Review of Economic Dynamics 33 (2019) 177–204

Contents lists available at ScienceDirect

Review of Economic Dynamics

www.elsevier.com/locate/red

A model of the federal funds market: Yesterday, today, and tomorrow ☆, ☆☆

Gara Afonso [a], Roc Armenter [b,*], Benjamin Lester [b]

[a] Federal Reserve Bank of New York, United States
[b] Federal Reserve Bank of Philadelphia, United States

ARTICLE INFO

Article history:
Received 29 May 2018
Received in revised form 29 March 2019
Available online 18 April 2019

JEL classification:
E42
E43
E44
E52
E58

Keywords:
Monetary Policy Implementation
Federal Funds Market
Over-the-Counter Markets

ABSTRACT

The landscape of the federal funds market changed drastically in the wake of the Great Recession as large-scale asset purchase programs left depository institutions awash with reserves, and new regulations made it more costly for these institutions to lend. As traditional levers for implementing monetary policy became less effective, the Federal Reserve introduced new tools to implement the target range for the federal funds rate, changing this landscape even more. In this paper, we develop a model that is capable of reproducing the main features of the federal funds market, as observed before and after 2008, in a single, unified framework. We use this model to quantitatively evaluate the evolution of interest rates and trading volume in the federal funds market as the supply of aggregate reserves shrinks. We find that these outcomes are highly sensitive to the dynamics of the distribution of reserves across banks.

1. Introduction

In response to the Great Recession, the Federal Reserve resorted to a number of unconventional policies that drastically changed the landscape of the federal funds (FF) market. Prior to 2008, depository institutions actively relied on the FF market, borrowing to satisfy their reserve requirements and payments needs, or lending to avoid holding unremunerated excess reserves. Trading volume in the FF market was robust, averaging more than $250 billion per day, and the majority of trades occurred between banks.[1] In this environment with *scarce reserves*, monetary policy implementation was fairly straightforward: The Open Market Trading Desk at the Federal Reserve Bank of New York would implement the desired target for the effective federal funds rate (EFFR) by adjusting the supply of reserves via open market operations.

In the wake of the 2008 financial crisis, the large-scale asset purchase programs left most depository institutions awash with excess reserves. As a result, trading activity between banks became rare, and volume in the FF market dropped sub-

☆ We thank Todd Keister as well as James Clouse, Huberto Ennis, Julie Remache, and Pierre-Olivier Weill for fruitful discussions, and Michael Blank for exceptional research assistance.

☆☆ Disclaimer: The views expressed in this paper are solely those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Philadelphia, the Federal Reserve Bank of New York, or the Federal Reserve System. Any errors or omissions are the responsibility of the authors.

* Corresponding author.
E-mail address: roc.armenter@phil.frb.org (R. Armenter).

[1] Market size was estimated by aggregating fed funds activity from quarterly regulatory filings. For details on how this estimate is produced, see, for example, Afonso et al. (2013b) and Afonso et al. (2013a).

https://doi.org/10.1016/j.red.2019.04.004
1094-2025/

stantially, to $75 billion or less per day. With few trades occurring between banks, activity in the FF market has been dominated by government-sponsored enterprises (GSEs) looking for some yield on their overnight balances. In this environment with *abundant reserves*, the Fed came to rely on two new policy levers to implement its desired target range for the EFFR: the rate of interest on reserves (IOR), offered exclusively to eligible depository institutions, was set at the top of the target range; and the rate of return at the overnight reverse repurchase (ON RRP) facility, which is available to an expanded set of counterparties including GSEs and some money market funds, was set at the bottom of the range.[2] The interest rate at the discount window—where depository institutions are able to borrow—became essentially irrelevant.[3]

In addition to the large-scale asset purchase programs and the advent of these new policy levers, the FF market has also changed as a result of enhanced regulatory requirements. In particular, FDIC insurance fees have made banks even more reluctant to borrow funds, and liquidity requirements have created incentives for banks to maintain substantial buffers of reserves (and other high-quality liquid assets). Hence, these regulations have only reinforced the shift in the FF market away from its pre-2008 landscape, in which robust bank-to-bank lending prevailed at rates above the interest rate available on overnight reserves (which was zero at the time).

In September 2014, however, the Federal Open Market Committee (FOMC) presented a strategy to shrink or "normalize" the Fed's balance sheet from its current exceptional levels. Given the tight link between the asset holdings of the Fed and the supply of reserves held by banks, a number of crucial questions emerge. For example, as the Fed's balance sheet shrinks and reserves become more scarce, will bank-to-bank lending in the FF market resume? If so, how much will the balance sheet have to shrink before this happens? What policy tools will be needed to ensure interest-rate control, both in the long run and during the transition? How do changes in regulatory requirements affect the answers to all of these questions?

What we do. We develop a simple model that is capable of reproducing the main features of the federal funds market in regimes with either scarce or abundant reserves, as observed before and after 2008, respectively. We use this model as a laboratory to quantitatively evaluate the future conditions in the federal funds market in response to changes in the supply of reserves, policy rates, and regulatory requirements.

We capture the over-the-counter nature of the FF market using a random-search model with two types of market participants: depository institutions (or "banks") and non-depository institutions (or "GSEs"). We assume that GSEs are homogeneous and always looking to lend in the overnight market. However, we allow for relatively rich heterogeneity across banks, and ascribe a central role to the decision of each bank to approach the FF market as a lender or a borrower.

In an environment with scarce reserves, banks with excess balances look to lend out funds to those banks with temporary shortfalls in their reserve holdings. Naturally, banks with large balances are willing to lend out funds at a rate above their outside option—the IOR rate—and banks with temporary shortfalls are willing to borrow at a rate below their outside option—the discount window rate. Hence, in an environment with scarce reserves, there are gains from trade between banks. In equilibrium, this implies robust trading volume in the FF market, driven by bank-to-bank trades; the median traded rate exceeds the IOR rate; and the EFFR is sensitive to small adjustments to the aggregate supply of reserves.

In contrast, when reserves are abundant, few (if any) banks find it profitable to lend, as there are little gains from trading with other banks. Instead, banks look to borrow from a GSE and realize arbitrage profits between the ON RRP rate and the IOR rate. Since there is little or no trade between banks, volume in the FF market is almost reduced to the funds provided by GSEs, and the EFFR typically trades below the IOR rate.

Naturally, each bank's decision to borrow or lend in the FF market depends on its own level of reserves, along with the supply of liquidity coming from GSEs and the spread between policy rates, which determines the potential profits from arbitrage. Moreover, because of the over-the-counter nature of the FF market, a bank's decision to borrow or lend also depends on the distribution of reserves across other banks looking to lend and looking to borrow. Banks' decisions, in turn, determine the market composition as well as the traded rates and market volume. This reinforcing mechanism is what allows the model to reproduce *qualitatively* the varying landscape of the FF market as a function of the aggregate supply of reserves, policy rates, and other factors.

However, the important questions posed above ultimately require *quantitative* answers. To meet this challenge, we start with a careful calibration of our model in an environment with abundant reserves. We use publicly available data from Call Reports for the period 2015-2016 to estimate the empirical distribution of excess reserves across banks, and a host of other observations and existing estimates to discipline the remaining parameters. We find that the model is able to match the observed distribution of market rates and trading patterns quite well. We also check the calibrated model's implications for an environment with scarce reserves by shifting the aggregate supply of reserves to the levels observed in 2002-2006.[4] We confirm that the model reproduces the hallmarks of the scarce-reserves or classic "corridor" regime: FF rates, largely determined by bank-to-bank trades, lie between the IOR rate and the discount window rate; trading volume is elevated; and small open market operations are an effective instrument for controlling market rates.

For our main policy exercise we trace the evolution of the FF market as we reduce aggregate excess reserves from its current levels down to $200 billion. Doing so requires us to specify the complete dynamics of the distribution of excess

[2] In June of 2018, the FOMC implemented a technical adjustment that set the IOR 5 basis points below the top of the target range. A second adjustment was implemented in December 2018.

[3] This rate is currently set 50 basis points above the top of Federal Open Market Committee's (FOMC) target range.

[4] There were also marked differences regarding policy rates and FDIC fees that we incorporate. We do not, however, perform a complete re-calibration.

reserves across banks along the path. In our baseline analysis, we find that the banks with the largest balances return to lending funds when aggregate excess reserves reach about $800 billion, kick-starting a resurgence in FF volume. As excess reserves continue to decline, the EFFR quickly rises above the IOR rate—somewhere between $700 billion and $800 billion—as bank-to-bank trades necessarily execute at rates above the IOR rate. This is an important event, as it marks the end of the implementation framework with abundant reserves that equates the IOR rate to the top of the target range for the EFFR. However, the level of aggregate excess reserves must decrease an additional $350 billion or so before the EFFR rises more than 5 basis points above the IOR rate and becomes responsive to small open market operations—what would be the hallmark of a classic corridor system.

Importantly, we find that the evolution of the FF market is sensitive to the dynamics of the distribution of excess reserves across banks, which are difficult to anticipate. In particular, the extent to which the largest banks hoard reserves is crucial to determine when the EFFR rises above the IOR rate. By varying the rate at which banks with higher balances reduce their holdings of reserves, relative to those banks with lower balances, the EFFR can first drift above the IOR rate with as few as $400 billion in aggregate excess reserves, or as much as $1 trillion.

As far as we know, Kim et al. (2017) is the only other work to attempt a similar exercise, though their approach is exclusively theoretical and based on a centralized market. As we do, Kim et al. (2017) carefully model the borrowing costs imposed by the FDIC fees. However, they choose to emphasize the possibility that interbank trading never returns. Such a scenario is actually possible in our model, but appears extremely unlikely: It would require either very high balance sheet costs, such that no bank wants to borrow, or a near-degenerate distribution of reserves across banks, such that there are no gains from trade among them.

Related literature. There is a long tradition of research on the FF market starting with Poole (1968). Most existing work models the FF market as a centralized, Walrasian market and studies regimes with scarce reserves. Recent contributions with a focus on the interbank market include Furfine (1999) and Whitesell (2006), among many others. Assuming a centralized market also allows for embedding the FF market into a general equilibrium model, as in Martin et al. (2013), Ennis (2014), and Bech and Keister (2017).

Starting with Ashcraft and Duffie (2007), several recent models have aimed to capture the over-the-counter nature of the FF market. Given the historical precedence, however, most of these papers focus on regimes with scarce reserves; see, among others, Berentsen and Monnet (2008), Ennis and Weinberg (2013), and Afonso and Lagos (2015).[5] The current regime with abundant reserves, and its implications for the federal funds rate, has only recently been studied in Bech and Klee (2011), Armenter and Lester (2017), and Williamson (2018). Given this context, the current paper can be viewed as bridging existing work on the FF market as an over-the-counter market under both scarce and abundant reserves.[6]

2. Model

2.1. Agents

There are three types of agents in the economy: a central bank, which we will refer to as "the Fed"; financial institutions that are eligible to earn interest on overnight reserves (IOR) at the Fed, which we will refer to as "banks"; and financial institutions that are not eligible to earn interest on overnight reserves at the Fed, which we will refer to as government-sponsored enterprises or "GSEs." We describe each of these agents in greater detail below.

The Fed. Consistent with the current operating framework in the U.S., we assume that the Fed manages three distinct facilities. First, as noted above, the Fed pays an interest rate i^{or} to banks that deposit overnight reserves.[7] Second, the Fed operates an overnight reverse repo (ON RRP) facility that offers an overnight interest rate on deposits $i^{rr} < i^{or}$ that is available to all financial institutions, including GSEs. Lastly, the Fed lends to banks at the discount window (DW) at an overnight interest rate $i^{dw} > i^{or}$.

GSEs. There is a mass γ of GSEs, each with y units of excess cash. GSEs can always deposit these funds at the ON RRP facility, but they would prefer to lend to a bank that is willing to pay a rate greater than i^{rr}. However, there are frictions in the interbank or "fed funds" market, and not every GSE will meet with a bank. A GSE that is matched with a bank earns an overnight rate ρ that is negotiated in the bilateral match. A GSE that is not matched can access the ON RRP facility and earn the overnight repo rate i^{rr}.

Banks. There is a mass of banks, which we normalize to 1, that are heterogeneous across several dimensions. In particular, a bank can be characterized by the vector $\omega \equiv (r, \ell, e, d, R, \kappa)$. The first four components of the vector ω describe the bank's

[5] See also Bianchi and Bigio (2017) for a general equilibrium perspective.

[6] Focusing on the Euro area interbank market, Bech and Monnet (2016) also deploy a two-sided search model where banks choose to be lenders or borrowers, building on the original contribution by Matsui and Shimizu (2005).

[7] We assume that all overnight deposits earn the same interest rate. Though the Fed has discretion to pay different rates on required and excess reserves, they are currently set at the same rate.

balance sheet: r, ℓ, e, and d represent the values of the bank's reserve balances, outstanding loans, equity, and deposits at the beginning of the period, respectively. The last two components of this vector relate to regulatory requirements: R denotes the bank's required reserve balances and κ denotes the balance sheet costs that a bank has to pay on its total assets. We describe these regulatory requirements, and associated costs, in more detail below, when we spell out the timing of events. Afterwards, we will show that it is sufficient to consider just two dimensions of heterogeneity: a bank's *excess* reserves at the beginning of the period, which we denote $x \equiv r - R$, and its balance sheet costs, κ.

2.2. Sequence of events

The game proceeds over a single day, which is broken into three sub-periods that we conveniently refer to as the morning ($t = 0$), afternoon ($t = 1$), and evening ($t = 2$).

Morning. At the beginning of $t = 0$, each bank decides whether to be a lender (L) or a borrower (B) by comparing the payoffs to lending and borrowing in the fed funds market. As noted above, GSEs always look to lend in the fed funds market. After banks decide whether to be a lender or a borrower, matching and trading occur in the fed funds market. In particular, we assume that the fed funds market is a decentralized or "over-the-counter" market, where lenders and borrowers are matched in bilateral pairs according to a standard matching function. More specifically, if a mass μ_L of lenders are searching for a borrower and a mass μ_B of borrowers are searching for a lender, then $m(\mu_L, \mu_B) \leq \min\{\mu_L, \mu_B\}$ matches are formed. We assume that matching is random, so that each lender is matched with probability $\frac{m(\mu_L,\mu_B)}{\mu_L}$ and each borrower is matched with probability $\frac{m(\mu_L,\mu_B)}{\mu_B}$. We also assume that the matching function exhibits constant returns to scale, so that the probability of matching is determined by the "market tightness," or ratio of borrowers to lenders.

Once matches are formed, the lender and the borrower trade if the gains from trade are positive. We assume that the terms of trade—the amount that the bank will borrow, f, and the interest rate, ρ—are determined by Nash bargaining. For simplicity, we assume that lenders and borrowers have equal bargaining power in bank-to-bank trades, while banks have bargaining power θ when borrowing from a GSE.

