FEDERAL RESERVE BANK OF ST. LOUIS
REVIEW

FIRST QUARTER 2018
VOLUME 100 | NUMBER 1

REVIEW

Volume 100 • Number 1

FEDERAL RESERVE BANK of ST. LOUIS

CENTRAL TO AMERICA'S ECONOMY®

Review
Review is published four times per year by the Research Division of the Federal Reserve Bank of St. Louis. Complimentary print subscriptions are available to U.S. addresses only. Full online access is available to all, free of charge.

Online Access to Current and Past Issues

The current issue and past issues dating back to 1967 may be accessed through our Research Division website: **http://research.stlouisfed.org/publications/review**. All nonproprietary and nonconfidential data and programs for the articles written by Federal Reserve Bank of St. Louis staff and published in *Review* also are available to our readers on this website.

Review articles published before 1967 may be accessed through our digital archive, FRASER: **http://fraser.stlouisfed.org/publication/?pid=820**.

Review is indexed in Fed in Print, the catalog of Federal Reserve publications (**http://www.fedinprint.org/**), and in IDEAS/RePEc, the free online bibliography hosted by the Research Division (**http://ideas.repec.org/**).

Authorship and Disclaimer

The majority of research published in *Review* is authored by economists on staff at the Federal Reserve Bank of St. Louis. Visiting scholars and others affiliated with the St. Louis Fed or the Federal Reserve System occasionally provide content as well. *Review* does not accept unsolicited manuscripts for publication.

The views expressed in *Review* are those of the individual authors and do not necessarily reflect official positions of the Federal Reserve Bank of St. Louis, the Federal Reserve System, or the Board of Governors.

Subscriptions and Alerts

Single-copy subscriptions (U.S. addresses only) are available free of charge. Subscribe here: **https://research.stlouisfed.org/publications/review/subscribe/**.

Our monthly email newsletter keeps you informed when new issues of *Review*, *Economic Synopses*, *Regional Economist*, and other publications become available; it also alerts you to new or enhanced data and information services provided by the St. Louis Fed. Subscribe to the newsletter here: **http://research.stlouisfed.org/newsletter-subscribe.html**.

Copyright and Permissions

Economic Data

General economic data can be obtained through FRED® (Federal Reserve Economic Data), our free database with over 500,000 national, international, and regional data series, including data for our own Eighth Federal Reserve District. You may access FRED through our website: **https://fred.stlouisfed.org**.

ISSN 0014-9187

A Short Introduction to the World of Cryptocurrencies

Aleksander Berentsen and Fabian Schär

In this article, we give a short introduction to cryptocurrencies and blockchain technology. The focus of the introduction is on Bitcoin, but many elements are shared by other blockchain implementations and alternative cryptoassets. The article covers the original idea and motivation, the mode of operation and possible applications of cryptocurrencies, and blockchain technology. We conclude that Bitcoin has a wide range of interesting applications and that cryptoassets are well suited to become an important asset class. (JEL G23, E50, E59)

Federal Reserve Bank of St. Louis *Review*, First Quarter 2018, *100*(1), pp. 1-16.
https://doi.org/10.20955/r.2018.1-16

1 INTRODUCTION

Bitcoin originated with the white paper that was published in 2008 under the pseudonym "Satoshi Nakamoto." It was published via a mailing list for cryptography and has a similar appearance to an academic paper. The creators' original motivation behind Bitcoin was to develop a cash-like payment system that permitted electronic transactions but that also included many of the advantageous characteristics of physical cash. To understand the specific features of physical monetary units and the desire to develop digital cash, we will begin our analysis by considering a simple cash transaction.

1.1 Cash

Cash is represented by a physical object, usually a coin or a note. When this object is handed to another individual, its unit of value is also transferred, without the need for a third party to be involved (Figure 1). No credit relationship arises between the buyer and the seller. This is why it is possible for the parties involved to remain anonymous.

The great advantage of physical cash is that whoever is in possession of the physical object is by default the owner of the unit of value. This ensures that the property rights to the units

Aleksander Berentsen is a professor of economic theory and Fabian Schär is managing director of the Center for Innovative Finance at the Faculty of Business and Economics, University of Basel.

Figure 1

Cash Transaction

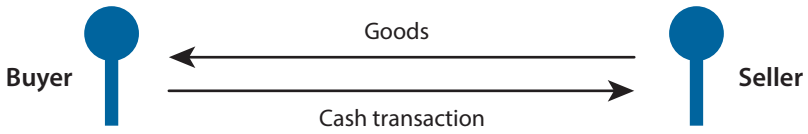
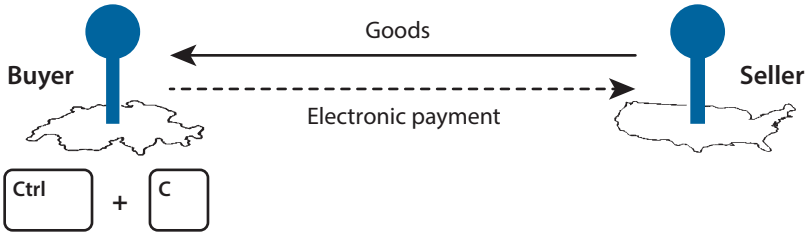


Figure 2

Electronic Payment



of value circulating in the economy are always clearly established, without a central authority needing to keep accounts. Furthermore, any agent can participate in a cash payment system; nobody can be excluded. There is a permissionless access to it. Cash, however, also has disadvantages. Buyers and sellers have to be physically present at the same location in order to trade, which in many situations makes its use impracticable.

1.2 Digital Cash

An ideal payment system would be one in which monetary value could be transferred electronically via cash data files (Figure 2). Such cash data files retain the advantages of physical cash but would be able to circulate freely on electronic networks.[1] A data file of this type could be sent via email or social media channels.

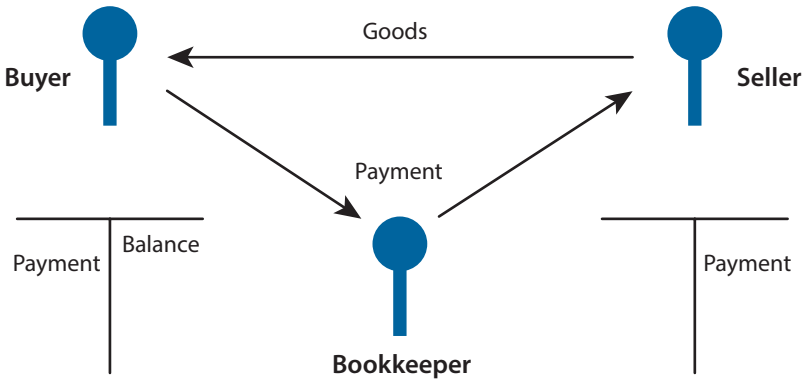
A specific feature of electronic data is that it can be copied any number of times at negligible cost. This feature is highly undesirable for money. If cash data files can be copied and the duplicates used as currency, they cannot serve as a payment instrument. This problem is termed the "double spending problem."

1.3 Electronic Payment Systems

To counteract the problem of double spending, classical electronic payment systems are based on a central authority that verifies the legitimacy of the payments and keeps track of the current state of ownership. In such systems, a central authority (usually a bank) manages the accounts of buyers and sellers. The buyer initiates a payment by submitting an order. The

Figure 3

Payment System with a Central Authority



central authority then ensures that the buyer has the necessary funds and adjusts the accounts accordingly (Figure 3).

Centralized payment systems solve the double spending problem, but they require trust. Agents must trust that the central authority does not misuse the delegated power and that it maintains the books correctly in any state of the world—that is, that the banker is not running away with the money. Furthermore, centralized systems are vulnerable to hacker attacks, technical failures, and malicious governments that can easily interfere and confiscate funds.

1.4 Stone Money of Yap

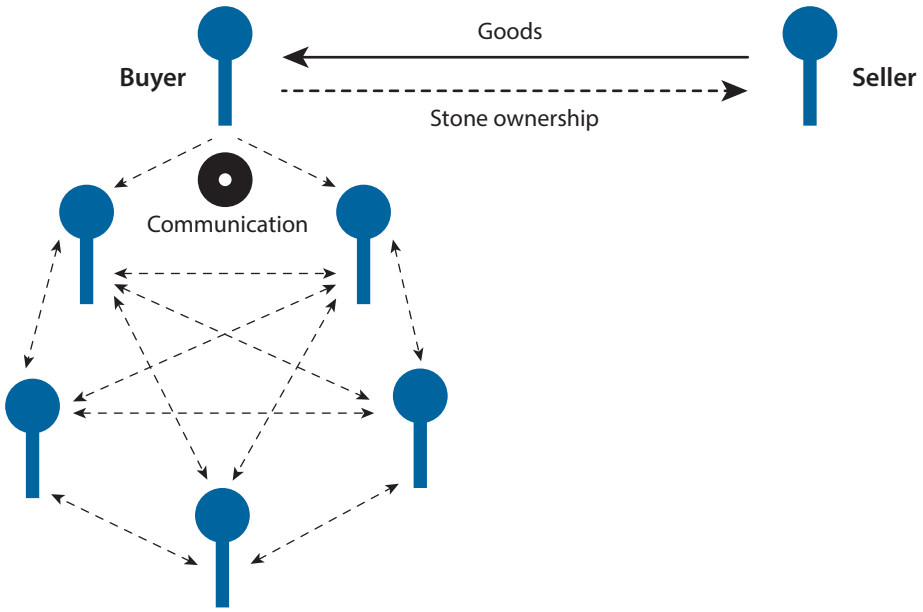
The key feature of the Bitcoin system is the absence of a centrally managed ledger. There is no central authority with an exclusive right to keep accounts. In order to understand how this is possible, we will first discuss a historical payment system that has certain similarities with the Bitcoin system.

On Yap Island, large millstone-like stones were used as a medium of exchange.[2] The stones were quarried almost 280 miles away on the island of Palau and brought to Yap by small boats. Every inhabitant could bring new stone money units into the system. The money creation costs, in the form of labor effort and equipment such as boats, protected the economy from inflation.

Instead of having to laboriously move the stones, which are up to 13 feet in diameter, with every transaction from a buyer's front yard to a seller's front yard, the ownership rights were transferred *virtually*. A stone remained at its original location, and the unit of value could be detached from it and circulated irrespective of the stone's whereabouts. It was sufficient that all the inhabitants knew who the owner of every stone was. The separation between the unit of value and the stone went so far that even the unit of value for stones that were lost at sea remained in circulation. The stone money of Yap can therefore be described as a quasi-virtual currency, as each unit of value was only loosely linked to a physical object.

Figure 4

Payment System with a Distributed Ledger



The Yap system was based on a distributed ledger, in which every inhabitant would keep track of a stone's ownership. When a buyer made a purchase, this person told his or her neighbors that the stone now belonged to the seller. The neighbors then spread the news until finally all of the island's inhabitants had been informed about the change in ownership (Figure 4). Through this communication, every islander had a precise idea of which unit of value belonged to which person at any point in time.

