_h} \sum_h \mathcal{P}_{ih}$, where N_h denotes the number of maintenance periods in the year.

Reallocation index by bank type. We computed the reallocation index for bank type $i \in \{F, M, S, G\}$ during maintenance period h as $\mathcal{R}_{ih} = \frac{\sum_{m \in i} v^e_{mh} - \sum_{m \in i} v^r_{mh}}{\sum_{m \in i} v^e_{mh} + \sum_{m \in i} v^r_{mh}}$. We then computed the reallocation index of group i in a given year as $\mathcal{R}_i = \frac{1}{N_h} \sum_h \mathcal{R}_{ih}$, where N_h denotes the number of maintenance periods in the year.

As explained in Section A.1, the arrows from one node to another in Figure 2 represent loans extended from banks of that type to the other. The arrow width is proportional to the volume of trade between the bank types connected by the arrow. The node size is proportional to the volume of trade between banks of a given type. The arrow widths and node sizes are defined relative to the trades within a year, so they are not comparable across years. The colors of the arrows and nodes are: light blue, dark blue, light red, or dark red, depending on whether the volume-weighted average interest rate on the loans between the two types of banks, expressed as a spread over the OFR, falls in the first, second, third, or fourth quartile, respectively.

C.5.3 Kernel density estimations

We use Gaussian kernel densities to estimate the distributions of payment shocks, beginning-of-day reserve balances, and aggregate reserve-supply shocks. For the distributions of payment shocks, and the distribution of reserve-draning shocks, we set the smoothing parameter, h, using a standard "rule of thumb", namely $h = 0.9 \min\left(\hat{\sigma}, \frac{\text{IQR}}{1.34}\right) \text{n}^{-1/5}$, where n, $\hat{\sigma}$, and IQR denote the number of observations, standard deviation, and interquartile range of the sample, respectively. For the distributions of beginning-of-day reserves we use the [iterative] methodology described in Botev et al. (2010) to set the smoothing parameter (since they may be multimodal, as seen in Figures 13-16).

C.5.4 Reduced-form estimation of the reserve demand (6)

As in Section 6.3, let s_t denote the OFR–IOR spread on day t, and Q_t be a measure of the quantity of reserves at the end of day t. We estimated equation (6) using a nonlinear least-squares procedure. For each sample period, we estimated the vector of parameters, $\nu \equiv \{\underline{s}, \tilde{s}, \xi, Q_0\}$,

with $\tilde{s} \equiv \overline{s} - \underline{s}$, to solve

$$\Xi = \min_{\nu} \left\{ \sum_{t} \left(s_t - D(Q_t) \right)^2 \right\} \text{ s.t. } 0 \leq \tilde{s}, 0 \leq \xi, \tag{25}$$

where $D(\cdot)$ is as defined in equation (6). We found the solution to (25) by following two steps. In the first step, we did a thorough grid search: we set equally spaced grids for each parameter in ν, computed the hypercube combining all these grids, and then evaluated Ξ for each entry in this hypercube.[107] Let ν_{grid} be the vector or parameters that delivered the lowest value of Ξ. In the second step, we used a Nelder-Mead optimization starting from ν_{grid}.

C.5.5 A mapping between reserves of all banks and reserves of active banks

Let Q_t^D denote the quantity of *total reserves* on day t in the sample of all banks in the data (e.g., the quantity of reserves shown in Figure 20). Let Q_t^M denote the quantity of *active excess reserves* on day t that we use to calibrate our model to the year 2019 (and in the interpolation procedure described in Appendix A.6, which also uses the year 2017 as an endpoint).[108] Let $\mathbb{T}$ denote a subset of trading days and let $\mathbf{t}$ be the cardinality of this set. For any sample $\left\{Q_t^D\right\}_{t\in\mathbb{T}}$, define $\bar{Q}_{\mathbb{T}}^D \equiv \frac{1}{\mathbf{t}} \sum_{t\in\mathbb{T}} Q_t^D$. Similarly, for any sample $\left\{Q_t^M\right\}_{t\in\mathbb{T}}$, define $\bar{Q}_{\mathbb{T}}^M \equiv \frac{1}{\mathbf{t}} \sum_{t\in\mathbb{T}} Q_t^M$.

Our model output, e.g., the aggregate demand for reserves, is computed using a quantity of reserves $Q \in \mathbb{R}$ constructed with the interpolation procedure described in Section A.6, which uses $\bar{Q}_{2017}^M$ and $\bar{Q}_{2019}^M$, i.e., the average quantity of reserves in excess of LCR and Regulation D *in our subsample of active banks* for the two base years. For some exercises (e.g., the top-right panel of Figure 21) we want to show—in the same graph—the model output along with actual daily data observations of total reserves and interest rates, but the observations that we have available at a daily frequency are $\left\{Q_t^D\right\}_{t\in\mathbb{T}}$, not $\left\{Q_t^M\right\}_{t\in\mathbb{T}}$. So we need a way to "transform" each daily observation, Q_t^D, into an estimate of Q_t^M.

We adopt a transformation $\mathcal{G}$, such that $Q_t^M = \mathcal{G}\left(Q_t^D; \mathbb{T}\right)$ for all $t \in \mathbb{T}$, which satisfies two properties for any sample $\mathbb{T}$: (1) daily variation in reserves in the full sample of banks is the

[107] We used grid sizes of: 50 points for $\underline{s}$, 50 points for $\tilde{s}$, 123 points for Q_0, and 63 points for ξ. This gave a combination of 19,372,500 values for ν. The bounds for each grid were: -0.50 and 0.01 for $\underline{s}$, 0.00 and 1.00 for $\tilde{s}$, $-3 \times Q_{2019}$ and $3 \times Q_{2019}$ for Q_0, and $1e-6$ and 0.10 for ξ. We always found the the optimal value for ν well within our bounds.

[108] That is, $\{Q_t^M\}$ is the time series for the aggregate quantity of reserves for the subsample of banks that were active in interbank overnight trading during the years under study, net of Regulation D and LCR requirements, as explained in Section A.3. 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 interbank market are inconsequential.

same as daily variation in reserves in the subsample of banks, i.e., $Q_{t+1}^M - Q_t^M = Q_{t+1}^D - Q_t^D$ for all $t \in \mathbb{T}$ (this is consistent with our strategy of calibrating the slope of our model-generated reserve demand to match the liquidity effect associated with variation in the quantity of reserves of the full sample of banks); and (2) $\bar{Q}_{\mathbb{T}}^M = \mathcal{F}\left(\bar{Q}_{\mathbb{T}}^D\right)$, where $\mathcal{F}$ is a linear function that satisfies $\mathcal{F}\left(\bar{Q}_{2017}^D\right) = \bar{Q}_{2017}^M$ and $\mathcal{F}\left(\bar{Q}_{2019}^D\right) = \bar{Q}_{2017}^M$ (the subscript "2017" denotes the sample of all trading days in the year 2017, and the subscript "2019" denotes the sample of all trading days in the year 2019). For any sample $\mathbb{T}$ of trading days, we posit

$$\begin{aligned} Q_t^M &= \mathcal{G}\left(Q_t^D; \mathbb{T}\right) \\ &\equiv Q_t^D - \bar{Q}_{\mathbb{T}}^D + \bar{Q}_{\mathbb{T}}^M, \end{aligned} \tag{26}$$

where

$$\bar{Q}_{\mathbb{T}}^M \equiv \omega_{\mathbb{T}}^D \bar{Q}_{2019}^M + \left(1 - \omega_{\mathbb{T}}^D\right) \bar{Q}_{2017}^M, \tag{27}$$

with

$$\omega_{\mathbb{T}}^D \equiv \frac{\bar{Q}_{\mathbb{T}}^D - \bar{Q}_{2017}^D}{\bar{Q}_{2019}^D - \bar{Q}_{2017}^D}. \tag{28}$$

For each day $t \in \mathbb{T}$, the transformation (26) constructs Q_t^M from Q_t^D by first subtracting from Q_t^D its sample mean, $\bar{Q}_{\mathbb{T}}^D$, and then recentering the resulting quantity by adding an *imputed* sample mean, $\bar{Q}_{\mathbb{T}}^M$, corresponding to the *subset of active banks*. The imputed sample mean $\bar{Q}_{\mathbb{T}}^M$ is defined by (27) and (28) as a convex combination of $\bar{Q}_{2017}^M$ and $\bar{Q}_{2019}^M$ (the observed sample means for the subset of active banks in the baseline years 2017 and 2019).

Next, we verify that the mappings $\mathcal{G}$ and $\mathcal{F}$ defined by (26)-(28) satisfy the desired properties. First, notice that for any sample $\mathbb{T}$, (26) implies $Q_{t+1}^M - Q_t^M = Q_{t+1}^D - Q_t^D$ for all $t \in \mathbb{T}$, so property (1) is satisfied. Second, notice that (27)-(28) define a linear transformation, $\mathcal{F}$, such that $\bar{Q}_{\mathbb{T}}^M = \mathcal{F}\left(\bar{Q}_{\mathbb{T}}^D\right)$, with

$$\mathcal{F}(\bar{Q}_{\mathbb{T}}^D) \equiv \frac{\bar{Q}_{2017}^D \bar{Q}_{2019}^M - \bar{Q}_{2019}^D \bar{Q}_{2017}^M}{\bar{Q}_{2017}^D - \bar{Q}_{2019}^D} + \frac{\bar{Q}_{2017}^M - \bar{Q}_{2019}^M}{\bar{Q}_{2017}^D - \bar{Q}_{2019}^D} \bar{Q}_{\mathbb{T}}^D,$$

which satisfies the desired property (2), i.e., $\mathcal{F}\left(\bar{Q}_{2017}^D\right) = \bar{Q}_{2017}^M$ and $\mathcal{F}\left(\bar{Q}_{2019}^D\right) = \bar{Q}_{2019}^M$.

The linear mapping $\bar{Q}^D = \mathcal{F}^{-1}\left(\bar{Q}^M\right)$ from the quantity of *active excess reserves* to the quantity of *total reserves* is a reasonable approximation for relatively narrow ranges of reserve balances (e.g., for quantities of reserves between $\bar{Q}_{2019}^M$ and $\bar{Q}_{2017}^M$). However, for some of our quantitative exercises (e.g., Figure 8, Figure **??**, and the right panels of Figure 21) we want a mapping to transform values of $\bar{Q}^M$ into values of $\bar{Q}^D$ that performs well *globally* (i.e., for

quantities of reserves balances that are far from $\bar{Q}^M_{2019}$ and $\bar{Q}^M_{2017}$). For this reason, whenever a figure includes a secondary horizontal axis for *total reserves* (i.e., $\bar{Q}^D_{\mathbb{T}}$) that "translates" the *active excess reserves* (i.e., $\bar{Q}^D_{\mathbb{M}}$) on the primary axis, we obtain $\bar{Q}^D_{\mathbb{T}}$ from the following *quadratic* mapping:

$$\bar{Q}^D_{\mathbb{T}} = \mathcal{T}\left(\bar{Q}^M_{\mathbb{T}}\right) \equiv AQ^M_{\mathbb{T}} + B\left(Q^M_{\mathbb{T}}\right)^2,$$

with

$$A \equiv \frac{\left(\bar{Q}^M_{2017}\right)^2 \bar{Q}^D_{2019} - \left(\bar{Q}^M_{2019}\right)^2 \bar{Q}^D_{2017}}{\left(\bar{Q}^M_{2017} - \bar{Q}^M_{2019}\right) \bar{Q}^M_{2017}\bar{Q}^M_{2019}}$$
$$B \equiv \frac{\bar{Q}^M_{2019}\bar{Q}^D_{2017} - \bar{Q}^M_{2017}\bar{Q}^D_{2019}}{\left(\bar{Q}^M_{2017} - \bar{Q}^M_{2019}\right) \bar{Q}^M_{2017}\bar{Q}^M_{2019}}.$$

The mapping $\mathcal{T}$ satisfies $\bar{Q}^D_{2017} = \mathcal{T}\left(\bar{Q}^M_{2017}\right)$, $\bar{Q}^D_{2019} = \mathcal{T}\left(\bar{Q}^M_{2019}\right)$, and $\mathcal{T}(0) = 0$, and it is consistent with the linear mapping $\mathcal{F}^{-1}$ (as defined by (27) and (28)) in the sense that for all practical purposes, the difference between the quadratic mapping $Q^D = \mathcal{T}\left(Q^M\right)$ and the linear mapping $Q^D = \mathcal{F}^{-1}\left(Q^M\right)$ is very small for all $Q^M \in [Q^M_{2019}, Q^M_{2017}]$.[109]

[109]Notice that $Q^D = \mathcal{F}^{-1}\left(Q^M\right)$ is the secant line to the quadratic mapping $Q^D = \mathcal{T}\left(Q^M\right)$ through the points $\left(Q^M_{2019}, Q^D_{2019}\right)$ and $\left(Q^M_{2017}, Q^D_{2017}\right)$. To see that the values of the mappings $\mathcal{F}^{-1}\left(Q^M\right)$ and $\mathcal{T}\left(Q^M\right)$ are indeed very close for $Q^M \in [Q^M_{2019}, Q^M_{2017}]$, we verify that

$$\arg \max_{Q \in [Q^M_{2019}, Q^M_{2017}]} \left[\mathcal{F}^{-1}(Q) - \mathcal{T}(Q)\right] = \frac{Q^M_{2019} + Q^M_{2017}}{2} \equiv Q^*_M,$$

and

$$\frac{\mathcal{F}^{-1}\left(Q^*_M\right) - \mathcal{T}\left(Q^*_M\right)}{\mathcal{T}\left(Q^*_M\right)} = \frac{14.21}{1897.05} \approx 0.0075.$$

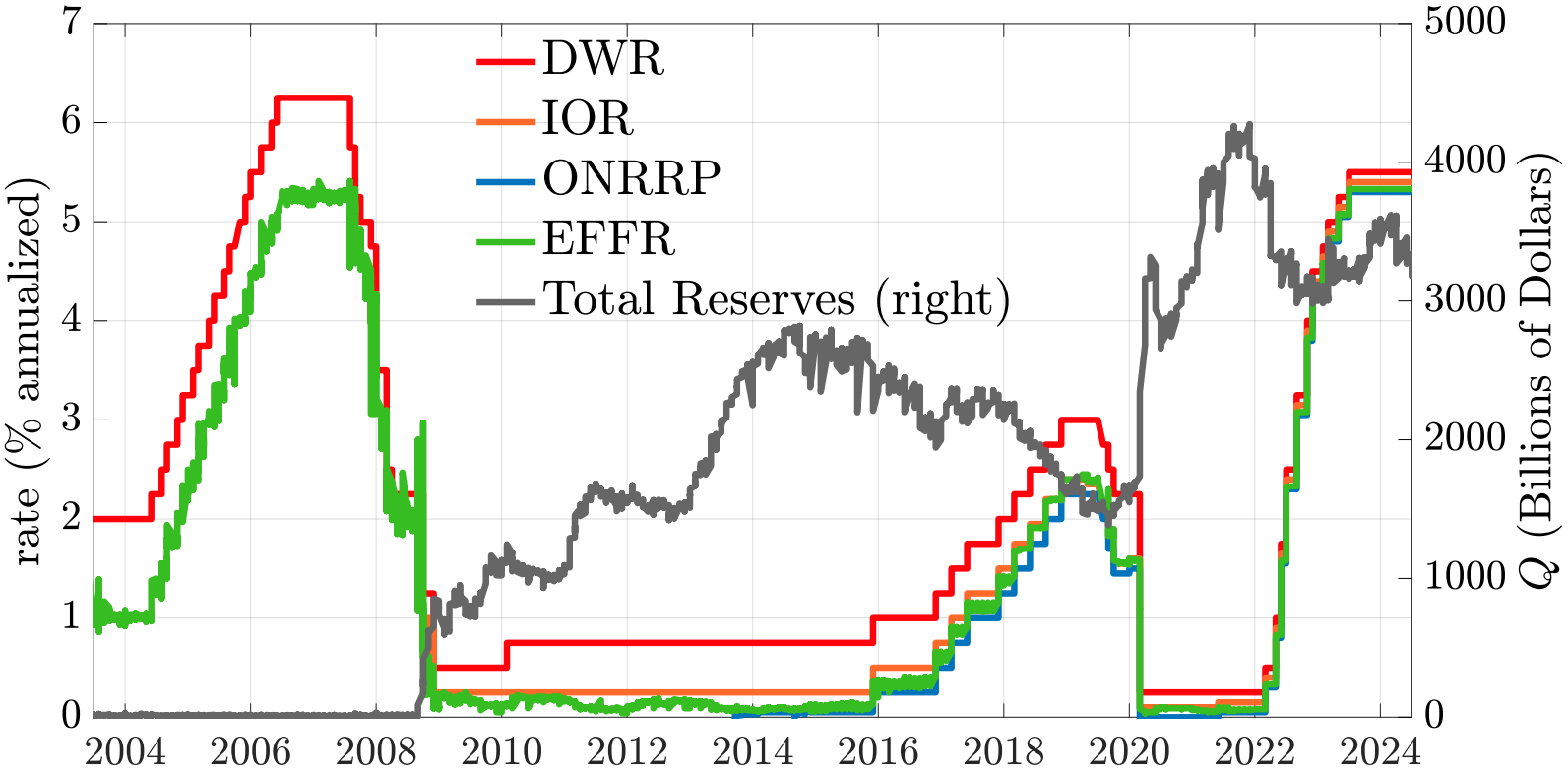

Figure 20: Time series of Total Reserves, and administered rates: Discount-Window rate (DWR), interest on reserves (IOR), and overnight reverse repo rate (ONRRP). Reserves are in billions of dollars. Rates are in percent per annum.