Afternoon. After trading in the fed funds market, banks incur costs based on their balance sheets at $t = 1$. While there are several sources of these so-called balance sheet costs, we focus on the costs associated with FDIC fees, which are assessed on a bank's total assets. We do so for several reasons. First, these are arguably the largest source of balance sheet costs. Second, other commonly cited regulatory costs—such as the liquidity coverage ratio—typically do not play a crucial role in determining a bank's decision to borrow or lend in the fed funds market, and hence are unlikely to affect our main results.[8] Lastly, since most branches of foreign banks are exempt from FDIC fees, these costs constitute an important source of heterogeneity across banks for understanding the data, as we discuss in Appendix B, where we describe the data in greater detail.

To calculate the balance sheet costs that a bank incurs after trading in the fed funds market, consider the two tables below, which illustrate the balance sheets of a bank at $t = 1$ that either borrowed ($f > 0$) or lent ($f < 0$) at $t = 0$.

Balance Sheet at $t = 1$ After Borrowing

Assets	Liabilities
$r + f$	e
ℓ	d
	f

Balance Sheet at $t = 1$ After Lending

Assets	Liabilities
$r - f$	e
ℓ	d
f	

We assume that the balance sheet cost κ is assessed on each unit of assets held at $t = 1$, so that the total costs can be written

$$\kappa \left[\ell + r + \max\{0, f\}\right]. \tag{1}$$

Before proceeding, it is worth highlighting the asymmetric treatment of funds in the determination of balance sheet costs: borrowing $f > 0$ units of reserves increases balance sheet costs by κf, but lending $-f > 0$ units of reserves does not decrease balance sheet costs.

Evening. In the last stage of the game, as in Poole (1968), each bank receives a late payment shock z, where we assume that z is distributed according to a cumulative distribution function $G(z)$ with mean zero. We adopt the convention that $z > 0$ is an outflow and $z < 0$ is an inflow.

[8] In Appendix A, we provide a more formal discussion of why liquidity coverage ratios do not affect banks' decision to borrow or lend funds in an environment with abundant reserves. However, it is worth noting that, as reserves decline, liquidity regulations *could* play a more significant role in banks' portfolio management decisions.

After the late payment shocks arrive, banks may have to borrow at the discount window in order to satisfy reserve requirements. In particular, let

$$r' = r + f - z \tag{2}$$

denote the end-of-day reserve balances of a bank that borrows an amount $f > 0$ in the fed funds market and realizes a late payment shock z. Then, letting

$$\delta = \max\{R - r', 0\} \tag{3}$$

denote the amount that it would have to borrow from the discount window, the bank would incur cost $i^{dw}\delta$ from borrowing at the discount window.

3. Equilibrium construction and analysis

To construct an equilibrium, we first derive the outcome of a match between two randomly selected banks, and the outcome of a match between a GSE and a randomly selected bank. Then, given the terms of trade in each type of match, we derive the (expected) payoffs that a bank receives from choosing to enter the fed funds market as a borrower or lender. We show that the equilibrium can be summarized by simple cutoff strategies, which determine the fraction of banks that ultimately borrow or lend in the fed funds market. Finally, we exploit the equilibrium characterization to study how interest rates and trading volume in the fed funds market depend on policy rates, balance sheet costs, and the distribution of reserves across banks.

3.1. Payoffs and gains from trade

As a first step, we derive the gains from trade that a bank or a GSE realizes from borrowing or lending some amount f at an interest rate ρ.

Banks. The payoffs to a bank from borrowing ($f > 0$) or lending ($f < 0$) at rate ρ can be written

$$\pi(f, \rho; \omega) = i^{or} \int_{-\infty}^{\infty} (r' + \delta) dG(z) - i^{dw} \int_{-\infty}^{\infty} \delta dG(z) - \rho f - \kappa [\ell + r + \max\{0, f\}]. \tag{4}$$

The first term in equation (4) is the expected interest the bank earns on its overnight reserves, and the second term is the expected interest it pays from borrowing at the discount window. The third term is the interest it pays or receives from borrowing or lending, respectively, in the fed funds market. Finally, the last term is the balance sheet cost assessed on its assets at $t = 1$, derived above in equation (1).

Equation (4) can be rewritten as

$$\pi(f, \rho; \omega) = f\left[i^{or} - \rho - \mathbf{1}_{\{f>0\}}\kappa\right] - \left(i^{dw} - i^{or}\right) \int_{-R+r+f}^{\infty} [R - r - f + z] dG(z) + C, \tag{5}$$

where $C = ri^{or} - \kappa(r + \ell)$ is a constant that is unaffected by the bank's decision to borrow or lend. Equation (5) highlights the (potential) costs and benefits from borrowing or lending. For example, borrowing unambiguously decreases the expected costs of borrowing from the discount window: as is evident from the second term in (5), both the probability of visiting the discount window and the amount borrowed (conditional on visiting the discount window) decreases in f. If $i^{or} - \rho - \kappa > 0$, there is an additional benefit to borrowing, as the bank takes advantage of an arbitrage opportunity in the fed funds market. Alternatively, if $i^{or} - \rho - \kappa < 0$, then borrowing in the fed funds market is costly.

To derive the gains from trade that a bank realizes by borrowing or lending, let $x \equiv r - R$ denote a bank's *excess reserves*, i.e., the bank's initial reserves in excess of its required reserves. The gains from trade, then, from borrowing an amount $f > 0$ at an interest rate ρ, relative to not trading ($f = 0$), can be written[9]

$$\Delta\pi_B(f, \rho; \omega) = f\left[i^{or} - \rho - \kappa\right] + \left(i^{dw} - i^{or}\right)\left[f[1 - G(x + f)] + \int_{x}^{x+f} (z - x) dG(z)\right]. \tag{6}$$

Looking forward, the two terms in equation (6) highlight some of the key economic forces in the model. In particular, the first term represents the dominant benefit from borrowing when reserves are abundant and banks' motive for borrowing is

[9] We provide a more detailed derivation of equation (6) in the Appendix.

to exploit an arbitrage opportunity, while the second term represents the dominant benefit from borrowing when reserves are scarce and banks' excess reserves are sufficiently close to zero.

The gains from trade that a bank realizes from lending an amount $-f > 0$ at an interest rate ρ can be written

$$\Delta\pi_L(f,\rho;\omega) = -f\left[\rho - i^{or}\right] + \left(i^{dw} - i^{or}\right)\left[f[1-G(x+f)] + \int_{x+f}^{x}(x-z)dG(z)\right]. \tag{7}$$

Note that the gains from trade that accrue to a bank from borrowing or lending depend on only two dimensions of heterogeneity across banks, x and κ. Hence, for the remainder of the analysis, it will be sufficient to consider a bank's type $\omega \equiv (x,\kappa)$.

GSEs. The payoffs to a GSE are much simpler: a GSE that lends an amount $-f \in [0, y]$ at rate ρ receives a payoff[10]

$$\tilde{\pi}(f,\rho) = (y+f)i^{rr} - f\rho. \tag{8}$$

In words, a GSE earns a return ρ on the amount that it lends, and deposits the remainder, $y+f$, at the ON RRP facility. Hence, the gains from trade that accrue to a GSE from lending are simply

$$\Delta\tilde{\pi}_L(f,\rho) = -f\left(\rho - i^{rr}\right). \tag{9}$$

3.2. Trade outcomes

We now derive the outcomes in a bank-to-bank match and a GSE-to-bank match.

Bank-to-bank trades. Given the gains from trade, the optimal trade size and corresponding interest rate are determined by solving the standard Nash bargaining problem, where we have assumed that each side has equal bargaining power. The solution specifies an optimal trade size, $\tau \geq 0$, that maximizes the joint surplus from the match,[11]

$$\begin{aligned} S(\tau;\omega,\omega') &= \Delta\pi_B(\tau,\rho;\omega) + \Delta\pi_L(-\tau,\rho;\omega') \\ &= -\kappa\tau + \\ &\quad \left(i^{dw} - i^{or}\right)\left[f[G(x'-\tau) - G(x+\tau)] + \int_{x}^{x+\tau}(z-x)dG(z) + \int_{x'-\tau}^{x'}(x'-z)dG(z)\right]. \end{aligned}$$

The corresponding interest rate ensures that the gains from trade that accrue to the borrower and lender are equal to half of the joint surplus, evaluated at the optimal trade size. We summarize the solution in the following lemma.

Lemma 1. *In a meeting between a borrowing bank of type $\omega = (x,\kappa)$ and a lending bank of type $\omega' = (x',\kappa')$, trade occurs if, and only if,*

$$\kappa < \left(i^{dw} - i^{or}\right)\left[G(x') - G(x)\right]. \tag{10}$$

In this case, the optimal trade size, $\tau^\star(\omega,\omega')$, is the value of $\tau > 0$ that satisfies

$$\kappa = \left(i^{dw} - i^{or}\right)\left[G(x'-\tau) - G(x+\tau)\right] \tag{11}$$

and the corresponding interest rate, $\rho^\star(\omega,\omega')$, is the value of ρ that satisfies

$$\Delta\pi_B(\tau^\star(\omega,\omega'),\rho;\omega) = \Delta\pi_L(-\tau^\star(\omega,\omega'),\rho;\omega') = \frac{1}{2}S(\tau^\star(\omega,\omega');\omega,\omega'). \tag{12}$$

If (10) is violated, then there is no trade, i.e., $\tau^\star(\omega,\omega') = 0$.

[10] We adopt the convention of using the tilde to denote variables corresponding to trades involving GSEs. Also, note that we focus exclusively on the payoffs to GSEs from lending, anticipating the result that GSEs never borrow in equilibrium.

[11] That is, the solution to the bargaining problem specifies that the lender transfers $-f = \tau$ dollars and the borrower receives $f = \tau$ dollars.

The inequality in (10) ensures that the balance sheet costs are sufficiently small, relative to the gains associated with a reduction in the likelihood of a bank visiting the discount window. Given, e.g., the excess reserves of the lender, x', these gains from trade increase as the borrower's excess reserves, x, get closer to zero.

In what follows, it will be helpful to define the total surplus created in a trade between a borrower of type ω and a lender of type ω', given the optimal quantity is traded, which we denote by $S^\star(\omega,\omega') \equiv S(\tau^\star(\omega,\omega');\omega,\omega')$. Using (11), this can be written

$$S^\star(\omega,\omega') = \left(i^{dw} - i^{or}\right)\left[\int_{x}^{x+\tau^\star(\omega,\omega')} (z-x)dG(z) + \int_{x'-\tau^\star(\omega,\omega')}^{x'} (x'-z)dG(z)\right] \tag{13}$$

if (10) holds, and zero otherwise.

GSE-to-bank trades. Again, the optimal trade size and corresponding interest rate are determined by Nash bargaining, where we assume that the bank has bargaining power $\theta \in [0,1]$. As in the analysis above, the optimal trade size maximizes the joint surplus when the bank borrows an amount $\tilde{\tau} \in [0,y]$ from the GSE. This joint surplus satisfies

$$\begin{aligned}\tilde{S}(\tilde{\tau};\omega) &= \Delta\pi_B(\tilde{\tau},\rho;\omega) + \Delta\tilde{\pi}_L(-\tilde{\tau},\rho)\\ &= \tilde{\tau}\left[i^{or} - i^{rr} - \kappa\right] + \left(i^{dw} - i^{or}\right)\left[\tilde{\tau}[1-G(x+\tilde{\tau})] + \int_{x}^{x+\tilde{\tau}} (z-x)dG(z)\right].\end{aligned}$$

We summarize the solution in the following lemma.

Lemma 2. *If*

$$\kappa < \left(i^{or} - i^{rr}\right), \tag{14}$$

then the optimal trade size in a meeting between a borrower of type $\omega = (x,\kappa)$ and a GSE with y units of cash is $\tilde{\tau}^\star(\omega) = y$. The corresponding interest rate, $\tilde{\rho}^\star(\omega)$, is the value of ρ that satisfies

$$\Delta\pi_B(y,\rho;\omega) = \theta\tilde{S}(y;\omega). \tag{15}$$

If (14) is violated, then there is no trade, i.e., $\tilde{\tau}^\star(\omega) = 0$.

In words, under condition (14), the surplus is strictly increasing in $\tilde{\tau}$, and hence it is optimal for the GSE to lend the bank all of its funds. In what follows, we will restrict attention to the case where (14) holds for all banks. Given this restriction, it is convenient to introduce the notation $\tilde{S}^\star(\omega) = \tilde{S}(y;\omega)$. Comparing $S^\star(\omega,\omega')$ and $\tilde{S}^\star(\omega)$ shows that the gains from trade between two banks come exclusively from the differences in their reserve balances, x, while the gains from trade between a bank and a GSE also stem from the spread between i^{or} and i^{rr}.

3.3. Equilibrium

We established above that a sufficient statistic for a bank's type, ω, is its level of excess reserves, x, and its balance sheet cost, κ. Anticipating our empirical exercise, below, let us assume that there is a finite number J of balance sheet costs. We let β_j denote the fraction of banks with balance sheet cost $\kappa_j \in \mathcal{J} \equiv \{1,2,...,J\}$, and $F_j(x)$ denote the conditional distribution of excess reserves among banks with balance sheet cost κ_j.