In its essential features, the Yap payment system is very similar to the Bitcoin system. A major difference is that in the Yap system false reports could not be immediately identified, so conflicts regarding the current state of the implicit ledger would have to be argued and settled by the group. The Yap system therefore was restricted to a group of manageable size with close relationships, in which misconduct could be punished by the group. In contrast, the Bitcoin system is designed to function in a network where no participant can trust any other participant. This feature is necessary because it is a permissionless payment system in which participants can remain anonymous through the use of pseudonyms.

1.5 Bitcoin and the Bitcoin Blockchain

Bitcoin is a virtual monetary unit and therefore has no physical representation. A Bitcoin unit is divisible and can be divided into 100 million "Satoshis," the smallest fraction of a Bitcoin. The Bitcoin Blockchain is a data file that carries the records of all past Bitcoin transactions, including the creation of new Bitcoin units. It is often referred to as the ledger of the Bitcoin

system. The Bitcoin Blockchain consists of a sequence of blocks where each block builds on its predecessors and contains information about new Bitcoin transactions. The average time between Bitcoin blocks is 10 minutes. The first block, block #0, was created in 2009; and, at the time of this writing, block #494600 was appended as the most recent block to the chain. Because everyone can download and read the Bitcoin Blockchain, it is a public record, a ledger that contains Bitcoin ownership information for any point in time.

The word "ledger" has to be qualified here. There is no single instance of the Bitcoin Blockchain. Instead, every participant is free to manage his or her own copy of the ledger. As it was with the stone money, there is no central authority with an exclusive right to keep accounts. Instead, there is a predefined set of rules and the opportunity for individuals to monitor that other participants adhere to the rules. The notion of "public record of ownership" also has to be qualified because the owners of Bitcoin units usually remain anonymous through the use of pseudonyms.

To use the Bitcoin system, an agent downloads a Bitcoin wallet. A Bitcoin wallet is software that allows the receiving, storing, and sending of (fractions of) Bitcoin units.[3] The next step is to exchange fiat currencies, such as the U.S. dollar, for Bitcoin units. The most common way is to open an account at one of the many Bitcoin exchanges and to transfer fiat currency to it. The account holder can then use these funds to buy Bitcoin units or one of the many other cryptoassets on the exchange. Due to the widespread adoption of Bitcoin, the pricing on large exchanges is very competitive with relatively small bid-ask spreads. Most exchanges provide order books and many other financial tools that make the trading process transparent.

A Bitcoin transaction works in a way that is similar to a transaction in the Yap payment system. A buyer broadcasts to the network that a seller's Bitcoin address is the new owner of a specific Bitcoin unit. This information is distributed on the network until all nodes are informed about the ownership transfer. We will examine some technical details of this step in Section 2.

For a virtual currency to function, it is crucial to establish at every point in time how many monetary units exist, as well as how many new units have been created. There must also be a consensus mechanism that ensures that all participants agree about the ownership rights to the virtual currency units. In small communities, as with the Yap islanders, everyone knows everyone else. The participants care about their reputation, and conflicts can be disputed directly. In contrast, within the Bitcoin system the number of participants is substantially larger, and network participants can remain anonymous. Consequently, reputation effects cannot be expected to have a significant positive impact, and coordination becomes very difficult. Instead, there is a consensus mechanism that allows the Bitcoin system to reach an agreement. This consensus mechanism is the core innovation of the Bitcoin system and allows consensus to be reached on a larger scale and in the absence of any personal relations.

1.6 Bitcoin Mining

To understand the consensus mechanism of the Bitcoin system, we first have to discuss the role of a miner. A miner collects pending Bitcoin transactions, verifies their legitimacy, and assembles them into what is known as a "block candidate." The goal is to earn newly cre-

ated Bitcoin units through this activity. The miner can succeed in doing this if he or she can convince all other network participants to add his or her block candidate to their copies of the Bitcoin Blockchain.

Bitcoin mining is permissionless. Anyone can become a miner by downloading the respective software and the most recent copy of the Bitcoin Blockchain. In practice, however, there are a few large miners that produce most of the new generally accepted blocks. The reason is that competition has become fierce and only large mining farms with highly specialized hardware and access to cheap electricity can still make a profit from mining.

For a block candidate to be generally accepted, it must fulfill a specific set of predefined criteria. For instance, all included transactions must be legitimate. Another important criterion is the so-called "fingerprint" of the block candidate. A miner obtains this fingerprint by computing the block candidate's hash value using the hash function dSHA256.

For example, we will look at the hash value for the text, "Federal Reserve Bank of Saint Louis." The fingerprint of this text, which was calculated using the hash function dSHA256, is

72641707ba7c9be334f111ef5238f4a0b355481796fdddfdaac4c5f2320eea68.

Now notice the small change in the original text to "federal Reserve Bank of Saint Louis." It will cause an unpredictable change of the fingerprint, which can be seen from the corresponding new hash value:

423f5dd7246de6faf8b839c41bf46d303014cffa65724ab008431514e36c4dba.

As suggested by this example, a data file's hash value cannot be prognosticated.

This characteristic is employed in the mining process as follows. For a block candidate to be accepted by all miners, its fingerprint must possess an extremely rare feature: The hash value must be below a certain threshold value—that is, it must display several zeroes at the beginning of the fingerprint. An example of a fingerprint of a block that was added to the Bitcoin Blockchain in 2010 is given in the following example:

Block #69785 (July 23rd, 2010, 12:09:36 CET)

$\underbrace{0000000000}_{\text{Need to be zero}}293b78a2833b45d78e97625f6484ddd1accbe0067c2b8f98b57995$

Miners are continuously trying to find block candidates that have a hash value satisfying the above mentioned criterion. For this purpose, a block includes a data field (called the nonce) that contains arbitrary data. Miners modify this arbitrary data in order to gain a new fingerprint. These modifications do not affect the set of included transactions. Just as with our example, every modification results in a new hash value. Most of the time, the hash value lies above the threshold value, and the miner discards the block candidate. If, however, a miner succeeds in creating a block candidate with a hash value below the current threshold value, he or she broadcasts the block candidate as quickly as possible to the network. All the other network participants can then easily verify that the fingerprint satisfies the threshold criterion by computing it themselves.

1.7 Consensus Mechanism

The consensus among miners is that every miner who receives a block candidate with a valid fingerprint adds it to his or her own copy of the Bitcoin Blockchain. From a game theoretical perspective, a strategy profile where all miners add valid blocks to their own copies of the Bitcoin Blockchain is a Nash equilibrium. If a miner believes that all other miners are acting accordingly, then it is a best response for that miner to add a valid block candidate to his or her own copy of the Bitcoin Blockchain. A deviation is not worthwhile, because it is not profitable to work on a version of the Bitcoin Blockchain that is not generally accepted. Any reward for finding blocks on a version of the chain that is not accepted by anyone else is worthless. Thus, although there is no authority enforcing this rule and miners are free to modify their copy of the Blockchain as they wish, there is a strong incentive to follow this rule. This self-enforcing rule allows the network to maintain consensus about the ownership of all Bitcoin units.[4]

Mining is expensive, as the computations use large amounts of electricity and are increasingly dependent on highly specialized hardware. Moreover, valid block candidates can be found only through a trial-and-error procedure. The consensus mechanism is therefore called "proof of work." If a miner finds a valid fingerprint for a block candidate, then this is proof that he or she has, on average, performed a large number of costly computations. Adding false information (e.g., illegitimate transactions) to a block candidate would render the block candidate invalid and essentially waste all the computations. Finding a valid fingerprint is therefore proof that the miner helped to maintain the Bitcoin system.

1.8 Monetary Policy

Every payment system needs rules that regulate how new monetary units are produced (or destroyed). The Bitcoin network is calibrated in such a way that, on average, a block candidate with a valid hash value is found every 10 minutes. The winner of the mining contest receives a predefined number of newly created Bitcoin units. The number currently is 12.5.

In the Bitcoin system, money creation is scheduled so that the number of Bitcoin units will converge to 21 million units (Figure 5). This limit exists because the reward for the miners is halved every 210,000 blocks (approximately every four years). Correspondingly, miners will be increasingly rewarded through transaction fees. But even today, the quick processing of a transaction can be guaranteed only if an adequate fee is paid to incentivize the miners to include the transaction in their block candidates.

Most Bitcoin users believe that Bitcoin's limited supply will result in deflation. That is, they are convinced that its value will forever increase. Indeed, up to this point we have witnessed a spectacular price increase from essentially a value of $0 for one Bitcoin unit in 2009 to a value of $7,000 at the time of this writing (Figure 6).

Nonetheless, these beliefs need to be challenged. Bitcoin units have no intrinsic value. Because of this, the present price of the currency is determined solely by expectations about its future price. A buyer is willing to buy a Bitcoin unit only if he or she assumes that the unit will sell for at least the same price later on. The price of Bitcoin, therefore, reacts highly elas-

Figure 5

Bitcoins in Circulation: Scheduled to Converge to 21 Million Units

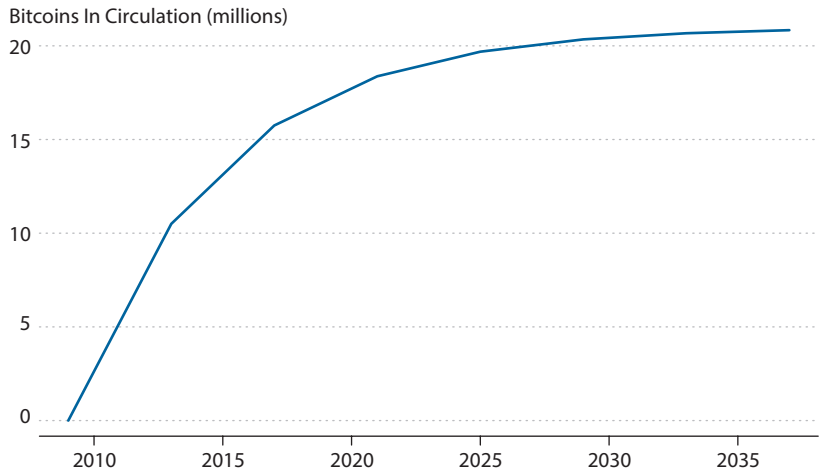


Figure 6

Market Price in U.S. Dollars (USD) for One Bitcoin Unit

SOURCE: Blockchain.info.

tically to changes in the expectations of market participants and is reflected in extreme price volatility. From monetary theory, we know that currencies with no intrinsic value have many equilibrium prices.[5] One of them is always zero. If all market participants expect that Bitcoin will have no value in the future, then no one is willing to pay anything for it today.