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".

D On reduced-form estimates of the demand for reserves

This section provides a more detailed comparison between our quantitative-theoretic approach and alternative reduced-form econometric approaches to estimating the demand for reserves. All the reduced-form estimation strategies we consider support the main lessons of Section 6.3.1.

D.1 Logistic demand

A common econometric approach is to posit a flexible reduced-form model of the demand for reserves, e.g., $s_t = D(Q_t)$ with $D(Q_t)$ given by (6), where s_t denotes an interbank rate and Q_t the quantity of reserves on day t, and to estimate the parameters $(\underline{s}, \overline{s}, \xi, Q_0)$. This generalized logistic specification is often deemed a natural choice for $D(\cdot)$ because it can fit the stylized logistic sigmoid shape of the reserve demand in the Poole (1968) model.

The top-left and bottom-left panels of Figure 21 display pairs of empirical observations of the total quantity of reserves and the corresponding OFR–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 represents 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 at each level of reserves over a wide range of reserve quantities.

The top-left panel of Figure 21 displays the fitted demand curve obtained by estimating (6) on the full sample (2017/01/20–2019/09/13) using nonlinear least squares (NLS).[110] This estimation presumes all observations lie on a single demand curve.[111] The estimated slope evaluated at the mean quantity of total reserves for the full sample (about \$1,974.69 bn) is -0.016, implying that a \$1 bn decrease in total reserves increases the OFR by 0.016 bps when total reserves are around \$2 tn.[112]

In Section 6.2, we showed that, holding the DWR–ONRRP spread constant, changes in the IOR–ONRRP spread shift and rotate the demand for reserves. This simple theoretical insight

[110] See Appendix C (Section C.5.4) for details.

[111] In a similar estimation exercise, Afonso et al. (2022, Sec. 6) 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 21 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 21.

[112] This local estimate (at about \$2 tn) is similar to the linear estimates in Appendix A (Section A.5).

implies that the data points from the sample period 2017/01/20–2019/09/13 plotted in the top-left panel of Figure 21 do not all lie on the same demand curve, contrary to the implicit assumption made when estimating (6) on the full sample without controlling for the spreads in the administered rates. The bottom-left panel of Figure 21 displays the same data points as the top-left panel, but partitioned into four subsamples, each defined 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).[113] The bottom-left panel also displays the four fitted demand curves that result from estimating (6) on each subsample.

To illustrate the pitfalls of the atheoretical demand estimation in the top-left panel of Figure 21, consider the estimation for the policy regime with an IOR–ONRRP spread of 10 bps in the bottom-left panel, and note two discrepancies relative to the top-left panel. First, the liquidity effect at about \$1,974.69 bn (the mean of *total reserves* for the full sample) is -0.0001 bps, whereas it is estimated at -0.016 bps when fitting a single demand curve to the full sample—an order of magnitude larger in absolute value.[114]

Second, suppose we want to use the estimated demand to identify the quantity of reserves that marks the transition from the "ample" to the "abundant" range, i.e., to estimate a quantity such as Q_1 in the top-right panel of Figure 1, above which the slope of the demand is virtually zero. For practical purposes, we adopt the convention that a supply of reserves Q is defined as "abundant" if reducing Q by \$1 bn increases the OFR by no more than one hundredth of a basis point. Given this definition, the demand estimated for the subsample with IOR–ONRRP = 10 bps implies $Q_1 =$ \$1,300 bn, while the demand estimated on the full sample implies $Q_1 =$ \$2,943 bn. Discrepancies of this magnitude should make central bankers wary of relying on atheoretical estimations of this kind to guide monetary-policy design.

A shortcoming of the atheoretical reduced-form econometric approach to estimating the global demand for reserves is that empirical extrapolations in sparsely observed regions of Q (e.g., very low values of Q) can be highly sensitive to our ability to identify the structural parameters that shift the aggregate demand being estimated. It is sensible to control for these "policy regimes," and the sample split in the bottom-left panel of Figure 21 is an attempt to do so. But is this the right way to split the sample? Can variation in other policy or

[113] The DWR–ONRRP spread was constant (equal to 75 bps) throughout the full sample (see Figure 20 in Appendix C).

[114] 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.

microstructure parameters shift or rotate the aggregate demand for reserves? We propose a quantitative, theory-based approach to address these questions.

The top-right and bottom-right panels of Figure 21 are the same as the top and middle panels of Figure 7, respectively. As pointed out in Section 6.3, these two estimation approaches yield very different results. To illustrate this, 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 of active excess reserves (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 OFR–IOR spread.[115]

In contrast, our theory estimates the steepest point on the demand at about $500 bn (or about $662 bn of total reserves), and predicts that reductions in the supply of reserves start to cause significant increases in the OFR–IOR spread for levels of reserves roughly below $700 of active excess reserves (below $1,064 bn of total reserves). For the reduced-form approach, the extrapolation to out-of-sample levels of Q is essentially driven by the assumed functional form. In contrast, our quantitative-theoretic extrapolation is based on the explicit equilibrium borrowing-and-lending activity that underlies 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 interbank market.

D.2 Semi-log demand

Another common reduced-form estimation strategy is to replace (6) with $D(Q_t)$ given by (7), and estimate the semi-log specification $s_t = D(Q_t)$ using ordinary least squares. Figure 22, which is analogous to Figure 21, contrasts the result from this approach with our quantitative-theoretic estimation.

First, by comparing the top-left and bottom-left panels of Figure 22 we see that—as in the logistic-demand estimation—the global estimates of the demand for reserves change substantially when incorporating the minimal theoretical insight that changes in the IOR–ONRRP spread act like demand shifters.

Second, even after we control by IOR–ONRRP regime, as we do in the bottom-left and bottom-right panels of the figure, the estimated demands are still quite different from our

[115] Note that the reduced-form demands estimated with active excess reserves (reported in the bottom-right panel of Figure 21) are very similar to the ones estimated with total reserves (reported in the bottom-left panel).

quantitative-theoretic estimates. To illustrate, compare the quantitative-theoretic estimate in the top-right panel with the reduced-form estimate in the bottom-right panel for the sample with IOR–ONRRP spread equal to 10 bps. According to the former, the slope of the demand becomes flat somewhere above \$1.3 tn of total reserves, while the slope of the latter remains positive even if total reserves exceed \$2.5 tn. The demands also differ for relatively low levels of reserves: the model-generated demand becomes quite steep at about \$600 bn of total reserves, but then flattens at about \$340 bn. In contrast, the slope of the semi-log estimate increases exponentially as total reserves decrease, and eventually becomes implausibly large.

D.3 Other reduced-form specifications

In this section we consider variants of the logistic and semi-log specifications. Section D.3.1 tries to improve on the logistic model by imposing theoretically-motivated constraints on the estimation. Section D.3.2 considers a variant of the semi-log specification that controls for bank deposits.

D.3.1 Logistic demand estimation with constraints motivated by theory

The first reduced-form approach we discussed in Section 6.3 consisted of estimating the parameters $(\underline{s}, \overline{s}, \xi, Q_0)$ in (6) by nonlinear least squares. The estimated demand fits the data well, but performs poorly for out-of-sample levels of reserves: Notably, the estimated demand predicts the OFR–IOR spread would remain unchanged if total reserves were drained from \$1 tn to zero.

In this section we try to improve upon this model by imposing two constraints on the estimation that are grounded on elementary theory. Specifically, we redo the NLS estimation of (6) but imposing that $\overline{s}$ should equal the largest value of $\bar{\iota}_w - \iota_r$ in the relevant sample, and that $\underline{s}$ should equal the lowest value of $\bar{\iota}_o - \iota_r$ in the relevant sample (with $\bar{\iota}_w$, $\bar{\iota}_o$, and ι_r as defined in Section 4). The results are reported in Figure 23, which is analogous of Figure 21.

The global fit of this reduced-form approach looks somewhat more credible than the unconstrained version in Figure 21; at least the OFR–IOR spread now rises as the quantity of reserves falls below \$1 tn. However, even after we control by IOR–ONRRP regime, as we do in the bottom-left and bottom-right panels of the figure, the estimated demands are still quite different from our quantitative-theoretic estimates. To illustrate, compare the quantitative-theoretic estimate in the top-right panel with the reduced-form estimate in the bottom-right

panel for the sample with IOR–ONRRP spread equal to 10 bps. In the former, the slope of the demand becomes flat somewhere above \$1.3 tn of total reserves, while the slope of the latter remains positive even if total reserves exceed \$2.5 tn. The behavior is also quite different for relatively low levels of reserves: the model-generated demand becomes quite steep at about \$600 bn of total reserves, while the slope of the reduced-form estimate does not vary much with the quantity of reserves (even as the quantity of total reserves approaches zero).

D.3.2 The López-Salido and Vissing-Jorgensen (2023) semi-log specification

In this section we consider the reduced-form specification for the demand for reserves proposed in López-Salido and Vissing-Jorgensen (2023), who assume

$$s_t = a + b \ln(Q_t) + c \ln(D_t), \tag{29}$$

where s_t denotes the OFR–IOR spread in period t, Q_t denotes the aggregate quantity of reserves in period t, and D_t is a measure of bank deposits in period t.[116] We estimate the parameters a, b, and c by OLS; the results are reported in Figure 24. The demand estimates are very similar to the ones we obtained in Section D.2 and reported in Figure 22, so the main points we made about the specification (7) also hold for (29).

[116] The baseline measure of D_t in López-Salido and Vissing-Jorgensen (2023) is "DPSACBW027SBOG" (*Deposits, All Commercial Banks*) from https://fred.stlouisfed.org.

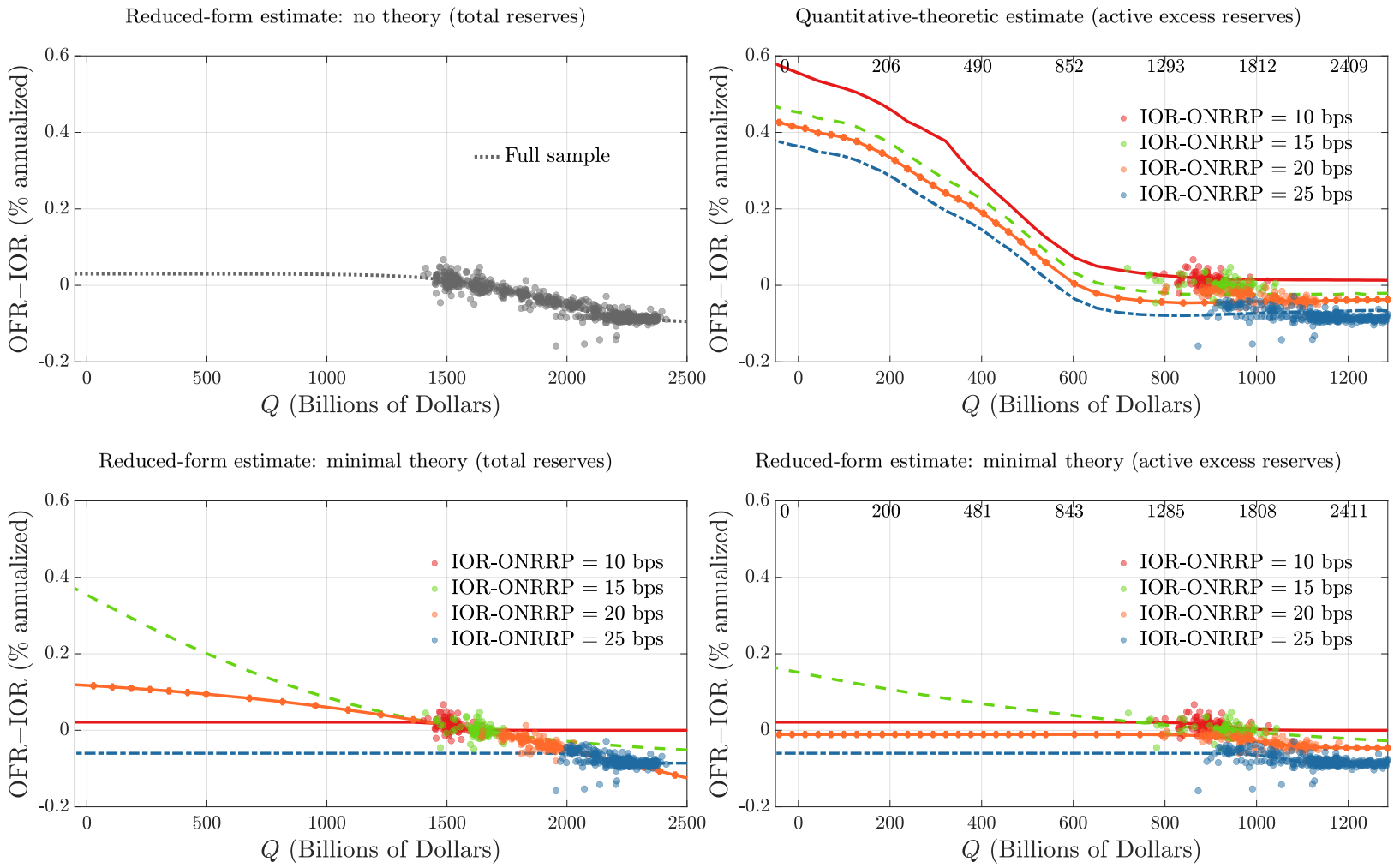

Figure 21: Reserve demand estimation: model vs. NLS fit of $s_t = \underline{s} + \frac{\bar{s}-\underline{s}}{1+e^{(Q_t-Q_0)\xi}}$.