In equilibrium, banks choose to be borrowers or lenders according to a simple threshold rule: a bank facing cost κ_j chooses to lend if its level of excess reserves exceeds a threshold, $x_j^\star \in \mathbb{R} \cup \infty$, and borrows otherwise. Taking as given these decision rules for all other banks, the expected gains from borrowing for an individual bank with balance sheet cost κ and excess reserves x can be written

$$\begin{aligned}\Pi_B(x,\kappa) = \frac{m(\mu_L,\mu_B)}{\mu_B}\Bigg\{&\frac{\gamma}{\mu_L}\theta\tilde{S}^\star(x,\kappa) + \\ &\sum_{j\in\mathcal{J}} \frac{\beta_j\left[1-F_j\left(x_j^\star\right)\right]}{\mu_L}\int_{x_j^\star}^{\infty}\frac{1}{2}S^\star\left((x,\kappa),(x',\kappa_j)\right)\frac{dF_j(x')}{1-F_j\left(x_j^\star\right)}\Bigg\},\end{aligned} \tag{16}$$

where μ_L and μ_B denote the equilibrium measures of borrowers and lenders implied by the threshold rules, i.e.,

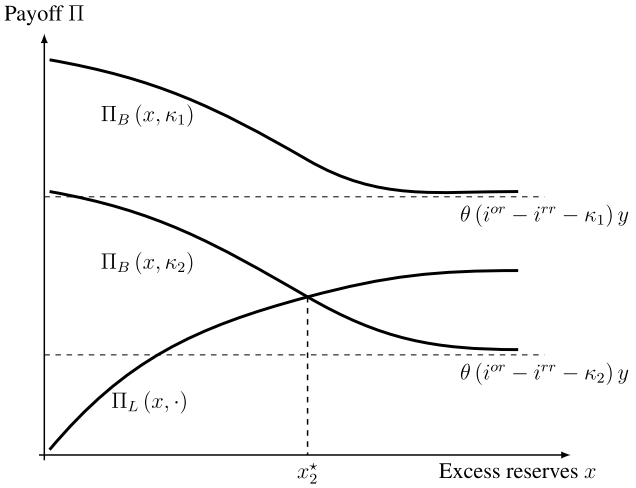

Fig. 1. Borrowing and Lending.

$$\mu_L = \gamma + \sum_{j\in\mathcal{J}} \beta_j \left[1 - F_j\left(x_j^\star\right)\right] \tag{17}$$

$$\mu_B = \sum_{j\in\mathcal{J}} \beta_j F_j\left(x_j^\star\right). \tag{18}$$

Intuitively, the expected gains from borrowing depend on the probability of meeting a lender, $\frac{m(\mu_L,\mu_B)}{\mu_B}$, and the expected gains from trade conditional on meeting various types of lenders. With probability $\frac{\gamma}{\mu_L}$, the borrower will meet a GSE, in which case the gains from trade are a fraction θ of the surplus $\tilde{S}^\star(x,\kappa)$, where we have expanded the vector $\omega=(x,\kappa)$ for expositional purposes. Alternatively, the borrower meets a bank of type j with probability $\frac{\beta_j\left[1-F_j\left(x_j^\star\right)\right]}{\mu_L}$, in which case the two banks split the surplus if it is positive.[12]

Similar logic reveals that the expected gains from lending for an individual bank with excess reserves x' and balance sheet cost κ' can be written

$$\Pi_L(x',\kappa') = \frac{m(\mu_L,\mu_B)}{\mu_L} \sum_{j\in\mathcal{J}} \frac{\beta_j F_j\left(x_j^\star\right)}{\mu_B} \int_{-\infty}^{x_j^\star} \frac{1}{2} S^\star\left((x,\kappa_j),(x',\kappa')\right) \frac{dF_j(x)}{F_j\left(x_j^\star\right)}. \tag{19}$$

An equilibrium, then, can be summarized by a vector of thresholds, $\mathbf{x}^\star = \left(x_1^\star, x_2^\star, ..., x_J^\star\right)$, with each $x_j^\star \in \mathbb{R}\cup\{\infty\}$. From (16) and (19), one can easily show that, taking others' behavior as given, the expected payoff to borrowing is decreasing in x while the expected payoff to lending is increasing in x. Hence, an interior solution $x_j^\star < \infty$ is the unique solution to $\Pi_B\left(x_j^\star,\kappa_j\right) = \Pi_L\left(x_j^\star,\kappa_j\right)$. Alternatively, if

$$\lim_{x\to\infty} \Pi_B\left(x,\kappa_j\right) \geq \lim_{x\to\infty} \Pi_L\left(x,\kappa_j\right),$$

then $x_j^\star = \infty$ and all banks with cost κ_j choose to borrow.

Since the joint surplus is independent of the balance sheet costs of the lending bank, it is immediate from (19) that the payoff to lending is independent of a bank's balance sheet cost, κ, while the payoff to borrowing is naturally decreasing in κ. Given these properties, it is immediate that

$$\kappa_1 \leq \kappa_2 \leq \cdots \leq \kappa_J \Rightarrow x_1^\star \leq x_2^\star \leq \cdots \leq x_J^\star,$$

as banks with higher balance sheet costs find lending relatively more attractive than borrowing. Fig. 1 illustrates a simple example in which $J=2$, $x_1^\star < \infty$, and $x_2^\star = \infty$. The definition below formalizes the equilibrium concept.

[12] Recall that $S^\star(\omega,\omega')=0$ if x is not sufficiently large relative to x' or if κ is too large, so that (10) is violated.

Definition 1. An equilibrium is a vector $\mathbf{x}^\star = \left(x_1^\star, x_2^\star, ..., x_J^\star\right) \subset [\mathbb{R} \cup \infty]^J$ such that, for each $j \in \mathcal{J}$,

$$\gamma\theta\tilde{S}^\star(x_j^\star, \kappa_j) + \sum_{j' \in J} \beta_{j'} \int_{x_{j'}^\star}^{\infty} \frac{1}{2} S^\star\left((x_j^\star, \kappa_j), (\hat{x}, \kappa_{j'})\right) dF_{j'}(\hat{x})$$

$$\geq \sum_{j' \in J} \beta_{j'} \int_{-\infty}^{x_{j'}^\star} \frac{1}{2} S^\star\left((\hat{x}, \kappa_{j'}), (x_j^\star, \kappa_j)\right) dF_{j'}(\hat{x}), \tag{20}$$

with equality if $x_j^\star < \infty$.

3.4. Aggregate reserves and the fed funds market

From the definition of equilibrium above, one can see that the landscape of the fed funds market—in particular, trading volume and the distribution of interest rates—depends heavily on which banks choose to enter the market as borrowers or lenders. For example, when all banks choose to borrow, trade occurs exclusively between GSEs and banks. Though there will naturally be dispersion in the interest rates that are traded, since the joint surplus depends on the banks' excess reserves, all such trades occur at a rate $\rho \in \left(i^{rr}, i^{or}\right)$. As a result, the EFFR, which is calculated as a volume-weighted median across trades in the fed funds market, must lie below i^{or}. Moreover, if the measure of GSEs, γ, is small relative to the measure of banks, which we normalized to 1, then the number of trades in the fed funds market will be relatively low. The following lemma reports a sufficient condition that ensures all banks choose to borrow, helping to link the key features of the economic environment—namely, the distribution of excess reserves and balance sheet costs, along with the policy rates—to rates and volume in the fed funds market.

Lemma 3. *If*

$$\gamma\theta\left[i^{or} - i^{rr} - \kappa_J\right] y > \frac{1}{2} \sum_{j \in \mathcal{J}} \beta_j \int_{-\infty}^{\infty} \left[\left(i^{dw} - i^{or}\right) \int_{x}^{\infty} (z - x) dG(z)\right] dF_j(x), \tag{21}$$

then the unique equilibrium is $x_j^\star = \infty$ for all $j \in \mathcal{J}$, i.e., all banks borrow.

The sufficient condition in (21) is derived by looking for conditions under which even the bank that has the most incentive to lend—a bank with balance sheet costs κ_J and arbitrarily large excess reserves—prefers to borrow, taking as given that all other banks are borrowing as well. The left-hand side of (21) is proportional to the expected surplus such a bank receives from borrowing from a GSE, which depends on the mass of GSEs, the bank's bargaining power, and the size of the surplus available from interest rate arbitrage. By similar logic, the right-hand side is proportional to the expected gains from lending to a bank. These gains depend on the mass of banks (normalized to 1), the bargaining weights (set to $\frac{1}{2}$), and the surplus created by our candidate bank lending enough funds to *spare* its trading partner from potentially having to use the discount window. In particular,

$$\bar{C}^{dw} \equiv \sum_{j \in \mathcal{J}} \beta_j \int_{-\infty}^{\infty} \left[\left(i^{dw} - i^{or}\right) \int_{x}^{\infty} (z - x) dG(z)\right] dF_j(x)$$

is the unconditional expected cost (across all banks) of borrowing from the discount window after late-payment shocks arrive. Since this cost converges to zero as banks' excess reserves grow large, equation (21) conveys the idea that all banks choose to borrow when they are sufficiently satiated in reserves, and when the gains from borrowing from GSEs are sufficiently large.

As reserves are withdrawn from the system, and the distribution of excess reserves shifts left (in a first-order stochastic dominance sense), then $\bar{C}^{dw}$ will increase and this condition will fail. Indeed, when the mass of banks that need to borrow (to avoid the discount window) becomes sufficiently large, those banks with large levels of excess reserves and large balance sheet costs will find it optimal to lend instead of borrow. Since bank-to-bank trades occur at rates above i^{or}, the EFFR will rise as a greater fraction of fed funds trades will be executed between banks. Moreover, if the measure of banks is large relative to the measure of GSEs, the entry of some banks as lenders will result in more trades in the fed funds market. In other words, as the supply of reserves in the market declines, the fed funds market will transition from what it looks like "today"—with relatively few trades, mostly between GSEs and banks, executed at rates in the interval $\left(i^{rr}, i^{or}\right)$—to what it looked like before the crisis, or "yesterday"—with more trades, often between two banks, executed at rates above i^{or}.

Hence, our model is capable of qualitatively reproducing the key features of the fed funds market before and after the crisis. However, many of the most important questions are *quantitative* in nature, and pertain to the transition from the current environment, with abundant reserves, to a future state of the world with fewer reserves. For example, policymakers may want to know the quantity of aggregate reserves that must be drained before the FF market returns to an active interbank market, or at least to ensure that the EFFR rises *above* the IOR. To answer these questions, and many more, we now turn to the quantitative predictions of our model.

4. Calibration

In this section, we solve the model computationally and calibrate the parameters using data from the fed funds market before the crisis (when reserves were *scarce*) and after the crisis (when reserves were *abundant*).[13] We start with the latter, taking advantage of superior data availability to discipline many of the model's parameters.

4.1. Today: abundant reserves

We take the period of 2015–2017 to be representative of an environment with "abundant" reserves. Between October 2014, when the last large-scale asset purchase program was completed, and June 2017, when the FOMC announced its plans to normalize the balance sheet, the size of the total System Open Market Account (SOMA) remained stable at approximately $4.5 trillion. Moreover, total reserves exceeded $2.1 trillion throughout this time span, with about 95 percent of all reserves being held in excess of reserve requirements. Lastly, lending in the FF market during this period was completely dominated by GSEs, which is a key implication of our model in an environment of abundant reserves.

4.1.1. Parameter choices

Whenever possible, we assign parameter values using either direct observations from the data or estimates from existing studies: We choose the administered rates offered by the Fed, the balance sheet costs associated with FDIC fees, and the distribution of excess reserves in this manner. The choice of the matching function and parameters associated with GSEs can also be tied reasonably tightly to direct observations. For the remaining parameters we rely on indirect inference, matching the model implications to the data.

Throughout the section we make extensive use of quarterly Call Report data, consolidated at the bank-parent level, which are informative about a number of the characteristics of banks—most notably regarding the distribution of reserves. Regarding the distribution of rates within the FF market, we rely on data provided by the Federal Reserve Bank of New York.[14] See Appendix B for details regarding the data documentation and construction.

Administered rates. The FOMC has increased its target range for the EFFR several times since lifting off from the zero lower bound in December 2015, but it has kept the implementation framework intact. In particular, the ON RRP and IOR rates have been set at the bottom and top of the target range, respectively, which is 25 basis points wide.[15] The discount window (or primary credit) rate has been set 50 basis points above the top of the target range for federal funds. Fed funds rates have moved in virtual lockstep with the target range—as our model would predict—so in our calibration we are somewhat free to choose any time frame within this period.[16] We pick the rates prevalent between March and June of 2017, when the ON RRP rate was 75 basis points, the IOR rate was 1 percent, and the discount window rate was 1.5 percent. The same period will inform the values for market rates.

Balance sheet costs. Consistent with our focus on FDIC fees, we split depository institutions between those that are insured by the FDIC and those that are not. The latter include most U.S. branches and agencies of foreign banking organizations (FBOs) and credit unions. We set the share of FDIC-insured institutions among banks to 87 percent, matching their share of total assets according to the Call Report data.

For FDIC-insured institutions, we pin their balance sheet costs to the latest estimate of the effective FDIC rate by Banegas and Tase (2016), 7 basis points. Two additional observations from Banegas and Tase (2016) are worth noting. First, effective FDIC rates have been declining as banks have been able to reduce their FDIC assessment fees, so our choice may be an upper bound of balance sheet costs, especially when we project into the future. Second, there is substantial variation in

[13] Note that we did not provide a formal proof of uniqueness in Section 3. However, when solving the model numerically, the equilibrium was unique for all parameter values that we tested. Later in this section, however, we introduce an additional dimension of heterogeneity: we allow the distribution of late payment shocks to depend on the size of banks' balance sheets. Under this specification, a broad search of the parameter space revealed the potential for multiplicity, though we did not find multiplicity for our calibrated parameter values.

[14] See https://www.newyorkfed.org/markets.

[15] At the June 12-13, 2018 FOMC meeting, the Committee raised the target range for the federal funds rate to 1 3/4 to 2 percent, and the Board of Governors raised the interest rates on required and excess reserve balances (IOR) to 1.95 percent. This technical adjustment placed IOR at a level 5 basis points below the top of the FOMC's target range for the federal funds rate. The FOMC implemented a second technical adjustment in December 2018. Our analysis ends at the end of June 2017 and abstracts from these dynamics.

[16] Indeed there is very little variation in rates day to day, with the exception of month ends, when some financial institutions engage in window dressing and put substantial downward pressure on FF rates.

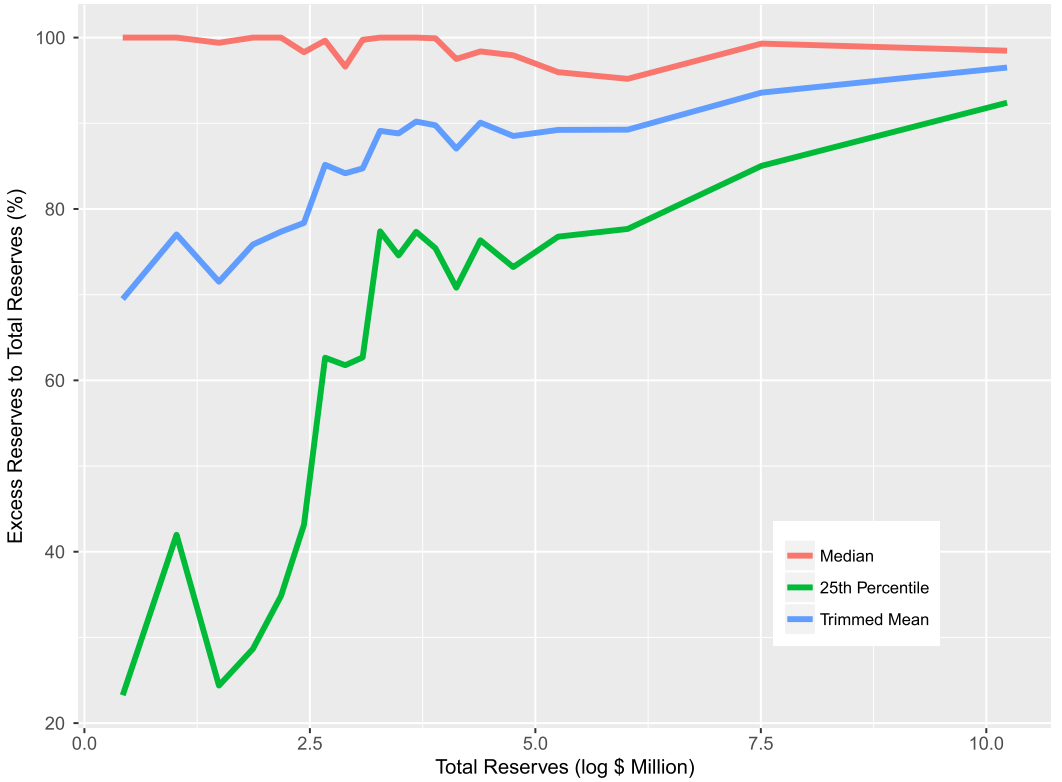

Fig. 2. Ratios of Excess Reserves to Total Reserves, by Total Reserves.

effective FDIC rates that depend on bank size. It is thus possible that our calibration understates the dispersion in rates and the level of borrowing in the data.[17] Naturally, non-insured institutions have zero balance sheet costs.