However, Bitcoin is not the only currency that has no intrinsic value. State monopoly currencies, such as the U.S. dollar, the euro, and the Swiss franc, have no intrinsic value either. They are fiat currencies created by government decree. The history of state monopoly currencies is a history of wild price swings and failures. This is why decentralized cryptocurrencies are a welcome addition to the existing currency system.

In the Bitcoin system, the path for the money supply is predetermined by the Bitcoin protocol written in 2008 and early 2009. Since then, many changes have been applied to the Bitcoin protocol. Most of these changes are not controversial and have improved the functioning of the Bitcoin system. However, in principle all aspects of the Bitcoin protocol can be amended, including the money supply. Many Bitcoin critics see this as a major shortcoming. Theoretically speaking, this is correct. Any network participant can decide to follow a new set of rules and, for example, double the amount of newly created "Bitcoin" units in his or her version of the ledger. Such a modification, however, is of no value because convincing all the other network participants to follow this new set of rules will be almost impossible. If the change of the protocol is not supported unanimously, there will be a so-called fork, a split in the network, which results in two co-existing blockchains and essentially creates a new crypto-asset. In this case, there would be Bitcoin (the original) and Bitcoin42 (a possible name for an alternative implementation with an upper bound of 42 million Bitcoin42 units). The market would price the original and the newly created Bitcoin42 assets according to the community's expectations and support. Therefore, even though in theory it is possible to increase the Bitcoin supply, in practice, such a change is very unlikely because a large part of the Bitcoin community would strongly oppose such an attempt.

Moreover, the same critique can be raised against any current government-operated fiat currency system. For example, since the Second World War, many central banks have become independent in order to shield them from political interference that yielded some undesirable outcomes. This independence has been given to them by the respective parliaments or related institutions and can be taken away if politicians decide accordingly. Political interference in the fiat currency system can be interpreted as a change in the "fiat currency protocol." Undesirable changes in fiat currency protocols are very common and many times have led to the complete destruction of the value of the fiat currency at hand. It could be argued that, in some ways, the Bitcoin protocol is more robust than many of the existing fiat currency protocols. Only time will tell.

2 BITCOIN TRANSACTIONS

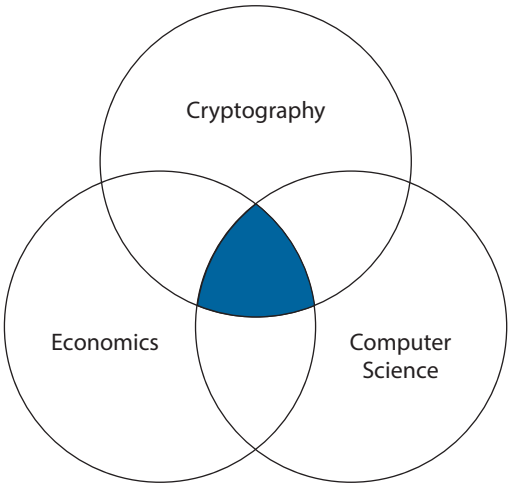
The complexity of the present material is due to interdisciplinarity. To understand the Bitcoin system, it is necessary to combine elements from the three disciplines of economics, cryptography, and computer science (Figure 7).

Having presented a broad overview of the Bitcoin system, we will explain a few technical elements of the system in greater detail. Blockchain uses proven technologies and links these in an innovative way. This combination has made the decentralized management of a ledger possible for the first time.

Berentsen and Schär (2017) argue that transaction processing demands that three requirements are satisfied: (1) transaction capability, (2) transaction legitimacy, and (3) transaction consensus. These three requirements will now be considered. In particular, we will explain how these conditions can be satisfied in the absence of a central authority.

Figure 7

Interdisciplinarity



2.1 Transaction Capability

What has to be resolved is how transactions can be initiated if there is no central authority. In a classical banking system, a client talks to his or her advisor or submits his or her payment instructions via the bank's online banking service. The infrastructure provided by the commercial bank and other central service providers ensures that the transaction will be communicated for execution. In the absence of a central authority, communicating a payment order in this traditional sense is not possible.

In the Bitcoin system, a payment order can be communicated to any number of network nodes. The network nodes are linked together in a loose network and forward the message until all nodes have been informed about the transaction (Figure 8).

The decentralization of the system has many advantages. In particular, it makes the system extremely robust. There is neither a central point of failure that can be attacked nor any system-relevant nodes that could cause the system to collapse. Therefore, the system functions even when some network nodes are unreachable, and it can always establish new connections and communication channels.

2.2 Transaction Legitimacy

Every participant can generate new payment orders and spread them across the network. This feature carries the risk of fraudulent messages. In this respect, there are two important questions that arise:

1. How do the nodes know that the initiator of the transaction is the rightful owner and that he or she is thereby entitled to transfer the Bitcoin units?
2. How can one ensure that the transaction message will not be tampered with before it is passed from one node to the next?

Figure 8

Bitcoin Transaction Communicated to Network Nodes

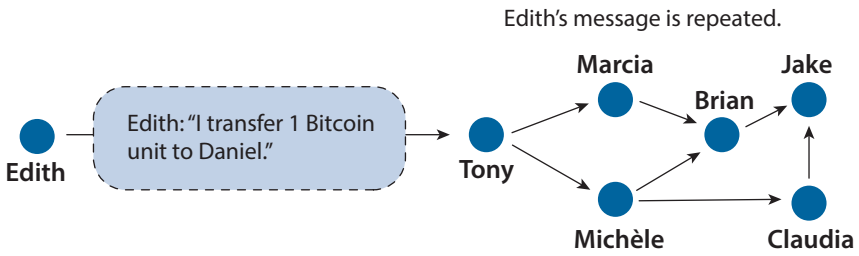


Figure 9

Bitcoin Transaction Manipulation Attempt

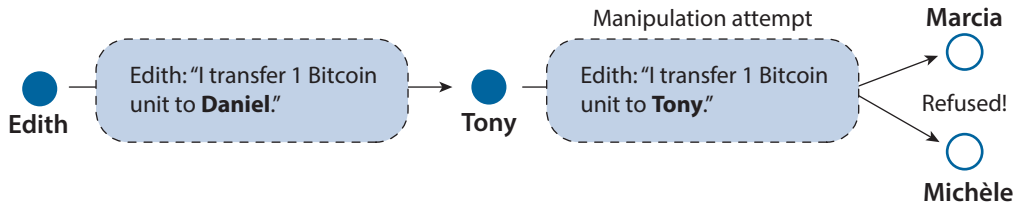
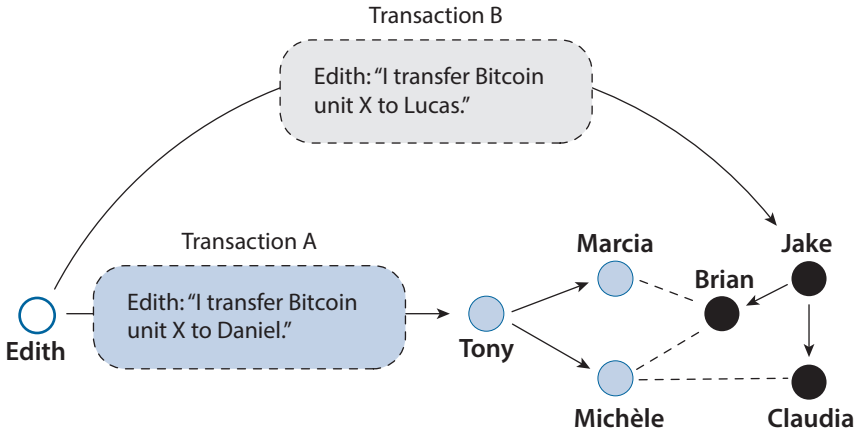


Figure 10

First Bitcoin Transaction Added to a Valid Block Candidate Is Confirmed



In the Bitcoin system, transaction legitimacy is guaranteed using asymmetric cryptography.[6] The idea is based on using pairs of keys consisting of a private and a public key. A private key should not be shared. It corresponds to a random value from an incredibly large set of numbers. A public key, on the other hand, is derived from that number and can be shared freely. It serves as a pseudonym in the Bitcoin network.[7]

A private key is used to encrypt a message that can be decrypted only by using its corresponding public key. This type of encryption is also known as a "signature." The signature clarifies that this approach is not used to hide any of the information in the encrypted message. Anyone can simply decrypt a message using its public key, but the signature serves as proof that the message has been previously encrypted using its corresponding private key; it's like a handwritten signature but much more secure.

For example, consider Edith, who wants to send a Bitcoin payment to Daniel over the Bitcoin network. She uses her private key to encrypt the message. The other network participants can only decrypt this message using Edith's public key. If an attempt is successful, it ensures that the message was encrypted using the corresponding private key. Because no one else has access to Edith's private key, this approach can be used to validate the transaction's origin (Figure 9).

When the transaction circulates in the network, any network participant can decrypt this message and is in the position to subsequently change the payment instructions. However, because the participant does not possess Edith's private key, he or she cannot re-encrypt the manipulated message. The tampered transaction will therefore be identified and rejected by the rest of the network.

2.3 Transaction Consensus

We have now discussed how a transaction message is communicated and how its legitimacy and origin can be verified. We have also explained how consensus regarding ownership of the Bitcoin units is achieved in the Bitcoin network by using the proof-of-work consensus protocol.

However, Edith would be able to generate two transactions that both reference the same Bitcoin units. Both transactions could be propagated simultaneously over the network (transaction capability), and both would display a valid origin (transaction legitimacy). Because of differences in the propagation of these two messages in the Bitcoin network, some of the nodes would first receive a message for transaction A while others would first receive a message for transaction B (Figure 10). In order to avoid double spending, it is important that only one of the two transactions finds its way into the Bitcoin Blockchain. A mechanism that decides which of the two transactions gets included in the Blockchain is therefore necessary.

The Bitcoin system solves this double spending problem in a clever way. The transaction that is first added to a valid block candidate, and therefore added to the Blockchain, is considered confirmed. The system ceases to process the other one—that is, miners will stop adding the conflicting transaction to their block candidates. Moreover, it is not possible for a miner to add conflicting transactions to the same block candidate. Such a block would be illegitimate and thus be rejected by all the other network participants.

3 OUTLOOK

As with any fundamental innovation, the true potential of blockchain technology will become apparent only many years, or possibly decades, after it becomes generally adopted. Forecasting the areas in which blockchain technology will be used to the greatest effect is therefore not possible. We nevertheless would like to mention a few areas where blockchain technology serves as an infrastructure platform that facilitates a variety of promising applications.

3.1 Cryptoassets

The most apparent application is Bitcoin as an asset. It is likely that cryptoassets such as Bitcoin will emerge as their own asset class and thus have the potential to develop into an interesting investment and diversification instrument. Bitcoin itself could over time assume a similar role as gold. Moreover, the potential for trading securities on a public blockchain is large. So-called colored coins can be traded on the Bitcoin (or similar) Blockchain and used in smart contracts, as described below.