Notes: In each panel: vertical axis is OFR–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 NLS fit of (6) on full sample. Bottom-left panel: total reserves and NLS fits of (6) 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 NLS fits of (6) on each subsample. Secondary horizontal axis (above top-left and bottom-left panels) translates active excess reserves into total reserves.

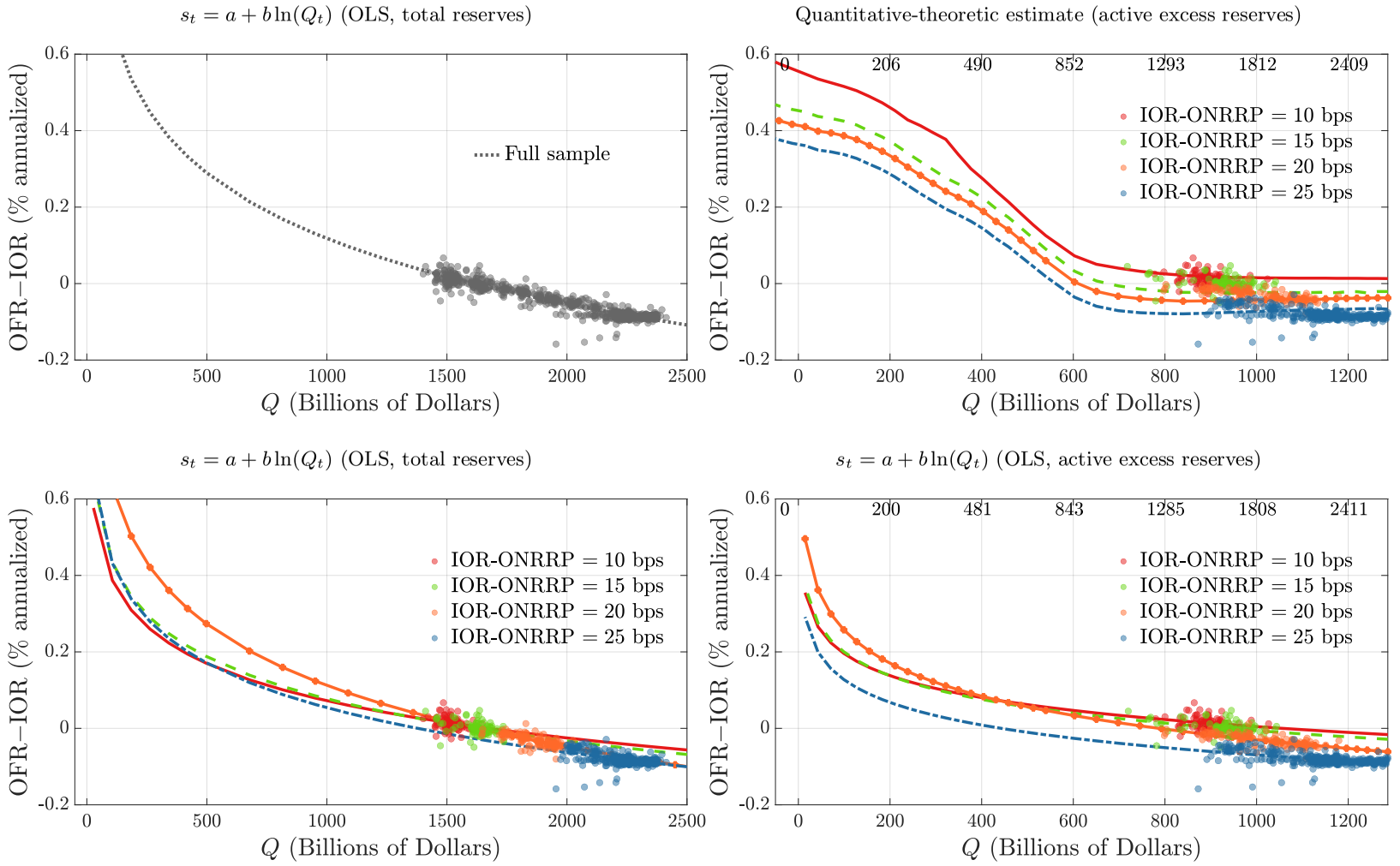

Figure 22: Reserve demand estimation: model vs. OLS fit of $s_t = a + b\ln(Q_t)$.

Notes: In each panel: vertical axis is OFR–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 OLS fit of (7) on full sample. Bottom-left panel: total reserves and OLS fits of (7) 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 OLS fits of (7) on each subsample. Secondary horizontal axis (above top-left and bottom-left panels) translates active excess reserves into total reserves.

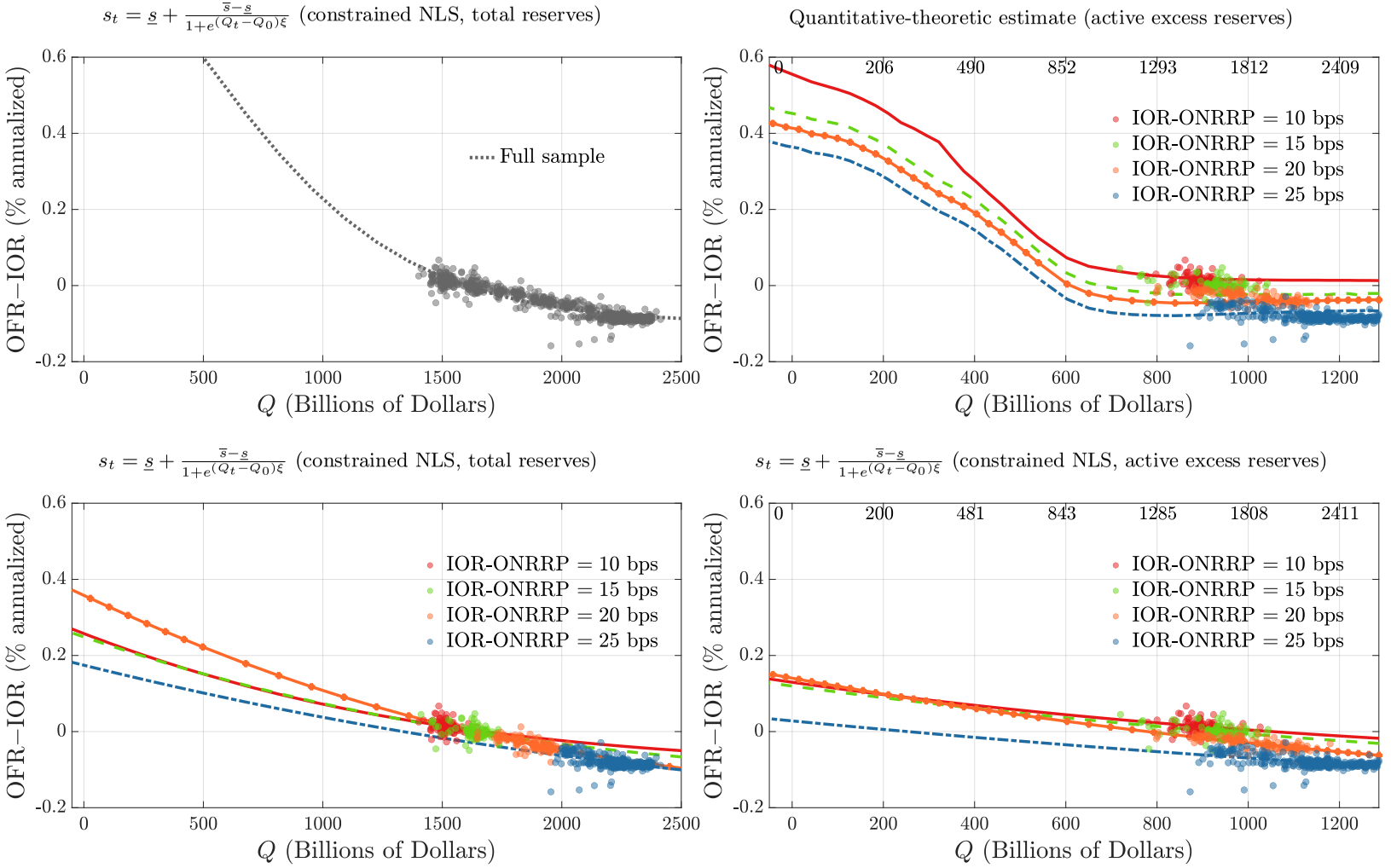

Figure 23: Reserve demand estimation: model vs. NLS fit of (6) with theoretically motivated constraints on $\overline{s}$ and $\underline{s}$.

Notes: In each panel: vertical axis is OFR–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 constrained NLS fit of (6) on full sample. Bottom-left panel: total reserves and constrained NLS fits of (6) 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 constrained NLS fits of (6) on each subsample. Secondary horizontal axis (above top-left and bottom-left panels) translates active excess reserves into total reserves.

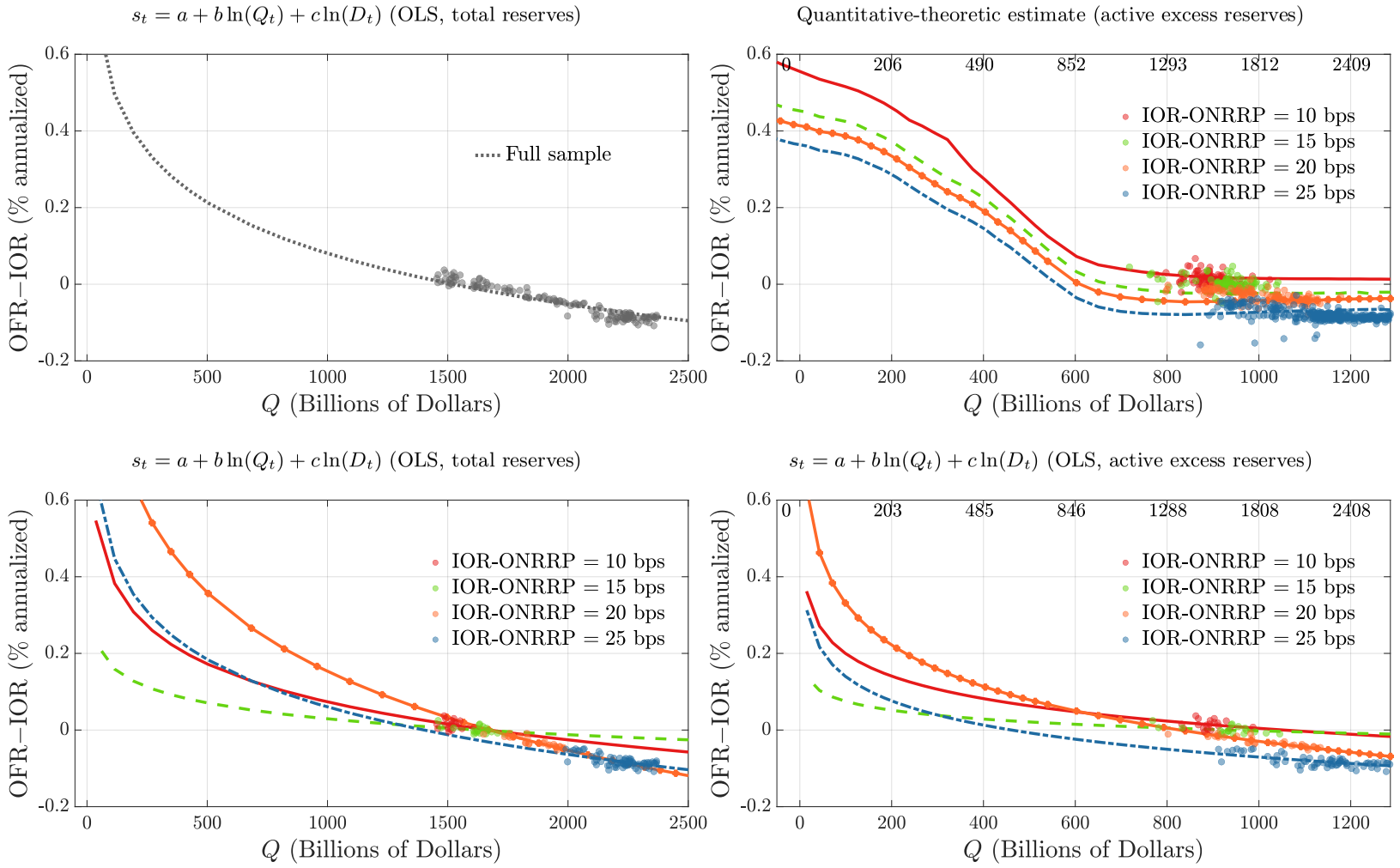

Figure 24: Reserve demand estimation: model vs. OLS fit of (29).

Notes: In each panel: vertical axis is OFR–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 OLS fit of (29) on full sample. Bottom-left panel: total reserves and OLS fits of (29) 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 OLS fits of (29) on each subsample. Secondary horizontal axis (above top-left and bottom-left panels) translates active excess reserves into total reserves.

E Robustness and alternative calibrations

In this section we explore alternative calibrations and verify the robustness of our quantitative-theoretic results. Section E.1 considers an alternative calibration strategy that uses different definitions of bank types, and targets empirical moments from Fedwire data adjusted using FR 2420 filings (as discussed in Section C.4.1). Section E.2 validates the shape of the reserve demand generated by the model with data from the post-COVID period with abundant reserves. Section E.3 recalibrates the model and estimates the reserve demand for the pre-GFC monetary-policy framework.

E.1 Alternative calibration: an FBO bank type and FR 2420 corrections

In this section, we consider an alternative calibration strategy that relies on different definitions of bank types and targets empirical moments from Fedwire data cross-referenced with FR 2420 filings (see Section C.4.1).

First, we compute the calibration targets using the payments and overnight loans data obtained from the two-step procedure described in Section C.4.1. Second, we sort banks into types using criteria that differ from those in Section 3: Type F comprises the eight domestic banks with the highest participation rates; Type M includes all FBOs with participation rates above 0.005; Type S includes all remaining domestic banks and FBOs not assigned to groups F or M; and Type G continues to denote the GSEs, as in the baseline calibration.

Table 2 reports the targeted moments, the corresponding theoretical moments, and the parameter values implied by this alternative 2019 calibration. Figure 26 plots the reserve demand generated by the model under this calibration. The result is very similar to the baseline calibration shown in the top-left panel of Figure 8.

E.2 Post-COVID policy regime

In this section, we extend our time-series validation exercise using data from 2023, a representative post-COVID year with abundant reserves, during which the DWR–ONRRP spread averaged 20 basis points and the IOR–ONRRP spread 10 basis points (per annum).

Figure 26 plots five reserve demand curves. The first four are implied by the quantitative model under the baseline calibration, each corresponding to a different IOR–ONRRP spread (10, 15, 20, and 25 basis points). The fifth curve, labeled "2023 data," corresponds to the baseline

calibration with a recalibrated value of $\iota_\ell = -0.00055/360$ (the baseline had $\iota_\ell = 0.00022/360$). This adjustment matches the average IOR–ONRRP spread observed in 2023. It may reflect that Regulation D was still operative in 2019 but not in 2023, and possibly that inflation was lower in 2019——two factors outside the model that could have shifted the demand for reserves downward.

The figure shows that the model-generated demand fits the data well across the full range of reserves observed in the extended sample. Moreover, the shape of the curve is very similar to that estimated in Section 6.3, indicating that our estimates of the reserve quantity consistent with an "ample" regime are robust to the inclusion of this additional data.