Distribution of excess reserves. The distribution of excess reserves is a key input to our model and, as we discuss later, the main determinant of the dynamics of the FF market as the aggregate supply of reserves decreases. Unfortunately, Call Reports include neither data on excess reserves nor all the information necessary to compute required reserves. We instead impute a reserve requirement value for each bank based on available information, including various transaction accounts, cash and currency, and demand deposits.[18] We then subtract the required reserves from total reserves—which *are* observed in the data—to obtain our estimate of excess reserves. To reduce some of the remaining measurement error in our imputed values, we average each bank's excess reserves holdings across 2015-2016.[19] We then simply use the resulting empirical cumulative distribution function (CDF), separately for FDIC-insured banks and other institutions, in order to obtain the conditional distributions $F_1(x)$ and $F_2(x)$.

It is worth noting that excess reserves are heavily concentrated at the top. Fig. 2 sorts banks into bins according to their total reserves, and then displays several moments of the distribution of excess reserves to total reserves within each bin. In every bin, about half of the banks have ratios at or very close to 100 percent, as shown by the median.[20] Both a 5 percent trimmed mean and the first quartile show a clearly upward profile: While a quarter of the banks with few total reserves have ratios below 40 percent, virtually all larger banks have ratios of 80 percent or more. Lower percentiles are very noisy, but we do find that the share of banks with very low ratios is decreasing in size as well.

Matching function. Daily data on outstanding amounts at the ON RRP facility shows very little take-up by GSEs outside dates surrounding month ends.[21] As lenders are clearly the short side of the market currently, we simply set $m(\mu_L, \mu_B) = \min\{\mu_L, \mu_B\}$, which implies that there are no unmatched GSEs. In any case, the choice of the matching function has no

[17] Though the model can certainly encompass richer variation in balance sheet costs, it does so at considerable computational cost. Moreover, data limitations would make it difficult to discipline the additional heterogeneity.

[18] See Appendix B for additional details.

[19] Another source of noise in the data is the fact that Call Reports provide the state of the balance sheet at quarter ends, which are not particularly representative of the average position during the quarter. See the Appendix for additional details.

[20] There are several reasons why this ratio may equal 100 percent. Depository institutions without transaction deposits have no reserve requirements—and thus a 100 percent ratio of excess reserves to total reserves. These institutions include both branches of foreign banks and some specialty domestic institutions.

[21] See Markets Group (2016), chart 6.

implications for traded rates in our model, as long as it maintains constant returns to scale: The number of realized matches simply scales up total traded volume in the market.

Distribution of late payment shocks. The calibration of $G(z)$, the distribution of payment shocks, poses several challenges. A direct estimate is impossible because data on reserve balances by bank are not publicly available at high frequencies. Moreover, despite the omnipresence of the Poole model, there are surprisingly few estimates in the literature that could provide guidance, especially because those that exist predate the era of abundant reserves. Lastly, our specification of additive (as opposed to proportional) payment shocks clashes with the enormous variation in total assets across banks.

To address these concerns, we introduce heterogeneity in the distribution of payment shocks, allowing the shocks to be correlated with the bank's balance sheet size, while simultaneously keeping the specification as parsimonious as possible. In particular, we assume that late payment shocks z follow a Laplace distribution, centered at zero, with CDF

$$G(z; x) = 1 - \frac{1}{2}\exp(-\xi(x)z)$$

for $z \geq 0$, where $\xi(x) > 0$ is the bank-specific scale parameter.[22] The choice of the Laplace distribution is mainly for analytic convenience, though we believe that it may be well suited to capture the banks' concern for rare but large payment shocks. We use a logistic function to link the bank's balances to the bank-specific scale parameter,

$$\xi(x) = \bar{\xi}(1 - \exp(-q_0 - q_1 x))$$

where $\bar{\xi}$ acts as the upper bound on the scale of payment shocks, and the parameters q_0 and q_1 determine the heterogeneity across banks.

Given this specification, we choose q_0 and q_1 to match the overall profile of total assets by excess reserves, with the idea that payment shocks are likely to be roughly proportional to the bank's assets. Then, for the maximum scale $\bar{\xi}$, we note that about 1 percent of FF trades occur at or above the IOR rate.[23] No bank would borrow at a rate above the IOR rate if it had no chance of resorting to the discount window. Hence, the top 1 percent of market rates are a natural target for our choice of $\bar{\xi}$.

Given these targets, we choose $\bar{\xi} = \$400$ million, $q_0 = 3 \times 10^{-3}$ and $q_1 = 1 \times 10^{-3}$. Table 2, below, illustrates that these choices do an excellent job of matching the top market rates. Our imputed values for the scale of payment shocks also trace the distribution of total assets closely. In Fig. 3 we plot the implied standard deviation (in logs) of the absolute payment shocks $|z|$ for each level of excess reserves (again in logs). In the same figure, we also plot 0.25 percent of each bank's total assets (in logs) against its excess reserves holdings.[24]

GSEs. The last set of parameters to calibrate are the cash holdings of GSEs, y, their bargaining power, $1 - \theta$, and the relative measure of GSEs to banks, γ. Since GSEs are the only lenders in an environment with abundant reserves, and they trade all of their cash in our candidate equilibrium, we can make a tight connection between their cash holdings y and the average transaction size in the FF market, which is about \$250 million.[25] We set the bargaining power parameter to match an EFFR of 91 basis points, as observed between March and June 2017. The resulting value, $1 - \theta = .9$, suggests that the GSEs are able to realize most of the profits from the arbitrage between the ON RRP rate and the IOR rate.

We cannot exactly identify the relative measure of GSEs γ in an environment of abundant reserves. As long as it is above .08, no bank seeks to lend—and as long as it is not above 1 (i.e., there are more GSEs than banks), the number of matches will be simply equal to γ. Thus any value of γ between .08 and 1 simply scales up the model. We set it to a placeholder of $\gamma = .2$.

Our choices of parameter values and specifications are summarized in Table 1.

4.1.2. Market FF rates

Table 2 summarizes the rates traded in the FF market between March 16, 2017 and June 14, 2017, excluding month ends, as reported by the Federal Reserve Bank of New York, along with the corresponding statistics generated by our calibrated model.[26] The model successfully matches the intended targets, the EFFR and the 99th percentile. The former is tightly linked to the bargaining power of the GSEs, as in Bech and Klee (2011). Balance sheet costs, however, also play an important role, reducing the arbitrage gains and putting downward pressure on rates; if balance sheet costs were zero, our parameters would imply an EFFR of 97 basis points.

The distribution of excess reserves plays a key role in matching the 99th percentile without generating undue dispersion in rates. While the vast majority of institutions have virtually no concerns regarding reserve requirements, the distribution of excess reserves displays a thin but long left tail. Through the model, this results in a *right* tail in market rates, keeping

[22] The Laplace distribution is symmetric over $\mathbb{R}$, so that $\Pr(z \geq 0) = .5$ and $|z|$ follows an exponential distribution. However, the exact distribution of negative payment shocks (that is, adding to the bank's balances) is irrelevant in our model.

[23] The 99th percentile of the volume-weighted FF rate distribution, as reported by the Federal Reserve Bank of New York, was at or above the IOR rate (100 basis points) for all but 5 days from March to June 2017 (excluding month ends).

[24] It is hard to judge whether 0.25 percent of the total assets is excessive for late payment shocks. Variation in aggregate total reserves is very large, with week-to-week changes occasionally in excess of \$150 billion, which is about 1 percent of aggregate total assets and 7 percent of aggregate excess reserves.

[25] Cipriani and Cohn (2015) report an average of 300 transactions per day.

[26] All statistics are volume-weighted and rounded to the nearest basis point.

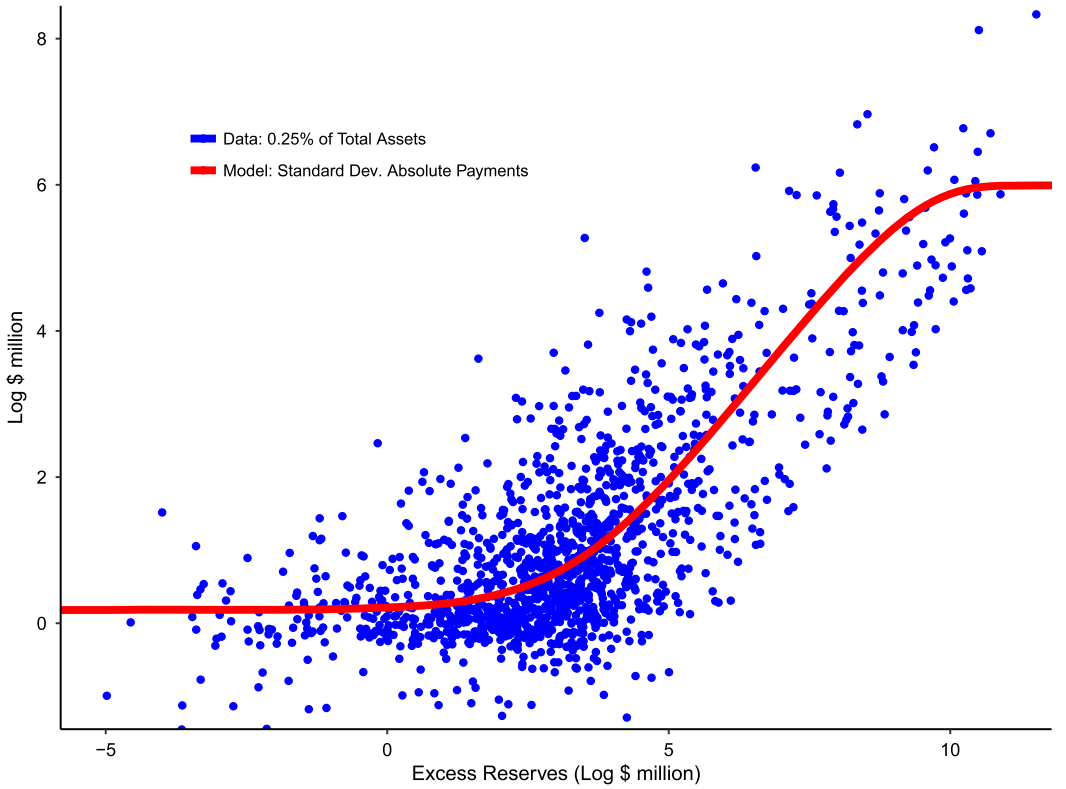

Fig. 3. Excess reserves, total assets, and scale of payment shocks.

Table 1
Calibration parameters.

Parameter description		Value
Administered rates		
Interest on reserves rate	i^{or}	100 b.p.
ON RRP rate	i^{rr}	75 b.p.
Discount window rate	i^{dw}	150 b.p.
GSEs		
Cash balances	y	$250 million
Bargaining power	$1-\theta$	0.9
Balance sheet costs		
FDIC insured	κ_1	7 b.p.
Other institutions	κ_2	0 b.p.
Balances distribution		
FDIC insured	$F_1(x)$	*Empirical CDF*
Other institutions	$F_2(x)$	*Empirical CDF*
Share FDIC insured	β_1	0.87
Payments distribution		
Maximum scale	$\bar{\xi}$	$400 million
Profile parameters	q_0	4×10^{-3}
	q_1	1×10^{-3}
Matching function	$m(\mu_l, \mu_b)$	$\min\{\mu_L, \mu_B\}$

the 75th percentile just one basis point above the EFFR while having the 99th percentile above the IOR rate. While the overall contour of the rates distribution is robust to the calibration of the payment shocks, the value of the 99th percentile is understandably sensitive.

In the data, as in the model, more than 50 percent of all trades are clustered at 91 and 92 basis points. Elsewhere, the model is not quite able to generate the short left tail in rates, predicting that the first percentile (and, as a matter of fact, the minimum) rate will be at 91 basis points, while in the data it averaged about 87 basis points. It is possible that those

Table 2
Market rates: Data and model.

	Data			
	Average	Min.	Max.	Model
Percentile 1th	87	77	91	91
Percentile 25th	91	91	91	91
EFFR (median)	91	91	91	91
Percentile 75th	92	91	92	92
Percentile 99th	101	97	105	101

All rates are volume-weighted and in basis points.

rates reflect the GSEs' preference for early return of funds and their willingness to pay a premium to those institutions able to provide them.[27]

4.2. Yesterday: scarce reserves

We now evaluate our model in an environment with scarce reserves. In particular, using the calibration of the model's "deep" parameters described above, we adjust the administered rates and the distribution of excess reserves to confirm that the model reproduces the hallmarks of the pre-crisis FF market: traded rates above the IOR rate; larger trading volume, driven by bank-to-bank trades; and a "demand" curve that slopes upward toward the discount window rate, making small open market operations effective at controlling market rates.

To reproduce the scarce reserves environment, we engineer a downward shift in the distribution of excess reserves, such that required reserves are approximately 96 percent of total reserves.[28] In particular, we posit a lognormal distribution such that the average excess reserves holdings is $175 million—roughly 0.2 percent of its level under the previous calibration—and approximately 40 percent of banks have less than $100 million in reserves.[29]

To assign values to administered rates, we note that the Fed did not pay interest on reserves before the crisis, and the ON RRP facility did not exist. We thus set both rates to zero. The Desk successfully implemented the FOMC's target by setting the discount window rate 100 basis points above the target for most of the 2000s: We set it at 200 basis points, consistent with a target of 1 percent.

We also note that the composition of GSEs present in the FF market was quite different during the previous regime, with Freddie Mac and Fannie Mae driving most of the lending, compared with today's prominent role for the Federal Home Loan Banks. However, it is not immediately obvious how to adjust the underlying parameters, and we prefer to carry over from the previous exercise as many parameter values as possible. Hence, we make no changes to the parameters governing the GSEs, even if the model could deliver a better fit of the data under scarce reserves if we were to recalibrate. It is also not obvious how to adjust the distribution of payment shocks to scale with aggregate reserves, and thus we leave the previous specification unchanged. Finally, since the FDIC fee was assessed on total deposits rather than total assets, we set balance sheet costs in the scarce reserves regime to zero.

We find that the model predicts an EFFR of 74 basis points—well above the (implicit) IOR rate of zero, but somewhat short of the target of 100 basis points. Traded FF volume is 78 percent higher than in the calibration for abundant-reserves regime, with bank-to-bank trades accounting for 44 percent of all volume. In the data, volume in 2006 was a bit more than twice the current level, and GSEs had only a 40 percent share of the borrowing.[30] We also note that small open market operations can easily shift the EFFR: It takes as little as a $60 million increase in aggregate excess reserves to implement a 1 basis point drop in the EFFR.[31]

We acknowledge that several elements are missing that could improve the model's quantitative performance in an environment with scarce reserves and, in particular, raise the predicted EFFR and FF volume. It is well known that banks are typically reluctant to resort to the discount window, occasionally preferring to borrow at rates above the discount window rate.[32] Another factor potentially driving rates up is the possibility of overdrafts, i.e., a negative reserve balance overnight that was traditionally penalized at highly punitive rates.[33] The model is also missing intermediation of trades throughout the day, which would naturally explain why total bank-to-bank trading is somewhat below the target in the data.[34]

[27] See Anderson and Huther (2016).