3.2 Colored Coins

A colored coin is a promise of payment that is linked to a Bitcoin transaction. This promise is possible because the communication protocol of the Bitcoin network allows additional information to be tied to a transaction. For example, promises for the delivery of an ounce of gold or a dividend payment can be added to a Bitcoin transaction and represented on the Bitcoin Blockchain. Any of these promises are of course subject to issuer risks and require some extent of trust. This is in sharp contrast to native cryptoassets such as Bitcoin units.

3.3 Smart Contracts

Smart contracts are self-executing contracts.[8] They can be used to stipulate that a Bitcoin payment will be executed only when a certain condition is met. The Ethereum network is currently the leader in the field of smart contracts. Similar to Bitcoin, it is based on blockchain technology and provides a native cryptoasset, called Ether. In contrast to Bitcoin, Ethereum provides a more flexible scripting language and is able to track contractual states. Potential applications include but are not limited to e-voting systems, identity management and decentralized organization, and various forms of fundraising (e.g., initial coin offerings).

3.4 Data Integrity

Another application for public blockchains is the potential to monitor data files. We have already shown how fingerprints of block candidates play an important role in the Bitcoin network. The same technology can be used to produce fingerprints for all kinds of data files and then store them in a blockchain. The entry of a fingerprint into a blockchain ensures that any manipulation attempt will become apparent because any change to the data file will lead to a completely different hash value. Because it is very difficult to change a blockchain retroactively, a fingerprint can serve as proof that a specific data file existed at a specific point in time and ensures the integrity of the data.

4 RISKS

Much like any other key innovation, blockchain technology introduces some risks. The following sections will consider some of these risks. As we mentioned in Section 3, we would like to note that this list is non-exhaustive.

4.1 Forks

As discussed in Section 1.8, the Bitcoin protocol can be altered if the network participants, or at least a sufficient number of them, agree on the suggested modification. It can happen (and in fact has happened) that a blockchain splits because various groups cannot agree about a modification. A split that persists is referred to as a "fork." The two best-known examples of persistent splits are the Bitcoin Cash fork and Ethereum's ideological dissent, which resulted in the split to Ethereum and Ethereum Classic.

4.2 Energy Wastage

Proof-of-work mining is expensive, as it uses a great deal of energy. There are those that criticize Bitcoin and assert that a centralized accounting system is more efficient because consensus can be attained without the allocation of massive amounts of computational power. From our perspective, however, the situation is not so clear-cut. Centralized payment systems are also expensive. Besides infrastructure and operating costs, one would have to calculate the explicit and implicit costs of a central bank. Salary costs should be counted among the explicit costs and the possibility of fraud in the currency monopoly among the implicit costs. Moreover, many cryptoassets use alternative consensus protocols, which do not (solely) rely on computational resources.

4.3 Bitcoin Price Volatility

The price of Bitcoin is highly volatile. This leads us to the question of whether the rigid predetermined supply of Bitcoin is a desirable monetary policy in the sense that it leads to a stable currency. The answer is no because the price of Bitcoin also depends on aggregate demand. If a constant supply of money meets a fluctuating aggregate demand, the result is fluctuating prices. In government-run fiat currency systems, the central bank aims to adjust the money supply in response to changes in aggregate demand for money in order to stabilize the price level. In particular, the Federal Reserve System has been explicitly founded "to provide an elastic currency" to mitigate the price fluctuations that arise from changes in the aggregate demand for the U.S. dollar. Since such a mechanism is absent in the current Bitcoin protocol, it is very likely that the Bitcoin unit will display much higher short-term price fluctuations than many government-run fiat currency units.

5 CONCLUSION

The Bitcoin creators' intention was to develop a decentralized cash-like electronic payment system. In this process, they faced the fundamental challenge of how to establish and transfer

digital property rights of a monetary unit without a central authority. They solved this challenge by inventing the Bitcoin Blockchain. This novel technology allows us to store and transfer a monetary unit without the need for a central authority, similar to cash.

Price volatility and scaling issues frequently raise concerns about the suitability of Bitcoin as a payment instrument. As an asset, however, Bitcoin and alternative blockchain-based tokens should not be neglected. The innovation makes it possible to represent digital property without the need for a central authority. This can lead to the creation of a new asset class that can mature into a valuable portfolio diversification instrument. Moreover, blockchain technology provides an infrastructure that enables numerous applications. Promising applications include using colored coins, smart contracts, and the possibility of using fingerprints to secure the integrity of data files in a blockchain, which may bring change to the world of finance and to many other sectors. ■

NOTES

1 An initial attempt was DigiCash in the 1990s; however, it was not able to establish itself.

2 See Furness (1910) who describes the Island of Stone Money.

3 Strictly speaking, Bitcoins are not "traveling" on the Bitcoin network. A Bitcoin payment is simply a message that is broadcasted to the network to communicate a change in ownership of the respective Bitcoin units.

4 In practice, a split in the Blockchain may occur if the network participants do not agree about changes in the Bitcoin protocol (i.e., the rule set). This issue is discussed further in this article.

5 See Kiyotaki and Wright (1993) for a search theoretic approach to money, Berentsen (1998) for a study of the acceptability of digital money, and Nosal and Rocheteau (2011) for a comprehensive introduction into the search theoretic approach to monetary economics.

6 Similar technologies are also used in traditional electronic payment systems and in many other fields, such as with online banking and shopping.

7 In fact, a public key is usually used to derive a so-called Bitcoin address. This address is then used as a pseudonym. We ignored this additional step to keep things as simple as possible. Both operations—that is, private key to public key and public key to Bitcoin address—are one-way functions. There is no known way to reverse these operations, so it is not feasible to obtain a private key from a corresponding pseudonym.

8 For an introduction to smart contracts and potential business applications, see Schär and Langer (2017).

REFERENCES

Berentsen, Aleksander. "Monetary Policy Implications of Digital Money." *Kyklos (International Review of Social Sciences)*, 1998, *51*(1), pp. 89-117; https://doi.org/10.1111/1467-6435.00039.

Berentsen, Aleksander and Schär, Fabian. *Bitcoin, Blockchain und Kryptoassets: Eine umfassende Einführung*. Books on Demand, Norderstedt, 2017.

Furness, William H. *The Island of Stone Money: Uap of the Carolines*. Philadelphia: J. B. Lippincott, 1910.

Kiyotaki, Nobuhiro and Wright, Randall. "A Search-Theoretic Approach to Monetary Economics." *American Economic Review*, 1993, *83*(1), pp. 63–77.

Nakamoto, Satoshi. "Bitcoin: A Peer-to-Peer Electronic Cash System." 2008; https://bitcoin.org/bitcoin.pdf.

Nosal, Ed and Rocheteau, Guillaume. *Money, Payments, and Liquidity*. Cambridge and London: The MIT Press, 2011; https://doi.org/10.7551/mitpress/9780262016285.001.0001.

Schär, Fabian and Langer, Dominik. "Smart Contracts – eine missverstandene Technologie mit hohem Potenzial." *Synpulse Magazin*, 2017, *3*(17), pp. 38-41.

Furnishing an "Elastic Currency": The Founding of the Fed and the Liquidity of the U.S. Banking System

Mark Carlson and David C. Wheelock

This article examines how the U.S. banking system responded to the founding of the Federal Reserve System (Fed) in 1914. The Fed was established to bring an end to the frequent crises that plagued the U.S. banking system, which reform proponents attributed to the nation's "inelastic" currency stock and dependence on interbank relationships to allocate liquidity and operate the payments system. Reform advocates noted that banking panics tended to occur at times of the year when the demands for currency and bank loans were normally at seasonal peaks and money markets were at their tightest. Moreover, they blamed the interbank system, upon which the banking system depended for seasonal accommodation and interregional payments, for transmitting shocks throughout the banking system. The article finds that after the Fed's founding, country national banks were much less dependent on correspondent banks for seasonal liquidity and that peaks in lending by individual Reserve Banks aligned with the liquidity needs of banks in their districts. Further, the article shows that after the Fed's founding, banks generally were less liquid and relied more heavily on deposits for funding, consistent with the idea that banks viewed the Fed as a reliable source of liquidity. The return of banking panics during the Great Depression, however, showed that the Fed was not, in fact, up to the challenge of serving as a full-fledged lender of last resort. (JEL E58, G21, N21, N22)

Federal Reserve Bank of St. Louis *Review*, First Quarter 2018, *100*(1), pp. 17-44.
https://doi.org/10.20955/r.2018.17-44

Financial crises often result in sweeping changes in financial regulation. The financial crises of the 1930s, for example, led to major changes in U.S. regulation of commercial banks and securities markets and the introduction of federal deposit insurance. Similarly, following the financial crisis of 2007-08, the Dodd-Frank Act of 2010 introduced the most far-reaching changes in U.S. bank regulation since the Great Depression, while U.S. and foreign authorities agreed on new capital and liquidity rules affecting large, internationally active banks.[1]

Financial crises—especially those involving the banking system—can also fundamentally alter the role of governments or central banks as lenders of last resort. Major changes in the

Mark Carlson is a senior economic project manager with the Board of Governors of the Federal Reserve System. David C. Wheelock is vice president and deputy director of research at the Federal Reserve Bank of St. Louis. Paul Morris provided research assistance. The authors thank Steve Williamson and Yi Wen for comments on a prior version of this article.

rules governing Federal Reserve (Fed) lending were enacted both during the Great Depression and following the crisis of 2007-08. In the 1930s, concerns that the Fed did too little to save the banking system or protect the economy prompted Congress to enact legislation that expanded the Fed's ability to lend to banks and other firms and restructured the Federal Reserve System in an effort to make it a more-responsive lender of last resort. In 2007-08, the Fed lent heavily to commercial banks and other financial institutions, in some cases using authorities granted during the Great Depression. Concerns that the Fed had too much latitude led Congress, in the Dodd-Frank Act, to rein in the Fed's ability to lend to distressed firms.

Economists and policymakers are interested in how banks respond to changes in regulation and the rules governing access to a lender of last resort to determine whether those changes have their intended effects. Such changes can affect banks' incentives to take risks, engage in certain activities, or grow in size, with implications for the broader economy.[2] New restrictions on a central bank's lending authority, for example, might cause banks to reduce the liquidity services they offer to their customers to lessen the chance they will need to borrow from the central bank, while an easing of restrictions might lead banks to take greater risks, knowing that the central bank will backstop them in a crisis.

This article examines how the U.S. banking system responded to the founding of the Federal Reserve System in 1914. The Federal Reserve was established primarily to bring an end to the recurring crises that plagued the U.S. banking system, which reform proponents saw as stemming from the nation's "inelastic" currency stock and dependence on interbank relationships to allocate liquidity and operate the payments system. The Federal Reserve Act was intended to solve these problems by creating a new currency—Federal Reserve notes—supplied by regional Reserve Banks through lending to their member banks. Member banks would hold reserve deposits with their Reserve Bank and acquire additional reserves or currency from the Reserve Bank as needed to accommodate the short-term credit and liquidity needs of local commercial and agricultural activity. Moreover, the Reserve Banks would provide check clearing and other payments services to their members. Although not stated as such in the Federal Reserve Act, the Fed was intended to perform the functions of a central bank, including serving as lender of last resort for the banking system.