E.3 Pre-GFC policy regime

Our quantitative analysis in the body of the paper focuses on the current post-GFC monetary-policy framework. For completeness, and because the pre-GFC period is of historical interest, in this section we also study the pre-GFC framework. The pre-GFC and post-GFC frameworks differ in two ways. First, the quantity of excess reserves was close to zero in the former, but is very large in the latter. Second, as discussed in Section A, regulations introduced after the GFC have affected banks' payoffs from interbank trading. For this reason, in this section we recalibrate the model for a base year before the GFC, which we choose to be 2006.[117]

E.3.1 Calibration

We set ι_w to match the prevailing DWR, and $\iota_o = 0$ (since there was no ONRRP facility in 2006). The remaining nine parameters, ι_r and $\{\beta_i, \kappa_i\}_{i\in\mathbb{N}}$, are calibrated so that the equilibrium of the model matches the following nine empirical moments: (i) effective fed funds rate[118];

[117] Our main motive for recalibrating the model is that the trading network, which in our theory is represented by the parameters $\{\beta_i, \kappa_i\}_{i\in\mathbb{N}}$, may not be stable across policy regimes. For example, it is reasonable to imagine that the trading patterns represented by the type-specific meeting rates may change in response to regulatory constraints, in particular those post-GFC regulations that increased the cost of borrowing, and therefore the cost of intermediating reserves. We use 2006 as the baseline year for the pre-GFC period for two reasons. First, policy rates and total reserves remained stable for most of that year, and it was the last "normal" year before the GFC that spurred the policy interventions that changed the landscape of the interbank market. Second, the 2006 calibration will allow us to assess the model fit in a pre-GFC-regulation environment, and it will also allow us to test the quantitative predictions of the theory as we vary the level of aggregate excess reserves from near zero (the level they had during 2006) to \$2.689 tn (the level they reached for all the banks in our sample in 2014, which was the last pre-GFC-regulation year).

[118] Our calibration strategy uses the EFFR as a calibration target unless the Federal Reserve pays interest on reserves (IOR) in the base year, in which case we simply set ι_r to match the IOR. For example, the IOR was 2.35% per annum in May-July 2019, so we set $\iota_r = 0.0235/360$ in the 2019 calibration. The Federal Reserve did

(ii)-(v) reallocation indices $\{\mathcal{R}_i\}_{i\in\mathbb{N}}$ (as defined in Section A.1); (vi)-(viii) participation rates $\{\mathcal{P}_i\}_{i\in\mathbb{N}\setminus\{F\}}$ (as defined in Section A.1)[119]; (ix) empirical estimates of the "liquidity effect" (at the average level of aggregate reserves outstanding in the base year, as reported in Section A.5).

Table 3 reports the parameter values, empirical targeted moments, and the corresponding theoretical moments for the 2006 calibration. Banks of type F, M, and S, accounted for about 0.5%, 3%, and 95%, of all the institutions that were active in the interbank market in 2006, respectively.[120] 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, $\lambda_F = 0.901$ implies a bank of type F receives a payment shock approximately every 1.1 seconds, on average. Similarly, $\lambda_M = 0.402$ implies a bank of type M receives a payment shock approximately every 2.5 seconds, and $\lambda_S = 0.007$ implies a bank of type S receives a payment shock approximately every 2.38 minutes, on average. The rate ι_w corresponds to a DWR equal to 6.25% per annum, which was in effect in the second half of 2006. The calibrated value of ι_r is 4.81% per annum.[121]

The frequency of trade, β_i, is the probability a bank of type i contacts a trading partner during a 42-second time interval. Thus, the calibrated values $\{\beta_i\}_{i\in\mathbb{N}}$ for 2006 imply that banks of type F, M, S, and G trade approximately every 1.75 minutes, 8 minutes, 20 minutes, and 3.5 minutes, respectively. The calibration also ensures that, when computed in the neighborhood of zero excess reserves, the magnitude of the "liquidity effect" in the theory is within the range of the empirical estimates reported in Section A.5 (i.e., about a 1.7 bp increase in the fed funds rate per \$1bn reduction in the aggregate quantity of reserves).[122] The borrowing costs $\{\kappa_i\}_{i\in\mathbb{N}}$,

not pay IOR before October 9, 2008, so in the 2006 calibration we regard ι_r as a proxy for a bank's unmodelled opportunity cost of lending reserves in the interbank market, and calibrate it internally so that the average (volume-weighted) interest rate in the model is equal to 5.25% per annum, which was the EFFR rate prevailing during the second half of 2006.

[119]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 types, since $\sum_{i\in\mathbb{N}} \mathcal{P}_i = 1$.

[120]The main change in the bank population between 2006 and 2019 is the reduction in the total number of active banks in our sample, mostly due to the fact that almost half of the banks of type S that were interbank market participants in 2006 did not trade during 2019.

[121]For comparison, 4.81% per annum is the 0.5 percentile of the volume-weighted distribution of rates observed in the second half of 2006. That is, only half of one percent of the loans traded in the second semester of 2006 had a rate below 4.81%, so we regard 4.81% as a reasonable proxy for the unmodelled opportunity cost of an alternative use of reserves. We focus on July–December because in that period the administered rates (i.e., the Discount-Window rate and the EFFR target) were constant and equal to the rates targeted in the 2006 calibration (the administered rates had been gradually increasing in the first half of 2006).

[122]Figure 27 shows the magnitude of the liquidity effect in the model calibrated to 2006 (extracting reserves using the procedure described in Section A.6), as well as the confidence bands for the estimates from Carpenter

which proxy for institutional and regulatory considerations that affect banks' incentives to borrow, are nill for banks of type F, M, and S in the 2006 calibration.[123]

E.3.2 Validation

In this section we report the model fit of empirical observations that were not targeted in the calibration. We organize the material in two sections: the first focuses on the cross-sectional distribution of loan rates, and the second on the main features of the interbank trading network.

Distribution of interest rates Figure 28 shows the empirical and theoretical cumulative distribution functions of bilateral interest rates in the year 2006 (expressed in percent per annum).[124] The model delivers a reasonable fit for the distribution of bilateral interest rates, which was not a calibration target.

Interbank trading network Figure 29 shows the empirical interbank trading network for the year 2006 (top panel) and the corresponding trading network generated by the model for the 2006 calibration (bottom panel). As explained in Section A.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 lending across and within bank types, as well as on the average interest rates on the underlying loans.[125]

The theoretical network matches several characteristics of the empirical one. For example, it replicates quite well the direction and volume of the loans between bank types (represented by the direction and width of the arrows between the nodes). In this regard, one shortcoming of

and Demiralp (2006) reported in Section A.5. The model-generated liquidity effect is within the range of empirical estimates.

[123] In every calibration the value of κ_G is set large enough to match the observation that GSEs essentially do not borrow in the interbank market.

[124] The empirical interest rates for 2006 are from the sample period July–December because throughout that period the Discount-Window rate and the EFFR target were constant and equal to the rates targeted in the 2006 calibration. To obtain the equilibrium rates for 2006, the model is calibrated as in Table 3.

[125] In comparing the top and bottom panels of Figure 29, note that the positions of the four nodes in $\mathcal{R}_i$-$\mathcal{P}_i$ space have been used as calibration targets, the remaining statistics that shape these network representations were not targeted in the calibration. This includes the node sizes (each of which is proportional to the volume of trade between banks of a given type), the direction of each arrow (which indicates which bank type lends), the width of each arrow (which is proportional to the volume of trade between the bank types connected by the arrow), the colors of the arrows and nodes (which 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).

the model is that it underpredicts the volume of loans within bank types S and M. The model is consistent with the empirical facts that banks of type S lend to each other at relatively high rates, while banks of type F can borrow at relatively low rates from banks of type M, S, and GSEs. In terms of shortcomings, the model predicts that banks of type S borrow at relatively high rates from GSEs, that loans between banks of type F carry relatively low rates, and that loans between banks of type M carry relatively high rates, as do loans from type F to type M, but these predictions do not match the empirical patterns.

E.3.3 Aggregate demand for reserves

Consider the model calibrated to the year 2006, as described in Table 3, but with $\iota_w = 0.0075/365$ and $\iota_r = 0.0025/365$, to match the DWR and IOR in the year 2014. Then, using the notation introduced in Section A.6, let $\mathrm{Y}_0 = 2006$ and $\mathrm{Y}_1 = 2014$, i.e., Y_0 and Y_1 represent the years 2006, and 2014, respectively, with $\bar{n}^i_{2014}$ and $\bar{F}^i_{2014}$ given by the estimates reported in Section A.3. Construct a grid, $\mathbb{G} \subset \mathbb{R}$ for ω, and for each $\omega \in \mathbb{G}$, use the interpolation procedure described by (10) and (11) 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 $(\bar{n}^i_{\mathrm{Y}_\omega}, \bar{F}^i_{\mathrm{Y}_\omega})$ of elements of this sample, use the model to compute the corresponding equilibrium value-weighted interest rate, which we denote $\iota^*_{\mathrm{Y}_\omega}$, and let $Q_{\mathrm{Y}_\omega} \equiv \sum_{i\in\mathbb{N}} \bar{n}^i_{\mathrm{Y}_\omega} \int a d\bar{F}^i_{\mathrm{Y}_\omega}(a)$. This procedure delivers a sample of pairs, $\left\{(Q_{\mathrm{Y}_\omega}, \iota^*_{\mathrm{Y}_\omega})\right\}_{\omega\in\mathbb{G}}$, which we represent with the mapping $\iota^*_{\mathrm{Y}_\omega} = \mathcal{D}(Q_{\mathrm{Y}_\omega}; \Pi)$. This mapping, which we interpret as the aggregate demand for reserves generated by the theory, is shown in Figure 30.[126]

[126] We use 2006 as an endpoint for our interpolation procedure since it was the last year of the scarce-reserve regime that prevailed until the GFC. We use 2014 as the other endpoint because it is the year when the quantity of reserves achieved its maximum historical level of the pre-2020 era. By varying ω on $[0, 1]$ we can use (12) to span any aggregate level of excess reserves between 0 (roughly the pre-GFC level prevailing in 2006) and \$2.7 tn (roughly the level achieved in 2014).

Parameter	Target	Moment	
		Data	Model
$n_F = 0.021$	proportion of financial institutions of type F	0.021	0.021
$n_M = 0.051$	proportion of financial institutions of type M	0.051	0.051
$n_S = 0.900$	proportion of financial institutions of type S	0.900	0.900
$n_G = 0.028$	proportion of financial institutions of type G	0.028	0.028
$\lambda_F = 0.873$	bank-level share of unexpected payments per second for type F	0.873	0.873
$\lambda_M = 0.024$	bank-level share of unexpected payments per second for type M	0.024	0.024
$\lambda_S = 0.013$	bank-level share of unexpected payments per second for type S	0.013	0.013
$\lambda_G = 0$	bank-level share of unexpected payments per second for type G	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
$\iota_\ell = 0$	average value-weighted rate	0.0237/360	0.0241/360
$\iota_s = 0.00752/360$	estimated liquidity effect for 2019 (bps per \$1 bn decrease in reserves)	$\in [-0.0145, -0.0035]$	-0.0128
$\underline{\theta} = 0.05$	conditional (below the IOR) average value-weighted overnight rate	0.0229/360	0.0227/360
$\beta_F = 0.1050$	number of loans of financial institutions of type F relative to average	14	16
$\beta_M = 0.0030$	participation rate of financial institutions of type M (i.e., $\mathcal{P}_M$)	0.25	0.26
$\beta_S = 0.0009$	participation rate of financial institutions of type S (i.e., $\mathcal{P}_S$)	0.13	0.17
$\beta_G = 0.0007$	participation rate of financial institutions of type G (i.e., $\mathcal{P}_G$)	0.17	0.20
$\kappa_F = 0.041e\text{-}3$	reallocation index of financial institutions of type F (i.e., $\mathcal{R}_F$)	0.29	0.30
$\kappa_M = 0$	reallocation index of financial institutions of type M (i.e., $\mathcal{R}_M$)	-0.91	-0.82
$\kappa_S = 0.002e\text{-}3$	reallocation index of financial institutions of type S (i.e., $\mathcal{R}_S$)	-0.56	-0.55
$\kappa_G = 1.25e\text{-}3$	reallocation index of financial institutions of type G (i.e., $\mathcal{R}_G$)	1	1

Table 2: Alternative 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, $r = 0$, $\mathbb{N} = \{F, M, S, G\}$ (as discussed in Section A.1), $\theta_{i,j} = 1/2$ for all $i, j \in \mathbb{N} \setminus \{G\}$, $\theta_{ij} = \underline{\theta}$ if $i \in \{G\}$ and $j \in \mathbb{N} \setminus \{G\}$, $\{G_{ij}\}_{i,j\in\mathbb{N}}$ are estimated as described in Section A.2, $\{F_0^i\}_{i\in\mathbb{N}}$ are estimated as described in Section A.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 A.6.

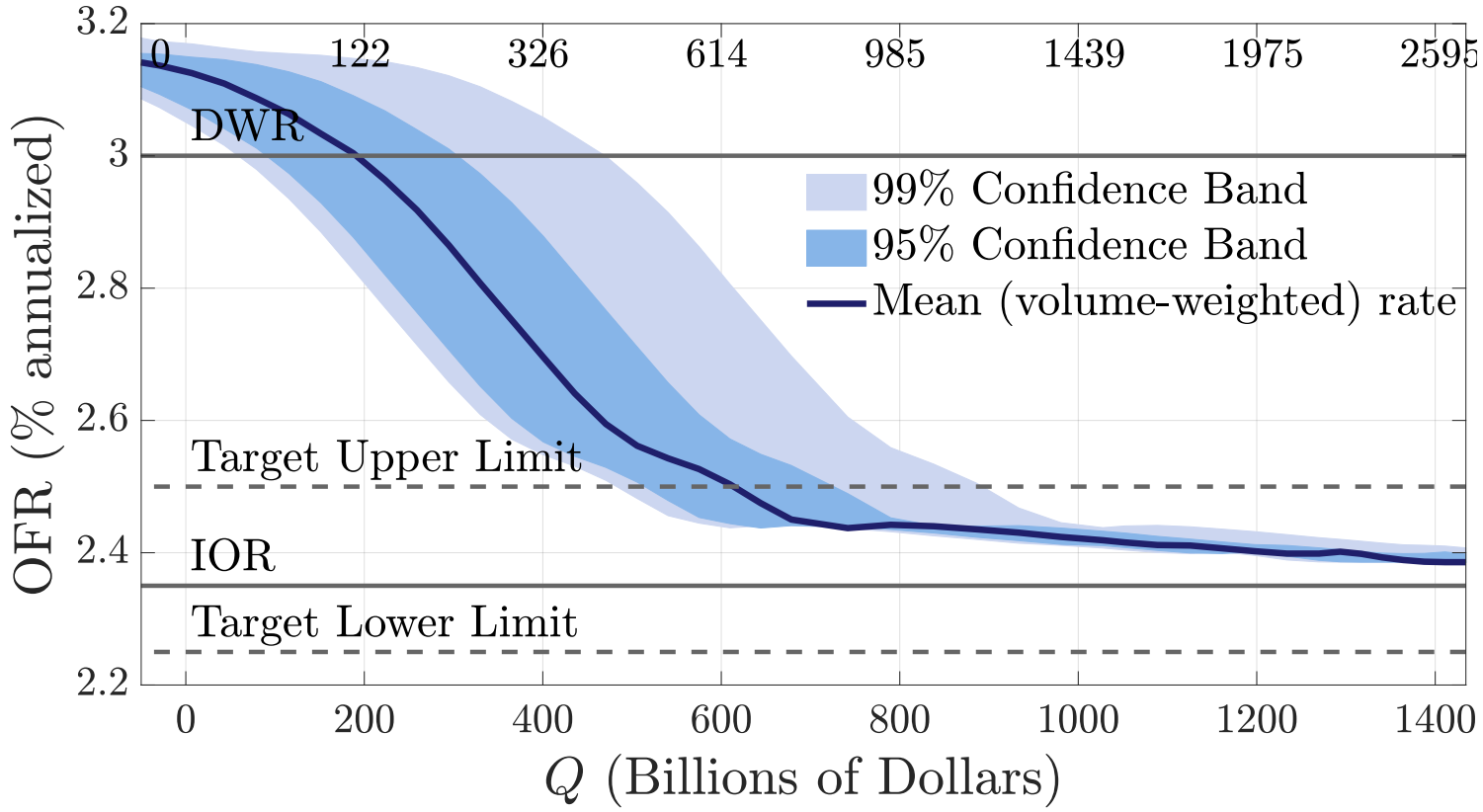

Figure 25: Monetary confidence bands for the alternative 2019 calibration.