[28] In the period 2002-2006 required reserves averaged between 95 and 97 percent of total reserves held by depository institutions, according to Table 1 of the H.3. release by the Board of Governors.

[29] Unfortunately we do not have any further detailed data regarding the distribution of reserves for dates prior to 2007 due to changes in the reporting forms for the Call Reports.

[30] See Afonso et al. (2013b) for both facts. There are important data collection differences with data prior to 2007. See Cipriani and Cohn (2015) for an extended discussion.

[31] Hamilton (1997) reports that a $30 million open market operation can move the EFFR by 10 basis points, while Goodfriend et al. (1986) and Bernanke and Mihov (1995) reported $24 million and $33 million, respectively. These estimates, though, are for data from 1990 or prior.

[32] See Ennis and Weinberg (2013) for a theoretical treatment of the discount window "stigma" in the context of a search model.

[33] Whitesell (2006) emphasizes this channel in his study of interest rate corridors and alternative frameworks.

[34] See Afonso and Lagos (2015).

Table 3
Normalization dynamics: Parametric distributions.

	Mean	Std. Dev.	Upper bound
Initial distribution	12	2.5	12
Midpoint distribution	8	0.8	8
Endpoint distribution	4	0.3	8

5. Tomorrow

In June 2017, the FOMC announced its intention to begin the process of normalizing the Fed's balance sheet. It did not, however, specify an endpoint for this process, beyond stating that "the Federal Reserve's securities holdings will continue to decline in a gradual and predictable manner until the Committee judges that the Federal Reserve is holding no more securities than necessary to implement monetary policy efficiently and effectively."[35] Given the novelty of this endeavor, there is considerable uncertainty regarding how the FF market will evolve as the aggregate supply of reserves decreases.

In what follows, we will use our calibrated model to study potential transition paths. We will focus on identifying the levels of aggregate reserves such that (*i*) the EFFR rises above the IOR rate, signaling the end of the current implementation framework; and (*ii*) the demand curve becomes steep enough that rates are firmly between the IOR rate and the discount window rate, signaling a return to a classic "corridor" system.

5.1. Normalization path

Studying the process of normalization requires that we specify not only the path of aggregate total reserves, but also the evolution of the *distribution* of excess reserves across banks at each point in time. As noted earlier, this distribution plays a key role in determining which banks actively lend or borrow in the FF market. Since there exists considerable uncertainty regarding the precise dynamics of this distribution, our hope here is to provide rough estimates for the effect of normalization on market outcomes.

There is some evidence that the draining of reserves initially occurs "from the top." In particular, Call Report data shows a $650 billion decline in total reserves between the first quarter of 2015 and the last quarter of 2016.[36] Over this period, total reserves within the top decile of banks (ranked by total assets) decreased $596 billion, while total reserves within the bottom 80 percent of the banks decreased a mere $2 billion. We capture these dynamics by assuming that, during an initial stage, virtually all of the decline in reserves is from the banks with the largest holdings. Eventually, though, large banks may want to hold on to their reserves, either to satisfy reserve requirements or as high-quality liquid assets to meet other regulations. When this point is reached, reserves would presumably drain more evenly across banks. To capture these dynamics, we posit a second stage where reserve holdings decrease proportionally across banks, with the exception of those banks with the lowest level of reserves (which we keep constant).[37]

To formalize these two stages, we specify three parametric distributions, with the first stage transitioning from the initial distribution of reserves to a "midpoint" distribution, and the second stage from this midpoint to an "endpoint" distribution.[38] We use a truncated log-normal distribution, as it captures the initial distribution of excess reserves well and allows us to easily implement the draining of reserves from the top. The initial distribution is fitted directly to the data, with the upper bound corresponding to the largest bank in the sample. For the midpoint distribution, we assume that banks with excess reserves above the 95th percentile of the initial distribution—roughly $3 billion—absorb most of the drop in aggregate reserves. We thus set the top censor in the midpoint distribution to $3 billion and, in order to have a single-tailed distribution, the mean at the same level. Finally, for the endpoint, we simply cut the mean by half. The standard deviations for the midpoint and endpoint distributions are set to have a (roughly) constant $20-80$ percentile range. Table 3 summarizes the parameter choices.

Fig. 4 plots the point density function of the three distributions, on a log support. We note the corresponding level of aggregate excess reserves as implied by each distribution. The inherent uncertainty in the dynamics of the distribution of reserves is not resolved by our admittedly arbitrary choices. In particular, we should expect banks to take a more active role in their liquidity management, especially when the EFFR rises above the IOR and reserves become an expensive way for banks to hold their desired liquidity buffer. We will later explore a wide range of possibilities and illustrate the implications for normalization.

[35] "Addendum to the Policy Normalization Principles and Plans," Federal Open Market Committee, June 14, 2017.

[36] Total reserves as reported in the H.3. table by the Board of Governors fell by an almost identical amount. The draining of reserves occurred due to autonomous factors, mainly currency growth and an increase in the Treasury general account.

[37] A more satisfactory approach would be to model explicitly the liquidity management of the banks and determine their reserve holdings as a function of the EFFR and rates for other liquid assets, e.g., Treasury bills. Such a model is, however, outside the scope of this paper.

[38] We bridge the distributions by specifying a probabilistic transition matrix for the banks' reserves. Using a mixing probability delivers similar results, but it implies that the distribution of excess reserves in between stages is occasionally bi-modal. We do not distinguish between FDIC-insured institutions and other institutions during these transitions, as we have no basis to specify different dynamics for the two types of banks.

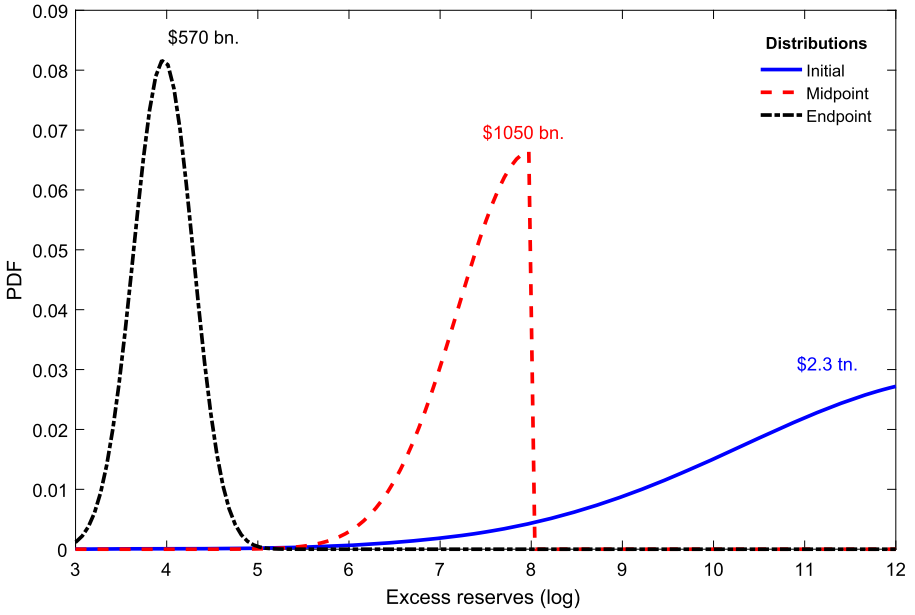

Fig. 4. Normalization dynamics: Distributions and aggregate total reserves.

Throughout the normalization scenarios we study, we hold administrative rates fixed, to facilitate comparison with the abundant-reserves scenario, and hold almost all structural parameters constant, with two exceptions. First, we postulate that the cash balances of GSEs are likely to decrease over time, though perhaps by less than total reserves. To account for this concern, we assume that y decreases steadily to 50% of its initial value, but not further. Second, we also conjecture that banks are bound to rein in some of the payment volatility as balances decline and, perhaps, some of the autonomous factors (like the Treasury general account) become more predictable. Once again, there is no obvious guide for how to adjust the distribution of payment shocks, so we opt for a simple solution and set the scale of payment shocks to 400 for all banks, saving us considerable computational time in the process.[39]

5.2. Baseline results

Given our specification for the process of normalization, we find that the EFFR drifts above the IOR rate when aggregate excess reserves reach approximately $750 billion. At this point, the FOMC would either have to halt the normalization process or reformulate its target range for the EFFR, since the IOR is set as the upper bound. However, aggregate reserves would have to decline an additional $350 billion or so before the EFFR begins sloping upward to the midpoint of the IOR and the discount window rates—the hallmark of a classic corridor system.

Fig. 5 plots a number of key statistics as a function of aggregate excess reserves.[40] The top-left panel is informative about the distribution of reserves, as it displays the median and the 20-80 percentile range. The dynamics are quite smooth, but one can appreciate that the center of the distribution is shifting down faster once excess reserves are a bit above $1 billion—when the shift from the initial to the midpoint distribution is about complete.

The top-right panel plots the EFFR rate along the transition path. It is markedly nonlinear, remaining at 92 basis points until excess reserves reach approximately $800 bn, and then quickly increasing to 101 basis points at $700 billion. Why do rates change so swiftly? Recall that the first banks to switch from borrowing to lending are those with the largest reserve holdings; these banks have plenty of funds to lend, and they do so aggressively, as their margins remain quite thin when rates are barely above the IOR. Moreover, since most banks remain borrowers, lenders remain on the short side of the market, and hence always match. As larger banks switch from borrowers to lenders, bank-to-bank trades make up a dominant share of the market (bottom-left) and total trading volume (bottom-right) rises to nearly $300 billion.

Returning to the top-right panel in Fig. 5, the path for the EFFR flattens out until excess reserves reach approximately $400 bn, at which point the slope steepens again. As reserves dwindle, lenders become more cautious, lending fewer funds and demanding higher rates. As a result the volume in the FF market starts dropping.[41] Bank-to-bank trades lose market share as the GSEs' balances decrease substantially less than the banks' balances.

[39] This is likely to overestimate the payment shocks for small banks, and underestimate them for the largest banks.
[40] There is virtually no effect of shrinking the balance sheet before excess reserves reach $1 trillion.
[41] Recall that in our model there is no intermediation.

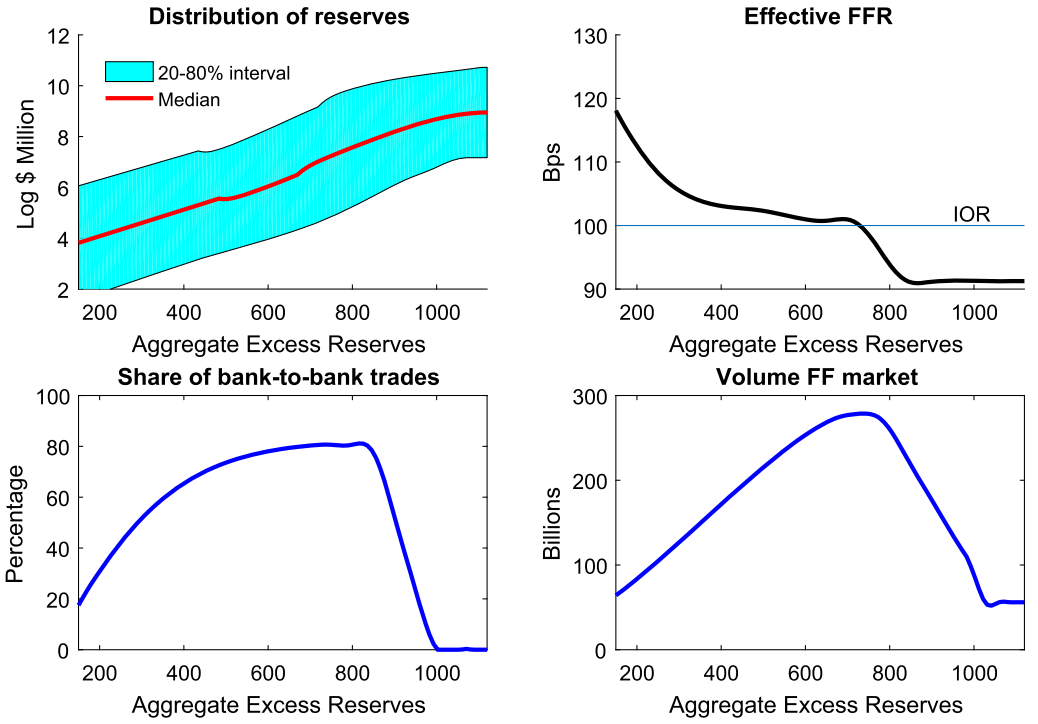

Fig. 5. Normalization dynamics: Baseline scenario.

To restore a corridor system akin to the implementation framework in place before 2007, excess reserves would need to be at or below $200 billion. Note that this is a considerably higher level of reserves than the "scarce reserves" scenario studied in Section 4.2. A key difference between the two environments is that, in the transition studied here, setting the IOR rate at 100 bp (and, to a lesser extent, the ON RRP rate at 75 bp) puts upward pressure on FF rates, while the absence of these two instruments implies a substantially lower EFFR rate, even at lower levels for total aggregate reserves. That said, our confidence in our baseline projection for the distribution of reserves lessens as conditions get tighter in the FF market and banks have stronger incentives to manage their reserve holdings by altering the composition of their liquid assets.

Lastly, reducing the balance sheet also has important implications for ON RRP take-up by GSEs, which is zero when lenders constitute the short side of the market. As more banks start lending, GSEs occasionally fail to find a counterparty and resort to the ON RRP facility. Take-up, though, does not exceed $60 billion at any point, held back in part by the decrease in GSEs balances.

5.3. Marginal cost of funds

The results above illustrate that activity in the fed funds market is often driven by arbitrage opportunities that are available to only some market participants. Moreover, the bilateral nature of trade implies that FF rates are in part determined by the bargaining power and the balance sheet costs of the banks that are currently active. Given these observations, a natural question arises: is the EFFR a good measure of the cost of short-term funding for all banks and, more generally, the stance of monetary policy?

To explore this question, we compute the marginal cost of funds (MCF) for every bank, defined as the marginal cost associated with financing an additional loan with initial reserve balances.[42] Let $V(\omega)$ be the expected payoff of a bank with balance sheet ω at $t = 0$, net of the return to loans, so that

$$V(\omega) = \pi(0, 0; \omega) + \max\{\Pi_B(\omega), \Pi_L(\omega)\}.$$

We define the marginal cost of funds as

$$\text{MCF}(\omega) = \frac{dV(\omega)}{dr}$$

[42] See Appendix A for a formal definition and derivations. We thank Todd Keister for suggesting this exercise.