In focusing on the need for an elastic currency, reform advocates noted that banking crises tended to occur at times of the year when the demands for currency and bank loans were normally at seasonal peaks and money markets were at their tightest. Moreover, they blamed the interbank system, upon which the banking system depended for seasonal accommodation and interregional payments, for transmitting shocks throughout the banking system. Researchers have shown that market interest rates exhibited much less seasonal variability after 1914 than before, suggesting that the Fed's founders accomplished their goal of eliminating seasonal strains in money markets (e.g., Miron, 1986).[3] Others have shown that the balances that national banks held with correspondent banks in major cities were also less seasonally variable after the Fed's founding, suggesting reduced seasonal pressures on the interbank system (Carlson and Wheelock, 2016a,b).[4] Further, Carlson and Wheelock (2016b) find that, as a percentage of their total assets, national banks held much lower levels of liquid assets, including cash and deposits with reserve agents (i.e., designated national banks during 1894-1914 and the Fed

during 1921-28), after the Fed's founding, suggesting that the Fed's presence made banks comfortable operating with less liquidity.

This article builds on the prior studies by comparing the seasonal volatility and average levels of key bank balance sheet ratios before and after the Fed was established. Although we do not test formally whether the Fed caused or contributed to changes in bank balance sheets, the article presents new evidence that is consistent with the objective of the Fed's founders to relieve seasonal pressures on the banking system and prior research indicating that the presence of the Fed allowed banks to operate with lower liquidity buffers.

The article begins by describing the defects of the U.S. banking system that the Fed's founders hoped to rectify. We then compare the seasonal variation in national bank balance sheets between the 20 years before the Fed's founding and the 1920s. Further, we show that differences in the seasonal patterns of lending by the individual Reserve Banks were consistent with regional differences in the timing of seasonal liquidity demands, especially between predominantly agricultural regions and regions with more-diversified economies. Finally, we compare the levels of key balance sheet ratios before and after the founding of the Fed to provide further evidence on how the presence of the Fed might have affected banks' willingness to assume greater liquidity risk or leverage.

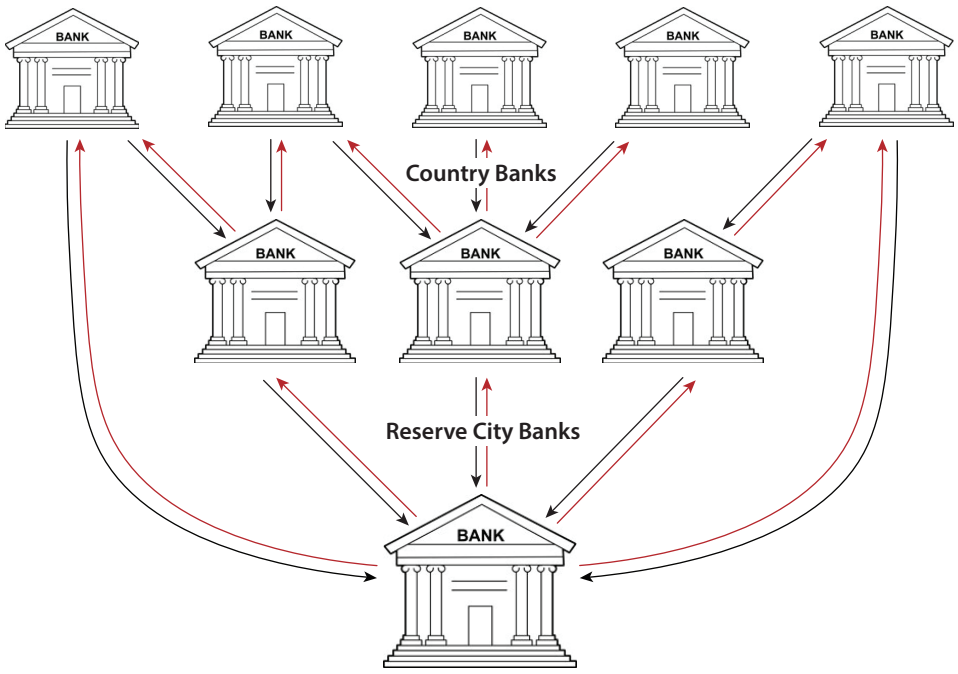
HISTORICAL BACKGROUND

The Federal Reserve System was established in 1914 to correct defects of the American banking system that reformers blamed for the banking panics that occurred every few years throughout the nineteenth and early-twentieth centuries. Panics were marked by widespread suspensions of cash withdrawals and payments, sharp increases in interest rates, and bank failures. They were widely attributed to the nation's inelastic currency, the concentration of the banking system's reserves in a small number of banks in New York City and a few other large cities, and the dependence of the system on interbank relationships to move funds between regions (see, e.g., Kemmerer, 1910, and Sprague, 1910).

Reform proponents, naturally, called for the creation of an "elastic currency," by which they meant a money stock whose supply adjusts to fluctuations in demand. At the time, the U.S. currency stock consisted primarily of national bank notes (notes issued by commercial banks with federal charters in proportion to the amount of U.S. government bonds they held in their portfolios), notes issued by the federal government during the Civil War ("greenbacks"), and various gold and other coins issued by the U.S. Treasury. Whereas the stock of currency (and coin) was relatively inflexible in the short run, the demands for money and credit, and the volume of payments, were highly variable. Reflecting the importance of agriculture in many parts of the country, money and credit demands fluctuated widely during the year, resulting in substantial intra-year variability in interest rates and money market conditions. Contemporary observers noted that banking panics tended to occur at the times of the year when the seasonal strains on money markets were most acute (see, e.g., Kemmerer, 1910, and Sprague, 1910).[5]

Figure 1

Structure of the National Banking System



NOTE: Arrows represent flows of funds between national banks in different tiers of cities. Typically, country banks maintained deposits with banks in reserve cities and central reserve cities, and reserve city banks maintained deposits with banks in central reserve cities. (Some banks also held deposits with banks in their own or a lower-tier city.) A deposit that one bank maintains in another bank (known as the correspondent) is an asset of the depositing bank and referred to as a deposit "due from" the correspondent bank. That deposit is a liability of the correspondent bank and is thus a deposit "due to" the first bank. Flows of funds between banks included withdrawals or additions to deposits with correspondents, interbank borrowing, and other payments. Banking panics were often characterized by suspensions of payments that interrupted interbank flows and cascaded through the different tiers of the system.

Observers also noted that the U.S. banking system was highly dependent on interbank connections. Unlike the banking systems of most countries, which were dominated by a few banks with nationwide branches, the U.S. banking system was composed of thousands of single-office "unit" banks that depended on correspondent relationships with banks in other cities and towns for payments and other services. Banks throughout the country maintained relationships with correspondent banks in larger cities to facilitate payments, invest surplus funds, and obtain additional funds needed to satisfy local demands for money or loans. The structure of reserve requirements imposed under the National Banking Acts of the 1860s further encouraged growth of the interbank system. Banks with federal charters, that is, national banks, were grouped into three reserve tiers. Those located in designated central reserve cities (originally just New York City, but later also Chicago and St. Louis) were required to maintain cash reserves equal to at least 25 percent of their deposit liabilities. National banks in desig-

nated reserve cities were also subject to a 25 percent requirement, but those banks could maintain as much as half of their required reserves in the form of deposits at national banks in central reserve cities. National banks in all other cities and towns, known as country banks, were subject to a 15 percent reserve requirement, three-fifths of which could be held as deposits at reserve city or central reserve city banks.[6] Most banks held a large portion of their reserves in the form of correspondent balances, which typically paid interest, rather than as vault cash. Indeed, many banks maintained correspondent balances well in excess of their statutory reserve requirement because of their usefulness for making payments and buffering seasonal liquidity demands. Banks would draw down their interbank deposits or borrow from their correspondents when local demands for cash and loans were high and deposit surplus funds with their correspondents when local demands were low. Figure 1 illustrates the structure of the national banking system.

Contemporaries viewed the interbank system as something of a necessary evil. Banks depended on the system to meet local demands for liquidity and for making interregional payments. However, the system seemed unable to supply enough funds to meet local needs (at least according to borrowers who complained about seasonal spikes in interest rates), and the system was vulnerable to disruptions caused by panics, especially in the central money markets. Major banking panics occurred in 1893 and 1907, for example, when banks throughout the country were unable to obtain funds from their New York City correspondents after those banks suspended withdrawals (Sprague, 1910). Calomiris and Carlson (2017) show that banks with substantial correspondent business were especially vulnerable to suspensions by New York City banks because of their inability to withdraw funds from New York City banks to satisfy the demands of their own respondents. Although the banks of New York City and other cities worked together under the auspices of their local clearinghouses to protect themselves and limit the fallout of panics, the system lacked a lender of last resort that could rapidly inject cash or other forms of liquidity into the banking system to halt a panic.

THE FED'S IMPACT

The Fed's founders believed that banking panics could be eliminated by solving the "inelastic currency" problem, which was reflected most obviously in wide seasonal fluctuations in interest rates and money market conditions. Interest rates displayed much less seasonal variability after the Fed's founding (Friedman and Schwartz, 1963, and Miron 1986), and the Fed's discount window lending added liquidity to the banking system at the times of the year when previously money markets had tightened and interest rates spiked (Carlson and Wheelock, 2016b). Further, both the average volume and seasonal variability of interbank deposits were much reduced after the Fed was established (Carlson and Wheelock, 2016b).

Following Carlson and Wheelock (2016a,b), we infer the importance of seasonal forces on national bank balance sheets from principal components analysis of intra-year changes in various balance sheet items across U.S. states. The principal components help reveal the main drivers of changes in item values over time. For example, if the first principal component exhibits a markedly seasonal pattern and explains a high percentage of the underlying cova-

riance of the data, we conclude that seasonal forces were an important influence on the given balance sheet item. Principal components analysis is also useful for detecting changes in the importance of seasonal forces on the item over time. Further, the analysis allows us to identify the states whose banks contribute the most to the components and, thus, whose balance sheets exhibit the greatest seasonal variability.

Our data consist of state-level balance sheet information for country national banks over the 20 years prior to the Fed's establishment, 1894-1914, and for 1921-28. We omit the years 1915-20 to avoid the three-year phase-in period for the Fed's member banks to adjust to reserve requirements specified in the Federal Reserve Act and the impact of World War I financing on Federal Reserve and member bank balance sheets.[7] The frequency of observations varies by year, reflecting the calls issued by the Comptroller of the Currency for banks to report their balance sheets. The Comptroller issued five calls per year during 1894-1914 and from three to five per year during 1921-28.