Notes: The curve labeled "Mean (volume-weighted) rate" is the theoretical money demand, $\mathcal{D}(Q)$ corresponding to the alternative 2019 calibration (Table 2). 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-supply 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.

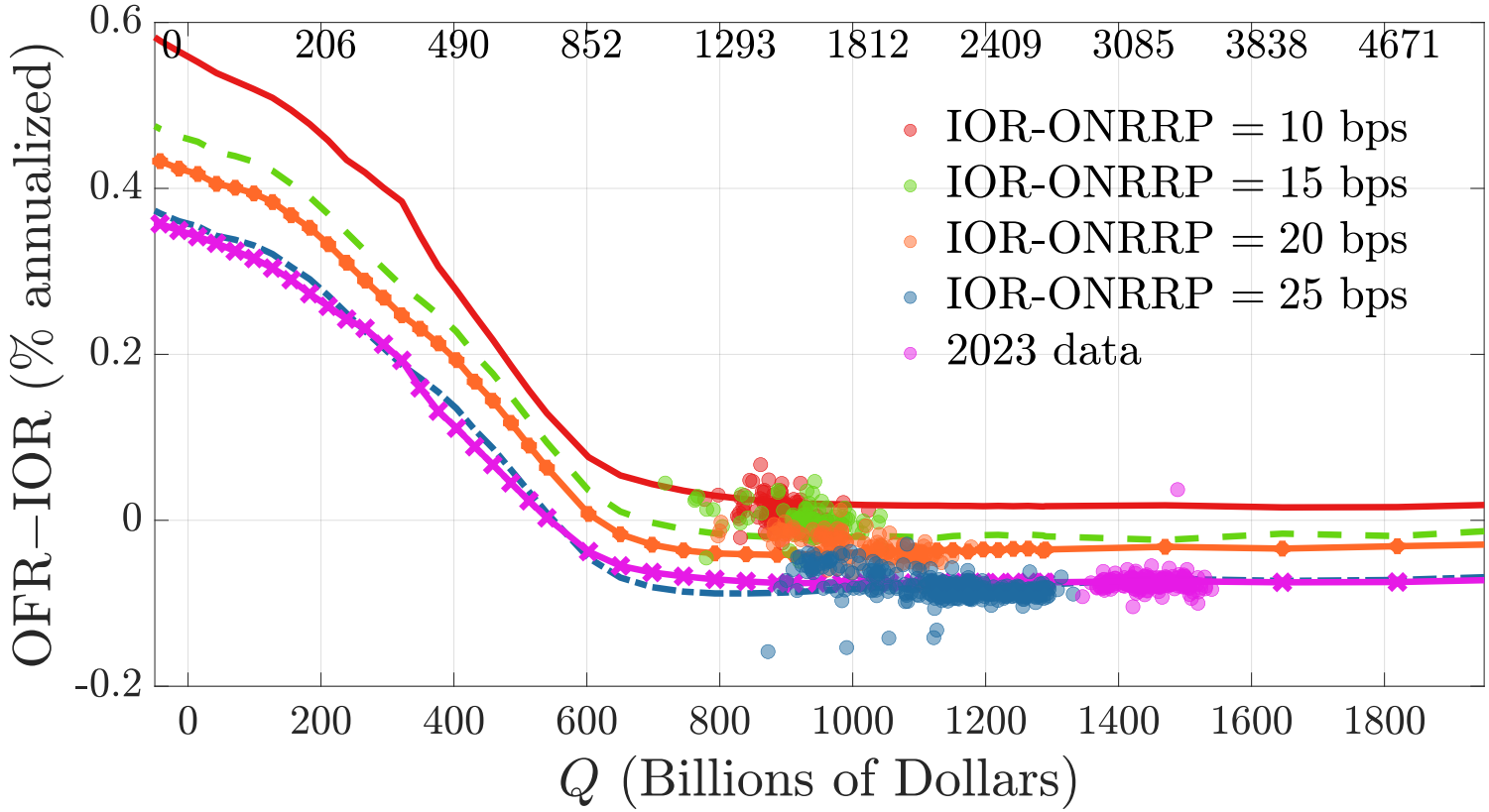

Figure 26: Reserve demand estimation: model fit for 2023.

Notes: Active excess reserves and aggregate demands implied by the theory for each IOR–ONRRP spread (all other parameters as in the baseline calibration, except ι_ℓ for the demand labeled "2023 data"). Secondary horizontal axis (above top-left and bottom-left panels) translates active excess reserves into total reserves.

Parameter	Target	Moment	
		Data	Model
$n_F = 0.005$	proportion of financial institutions of type F	4/754	0.005
$n_M = 0.029$	proportion of financial institutions of type M	22/754	0.029
$n_S = 0.950$	proportion of financial institutions of type S	716/754	0.950
$n_G = 0.016$	proportion of financial institutions of type G	12/754	0.016
$\lambda_F = 0.901$	bank-level share of unexpected payments per second for type F	0.901	0.901
$\lambda_M = 0.402$	bank-level share of unexpected payments per second for type M	0.402	0.402
$\lambda_S = 0.007$	bank-level share of unexpected payments per second for type S	0.007	0.007
$\lambda_G = 0$	bank-level share of unexpected payments per second for type G	0	0
$\iota_w = 0.0625/360$	Discount-Window rate (primary credit, 6.25% per annum)	0.0625/360	0.0625/360
$\iota_r = 0.0481/360$	effective fed funds rate (5.25% per annum)	0.0525/360	0.0525/360
$\beta_F = 0.401$	estimated liquidity effect around zero excess reserves (bps per \$1 bn)	$\in [1, 3]$	1.7
$\beta_M = 0.089$	participation rate of financial institutions of type M (i.e., $\mathcal{P}_M$)	0.27	0.27
$\beta_S = 0.033$	participation rate of financial institutions of type S (i.e., $\mathcal{P}_S$)	0.18	0.19
$\beta_G = 0.200$	participation rate of financial institutions of type G (i.e., $\mathcal{P}_G$)	0.07	0.07
$\kappa_F = 0$	reallocation index of financial institutions of type F (i.e., $\mathcal{R}_F$)	0.068	0.064
$\kappa_M = 0$	reallocation index of financial institutions of type M (i.e., $\mathcal{R}_M$)	-0.385	-0.268
$\kappa_S = 0$	reallocation index of financial institutions of type S (i.e., $\mathcal{R}_S$)	-0.268	-0.126
$\kappa_G = 1.25e\text{-}3$	reallocation index of financial institutions of type G (i.e., $\mathcal{P}_G$)	0.995	1

Table 3: Calibration for the year 2006.

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, $r = 0$, $\mathbb{N} = \{F, M, S, G\}$ (as discussed in Section A.1), $\theta_i = 1/2$ for all $i \in \mathbb{N}$, $\{G_{ij}\}_{i,j\in\mathbb{N}}$ are estimated as described in Section A.2, $\{F_0^i\}_{i\in\mathbb{N}}$ are estimated as described in Section A.3, $u_i = 0$ for all $i \in \mathbb{N}$, $\{U_i\}_{i\in\mathbb{N}}$ are as in Section 4).

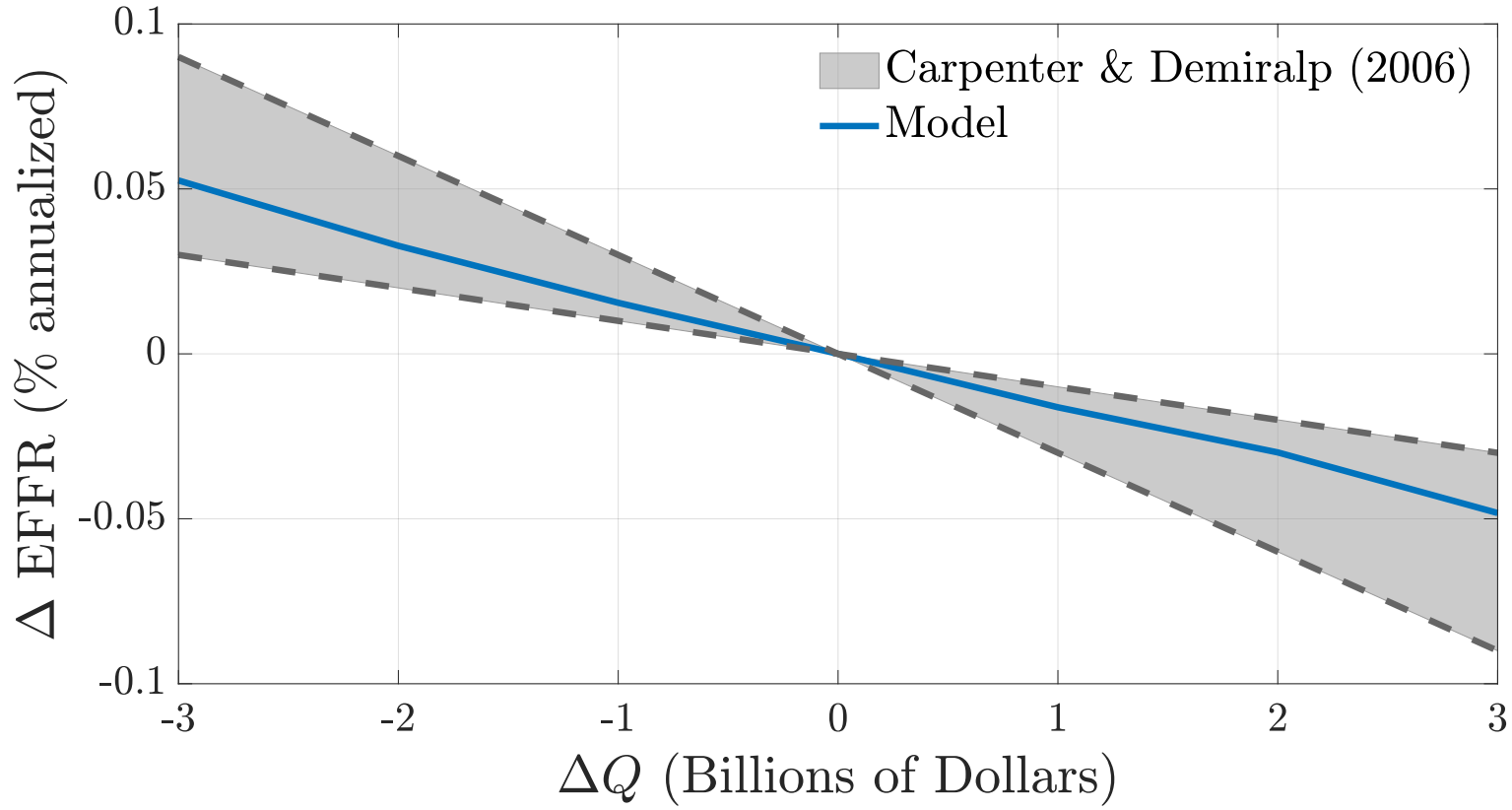

Figure 27: Liquidity effect: model and empirical estimates for the year 2006.

Notes: Rates in the vertical axis are in percent per annum. The shaded area represents the 95% confidence interval for the point estimate of the liquidity effect from Carpenter and Demiralp (2006). The solid line is the change in the equilibrium average value-weighted rate implied by the theory in response to changes in the total quantity of reserves (starting from the quantity of reserves corresponding to the 2006 calibration, and extracting reserves using the procedure described in Section A.6.)

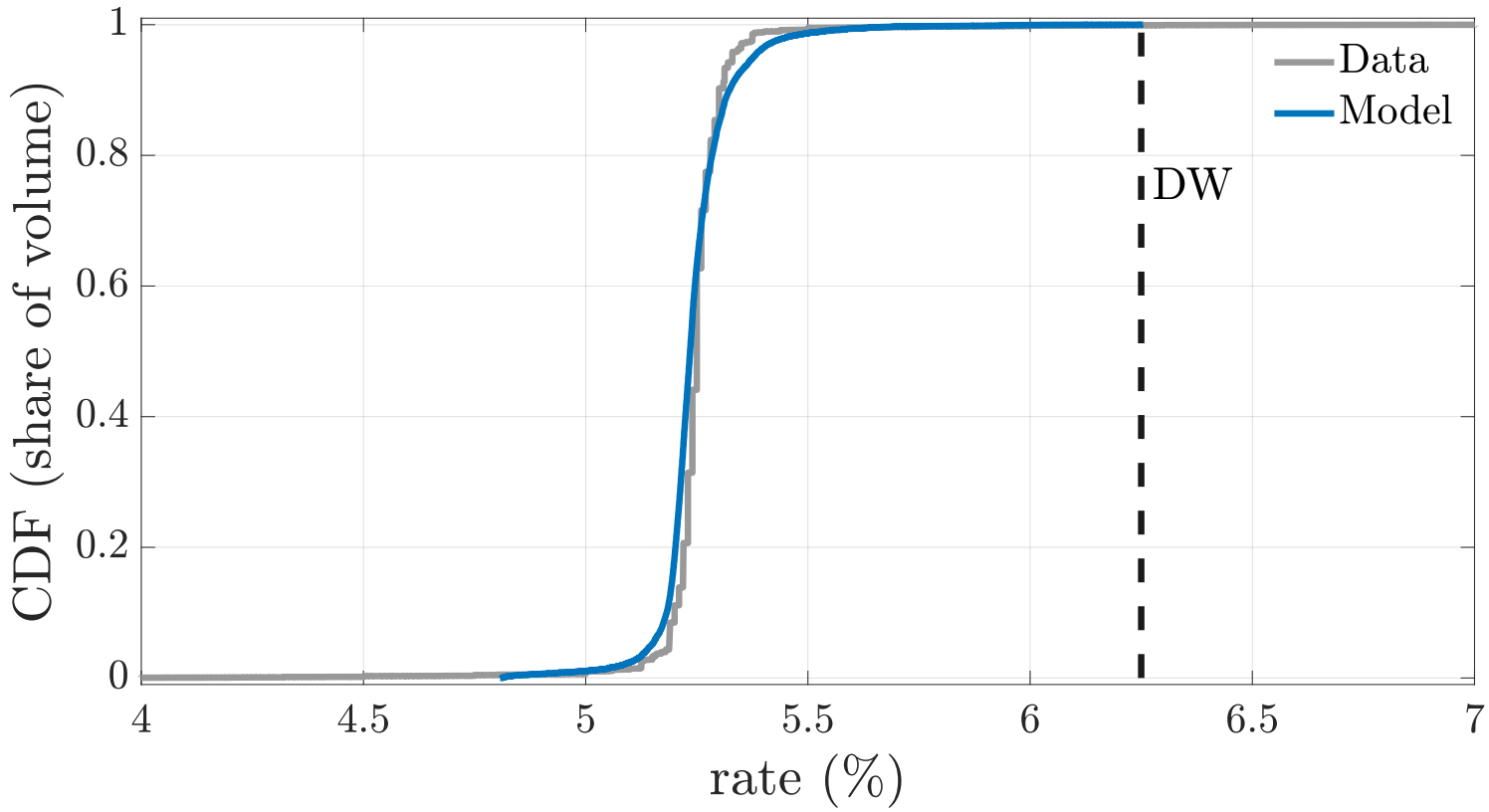

Figure 28: Empirical and theoretical cumulative distribution functions of bilateral rates for the year 2006.