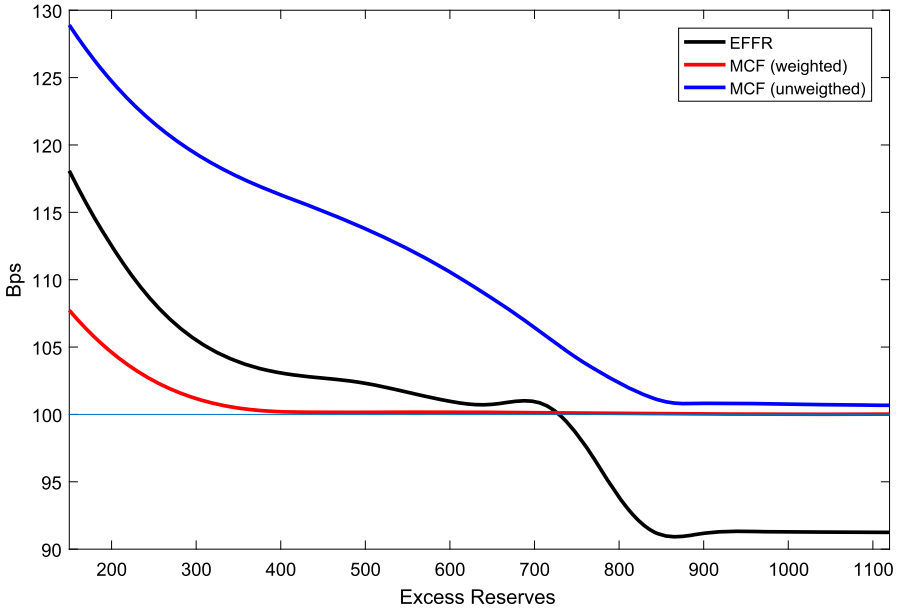

Fig. 6. Normalization dynamics: Marginal cost of funds.

with $dl = -dr$.[43] We then average across banks, with and without weighting by bank assets. Fig. 6 displays the weighted and unweighted average MCF, along with the EFFR, for our baseline scenario.

When reserves are abundant—i.e., in excess of $900 billion—the MCF (both weighted and unweighted) is driven by the opportunity cost of earning the IOR rate. Intuitively, all banks are borrowing and virtually none of them are at an elevated risk of visiting the discount window because of a late payment shock. Hence, the marginal cost of a reduction in initial reserves is roughly equal to the foregone IOR that the bank would earn. The EFFR in this region lies about 10 basis points below the MCF.

As excess reserves decline below $900 billion, the unweighted MCF starts to increase while the weighted MCF remains close to the IOR. It is now more likely for some small banks, with relatively low levels of excess reserves, to resort to the discount window, and as a result these small banks value their funds above the IOR. However, since the distribution of assets across U.S. banks is heavily concentrated at the largest banks who still have ample balances, it does not have an effect on the weighted MCF.

As the level of aggregate excess reserves continues to decline, the EFFR rises above the IOR. The unweighted MCF continues to increase but at a slow pace, tempered by the resurgence of the inter-bank market, which provides easy access to cheap funds. The weighted MCF remains close to the IOR. The growing differences between the weighted and unweighted averages reveal the large dispersion in the MCF across banks.

It is only when aggregate excess reserves drop below $300 billion that the weighted MCF—arguably, the closer measure to the monetary policy stance—starts to rise. By then the weighted MCF is *below* the EFFR, albeit by a modest amount. At this point, the volume-weighted median trade is between two banks, and thus the EFFR splits the surplus between a bank in need of funds (with a high MCF) and a bank willing to lend them (with a low MCF). Since large banks are the first to lend, it is the banks in the latter group that are almost exclusively driving the weighted MCF. It is worth noting that there remains significant dispersion in the MCF across banks that is not captured by the dispersion in FF rates, since the latter only involves banks that successfully matched and had an opportunity to adjust their desired reserve balances.

Returning to our initial question, we do see systematic deviations between the EFFR and our measures of the stance of monetary policy. Interestingly, both the sign and magnitude of these deviations vary with the aggregate supply of reserves. Moreover, there is considerable dispersion in the MCF across banks, which is only partly captured by dispersion in the FF rates. These results highlight the advantages of studying the effects of policy within a structural model, as it allows us to study the cost of short-term funding for both active *and inactive* banks. At the same time, it's debatable whether the magnitude of the difference between the EFFR and the MCF constitutes a serious deviation in terms of the stance of monetary policy.

[43] In our calculation, total assets remain constant and so does the FDIC fee. Since fed funds are considered high-quality assets for liquidity regulation purposes, there will typically be implications for other regulatory ratios, like the liquidity coverage ratio, that we do not impute to the marginal cost of funds.

5.4. Alternative scenarios

As noted above, the evolution of the FF market is quite sensitive to the dynamics of the distribution of excess reserves across banks. This is because rates and volume depend heavily on the incentives of the largest banks to become lenders, which in turn depend heavily on the distribution of reserves held by other banks. To illustrate this point more clearly, consider two thought experiments.

Thought Experiment 1. According to our Call Report data, the 5 percent of banks with the largest balances hold more than 90 percent of the aggregate supply of total reserves. It would thus be possible to reduce the balances of 95 percent of depository institutions to zero by draining just 10 percent of the current level of total reserves—a bit less than $250 billion. With the vast majority of banks in dire need of funds to satisfy reserve requirements, the handful of banks with excess reserves would act as lenders, and most trades in the FF market would be bank-to-bank. These trades would necessarily occur at a rate above the IOR rate; indeed, if borrowing banks have zero initial balances, rates would likely be even closer to the discount window rate than to the IOR rate. Thus, there would be ample excess reserves in the aggregate—more than $2 trillion—yet the EFFR would be substantially above the IOR and the FF market would resemble a classic corridor.

Thought Experiment 2. Now consider the opposite scenario. Since total reserve requirements add up to approximately $150 billion, it is feasible to assign reserves in such a way that all banks have excess reserves equal to 25 percent of their required reserves—a sizable buffer—with total aggregate reserves equal to just 200 billion. In this hypothetical scenario there would still be no gains of trade between banks: All trades in the FF market would be with GSEs as lenders, exploiting the arbitrage opportunity between the IOR and the ON RRP rates, and thus the EFFR would remain below the IOR.

While neither of these scenarios is realistic, they illustrate how the distribution of excess reserves can radically shape the FF market, *independently* of the aggregate supply of reserves. In particular, we find that the level of excess reserves at which the EFFR breaches the IOR is particularly sensitive to our choice of the midpoint distribution of excess reserves, as this choice determines the level of reserves being held by the largest banks, which are pivotal. If the banks with most reserves decrease their holdings at a slower pace, then the distribution of reserves becomes more concentrated at the top and, mechanically, there is a larger fraction of banks with low balances.[44] This is the basis of our first alternative scenario, which we name "high concentration." We also simulate a scenario in which banks take a much more active role in their liquidity management as FF rates increase: The largest banks shed reserves more quickly, as they acquire other high-quality, liquid assets like Treasuries as the opportunity cost of holding idle balances increases. The smaller banks would gladly absorb these additional reserves, as the cost of borrowing funds increases. The net effect is a decrease in the concentration of reserves at the top relative to the baseline. We refer to this scenario as "low concentration."

Under the three scenarios described above—baseline, high concentration, and low concentration—Fig. 7 plots the median holdings (in logs) and the percentage of banks with balances less than $500 million as the level of aggregate reserves shrinks from $1.2 trillion to $200 billion.[45] All three scenarios start with the same distribution of reserves and converge to the same endpoint distribution.[46] The top panel illustrates that the median holdings fall earlier in the normalization under the high concentration scenario, relative to the baseline, causing a rapid rise in the fraction of banks with relatively low balances (the bottom panel).

Turning to Fig. 8, this implies that the EFFR drifts above the IOR earlier under the high concentration scenario than in the baseline—when aggregate excess reserves shrink to approximately $1 trillion, as opposed to $750 billion. Intuitively, as the largest banks hoard balances, the demand for reserves (to satisfy requirements) increases quickly for medium and small banks, which triggers the largest institutions to start lending funds earlier. The combination of a higher demand for funds by borrowing institutions and more balances available at lending institutions results in a large increase in trading volume in the FF market, as seen in Fig. 9. Similarly, the share of trades that are bank-to-bank (not shown) rises earlier and by more than in the baseline scenario. The opposite is true, of course, in the low concentration scenario. Under this scenario, the EFFR drifts above the IOR when aggregate excess reserves reach approximately $400 billion.[47]

[44] This may occur because, e.g., large banks that are subject to more strict supervision might rely more heavily on reserves as high-quality liquid assets to satisfy regulations.

[45] All scenarios have the same average balances along the normalization path by construction.

[46] There is little scope for variation at either the start or the end of the normalization process. Regarding the former, it is hard to argue that the distribution would depart radically from what we observed from 2015. Regarding the latter, the distribution necessarily compresses as excess reserves become scarce.

[47] Note that our estimates for when the FF market returns to a corridor system is roughly in line with several external estimates and surveys. For example, in July 2017, the Federal Reserve Bank of New York published the Projections for the SOMA Portfolio and Net Income. In this report, they project that the long-run level of reserve balances could range from $406 billion to $1 trillion, with their "median scenario" projecting that reserve balances will be $613 bn. See https://www.newyorkfed.org/medialibrary/media/markets/omo/SOMAPortfolioandIncomeProjections_July2017Update.pdf In the June 2018 Survey of Market Participants, respondents were asked to indicate their expectations for the level of total reserves in 2025. The 25th percentile was $750 billion, the median was $1 trillion, and the 75th percentile was $1.37 trillion; see https://www.newyorkfed.org/medialibrary/media/markets/survey/2018/jun-2018-smp-results.pdf.

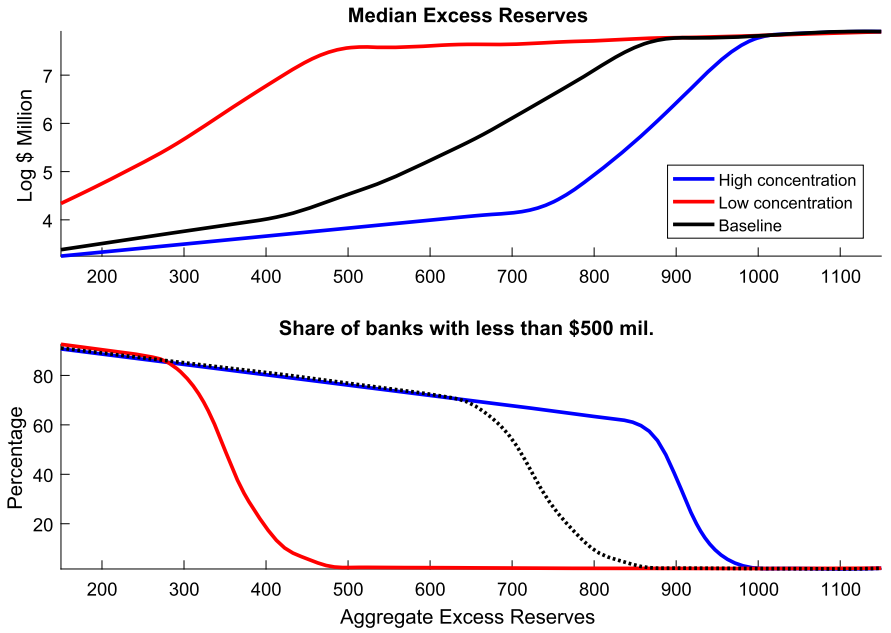

Fig. 7. Normalization dynamics: Alternative scenarios.

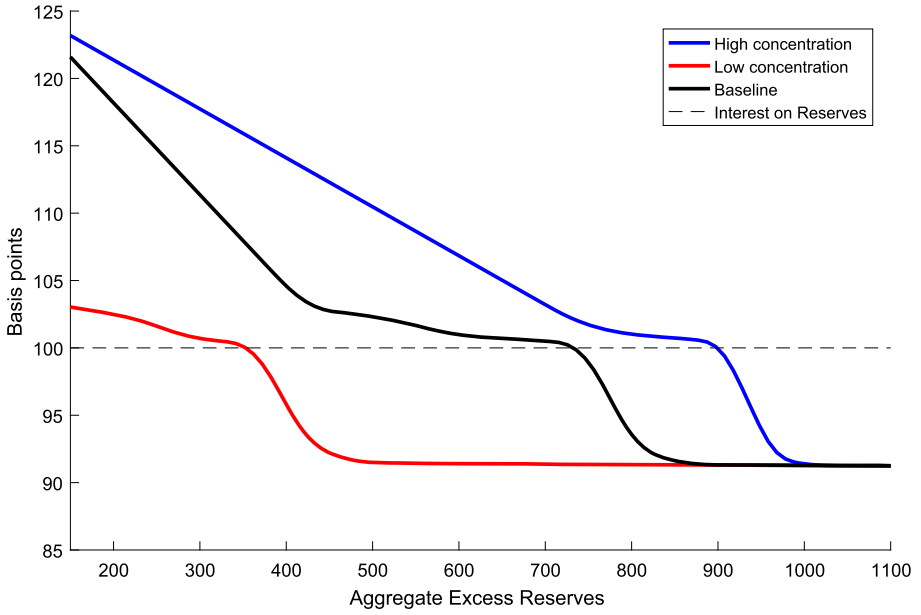

Fig. 8. Alternative scenarios: Effective FFR.

6. Conclusions

We developed a model that is capable of reproducing—both qualitatively and quantitatively—the main features of the fed funds market before the financial crisis, when reserves were scarce, and after the crisis, when reserves became abundant. We use this model to inform the evolution of the fed funds market as the FOMC normalizes the Fed's balance sheet and, as a consequence, the aggregate supply of reserves declines.

While we provide a baseline scenario for normalization, the overwhelming message is one of uncertainty. The precise dynamics of the distribution of excess reserves across banks can drive the EFFR above the IOR rate when aggregate excess reserves are as high as $1 trillion, or imply that interbank trading would not return even when aggregate excess reserves are as low as $400 billion. That is an uncomfortably large range for a key event: Once the EFFR drifts above the IOR rate,

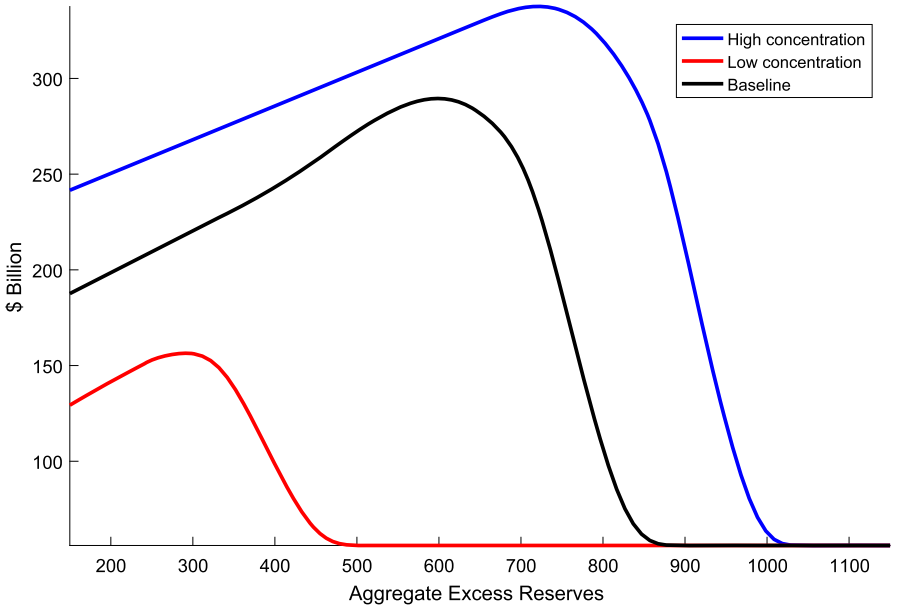

Fig. 9. Alternative scenarios: Federal Funds volume.

the FOMC will be forced to either halt the normalization process or to rethink the implementation that sets the IOR equal to the top of the target range for the EFFR.