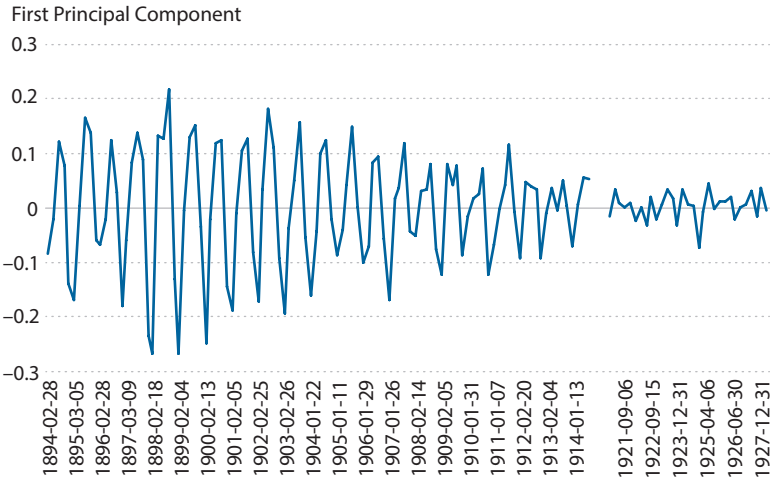
Figure 2 plots the first principal component of changes between reporting dates in net balances due from national banks for country banks, scaled by their initial-period total assets; that is, (Net Due From$_t$ – Net Due From$_{t-1}$)/Assets$_{t-1}$.[8] The first principal component explains 26 percent of the variation in the data and exhibits a decidedly seasonal pattern. The pattern appears less seasonal in the 1920s, with less intra-year variability, consistent with reduced seasonal pressure on the interbank system after the Fed was established.[9]

Besides drawing on their accounts with correspondent banks, country banks often borrowed from their correspondents for short periods, especially at times of the year when local demands for cash and loans were highest. Figure 3 displays the first principal component of changes in the short-term borrowing of country banks scaled by their total assets.[10] The first principal component explains 44 percent of the variation in the data and exhibits a decidedly seasonal pattern. The intra-year variation appears somewhat lower in the 1920s, but not markedly so. Although the balance sheet information for national banks does not indicate the source of their loans, in the 1920s national banks likely borrowed mainly from the Fed rather than from other national banks. We show that the seasonal pattern of Federal Reserve lending in the 1920s was similar to the seasonal pattern of borrowing by national banks.

Further evidence of reduced seasonal pressures on country bank balance sheets is provided in Figure 4, which plots the first principal component of changes in the reserves-to-assets (hereafter reserves/assets) ratio. (We define reserves as the sum of vault cash, cash items in the process of collection, and deposits with reserve agents.) The principal component accounts for 26 percent of the variance in the data and exhibits a highly seasonal pattern. Country banks, especially in farming regions, faced wide seasonal fluctuations in their customers' demands for cash and loans. In peak seasons, country banks experienced both high loan demand and cash withdrawals. To meet those demands, country banks drew down their reserves of cash and correspondent deposits, causing their reserves/assets ratios to fall. When country banks experienced slack demand for loans and cash, they built up their reserves, causing their reserves/assets ratios to rise. As Figure 4 shows, the intra-year variability in the first principal component of changes in the reserves/assets ratio was much lower in the 1920s, suggesting that Federal Reserve lending enabled banks to smooth their reserves/assets ratios across seasons, which made them less vulnerable to shocks.

Figure 2

Interbank Deposit Flows, 1894-1914 and 1921-28

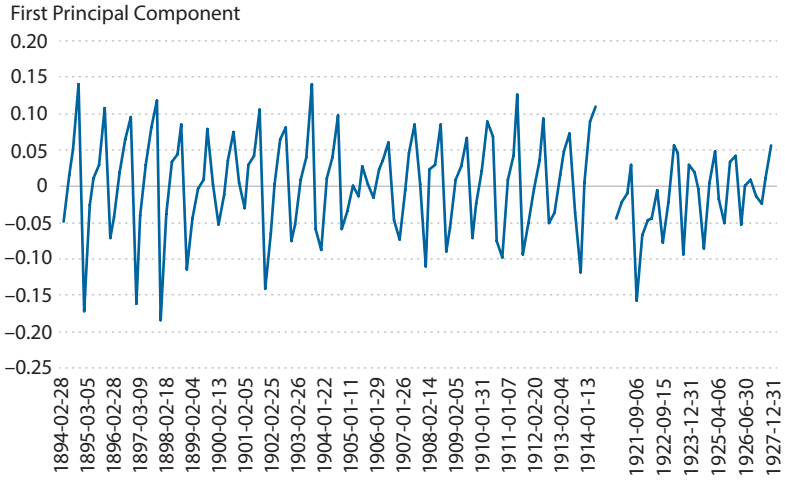


NOTE: The figure plots the first principal component of changes between reporting dates in the ratio of deposits due from other national banks to total assets for country national banks.

SOURCE: National bank data through 1910: Weber (2000). National bank data after 1910: U.S. Office of the Comptroller of the Currency (1917-28).

Figure 3

Short-Term Borrowing by Country National Banks, 1894-1914 and 1921-28

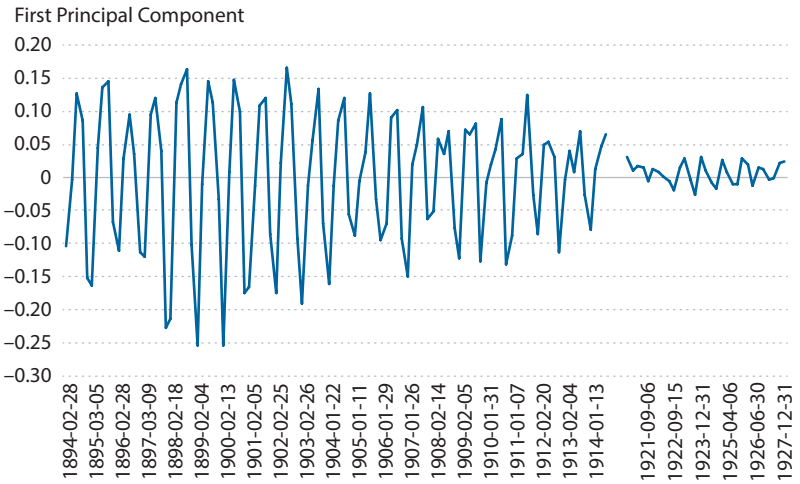


NOTE: The figure plots the first principal component of changes between reporting dates in the ratio of short-term borrowing ("bills discounted" and "bills payable") to total assets for country national banks.

SOURCE: National bank data through 1910: Weber (2000). National bank data after 1910: U.S. Office of the Comptroller of the Currency (1917-28).

Figure 4

Reserve Flows of Country National Banks, 1894-1914 and 1921-28

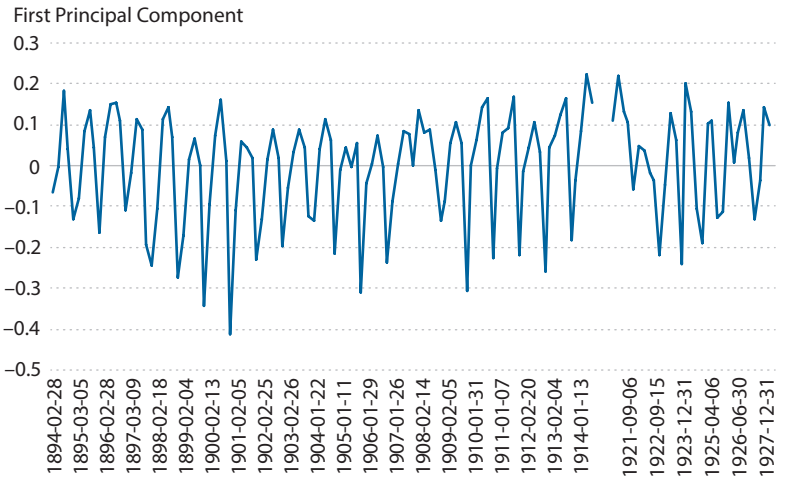


NOTE: The figure plots the first principal component of changes between reporting dates in the ratio of reserves to total assets for country national banks, where reserves is the sum of vault cash, cash items in the process of collection, and deposits with reserve agents.

SOURCE: National bank data through 1910: Weber (2000). National bank data after 1910: U.S. Office of the Comptroller of the Currency (1917-28).

Figure 5

Non-Bank Deposit Flows of Country National Banks, 1894-1914 and 1921-28

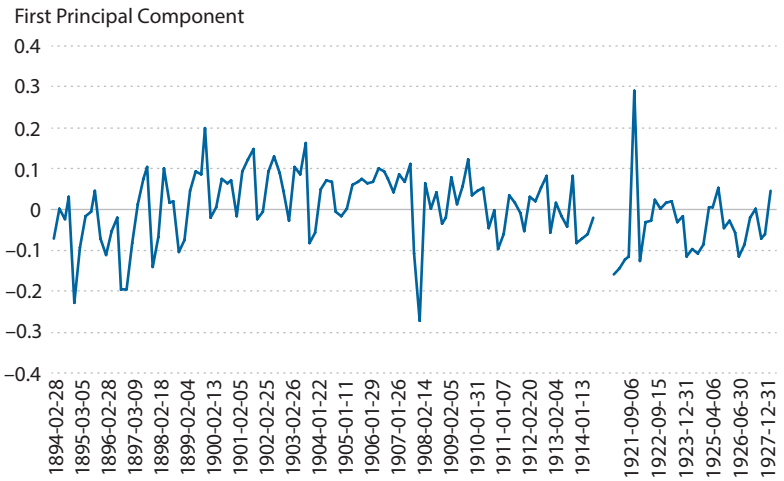


NOTE: The figure plots the first principal component of changes between reporting dates in the ratio of deposits of individuals to total assets for country national banks, where deposits of individuals includes deposits of firms, households, and state and local governments, but does not include federal government deposits or interbank deposits.

SOURCE: National bank data through 1910: Weber (2000). National bank data after 1910: U.S. Office of the Comptroller of the Currency (1917-28).

Figure 6

Changes in Total Loans of Country National Banks, 1894-1914 and 1921-1928



NOTE: The figure plots the first principal component between reporting dates in the ratio of loans to total assets for country national banks.

SOURCE: National bank data through 1910: Weber (2000). National bank data after 1910: U.S. Office of the Comptroller of the Currency (1917-28).