Notes: Rates in the horizontal axis are in percent per annum.

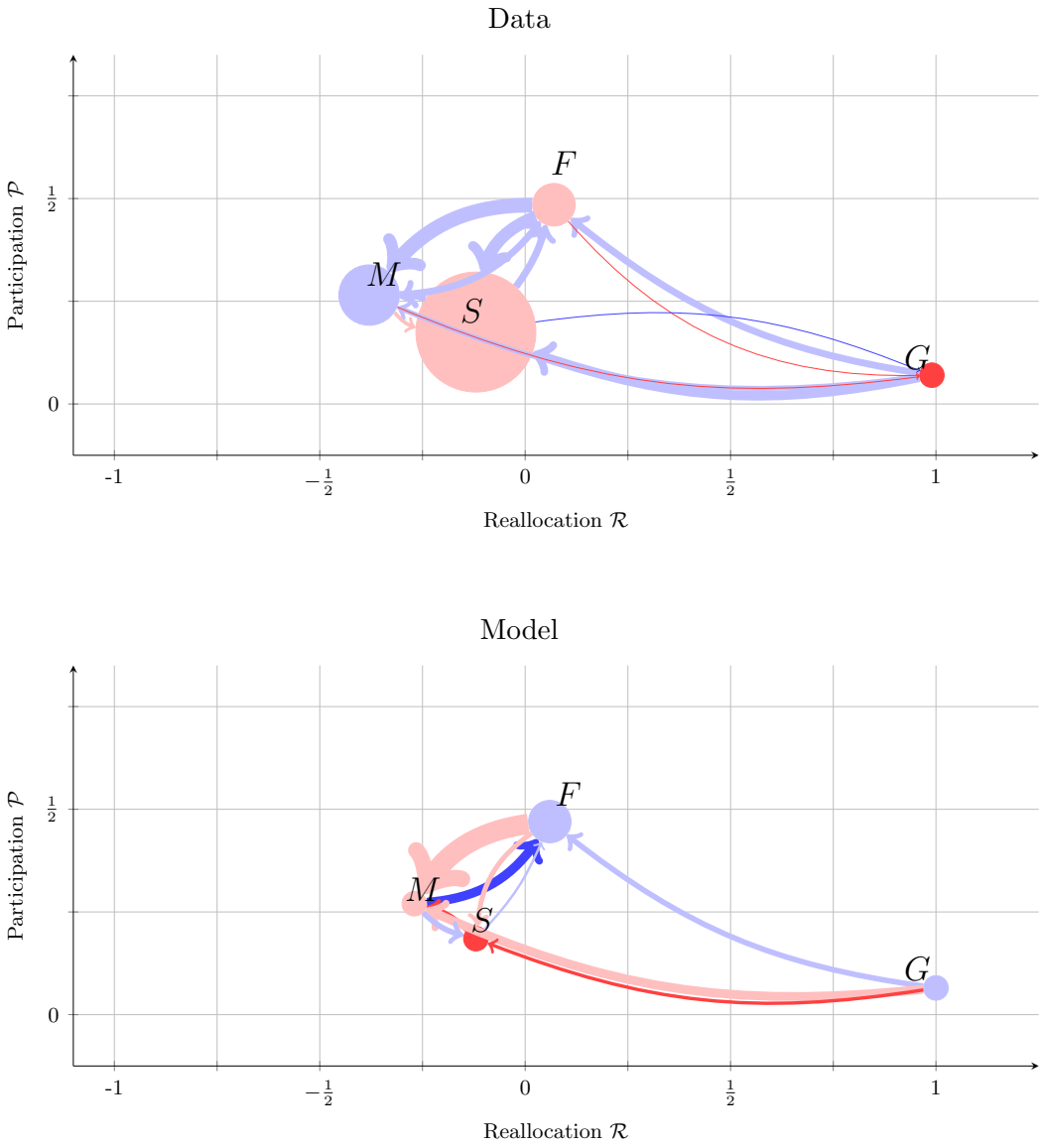

Figure 29: Empirical and theoretical interbank trading networks for 2006.

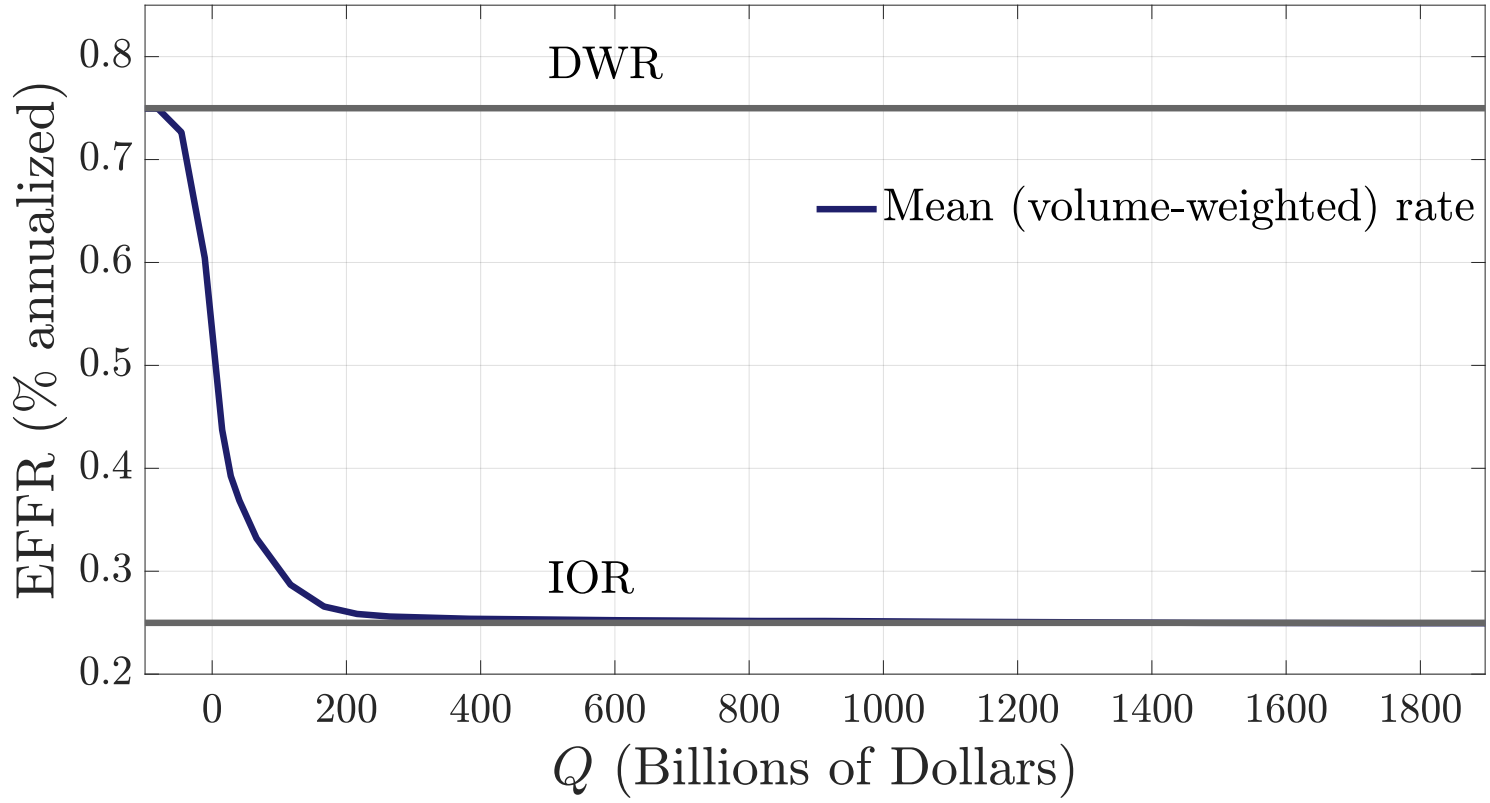

Figure 30: Theoretical aggregate demand for reserves for the year 2006 calibration.

Notes: Theoretical aggregate demand $\iota^*_{\mathrm{Y}_\omega} = \mathcal{D}(Q_{\mathrm{Y}_\omega}; \Pi)$ for the model calibrated as in Table 3, and with $\iota^*_{\mathrm{Y}_\omega}$ and Q_{Y_ω} computed with the interpolation procedure described in Section A.6, for $\mathrm{Y}_0 = 2006$ and $\mathrm{Y}_1 = 2014$.

F Theory: supplementary results

F.1 Value function

Let $J_t^i(a,c) : \mathbb{N} \times \mathbb{T} \times \mathbb{R}^2 \to \mathbb{R}$ denote 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 $c \in \mathbb{R}$. Then, $J_t^i(a,c)$ satisfies

$$\begin{aligned}
&J_t^i(a,c) \\
&= \mathbb{E}_t \Bigg\{ \mathbb{I}_{\{T-t \leq \min[\tau(\beta_i), \tau(\lambda_i)]\}} \left[\int_t^T e^{-r(s-t)} u_i(a)\, ds + e^{-r(T-t)} \left[U_i(a) + e^{-r(\bar{T}-T)} c \right] \right] \\
&+ \mathbb{I}_{\{\tau(\lambda_i) < \min[\tau(\beta_i), T-t]\}} \left[\int_t^{t+\tau(\lambda_i)} e^{-r(s-t)} u_i(a)\, ds \right. \\
&\left. + e^{-r\tau(\lambda_i)} \sum_{j \in \mathbb{N}} \pi_j \int J^i_{t+\tau(\lambda_i)}(a-z, c)\, dG_{ij}(z) \right] \\
&+ \mathbb{I}_{\{\tau(\beta_i) < \min[\tau(\lambda_i), T-t]\}} \left[\int_t^{t+\tau(\beta_i)} e^{-r(s-t)} u_i(a)\, ds \right. \\
&\left. + e^{-r\tau(\beta_i)} \sum_{j \in \mathbb{N}} \sigma_j \int J^i_{t+\tau(\beta_i)} \left[a - b^{ij}_{t+\tau(\beta_i)}(a, \tilde{a}), c + R^{ji}_{t+\tau(\beta_i)}(\tilde{a}, a) \right] dF^j_{t+\tau(\beta_i)}(\tilde{a}) \right] \Bigg\}, \qquad (30)
\end{aligned}$$

where $\tau(\zeta)$ denotes the exponentially distributed first passage time of the Poisson process with arrival rate ζ,

$$\begin{aligned}
\pi_j &\equiv \frac{\lambda_j n_j}{\sum_{i \in \mathbb{N}} \lambda_i n_i} \\
\sigma_j &\equiv \frac{\beta_j n_j}{\sum_{k \in \mathbb{N}} \beta_k n_k},
\end{aligned}$$

and

$$\begin{aligned}
&(b_t^{ij}(a, \tilde{a}), R_t^{ji}(\tilde{a}, a)) \\
&= \arg \max_{(b,R) \in \mathbb{R}^2} \left[J_t^i(a-b, c+R) - J_t^i(a,c) \right]^{\theta_{ij}} \left[J_t^j(\tilde{a}+b, c-R) - J_t^j(\tilde{a}, c) \right]^{\theta_{ji}}. \qquad (31)
\end{aligned}$$

Lemma 1 *The function*

$$J_t^i(a,c) = V_t^i(a) + e^{-r(\bar{T}-t)} c \qquad (32)$$

satisfies (30) if and only if $V_t^i(a)$ *satisfies*

$$
\begin{aligned}
V_t^i(a) = \mathbb{E}_t\Bigg\{ & \mathbb{I}_{\{T-t\leq \min[\tau(\beta_i),\tau(\lambda_i)]\}}\left[\int_t^T e^{-r(s-t)}u_i(a)\,ds + e^{-r(T-t)}U_i(a)\right] \\
& + \mathbb{I}_{\{\tau(\lambda_i)<\min[\tau(\beta_i),T-t]\}}\Bigg[\int_t^{t+\tau(\lambda_i)} e^{-r(s-t)}u_i(a)\,ds \\
& + e^{-r\tau(\lambda_i)}\sum_{j\in\mathbb{N}}\pi_j\int V^i_{t+\tau(\lambda_i)}(a-z)\,dG_{ij}(z)\Bigg] \\
& + \mathbb{I}_{\{\tau(\beta_i)<\min[\tau(\lambda_i),T-t]\}}\Bigg[\int_t^{t+\tau(\beta_i)} e^{-r(s-t)}u_i(a)\,ds \\
& + e^{-r\tau(\beta_i)}\sum_{j\in\mathbb{N}}\sigma_j\int\left[V^i_{t+\tau(\beta_i)}(a-b^{ij}_{t+\tau(\beta_i)}(a,\tilde{a})) + \bar{R}^{ji}_{t+\tau(\beta_i)}(\tilde{a},a)\right]dF^j_{t+\tau(\beta_i)}(\tilde{a})\Bigg]\Bigg\}, \quad (33)
\end{aligned}
$$

with

$$
\bar{R}^{ji}_{t+\tau(\beta_i)}(\tilde{a},a) \equiv e^{-r\left\{\bar{T}-[t+\tau(\beta_i)]\right\}}R^{ji}_{t+\tau(\beta_i)}(\tilde{a},a),
$$

and $(b_t^{ij}(a,\tilde{a}), R_t^{ji}(\tilde{a},a))$ *given by (1) and (2).*

Proof. With (32), (31) becomes equivalent to (1) and (2). Substitute (32) into (30) to get

$$
\begin{aligned}
& V_t^i(a) + e^{-r(\bar{T}-t)}c \\
& = \mathbb{E}_t\Bigg\{ \mathbb{I}_{\{T-t\leq \min[\tau(\beta_i),\tau(\lambda_i)]\}}\left[\int_t^T e^{-r(s-t)}u_i(a)\,ds + e^{-r(T-t)}\left[U_i(a) + e^{-r(\bar{T}-T)}c\right]\right] \\
& + \mathbb{I}_{\{\tau(\lambda_i)<\min[\tau(\beta_i),T-t]\}}\Bigg[\int_t^{t+\tau(\lambda_i)} e^{-r(s-t)}u_i(a)\,ds \\
& + e^{-r\tau(\lambda_i)}\sum_{j\in\mathbb{N}}\pi_j\int\left[V^i_{t+\tau(\lambda_i)}(a-z) + e^{-r\{\bar{T}-[t+\tau(\lambda_i)]\}}c\right]dG_{ij}(z)\Bigg] \\
& + \mathbb{I}_{\{\tau(\beta_i)<\min[\tau(\lambda_i),T-t]\}}\Bigg[\int_t^{t+\tau(\beta_i)} e^{-r(s-t)}u_i(a)\,ds + e^{-r\tau(\beta_i)}\sum_{j\in\mathbb{N}}\sigma_j\int\Bigg[V^i_{t+\tau(\beta_i)}(a-b^{ij}_{t+\tau(\beta_i)}(a,\tilde{a})) \\
& + e^{-r\{\bar{T}-[t+\tau(\beta_i)]\}}[c + R^{ji}_{t+\tau(\beta_i)}(\tilde{a},a)]\Bigg]dF^j_{t+\tau(\beta_i)}(\tilde{a})\Bigg]\Bigg\},
\end{aligned}
$$

which after cancelling the terms proportional to c, becomes identical to (33). ■

Lemma 2 *The Bellman equation (33) can be written as*

$$\begin{aligned} V_t^i(a) = &\left[1 - e^{-(r+\beta_i+\lambda_i)(T-t)}\right] \frac{u_i(a)}{r+\beta_i+\lambda_i} + e^{-(r+\beta_i+\lambda_i)(T-t)} U_i(a) \\ &+ \lambda_i \int_t^T e^{-(r+\beta_i+\lambda_i)(\tau-t)} \left[\sum_{j\in\mathbb{N}} \pi_j \int V_\tau^i(a-z)\, dG_{ij}(z)\right] d\tau \\ &+ \beta_i \int_t^T e^{-(r+\beta_i+\lambda_i)(\tau-t)} \left[V_\tau^i(a) + \sum_{j\in\mathbb{N}} \sigma_j \theta_{ij} \int \max_{b\in\mathbb{R}} S_\tau^{ij}(a,\tilde{a},b)\, dF_\tau^j(\tilde{a})\right] d\tau \end{aligned} \tag{34}$$

or equivalently, as (3) with boundary condition $V_T^i(a) = U_i(a)$.