Appendix A. Omitted proofs

Derivation of equation (6)

The gains from trade from borrowing an amount $f > 0$ at rate ρ are:

$$\begin{aligned}
\Delta\pi_B(f,\rho;\omega) &= \pi(f,\rho;\omega) - \pi(0,\cdot;\omega) \\
&= f\left[i^{or} - \rho - \kappa\right] - \left(i^{dw} - i^{or}\right)\int_{x+f}^{\infty}[-x - f + z]\,dG(z) + C \\
&\quad + \left(i^{dw} - i^{or}\right)\int_{x}^{\infty}[-x + z]\,dG(z) + C \\
&= f\left[i^{or} - \rho - \kappa\right] + \left(i^{dw} - i^{or}\right)\left[\int_{x+f}^{\infty} f\,dG(z) + \int_{x}^{x+f}(z - x)\,dG(z)\right] \\
&= f\left[i^{or} - \rho - \kappa\right] + \left(i^{dw} - i^{or}\right)\left[f[1 - G(x+f)] + \int_{x}^{x+f}(z - x)\,dG(z)\right].
\end{aligned}$$

Proof of Lemma 1. One can easily show that: (*i*) S is strictly concave in τ; (*ii*) $S(0;\omega,\omega') = 0$; and (*iii*) there exists $\tau \in \mathbb{R}_+$ sufficiently large such that $S(\tau';\omega,\omega') < 0$ for all $\tau' > \tau$. Hence, a necessary and sufficient condition to ensure that there exists a $\tau \in \mathbb{R}_+$ such that $S(\tau;\omega,\omega') > 0$ is

$$\left.\frac{\partial S(\tau;\omega,\omega')}{\partial\tau}\right|_{\tau=0} > 0 \quad \Leftrightarrow \quad \kappa < \left(i^{dw} - i^{or}\right)\left[G(x') - G(x)\right].$$

When this condition is satisfied, the first-order condition characterizing the value of τ that maximizes the surplus,

$$\frac{\partial S(\tau;\omega,\omega')}{\partial\tau} = -\kappa + \left(i^{dw} - i^{or}\right)\left[G(x' - \tau) - G(x + \tau)\right] = 0,$$

is necessary and sufficient. Finally, the interest rate ρ is such that each agent receives half the surplus, as in (12).

Proof of Lemma 3. Consider a candidate equilibrium in which $x_j^\star = \infty$ for all $j \in \mathcal{J}$. Then it is individually optimal for every bank to borrow if, under this candidate equilibrium, even the bank with the most incentive to lend finds it optimal to borrow, which is true if

$$\lim_{x_J^\star \to \infty} \Pi_B(x_J^\star, \kappa_J) \geq \lim_{x_J^\star \to \infty} \Pi_L(x_J^\star, \kappa_J). \tag{22}$$

The left-hand side of (22) reduces to

$$\gamma\theta\left[i^{or} - i^{rr} - \kappa_J\right] y.$$

The right-hand side of (22) is bounded above by the gains from lending when $x_J^\star \to \infty$ and $\kappa_1 = \kappa_2 = \cdots = \kappa_J = 0$, which reduces to

$$\frac{1}{2}\sum_{j\in\mathcal{J}} \beta_j \int_{-\infty}^{\infty} \left[\left(i^{dw} - i^{or}\right) \int_x^{\infty} (z-x) dG(z) \right] dF_j(x).$$

Liquidity coverage ratio

We model the liquidity coverage ratio (LCR) in a simplified manner. In particular, we assume that banks are assessed a fee ψ on the difference between 1 and the LCR, as a way to capture the regulatory fees and/or the internal costs associated with a liquidity shortfall. If the bank borrowed f, the LCR is given by

$$\frac{\eta\ell + r + f}{d + f}$$

where η is a haircut applied to the loans.[48] In this case, the fee would be

$$\psi\left(d + f - \eta\ell - r - f\right) = \psi\left(d - \eta\ell - r\right).$$

Alternatively, if the bank lent f, the LCR is

$$\frac{\eta\ell + R - f}{d - f}$$

so that the fee would be

$$\psi\left(d - f - \eta\ell - b + f\right) = \psi\left(d - \eta\ell - b\right).$$

Hence, the LCR does not affect the decision to borrow or lend in this simple environment. It is worth noting, however, that we have abstracted from several details in the actual calculation of LCR.[49] In particular, current regulations assign different haircuts on net outflows, depending on the counterparty. Hence, when aggregate reserves decline and the composition of lenders in the market changes, these regulations could have effects that we have not considered.

Marginal cost of funds

Let $V(\omega)$ be the expected payoff of bank ω, at the beginning of the period,

$$V(\omega) = \pi(0, 0; \omega) + \max\{\Pi_B(\omega), \Pi_L(\omega)\}.$$

We define the marginal cost of funds as

$$\text{MCF}(\omega) = \frac{dV(\omega)}{dr}$$

[48] The regulation has a three-tier classification for assets that qualify as high-quality liquid assets (HQLA), with different haircuts for each of them as well as caps on the maximum percentage of HQLA that can come from each class. The level and composition of the bank's asset portfolio, though, does not impact the marginal effect of borrowing or lending funds.

[49] See House et al. (2016) for a more detailed description of the calculation of the LCR. Ihrig et al. (2017) documents how banks have managed the composition of HQLA to date.

with $dl = -dr$.[50] Note that the expected payoff does not include the return from loans, but by setting $dl = -dr$ we keep total assets constant and avoid the balance-sheet costs to be imputed to the MCF.

We split the algebra in several steps. First, the marginal effect of not trading,

$$\begin{aligned}\frac{\partial \pi(0,0;\omega)}{\partial r} &= \frac{\partial}{\partial r}\left[r i^{or} - \kappa(r+\ell) - \left(i^{dw} - i^{or}\right) \int_{-R+r}^{\infty} [R - r + z]\, dG(z) \right] \\ &= i^{or} + \left(i^{dw} - i^{or}\right)(1 - G(x)).\end{aligned}$$

Next, turning to the expected gains from trade, we start with the surplus function $S^*(\omega, \omega')$. We first differentiate with respect to x (the excess reserves of the borrower) and then with respect to x' (the excess reserves of the lender). In both cases, the envelope theorem implies that there is no effect of a marginal change in excess reserves on the optimal transaction amount τ^*.

From the point of view of the borrower, x, when meeting a bank, we obtain

$$\begin{aligned}\left(i^{dw} - i^{or}\right)^{-1} \frac{\partial S^\star(\omega,\omega')}{\partial x} &= -\tau^\star g(x+\tau^\star) + \tau^\star g(x+\tau^\star) - \left(G(x+\tau^\star) - G(x)\right) \\ &= -\left(G(x+\tau^\star) - G(x)\right).\end{aligned}$$

Note that this expression is equal to zero if $\tau^\star(\omega,\omega') = 0$, and thus is well-defined even if there is no trade with the lender. When the borrower meets a GSE,

$$\left(i^{dw} - i^{or}\right)^{-1} \frac{\partial \tilde{S}^\star(\omega)}{\partial x} = -\left(G(x+y) - G(x)\right).$$

Finally, from the point of the lending bank, x',

$$\frac{\partial S^\star(\omega,\omega')}{\partial x'} = (i^{dw} - i^{or})\left(G(x') - G(x' - \tau^\star)\right).$$

The last step is to use these result to study the marginal effect on expected payoffs. For a bank ω seeking to borrow, i.e., $\Pi_B(\omega) > \Pi_L(\omega)$, we obtain

$$\begin{aligned}\frac{\partial V(\omega)}{\partial x} &= i^{or} + (i^{dw} - i^{or})(1 - G(x)) \\ &\quad - (i^{dw} - i^{or})\mathbb{P}(\text{GSE})\theta\,(G(x+y) - G(x)) \\ &\quad - (i^{dw} - i^{or})\mathbb{P}(\text{bank})\frac{1}{2}E_{\omega'}\left(G(x+\tau^\star(\omega,\omega')) - G(x)\right)\end{aligned}$$

where the probabilities are defined as we would expect, and independent of the bank type,

$$\begin{aligned}\mathbb{P}(\text{GSE}) \equiv \Pr(\text{match with a GSE}) &= \frac{m(\mu_L,\mu_B)}{\mu_B}\frac{\gamma}{\mu_L}, \\ \mathbb{P}(\text{bank}) \equiv \Pr(\text{match with a bank}) &= \frac{m(\mu_L,\mu_B)}{\mu_B}\frac{\mu_L - \gamma}{\mu_L},\end{aligned}$$

and the expectation operator $E_{\omega'}$ is defined over the distribution of trades with banks, i.e.,

$$E_{\omega'}\frac{\partial S^\star(\omega,\omega')}{\partial x} = \sum_{j\in\mathcal{J}} \frac{\beta_j}{\mu_L - \gamma} \int_{x_j^\star}^{\infty} \frac{1}{2}\frac{\partial S^\star(\omega,\omega')}{\partial x} dF_j(x').$$

Terms can be re-arranged for a more straightforward interpretation as

$$\begin{aligned}\frac{\partial V(\omega)}{\partial r} &= i^{or} \\ &\quad + (i^{dw} - i^{or})\mathbb{P}(\text{GSE})\theta\,(1 - G(x+y)) \\ &\quad + (i^{dw} - i^{or})\mathbb{P}(\text{bank})\frac{1}{2}E_{\omega'}\left(1 - G(x+\tau^\star(\omega,\omega'))\right) \\ &\quad + (i^{dw} - i^{or})(1 - \mathbb{P}(\text{GSE})\theta - \mathbb{P}(\text{bank})\frac{1}{2})(1 - G(x)).\end{aligned}$$

[50] The MCF is well defined for all banks but a set of measure zero, namely those banks indifferent between lending and borrowing.

The derivation for a bank ω seeking to lend follows similar steps to arrive at

$$\frac{\partial V(\omega)}{\partial r} = i^{or} + (i^{dw} - i^{or})\mathbb{P}(\text{bank})\frac{1}{2}E_{\omega'}\left(1 - G(x - \tau^{\star}(\omega', \omega))\right) + (i^{dw} - i^{or})(1 - \mathbb{P}(\text{bank})\frac{1}{2})(1 - G(x)).$$

Appendix B. Data

Data sources

Financial data for this paper come from several reporting forms. We first collect data from the Report of Condition and Income, commonly known as "Call Reports." Every national bank, state member bank, insured state nonmember bank, and savings association is required to file a Call Report on a quarterly basis. Reporting requirements vary according to an institution's size, the nature of its activities, and whether it has any foreign offices. In particular, we collect data from the following reports:

- **FFIEC 031** for banks with both domestic and foreign offices,
- **FFIEC 041** for banks with domestic offices only.

We augment these data with quarterly information on assets and liabilities of U.S. branches and agencies of foreign banks (**FFIEC 002**.) Data from March 31, 2001 are available for download at the Central Data Repository's Public Data Distribution by the Federal Financial Institutions Examination Council (FFIEC).[51] Sample forms can be obtained from the FFIEC or Federal Reserve Bank of Chicago.[52] A comprehensive data dictionary is also available from the Federal Reserve Board.[53]

We aggregate our measure of excess reserves up to the parent bank holding company-FDIC insurance level, such that for each quarter there is a unique observation for the subsidiaries of a bank holding company that are FDIC insured (e.g. commercial banks) and, if applicable, the subsidiaries that are not FDIC insured (e.g. the branches of foreign banks). To do this aggregation, we merge into our dataset an FDIC insurance indicator variable and a mapping from entity identifiers to parent identifiers. Both of these variables come from the National Information Center (NIC), a central repository of data about banks and other institutions that contains information on their organizational structures.[54]

To get financial information at the consolidated level for parent bank holding companies, we also merge in data from the Consolidated Financial Statements for Holding Companies, or FR Y-9C.[55] The FR Y-9C is filed quarterly by bank holding companies, savings and loan holding companies, and intermediate holding companies with total consolidated assets of $1 billion or more (prior to 2015, this threshold was just $500 billion).

Sample

We start from a sample of all entities that file Call Reports (FFIEC 031 or 041) or reporting form FFIEC 002 and report holding a positive amount of reserves (Call item RCFD 0090, "Balances due from Federal Reserve Banks") over 2015Q1-2016Q4, aggregated up to the parent-FDIC insurance level. Note that the schedule from which this item comes (Schedule RC-A) only needs to be filed by banks with foreign offices or with at least $300 million in assets.

We then restrict our sample to include institutions that engage in (some) fed funds activity. Specifically, we keep in our dataset every parent entity that accounts for at least 0.01 percent of total fed funds activity (in terms of fed funds sold plus fed funds purchased) on at least one quarter-end date over 2005Q1-2016Q4. We implement this participation condition at the consolidated parent level to avoid double-counting fed funds that occur between banks of the same bank holding company. In addition, the trading motives and terms of these "intra-bank" fed funds trades may different than those of other fed funds transactions.

To measure how intensely a parent entity participates in the fed funds market, we calculate for every parent entity-quarter the parent entity's percentage share of fed funds activity in the quarter, defined as the sum of fed funds sold and purchased by the parent entity divided by total fed funds activity (the sum of fed funds sold and fed funds purchased by every entity that holds reserves). Fed funds sold and purchased are from the FR Y-9C items BHDM B987 ("Federal funds sold in domestic offices") and BHDM B993 ("Federal funds purchased in domestic offices"). When FR Y-9C data are not available

[51] https://cdr.ffiec.gov/public/.
[52] https://www.chicagofed.org/banking/financial-institution-reports/commercial-bank-data.
[53] https://www.federalreserve.gov/apps/mdrm/.
[54] https://www.ffiec.gov/nic/.
[55] https://www.chicagofed.org/banking/financial-institution-reports/bhc-data.

(because, for instance, of a bank holding company not meeting the asset threshold for filing the Y-9C, or the entity being an uninsured foreign branch), these variables are estimated by summing the corresponding FFIEC 002/031/041 fields across all entities held by the parent.

After restricting our initial sample to institutions that meet our fed funds market participation criteria, only around half of the institutions remain. Still, these banks account for more than 95 percent of total assets and of reserve balances.

Excess reserves

We calculate excess reserves at the bank (Call Report filer) level as the difference between total reserve balances and required reserve balances.

Total reserve balances To calculate total reserves held at a Federal Reserve Bank, we simply take the number from item RCFD 0090 in the Call Report ("Balances due from Federal Reserve Banks"), which gives on the final day of the quarter the total amount of reserves that a bank has at the Fed.