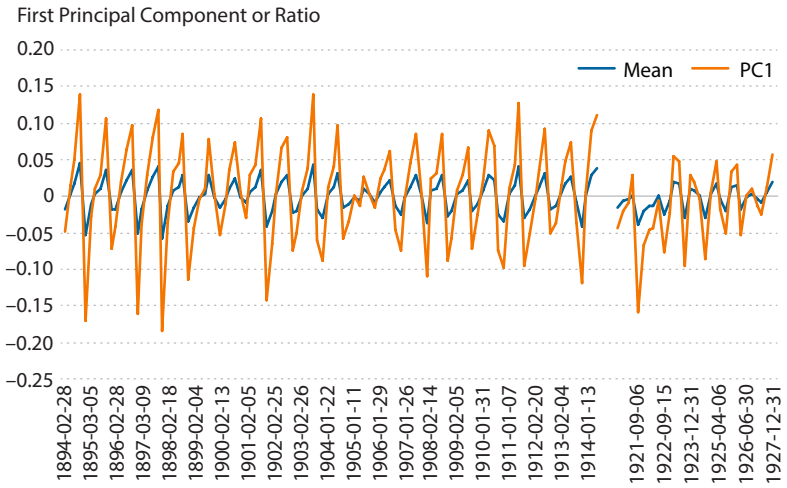
Figure 5 plots the first principal component of changes in the deposits of individuals (including firms and state and local governments, but not the federal government or interbank deposits), scaled by total assets—that is, deposits/assets, for country national banks. The principal component exhibits a highly seasonal pattern, with little evidence of a change in intra-year variation in the 1920s. Figure 6 plots the first principal component of changes in the loans-to-assets ratio for country national banks. This principal component appears somewhat less seasonal than the principal component of changes in deposits/assets, both before 1914 and during the 1920s. Miron (1986) argues that total loans in the economy (both private loans and Federal Reserve loans) should have exhibited more seasonal variability after the Fed's founding, but that private lending should have been less seasonal. However, the intra-year variation in the first principal component of changes in loans/assets exhibits no clear evidence of a change in seasonal pattern. The relative stability of reserves/assets in the face of continued seasonal variability in deposits/assets and loans/assets during the 1920s, as well as the substantial decline in the seasonal variability in market interest rates, is consistent with the Fed having provided seasonal liquidity to the banking system. The next section provides additional evidence linking the provision of seasonal liquidity to Federal Reserve lending.

Regional Patterns in Seasonal Demands and Federal Reserve Lending

In his study for the National Monetary Commission, Kemmerer (1910) documented distinct regional differences in seasonal demands for money and credit and their effects on the

Figure 7

Short-Term Borrowing by Country Banks in the U.S. South, 1894-1914 and 1921-28



NOTE: The figure plots the first principal component of changes between reporting dates in the ratio of short-term borrowing to total assets for all country national banks ("PC1") and the mean change in the ratio for country national banks in eight southern states (Alabama, Arkansas, Georgia, Louisiana, Mississippi, North Carolina, South Carolina, and Texas) ("Mean").

SOURCE: National bank data through 1910: Weber (2000). National bank data after 1910: U.S. Office of the Comptroller of the Currency (1917-28).

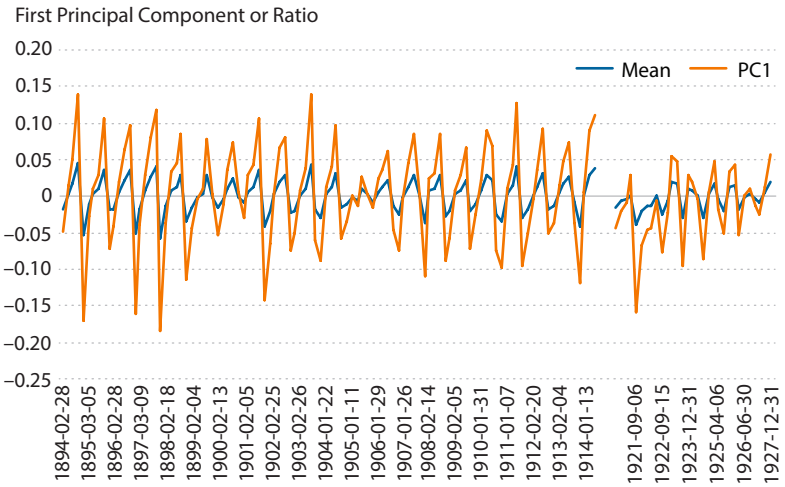
size and timing of currency and payments flows through the interbank system. Kemmerer showed that movements of currency and bank reserves, both within and between different regions of the country, were large and dominated by the demands of agriculture and other activities with regular seasonal patterns. Further, he showed that regional differences in the timing and size of seasonal demands reflected differences in the size of the agricultural sector and types of commodities produced and other seasonal activities.

This subsection focuses on the regional differences in the seasonal patterns of bank balance sheets and Federal Reserve Bank lending during the 1920s. Proponents of a geographically decentralized system argued that a monolithic central bank would not be responsive to the needs of different parts of the country. Hence, the Federal Reserve System was designed as a confederation of regional Reserve Banks that would each provide for the currency and credit requirements of its own district and thereby take pressure off of the interbank system to move funds between regions:

> The very essence of the new plan is intended to meet the condition which in the past has caused chief trouble by eliminating this necessity of interdependence between districts. The Federal Reserve Act will presumably afford a means of making each district self-supporting in a credit way so that assuming the plan to work as it is expected to work the need for mutual seasonal aid and shipments of currency will be minimized. (Reserve Bank Organization Committee, 1914, p. 15)

Figure 8

Changes in Total Loans of Country National Banks in the U.S. South, 1894-1914 and 1921-28

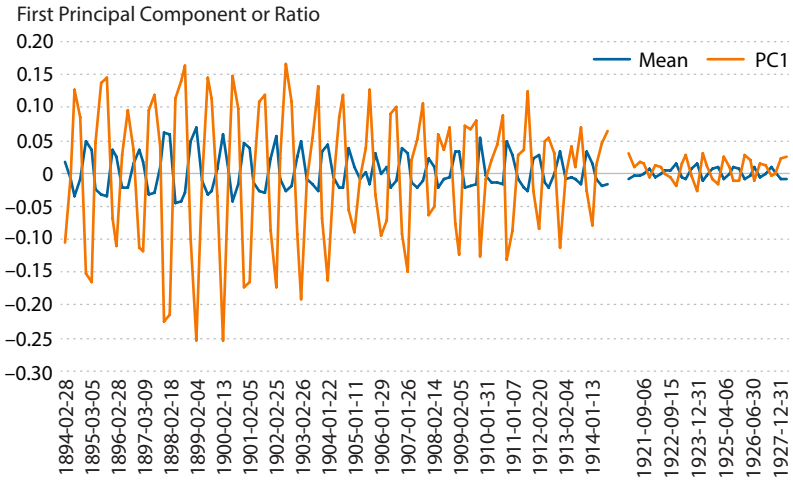


NOTE: The figure plots the first principal component of changes between reporting dates in the ratio of loans to total assets for all country national banks ("PC1") and the mean change in the ratio for country national banks in eight southern states (Alabama, Arkansas, Georgia, Louisiana, Mississippi, North Carolina, South Carolina, and Texas) ("Mean").

SOURCE: National bank data through 1910: Weber (2000). National bank data after 1910: U.S. Office of the Comptroller of the Currency (1917-28).

Figure 9

Reserve Flows of Country National Banks in the U.S. South, 1894-1914 and 1921-1928



NOTE: The figure plots the first principal component of changes between reporting dates in the ratio of reserves to total assets for all country national banks ("PC1") and the mean change in the ratio for country national banks in eight southern states (Alabama, Arkansas, Georgia, Louisiana, Mississippi, North Carolina, South Carolina, and Texas) ("Mean").

SOURCE: National bank data through 1910: Weber (2000). National bank data after 1910: U.S. Office of the Comptroller of the Currency (1917-28).

If the System worked as the organizers intended, the seasonal lending of the different Reserve Banks should have mimicked the seasonal patterns in liquidity demands of their districts. Further, evidence that the Fed's lending coincided with seasonal demands would seem to indicate that the Fed, rather than something else, was responsible for the easing of seasonal pressures on money markets and the interbank system.

Before the Fed was established, country banks borrowed short-term funds mainly from their correspondents. As shown in Figure 3, the short-term borrowing of country national banks exhibited a decidedly seasonal pattern. It was also highly concentrated geographically. The states that load most strongly on the first principal component of changes in borrowing (scaled by total assets)—that is, borrowing/assets—between reporting dates are eight southern states: Alabama, Arkansas, Georgia, Louisiana, Mississippi, North Carolina, South Carolina, and Texas. Figure 7 plots the average change in borrowing/assets across these eight states alongside the first principal component shown in Figure 3. Clearly, the seasonal pattern of the principal component mimicked the short-term borrowing by banks in those eight states. According to Redenius and Weiman (2011), the dominance of a single crop—cotton—and the particularly important role of local banks (and by extension their correspondents) in financing the marketing of the cotton crop gave the South a "voracious appetite" for interbank loans and an outsized impact on the "systemic seasonality" of U.S. liquidity demand. Citing data from Kemmerer (1910), Redenius and Weiman (2011) note that southern banks, despite accounting for less than 4 percent of total U.S. bank assets in 1900, accounted for 25 percent of the outflow of cash from the New York money market in the fall and some 30 percent of the flow into New York in the late winter and early spring. By contrast, in the Midwest and other farming regions, greater diversity in the mix of crops and animal production smoothed seasonal demands somewhat, while a more efficient distribution system placed less demand on local banks for financing crop marketing.

Figure 8 plots the first principal component of the change in loans/assets for country national banks alongside the average values for the eight southern states. In the South, loan demand peaked during the fall harvest and coincided with the national peak, as reflected in the seasonal pattern of the first principal component. Figure 9 plots the first principal component of the change in reserves/assets along with the averages for the eight southern states. Clearly, the diminished seasonal volatility evident in the principal component of changes in reserves/assets in the 1920s mirrors reduced seasonal volatility in the data for country national banks in the South.

Federal Reserve Bank Lending Patterns

The Federal Reserve Act specified that a committee, composed of the Secretaries of the Treasury and Agriculture and the Comptroller of the Currency, would set the boundaries of the Federal Reserve districts and choose the cities for Reserve Banks. Relative to the total volume of bank assets of the region, the South was awarded a disproportionate number of Reserve Banks.[11] The cotton states were divided among the Richmond, Atlanta, St. Louis, and Dallas districts.