Proof. With the bargaining outcomes (1) and (2), (33) can be rewritten as

$$\begin{aligned} V_t^i(a) = \mathbb{E}_t \Bigg\{ &\mathbb{I}_{\{T-t\le \min[\tau(\beta_i),\tau(\lambda_i)]\}} \left[\int_t^T e^{-r(s-t)} u_i(a)\, ds + e^{-r(T-t)} U_i(a)\right] \\ &+ \mathbb{I}_{\{\tau(\lambda_i)<\min[\tau(\beta_i),T-t]\}} \left[\int_t^{t+\tau(\lambda_i)} e^{-r(s-t)} u_i(a)\, ds + e^{-r\tau(\lambda_i)} \sum_{j\in\mathbb{N}} \pi_j \int V_{t+\tau(\lambda_i)}^i(a-z)\, dG_{ij}(z)\right] \\ &+ \mathbb{I}_{\{\tau(\beta_i)<\min[\tau(\lambda_i),T-t]\}} \Bigg[\int_t^{t+\tau(\beta_i)} e^{-r(s-t)} u_i(a)\, ds \\ &+ e^{-r\tau(\beta_i)} \sum_{j\in\mathbb{N}} \sigma_j \int \left[V_{t+\tau(\beta_i)}^i(a) + \theta_{ij} \max_{b\in\mathbb{R}} S_{t+\tau(\beta_i)}^{ij}(a,\tilde{a},b)\right] dF_{t+\tau(\beta_i)}^j(\tilde{a})\Bigg]\Bigg\}, \end{aligned}$$

where

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

The first term on the right side of $V_t^i(a)$ can be written as

$$\begin{aligned} &\mathbb{E}_t \left\{\mathbb{I}_{\{T-t\le \min[\tau(\beta_i),\tau(\lambda_i)]\}} \left[\int_t^T e^{-r(s-t)} u_i(a)\, ds + e^{-r(T-t)} U_i(a)\right]\right\} \\ &= e^{-(\beta_i+\lambda_i)(T-t)} \left\{\left[1 - e^{-r(T-t)}\right] \frac{u_i(a)}{r} + e^{-r(T-t)} U_i(a)\right\}. \end{aligned}$$

The second term on the right side of $V_t^i(a)$ can be written as

$$\mathbb{E}_t\left\{\mathbb{I}_{\{\tau(\lambda_i)<\min[\tau(\beta_i),T-t]\}}\left[\int_t^{t+\tau(\lambda_i)} e^{-r(s-t)}u_i(a)\,ds + e^{-r\tau(\lambda_i)}\sum_{j\in\mathbb{N}}\pi_j\int V^i_{t+\tau(\lambda_i)}(a-z)\,dG_{ij}(z)\right]\right\}$$
$$= \frac{\lambda_i}{\beta_i+\lambda_i}\frac{r\left[1-e^{-(\beta_i+\lambda_i)(T-t)}\right]-(\beta_i+\lambda_i)\,e^{-(\beta_i+\lambda_i)(T-t)}\left[1-e^{-r(T-t)}\right]}{r+\beta_i+\lambda_i}\frac{u_i(a)}{r}$$
$$+\int_0^{T-t}\lambda_i e^{-(r+\beta_i+\lambda_i)y}\left[\sum_{j\in\mathbb{N}}\pi_j\int V^i_{t+y}(a-z)\,dG_{ij}(z)\right]dy.$$

The third term on the right side of $V_t^i(a)$ can be written as

$$V_t^i(a) = \mathbb{E}_t\Bigg\{\mathbb{I}_{\{\tau(\beta_i)<\tau(\lambda_i)\}}\mathbb{I}_{\{\tau(\beta_i)<T-t\}}\Bigg[\int_t^{t+\tau(\beta_i)} e^{-r(s-t)}u_i(a)\,ds$$
$$+ e^{-r\tau(\beta_i)}\sum_{j\in\mathbb{N}}\sigma_j\int\left[V^i_{t+\tau(\beta_i)}(a)+\theta_{ij}\max_{b\in\bar{\mathbb{R}}}S^{ij}_{t+\tau(\beta_i)}(a,\tilde{a},b)\right]dF_t^j(\tilde{a})\Bigg]\Bigg\}$$
$$= \frac{\beta_i}{\beta_i+\lambda_i}\frac{r\left[1-e^{-(\beta_i+\lambda_i)(T-t)}\right]-(\beta_i+\lambda_i)\,e^{-(\beta_i+\lambda_i)(T-t)}\left[1-e^{-r(T-t)}\right]}{r+\beta_i+\lambda_i}\frac{u_i(a)}{r}$$
$$+\int_0^{T-t}\beta_i e^{-(r+\beta_i+\lambda_i)z}\left[\sum_{j\in\mathbb{N}}\sigma_j\int\left[V^i_{t+z}(a)+\theta_{ij}\max_{b\in\bar{\mathbb{R}}}S^{ij}_{t+z}(a,\tilde{a},b)\right]dF^j_{t+z}(\tilde{a})\right]dz.$$

Thus, we can write

$$V_t^i(a) = \left[1-e^{-(r+\beta_i+\lambda_i)(T-t)}\right]\frac{u_i(a)}{r+\beta_i+\lambda_i} + e^{-(r+\beta_i+\lambda_i)(T-t)}U_i(a)$$
$$+\lambda_i\int_0^{T-t}e^{-(r+\beta_i+\lambda_i)y}\left[\sum_{j\in\mathbb{N}}\pi_j\int V^i_{t+y}(a-s)\,dG_{ij}(s)\right]dy$$
$$+\beta_i\int_0^{T-t}e^{-(r+\beta_i+\lambda_i)z}\left[\sum_{j\in\mathbb{N}}\sigma_j\int\left[V^i_{t+z}(a)+\theta_{ij}\max_{b\in\bar{\mathbb{R}}}S^{ij}_{t+z}(a,\tilde{a},b)\right]dF^j_{t+z}(\tilde{a})\right]dz.$$

With a change of variables in the integrals with respect to time,

$$V_t^i(a) = \left[1-e^{-(r+\beta_i+\lambda_i)(T-t)}\right]\frac{u_i(a)}{r+\beta_i+\lambda_i} + e^{-(r+\beta_i+\lambda_i)(T-t)}U_i(a)$$
$$+\lambda_i\int_t^{T}e^{-(r+\beta_i+\lambda_i)(\tau-t)}\left[\sum_{j\in\mathbb{N}}\pi_j\int V^i_{\tau}(a-z)\,dG_{ij}(z)\right]d\tau$$
$$+\beta_i\int_t^{T}e^{-(r+\beta_i+\lambda_i)(\tau-t)}\left[V^i_{\tau}(a)+\sum_{j\in\mathbb{N}}\sigma_j\theta_{ij}\int\max_{b\in\bar{\mathbb{R}}}S^{ij}_{\tau}(a,\tilde{a},b)\,dF^j_{\tau}(\tilde{a})\right]d\tau. \qquad (35)$$

To obtain (3), simply differentiate (34) with respect to t. ■

F.2 Extension: regulatory borrowing costs

In this section we generalize the theory to allow for proportional borrowing costs to proxy for the effects of regulatory constraints that affect banks' incentives to borrow. Let

$$\Gamma_t^i(a,b,R) \equiv V_t^i(a-b) - V_t^i(a) + \left[1 + \mathbb{I}_{\{b<0\}}\kappa_i\right] e^{-r\left(\bar{T}-t\right)}R \tag{36}$$

denote payoff of a bank of type $i \in \mathbb{N}$, with pre-trade balance a, that at time t sells a loan of size b in exchange for a repayment of size R delivered at time $\bar{T}$, with $\kappa_i \in \mathbb{R}_+$. Intuitively, if $b, R \in \mathbb{R}_+$, then the bank is "selling funds" (i.e., lending) and the gain from trade is as in Section 2.2. Conversely, if $b, R \in \mathbb{R}_-$, then the bank is "buying funds" (i.e., *borrowing*), and κ_i captures the effects of policies that increase the shadow cost of the bank's liabilities. In all our calibrations we set κ_G large enough to make our theory consistent with the fact that the business model of a GSE consists of lending, but not borrowing in the interbank market. In our 2019 calibration we use κ_i for $i \in \{F, M, S\}$ to capture the effects of the prudential liquidty regulations discussed in Appendix B (Section B.2). With borrowing costs, the bargaining outcome at time t between two banks of type i and j, with respective balances a and $\tilde{a}$, denoted $\left(b_t^{ij}\left(a,\tilde{a}\right), R_t^{ji}\left(\tilde{a},a\right)\right)$, is the solution to

$$\max_{(b,R)\in\bar{\mathbb{R}}\times\mathbb{R}} \Gamma_t^i(a,b,R)^{\theta_{ij}}\Gamma_t^j(\tilde{a},-b,-R)^{\theta_{ji}}. \tag{37}$$

The correspondig first-order condition with respect to R is

$$\theta_{ij}\left[1 + \mathbb{I}_{\{b<0\}}\kappa_i\right]\Gamma_t^j(\tilde{a},-b,-R) = \theta_{ji}\left[1 + \mathbb{I}_{\{0<b\}}\kappa_j\right]\Gamma_t^i(a,b,R),$$

which implies $R_t^{ji}\left(\tilde{a},a\right)$ is given by

$$\begin{aligned} e^{-r\left(\bar{T}-t\right)}R_t^{ji}\left(\tilde{a},a\right) &= \frac{\theta_{ij}}{1+\mathbb{I}_{\{0<b_t^{ij}(a,\tilde{a})\}}\kappa_j}[V_t^j(\tilde{a}+b_t^{ij}\left(a,\tilde{a}\right)) - V_t^j(\tilde{a})] \\ &\quad + \frac{\theta_{ji}}{1+\mathbb{I}_{\{b_t^{ij}(a,\tilde{a})<0\}}\kappa_i}[V_t^i(a) - V_t^i(a-b_t^{ij}\left(a,\tilde{a}\right))], \end{aligned} \tag{38}$$

and

$$b_t^{ij}\left(a,\tilde{a}\right) \in \arg\max_{b\in\bar{\mathbb{R}}} \hat{S}_t^{ij}\left(a,\tilde{a},b\right), \tag{39}$$

where

$$\hat{S}_t^{ij}\left(a,\tilde{a},b\right) \equiv \hat{\Gamma}_t^{ij}(a,\tilde{a},b)^{\theta_{ij}}\hat{\Gamma}_t^{ji}(\tilde{a},a,-b)^{\theta_{ji}},$$

with

$$\begin{aligned}\hat{\Gamma}_t^{ij}(a,\tilde{a},b) &\equiv \theta_{ij}\left\{S_t^{ij}(a,\tilde{a},b) - \frac{\mathbb{I}_{\{0<b\}}\kappa_j - \mathbb{I}_{\{b<0\}}\kappa_i}{1+\mathbb{I}_{\{0<b\}}\kappa_j}[V_t^j(\tilde{a}+b) - V_t^j(\tilde{a})]\right\}\\ \hat{\Gamma}_t^{ji}(\tilde{a},a,-b) &\equiv \theta_{ji}\left\{S_t^{ij}(a,\tilde{a},b) - \frac{\mathbb{I}_{\{0<b\}}\kappa_j - \mathbb{I}_{\{b<0\}}\kappa_i}{1+\mathbb{I}_{\{b<0\}}\kappa_i}\left[V_t^i(a) - V_t^i(a-b)\right]\right\}.\end{aligned}$$

In summary, the bargaining solution, $(b_t^{ij}(a,\tilde{a}), R_t^{ji}(\tilde{a},a))$, is given by (39) and (38), and the value function $V_t^i(a)$ now satisfies

$$\begin{aligned}rV_t^i(a) - \dot{V}_t^i(a) = u_i(a) &+ \lambda_i \sum_{j\in\mathbb{N}} \pi_j \int \left[V_t^i(a-z) - V_t^i(a)\right] dG_t^{ij}(z)\\ &+ \beta_i \sum_{j\in\mathbb{N}} \sigma_j \int \Gamma_t^i(a, b_t^{ij}(a,\tilde{a}), R_t^{ji}(\tilde{a},a)) dF_t^j(\tilde{a}),\end{aligned} \tag{40}$$

with Γ_t^i as defined in (36). Notice that (39), (38), and (40) generalize (1), (2), and (3), respectively (and the former reduce to the latter if $\kappa_i = 0$ for all $i \in \mathbb{N}$).

G Model computation

In this section we discuss computational issues. Section G.1 outlines the solution algorithm. Section G.2 explains how we compute, in the quantitative theory, the statistics that we compare with their empirical counterparts.

G.1 Solution algorithm

The steps we use to solve for the equilibrium of the model are as follows.

Step 0: Set grids. We think of the time interval $[0, T]$ as corresponding to a trading day in the fed-funds market, which consists of 9.5 hours (from 9.00 AM to 5.30 PM). We divide the interval $[0, T]$ into $N_T + 1$ periods, denoted $t = 0, 1, \ldots, N_T$, and set $N_T = 799$. As we have 800 periods, each period represents approximately 42 seconds (i.e., $\frac{9.5 \times 60 \times 60}{800} = 42.75$ seconds).[127]

For each bank type $i \in \mathbb{N}$, we construct an equally spaced grid for reserve balances, $\mathbb{A}^i = \left\{a_1^i, a_2^i, \ldots, a_{N_a}^i\right\}$, with $N_a = 150$. We interpret each unit of reserves in the model as corresponding to \$10 bn in the data. For the benchmark years 2017 and 2019, we set a_1^i and $a_{N_a}^i$ equal to the 0.5^{th} and 99.5^{th} percentiles of the kernel estimate of the beginning-of-day distributions, respectively (see Section A.3). We use the interpolation procedure explained in Section A.6 to construct grids whenever we change the total quantity of balances. In all cases we add 5 additional points to the grid, $\{-0.2, -0.1,\ 0,\ 0.1,\ 0.2\}$.[128]

For each pair of bank types, $i, j \in \mathbb{N}$, we construct a grid for payment sizes, $\mathbb{Z}^{ij} =$

[127] 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 A.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.

[128] We add these grid points because the value functions are numerically close to having a kink around $a = 0$ towards the end of the trading day (i.e., as t gets closer to N_T).

$\left\{z_1^{ij}, z_2^{ij}, \ldots, z_{N_z}^{ij}\right\}$, with $N_z = 35$. The probability mass function for payment sizes, $\{G_{ij}(z)\}_{z\in\mathbb{Z}^{ij}}$, is constructed with the procedure described in Section A.2.