Required reserve balances Calculating required reserve balances is more complex, as it requires multiple Call Report items. Required reserve balances are defined as

$$Required\ reserve\ balances = Required\ reserves - Vault\ cash$$

where *Required reserves* is calculated as an increasing function of a bank's net transaction accounts. Net transaction accounts equal a bank's total transaction accounts (including demand deposits, ATS accounts, and NOW accounts) minus amounts due from other depository institutions and cash items in the process of collection. To calculate net transactions, we take from the Call item RCON 2215 (the bank's "Total Transaction Accounts" (including "Total Demand Deposit" in domestic offices, which also includes ATS and NOW accounts)) to estimate total transaction accounts, and subtract from it our estimate of amounts due from other depository institutions (the sum of item RCFD 0083 ("Balances due from depository institutions in the U.S.: U.S. branches and agencies of foreign banks (including their IBFs)") and RCFD 0085 ("Balances due from depository institutions in the U.S.: Other depository institutions in the U.S. (including their IBFs)") and cash in the process of collection, item RCON 0020 ("Cash items in process of collection and unposted debit"). This gives us an estimate of a bank's net transaction accounts. Given net transaction accounts, we calculate *Required reserves* using reserve requirement information from the Federal Reserve Board.[56]

To finally calculate the bank's *Required reserve balances* (i.e., how much reserves it must hold at the Fed), we subtract from the estimated *Required reserves* number our estimate of the bank's *Vault cash*, item RCON 0080 ("Currency and coin"). When this calculation of reserve balance requirements yields a negative number (either because of estimation error or some genuine feature of the bank, for instance when a bank holds more vault cash than its reserve requirement), we set the reserve balance requirement equal to zero, since negative requirements are not possible.

Excess reserves Excess reserves equal total reserves minus required reserve balances.

Summary statistics

After our sample selection and imposing a balanced panel over 2015-2016, we are left with 1,508 depository institutions. Relative to the H.3 release by the Board of Governors, our sample captures 81 percent of the aggregate total reserves in the system and 80.6 percent of the aggregate excess reserves. Because of the minor difference in coverage, the aggregate ratio of excess reserves to total reserves is slightly lower in our data than in the H.3 reported by the Board. The difference, though, is small and both series track each other closely over time (see Fig. 10).

To reduce some of the noise in our excess reserves measure, we average each bank's holdings of reserves over the period 2015-2016. Table 4 reports several statistics for assets, total and excess reserves, as well as some selected percentiles. The significant difference between means and medians already speaks to the large amount of skewness in the data.

Of special interest is the ratio of excess reserves to total reserves. This ratio captures the level of reserves that a bank holds beyond what it needs to meet its requirement. Not surprisingly in the current environment of abundant reserves, the vast majority of institutions have very large ratios (the median bank in our sample has a ratio close to 100 percent). Zooming in to the lower half of the distribution, Table 5 shows some selected percentiles both unweighted and weighted by total assets. Both are roughly similar outside the bottom percentiles: Smaller banks have systematically lower ratios. Note that there are some institutions (fewer than 60) that have negative excess reserves. This may be due to some measurement error in our measure of excess reserves, but it is also possible that for some institutions their balances on a given day are below requirements, as compliance is computed over an average of 15 days.

[56] https://www.federalreserve.gov/monetarypolicy/reservereq.htm.

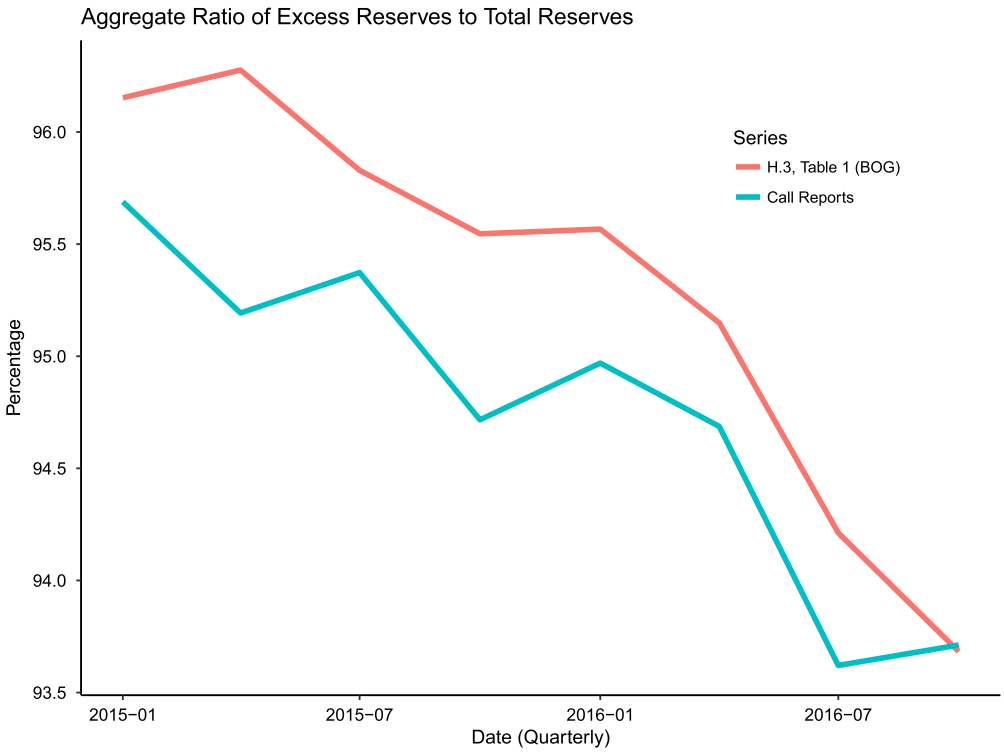

Fig. 10. Aggregate ratio of excess reserves to total reserves.

Table 4
Summary statistics ($ million).

	Assets	Total reserves	Excess reserves
Mean	11431.6	1272.4	1205.9
Median	803.5	24.3	19.9
Std. Dev.	93396.7	11221.5	10730.3
Percentiles			
5th	278.2	0.4	0
10th	334.8	1.6	0.6
25th	441.7	7.8	5.1
75th	2204.6	77.6	64.5
90th	9255	484	425.8
95th	28640.4	3272.3	3008.3

Table 5
Distribution of Excess Reserves to Total Reserves Ratios (%).

Percentiles	Unweighted	Weighted
1	−181.80	−2.20
5	12.20	59.30
10	42.50	76.10
15	59.70	76.10
20	71.10	81.30
25	80.30	85
50	98.50	94.40

An important distinction in our analysis is between those institutions that are insured by the FDIC and those that are not. In our sample 1,376 institutions are FDIC insured. They represent 87 percent of the total assets in the data, but only 63 percent of total reserves. This basically reflects that not FDIC-insured institutions (typically U.S. branches of foreign banks) tend to be larger, and that the group of FDIC-insured institutions includes smaller entities such as the smaller domestic banks. Table 6 collects the average reserve positions and total assets for these two groups of institutions.

Table 6
Average statistics for FDIC-insured and other institutions ($ million).

	Total reserves	Excess reserves	Assets	XR/TR(%)
FDIC insured	884	820	10886	93
Others	5320	5224	17117	98

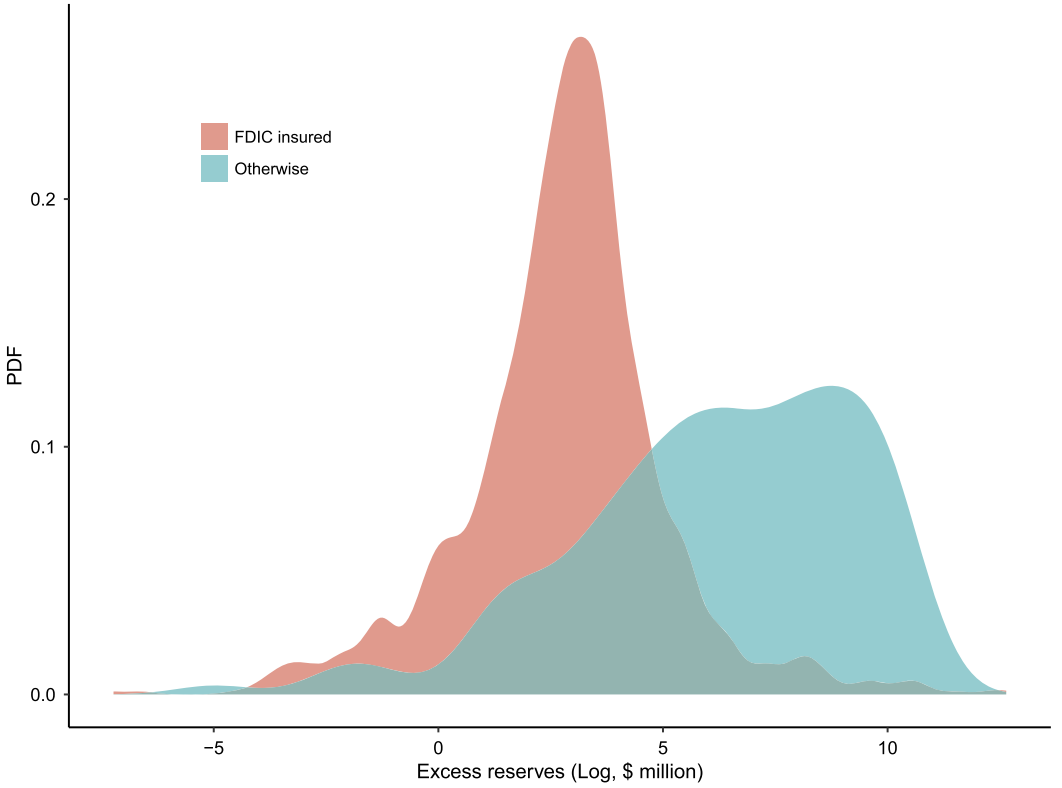

Fig. 11. Distribution of excess reserves (logs).

Finally we take a look at the full distribution of excess reserves that is a key input in the calibration. Given the enormous dispersion in holdings, we plot the log of excess reserves and drop those banks with negative holdings. Fig. 11 shows the kernel densities by institution type. The distribution for FDIC-insured banks could be reasonably approximated by a Lognormal distribution, though it is not exactly symmetric over logs. The distribution for other institutions is clearly skewed to the left and does not have an immediate parametric counterpart. We thus use the empirical CDF for each institution type rather than attempting a fitting exercise.

Appendix C. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.red.2019.04.004.

References

Afonso, Gara, Entz, Alex, LeSueur, Eric, 2013a. Who's Borrowing in the Fed Funds Market? Technical Report, Federal Reserve Bank of New York. Liberty Street Economics.
Afonso, Gara, Entz, Alex, LeSueur, Eric, 2013b. Who's Lending in the Fed Funds Market? Technical Report, Federal Reserve Bank of New York. Liberty Street Economics.
Afonso, Gara, Lagos, Ricardo, 2015. Trade dynamics in the market for federal funds. Econometrica 83 (1), 263–313.
Anderson, Alyssa G., Huther, Jeffrey W., 2016. Modelling overnight RRP participation. Technical Report 2016-023, Federal Reserve Board. Finance and Economics Discussion Series.
Armenter, Roc, Lester, Benjamin, 2017. Excess reserves and monetary policy implementation. Review of Economic Dynamics 23, 212–235.
Ashcraft, Adam B., Duffie, Darrell, 2007. Systemic illiquidity in the federal funds market. The American Economic Review 97 (2), 221–225.
Banegas, Ayelen, Tase, Manjola, 2016. Reserve Balances, the Federal Funds Market and Arbitrage in the New Regulatory Framework. Finance and Economics Discussion Series 2016-079, Board of Governors of the Federal Reserve System (U.S.).
Bech, Morten, Keister, Todd, 2017. Liquidity regulation and the implementation of monetary policy. Journal of Monetary Economics 92 (Supplement C), 64–77.

Bech, Morten L., Klee, Elizabeth, 2011. The mechanics of a graceful exit: interest on reserves and segmentation in the federal funds market. Journal of Monetary Economics 58 (5), 415–431.
Bech, Morten, Monnet, Cyril, 2016. A search-based model of the interbank money market and monetary policy implementation. Journal of Economic Theory 164, 32–67. Symposium Issue on Money and Liquidity.
Bernanke, Ben S., Mihov, Ilian, 1995. Measuring Monetary Policy. NBER Working Paper 5145.
Berentsen, Aleksander, Monnet, Cyril, 2008. Monetary policy in a channel system. Journal of Monetary Economics 55 (6), 1067–1080.
Bianchi, Javier, Bigio, Saki, 2017. Banks, Liquidity Management, and Monetary Policy. Federal Reserve Bank of Minneapolis, Research Department Staff Report 503.
Cipriani, Marco, Cohn, Jonathan, 2015. The FR 2420 Data Collection: a New Base for the Fed Funds Rate. Technical Report, Federal Reserve Bank of New York. Liberty Street Economics.
Ennis, Huberto M., 2014. A Simple General Equilibrium Model of Large Excess Reserves. Federal Reserve Bank of Richmond, Working paper WP 14-14.
Ennis, Huberto M., Weinberg, John A., 2013. Over-the-counter loans, adverse selection, and stigma in the interbank market. Review of Economic Dynamics 16 (4), 601–616.
Furfine, Craig H., 1999. The microstructure of the federal funds market. Financial Markets, Institutions & Instruments 8 (5), 24–44.
Goodfriend, Marvin, Anderson, Gary, Kashyap, Anil, Moore, George, Porter, Richard D., 1986. A weekly rational expectations model of the nonborrowed reserve operating procedure. Economic Review 72 (1), 11–28. Federal Reserve Bank of Richmond.
Hamilton, James D., 1997. Measuring the liquidity effect. The American Economic Review 87 (1), 80–97.
House, Mark, Sablik, Tim, Walter, John R., 2016. Understanding the New Liquidity Coverage Ratio Requirements. Technical Report. Federal Reserve Bank of Richmond.
Ihrig, Jane E., Kim, Edward, Kumbhat, Ashish, Vojtech, Cindy M., Weinbach, Gretchen C., 2017. How Have Banks Been Managing the Composition of High-Quality Liquid Assets?. Board of Governors of the Federal Reserve System (US). Finance and Economics Discussion Series 2017-092.
Kim, Kyungmin, Martin, Antoine, Nosal, Ed, 2017. Can the US Interbank Market Be Revived?. Federal Reserve Bank of Chicago.
Markets Group, 2016. Domestic Open Market Operations. Technical Report, Federal Reserve Bank of New York.
Martin, Antoine, McAndrews, James, Palida, Ali, Skeie, David, 2013. Federal Reserve Tools for Managing Rates and Reserves. Federal Reserve Bank of New York Staff Reports. p. 642.
Matsui, Akihiko, Shimizu, Takashi, 2005. A theory of money and marketplaces. International Economic Review 46 (1), 35–59.
Poole, William, 1968. Commercial bank reserve management in a stochastic model: implications for monetary policy. Journal of Finance 23 (5), 769–791.
Whitesell, William, 2006. Interest rate corridors and reserves. Journal of Monetary Economics 53 (6), 1177–1195.
Williamson, Stephen D., 2018. Interest on reserves, interbank lending, and monetary policy. Journal of Monetary Economics, hrig.