Step 1: Guess the distribution of balances. For each $a \in \mathbb{A}^i$, let $f_t^i(a)$ be the fraction of banks of type $i \in \mathbb{N}$ that hold a quantity of reserves equal to a at the beginning of period t, with $\sum_{a\in\mathbb{A}^i} f_t^i(a) = 1$. The beginning-of-day distribution, $f_0^i(\cdot)$, is given since $F_0^i(a) \equiv \sum_{x\in\mathbb{A}^i} f_0^i(x)\mathbb{I}_{\{x\leq a\}}$ is estimated from the data with the procedure described in Section A.3. Guess the distributions $\{f_t^i(a)\}_{a\in\mathbb{A}^i,i\in\mathbb{N}}$ for each $t \in \{1, 2, \ldots, N_T\}$.

Step 2: Compute the value function. Since for each $i \in \mathbb{N}$ and $a \in \mathbb{A}^i$ we have the terminal condition, $V_{N_T}^i(a) = U_i(a)$, where $U_i(\cdot)$ is the exogenous end-of-day payoff function, we can then solve backwards for the value function, $\{V_t^i(a)\}_{a\in\mathbb{A}^i,i\in\mathbb{N},t\in\{0,\ldots,N_T-1\}}$. Each of these backward iterations between period $t \in \{N_T, \ldots, 1\}$ and period $t-1$ consists of two steps. In the first step, for each pair of bank types $i, j \in \mathbb{N}$, we compute the bargaining outcomes, $\left\{b_t^{ij}(a^i, a^j), R_t^{ji}(a^j, a^i)\right\}_{a^i\in\mathbb{A}^i,a^j\in\mathbb{A}^j}$, taking $\{V_t^i(a)\}_{a\in\mathbb{A}^i,i\in\mathbb{N}}$ as given. In the second step we solve for the value function backwards, i.e., we solve for $\{V_{t-1}^i(a)\}_{a\in\mathbb{A}^i,i\in\mathbb{N}}$ given the one-period-ahead bargaining outcomes and values, i.e., $\left\{b_t^{ij}(a^i, a^j), R_t^{ji}(a^j, a^i), V_t^i(a^i)\right\}_{a^i\in\mathbb{A}^i,a^j\in\mathbb{A}^j,(i,j)\in\mathbb{N}^2}$. Next, we explain these two steps in detail.

Step 2.1: *Solve for* $b_t^{ij}(\cdot,\cdot)$ *and* $R_t^{ji}(\cdot,\cdot)$. Given the values $\{V_t^i(\cdot)\}_{i\in\mathbb{N}}$, we compute the bargaining outcome for the loan size, $b_t^{ij}(a, \tilde{a})$, as in (39), which can be written as:

$$b_t^{ij}(a, \tilde{a}) = \arg\max_b \left\{ \frac{1}{1 + \mathbb{I}_{\{b<0\}}\kappa_i} V_t^i(a - b) + \frac{1}{1 + \mathbb{I}_{\{0<b\}}\kappa_j} V_t^j(\tilde{a} + b) - \epsilon|b| \right\}, \tag{41}$$

where $\epsilon =$ 1e-9 is a small trading cost introduced to rule out loans with negligible gains from trade. Since a unit of reserves in the model corresponds to \$10 bn in the data, the value of ϵ implies a trading cost of \$10 for a loan of size \$1 bn. We use a Golden search routine to solve for $b_t^{ij}(a, \tilde{a})$ in (41), for each $a \in \mathbb{A}^i$, $\tilde{a} \in \mathbb{A}^j$, and $(i, j) \in \mathbb{N} \times \mathbb{N}$. Given the bargained loan sizes, $\{b_t^{ij}(a^i, a^j)\}_{a^i\in\mathbb{A}^i,a^j\in\mathbb{A}^j,(i,j)\in\mathbb{N}^2}$, we can compute the associated repayments, $R_t^{ji}(a^j, a^i)$, as in (38), and the gain from trade to the bank of type i and balance a^i, i.e., $\Gamma_t^i(a, b_t^{ij}(a^i, a^j), R_t^{ji}(a^j, a^i))$, as in (36).

Step 2.2: *Solve for* $V_t^i(a)$ *backwards.* We divide each period into two stages. Random payments between pairs of banks take place in the first stage. Trade between pairs of banks takes place in the second stage. The first stage is divided further into N_S subperiods, each

indexed by $s \in \{1, 2, \ldots, N_S\}$ with $N_S = 42$, so each of these subperiods corresponds to 1 second (since each full model period corresponds to approximately 42 seconds). We solve for the value function within each model period backwards: we start by solving for the value of trade decisions in the second stage, and then integrate the value of the payment shocks in the 42 subperiods of the first stage.

Let $\hat{V}_t^i(a)$ be the value of a bank of type $i \in \mathbb{N}$ with balance $a \in \mathbb{A}^i$ at the beginning of the second stage of period t. This value satisfies

$$(1 + \Delta r)\hat{V}_t^i(a) = \Delta u_i(a) + \Delta\beta_i \sum_{j\in\mathbb{N}} \sigma_j \sum_{\tilde{a}\in\mathbb{A}_j} \bar{\Gamma}_{t+1}^{ij}(a, \tilde{a}) f_{t+1}^j(\tilde{a}) + V_{t+1}^i(a), \tag{42}$$

where $\bar{\Gamma}_t^{ij}(a, \tilde{a}) \equiv \Gamma_t^i(a, b_t^{ij}(a, \tilde{a}), R_t^{ji}(\tilde{a}, a))$, and $\Delta = 1/\left[N_S(N_T + 1)\right]$ is the size of the time interval (including all trade and payment periods in the day). Let $\tilde{V}_{t,s}^i(a)$ be the value of a bank of type $i \in \mathbb{N}$ with balance $a \in \mathbb{A}^i$ at the beginning of subperiod s of the first stage of period t. This value satisfies

$$(1 + \Delta r)\tilde{V}_{t,s}^i(a) = \Delta u_i(a) + \Delta\lambda_i \sum_{j\in\mathbb{N}} \pi_j \sum_{z\in\mathbb{Z}^{ij}} \left[V_t^i(a - z) - V_t^i(a)\right] G_{ij}(z) + \tilde{V}_{t,s+1}^i(a), \tag{43}$$

for $s = 1, \ldots, N_S$, with boundary conditions $\tilde{V}_{t,N_S+1}^i(a) = \hat{V}_t^i(a)$, and $\tilde{V}_{t,1}^i(a) = V_t^i(a)$. Equations (42) and (43) are the discrete-time approximations to the Bellman equation (40).

We solve (42)-(43) backwards, as follows. Given $V_{t+1}^i(\cdot)$ (recall the terminal condition $V_{N_T}^i(\cdot) = U_i(\cdot)$), we compute $\bar{\Gamma}_{t+1}^{ij}(\cdot, \cdot)$, and given our guess of $\{f_{t+1}^i(\cdot)\}_{i\in\mathbb{N}}$, we compute $\hat{V}_t^i(\cdot)$ using (42). We then compute $\{\tilde{V}_{t,s}^i(\cdot)\}_{s\in\{1,2,\ldots,N_S\}}$ using (43) and the terminal condition $\tilde{V}_{t,N_S+1}^i(\cdot) = \hat{V}_t^i(\cdot)$ by iterating backwards, which delivers $V_t^i(\cdot) = \tilde{V}_{t,1}^i(\cdot)$.

Step 3: Compute the implied distribution of balances Given the negotiated loan sizes, $b_t^{ij}(\cdot, \cdot)$ and the distribution of random payments, we can solve for the distribution of balances forward from an initial condition, $f_0^i(a)$. As in **step 2**, we need to compute the evolution of balances for the two within-period stages (the payments stage, and the trading stage). Since we are solving for the distribution of reserves forward, we start with the first stage and then move to the second stage.

Let $\tilde{f}_{t,s}^{i,\text{new}}(a_m)$ be the fraction of banks of type $i \in \mathbb{N}$ with balance $a_m \in \mathbb{A}^i$, at the beginning of subperiod s of the first stage of period t. We use the superscript "new" to emphasize that this is the new distribution implied by the bargaining outcomes in **step 2** (rather than the

distribution that was used to derive those outcomes). Then,

$$\tilde{f}_{t,s}^{i,\text{new}}(a_m) = (1-\Delta\lambda_i)\tilde{f}_{t,s-1}^{i,\text{new}}(a_m) + \Delta\lambda_i \sum_{j\in\mathbb{N}}\sum_{a\in\mathbb{A}^i}\sum_{z\in\mathbb{Z}^{ij}} \pi_j \mathbb{L}\left(a_m, a-z\right) G_{ij}(z) \tilde{f}_{t,s-1}^{i,\text{new}}(a) \quad (44)$$

for $s = 1, \ldots, N_S$, with initial condition $\tilde{f}_{t,0}^{i,\text{new}}(a_m) = f_t^{i,\text{new}}(a_m)$, and where

$$\mathbb{L}(a_m, x) \equiv \mathbb{I}_{\{x\in(a_{m-1},\ a_m]\}} \frac{x - a_{m-1}}{a_m - a_{m-1}}$$

implements a linear interpolation. The recursion (44) is initialized with the exogenous time-0 distribution of balances, i.e., $f_0^{i,\text{new}}(a_m) = f_0^i(a_m)$.

Let $\hat{f}_t^{i,\text{new}}(a_m)$ be the fraction of banks of type $i \in \mathbb{N}$ with balance $a_m \in \mathbb{A}^i$ after the trades in the second stage of period t; it is given by

$$\begin{aligned}\hat{f}_t^{i,\text{new}}(a_m) = {} & (1-\Delta\beta_i)\tilde{f}_{t,N_S}^{i}(a_m) \\ & + \Delta\beta_i \sum_{j\in\mathbb{N}}\sum_{a\in\mathbb{A}^i}\sum_{\tilde{a}\in\mathbb{A}^j} \sigma_j \mathbb{L}\left(a_m, a - b_t^{ij}(\hat{a},\tilde{a})\right) \tilde{f}_{t,N_S}^{i,\text{new}}(a)\tilde{f}_{t,N_S}^{j,\text{new}}(\tilde{a}).\end{aligned}$$

Having solved for $\hat{f}_t^{i,\text{new}}(\cdot)$, set $f_{t+1}^{i,\text{new}}(\cdot) = \tilde{f}_{t+1,0}^{i,\text{new}}(\cdot) = \hat{f}_t^{i,\text{new}}(\cdot)$, and move to next period.

Step 4: Check for convergence. We use two criteria for convergence.

Criterion 1. We determine that the algorithm has converged if the probability distribution in **step 1** is close enough to the probability distribution obtained after **step 4**. Specifically, we consider the algorithm has converged if $\mathcal{E}(f) \equiv \max_{a,i,t}|f_t^i(a) - f_t^{i,\text{new}}(a)| <$ 1e-4.

Criterion 2. We determine that the algorithm has converged if some key theoretical moments have stabilized across iterations. In particular, we look at convergence in the distribution of interest rates and measures of trading activity.[129] Specifically, let ρ_t^p denote the p-percentile of the (volume weighted) distribution of interest rates at time t. Every 10 iterations of the algorithm, we compute the rate percentiles ρ_t^p for $p \in \{0.05, 0.10, 0.30, 0.50, 0.70, 0.90, 0.95\}$, and then compute the error $\mathcal{E}(\rho) \equiv \max_{p,t}|\rho_t^p - \rho_t^{p,\text{new}}|$. Every 10 iterations, we also compute: the value-weighted average interest rate, the participation for each bank, $\mathcal{P}_i$, and the reallocation for each bank, $\mathcal{R}_i$. We consider the algorithm has converged if after 10 iterations, we have: (i) $\mathcal{E}(f) <$ 1e-3, (ii) $\mathcal{E}(\rho) <$ 1e-4, and (iii) the errors for the value-weighted average interest rate, $\mathcal{P}_i$, and $\mathcal{R}_i$ all below 1e-4. For all these error computations, we check errors comparing results 10 iterations apart (e.g.: the value-weighted average interest rate this iteration compared with

[129] See Section G.2 for details on computation of theoretical moments.

the value-weighted average interest rate 10 iterations ago), which ensure that results are stable across algorithm iterations.

The reason we sometimes use **Criterion 2** is that, despite our using the trading cost ϵ in equation (41), we sometimes observe loans that entail very small gains from trade, but still affect the distributions $\{f_t^i(\cdot)\}$. These small-surplus trades may keep the error $\mathcal{E}(F)$ above the convergence tolerance, but have no significant effect on the distribution of rates nor in the relevant measures of trading activity. To ensure the algorithm has stabilized, we only start implementing **Criterion 2** after 25 iterations, and check errors $\mathcal{E}(\rho)$, value-weighted average interest rate, $\mathcal{P}_i$, and $\mathcal{R}_i$ every 10 iterations. We have found that using **Criterion 1** exclusively has no significant effect on the main results, but it typically takes longer for the algorithm to converge.

G.2 Computation of theoretical moments

Many of the statistics that we compute from model output are volume-weighted, which is the standard way many official statistics are computed (e.g., the value-weighted average interest rate). In this section we provide more details on how to perform these calculations in the theory.

Let $\omega_t^{ij}(a,\tilde{a})$ be the share of loans between banks type $i \in \mathbb{N}$ and $j \in \mathbb{N}$ with balances $a \in \mathbb{A}^i$ and $\tilde{a} \in \mathbb{A}_j$ at time t, relative to the total volume of loans in the trading-day, υ. That is,

$$\omega_t^{ij}(a,\tilde{a}) = \frac{\tilde{\upsilon}_t^{ij}(a,\tilde{a})}{\upsilon}$$

where

$$\tilde{\upsilon}_t^{ij}(a,\tilde{a}) = (\Delta\beta_i)\, n_i \sigma_j F_t^i(a) F_t^i(\tilde{a}) |b_t^{ij}(a,\tilde{a})|,$$

and

$$\upsilon = \sum_t \sum_{i,j\in\mathbb{N}} \sum_{(a,\tilde{a})\in\mathbb{A}^i\times\mathbb{A}^j} \tilde{\upsilon}_t^{ij}(a,\tilde{a})$$

is the total volume of loans in the trading day.

The OFR in the model is the volume-weighted mean of all daily traded rates, i.e.,

$$\text{OFR} = \sum_t \sum_{i,j\in\mathbb{N}} \sum_{(a,\tilde{a})\in\mathbb{A}^i\times\mathbb{A}^j} \omega_t^{ij}(a,\tilde{a}) \rho_t^{ij}(a,\tilde{a}). \tag{45}$$

Let υ_i^e and υ_i^r denote the values of all the loans that were extended and received, respectively, throughout the trading day by all banks of type $i \in \mathbb{N}$, i.e.,

$$\upsilon_i^e = \sum_t \sum_{i,j \in \mathbb{N}} \sum_{(a,\tilde{a}) \in \mathbb{A}^i \times \mathbb{A}^j} \omega_t^{ij}(a,\tilde{a}) b_t^{ij}(a,\tilde{a}) \mathbb{I}_{\left\{b_t^{ij}(a,\tilde{a})>0\right\}}$$
$$\upsilon_i^r = \sum_t \sum_{i,j \in \mathbb{N}} \sum_{(a,\tilde{a}) \in \mathbb{A}^i \times \mathbb{A}^j} \omega_t^{ij}(a,\tilde{a}) b_t^{ij}(a,\tilde{a}) \mathbb{I}_{\left\{b_t^{ij}(a,\tilde{a})<0\right\}}.$$

The participation and reallocation measures are $\mathcal{P}_i = (\upsilon_i^e + \upsilon_i^r)/2\upsilon$, and $\mathcal{R}_i = (\upsilon_i^e - \upsilon_i^r)/(\upsilon_i^e + \upsilon_i^r)$, respectively.