Table 1

Seasonal Patterns in Federal Reserve Credit, 1922-31

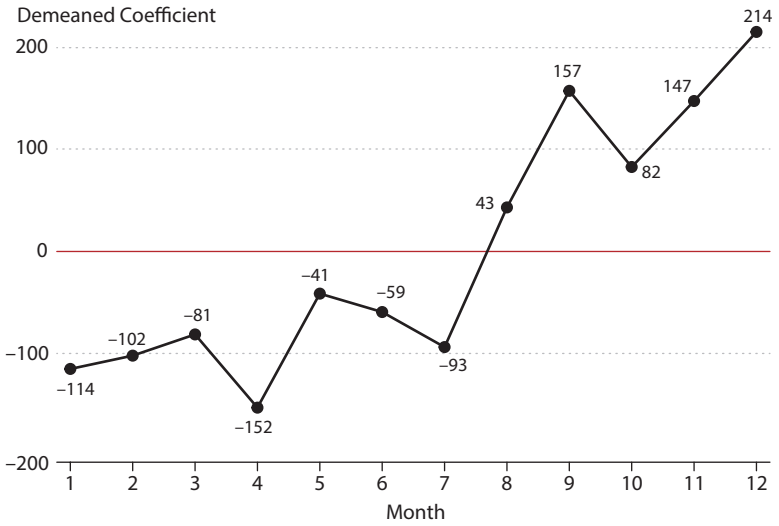
	System	Boston	New York	Philadelphia	Cleveland	Richmond	Atlanta	Chicago	St. Louis	Minneapolis	Kansas City	Dallas	San Francisco
January	179.6	2.2	–25.8	18.1	24.4*	41.0***	14.3	20.4	11.8*	14.7***	14.8**	27.4***	16.4
	(131.1)	(12.7)	(43.9)	(12.9)	(12.7)	(7.9)	(10.5)	(21.5)	(6.4)	(4.2)	(6.3)	(5.4)	(11.2)
February	192.6	2.1	–14.1	24.5*	25.6**	40.8***	7.2	17.6	14.6**	14.7***	13.1**	25.0***	21.5*
	(132.1)	(12.8)	(44.2)	(12.9)	(12.8)	(7.9)	(10.6)	(21.6)	(6.4)	(4.2)	(6.4)	(5.4)	(11.2)
March	213.3	3.3	7.6	23.3*	23.4*	42.1***	8.0	24.3	15.0**	14.7***	11.2*	19.8***	20.8*
	(129.5)	(12.6)	(43.3)	(12.7)	(12.6)	(7.8)	(10.4)	(21.2)	(6.3)	(4.2)	(6.2)	(5.3)	(11.0)
April	141.9	–13.5	–39.1	17.2	21.4*	44.9***	15.2	6.4	13.3**	19.3***	15.4**	20.3***	21.2*
	(131.0)	(12.7)	(43.8)	(12.8)	(12.7)	(7.9)	(10.5)	(21.5)	(6.4)	(4.2)	(6.3)	(5.4)	(11.2)
May	253.0**	10.6	–0.3	27.8**	30.1**	47.0***	19.0*	19.6	20.0***	17.7***	19.9***	22.5***	19.1*
	(121.4)	(11.8)	(40.6)	(11.9)	(11.8)	(7.3)	(9.7)	(19.9)	(5.9)	(3.9)	(5.9)	(5.0)	(10.3)
June	235.1*	4.1	7.8	24.8**	25.0**	46.7***	16.0	21.0	20.1***	16.6***	13.7*	25.3***	13.8
	(122.5)	(11.9)	(41.0)	(12.0)	(11.9)	(7.4)	(9.8)	(20.1)	(6.0)	(3.9)	(5.9)	(5.0)	(10.4)
July	200.9*	–2.2	–18.3	23.2*	20.7*	45.0***	17.3*	17.9	18.6***	17.4***	13.4**	26.3***	21.7**
	(120.6)	(11.7)	(40.3)	(11.8)	(11.7)	(7.2)	(9.6)	(19.8)	(5.9)	(3.9)	(5.8)	(5.0)	(10.3)
August	337.0***	10.2	28.5	32.3***	29.8***	49.2***	27.7***	26.2	28.4***	21.2***	15.4***	34.2***	33.9***
	(113.4)	(11.0)	(37.9)	(11.1)	(11.0)	(6.8)	(9.1)	(18.6)	(5.5)	(3.6)	(5.5)	(4.7)	(9.7)
September	450.7***	17.1	83.3**	33.8***	39.0***	54.4***	33.7***	35.6*	33.7***	23.2***	20.1***	33.2***	43.5***
	(115.0)	(11.2)	(38.5)	(11.3)	(11.2)	(6.9)	(9.2)	(18.8)	(5.6)	(3.7)	(5.5)	(4.7)	(9.8)
October	376.4***	2.5	33.1	28.6**	44.9***	53.1***	31.2***	40.3*	27.5***	21.0***	27.1***	33.8***	33.2***
	(126.3)	(12.3)	(42.3)	(12.4)	(12.3)	(7.6)	(10.1)	(20.7)	(6.2)	(4.1)	(6.1)	(5.2)	(10.8)
November	440.7***	26.0**	41.5	32.4***	54.1***	51.4***	37.5***	52.5**	24.4***	18.3***	29.6***	33.3***	39.7***
	(125.5)	(12.2)	(42.0)	(12.3)	(12.2)	(7.5)	(10.0)	(20.5)	(6.1)	(4.0)	(6.0)	(5.2)	(10.7)
December	508.1***	38.5***	116.6***	36.5***	60.0***	50.4***	26.2**	61.9***	21.2***	17.3***	20.5***	30.1***	28.8**
	(129.2)	(12.5)	(43.2)	(12.7)	(12.6)	(7.8)	(10.3)	(21.2)	(6.3)	(4.1)	(6.2)	(5.3)	(11.0)
Industrial Production	6.6***	0.8***	1.9***	0.6***	0.5***	0.2**	0.4***	1.0***	0.3***	0.1**	0.2***	0.1***	0.6***
	(1.1)	(0.1)	(0.4)	(0.1)	(0.1)	(0.1)	(0.1)	(0.2)	(0.1)	(0.0)	(0.1)	(0.0)	(0.1)
Time Trend Jan 1922 to Aug 1931	–3.0***	–0.3***	–0.2	–0.4***	–0.2***	–0.2***	–0.1**	–0.5***	–0.2***	–0.1***	–0.1*	–0.2***	–0.4***
	(0.8)	(0.1)	(0.3)	(0.1)	(0.1)	(0.0)	(0.1)	(0.1)	(0.0)	(0.0)	(0.0)	(0.0)	(0.1)
Obs.	116	116	116	116	116	116	116	116	116	116	116	116	116
Adj. R^2	0.91	0.88	0.86	0.88	0.90	0.91	0.87	0.86	0.90	0.86	0.86	0.89	0.91

NOTE: The dependent variable in each regression is the sum of Federal Reserve discount window loans and bankers' acceptance holdings for the System as a whole or indicated Federal Reserve district. Standard errors are in parentheses. ***, **, and * indicate statistical significance at the 1, 5, and 10 percent levels, respectively.

SOURCE: Federal Reserve Credit: Board of Governors of the Federal Reserve System (1943). Industrial Production: Miron and Romer (1989).

Figure 10

Seasonal Patterns in Federal Reserve Credit, 1922-31



NOTE: The figure plots the demeaned coefficients on monthly dummy variables from a regression of Federal Reserve credit (sum of discount window loans and bankers' acceptance holdings) on the dummy variables, an index of industrial production (relative to its level in January 1915), and a time trend.

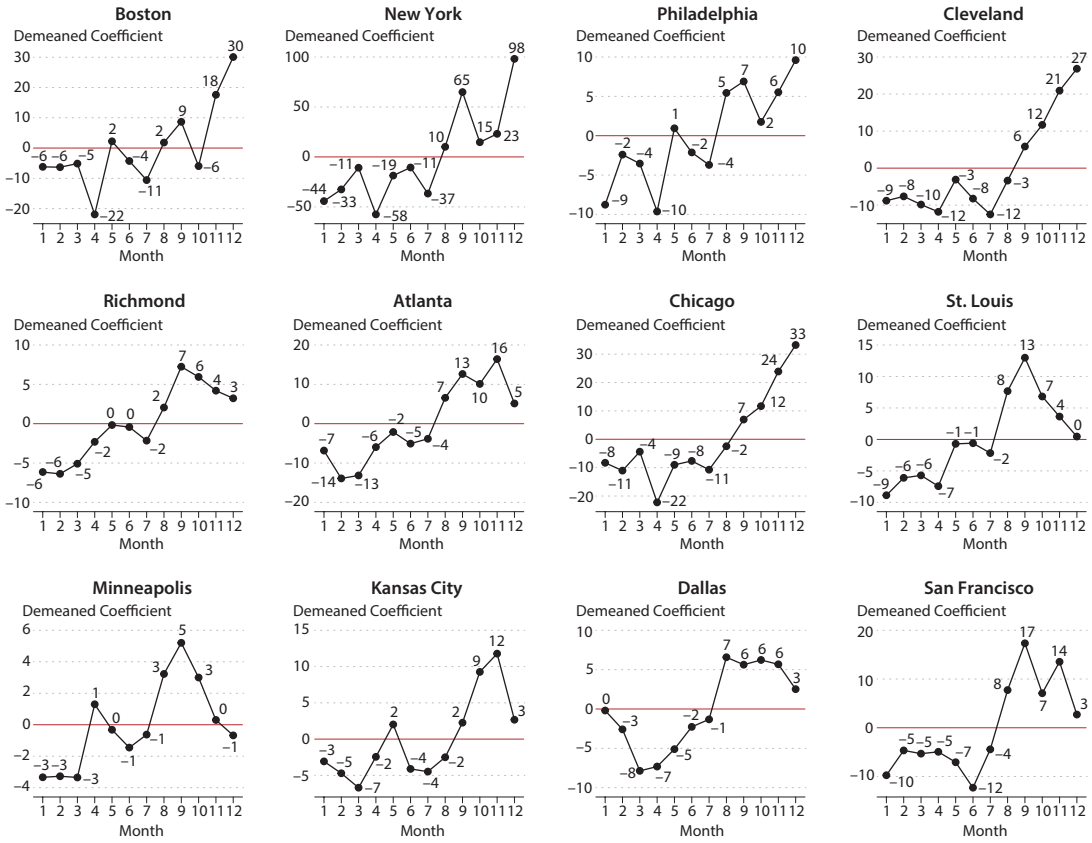
SOURCE: Federal Reserve Credit: Board of Governors of the Federal Reserve System (1943). Industrial Production: Miron and Romer (1989).

If the Federal Reserve Banks provided the seasonal accommodation that the Fed's founders intended, we would expect to find differences in the seasonal patterns of loans supplied by individual Reserve Banks reflected in differences in the timing of seasonal demands for loans and currency in the different regions of the country. The Federal Reserve Act authorized the Reserve Banks to rediscount short-term commercial and agricultural paper (i.e., bank loans) and to purchase bankers' acceptances (also known as "bills"). In practice, each Reserve Bank set a discount rate and schedule of bill buying rates and purchased the paper offered to them at those rates.[12] Discount and bill buying rates were not changed frequently or on a seasonal schedule. Hence, seasonal fluctuations in Federal Reserve credit outstanding were largely driven by demand (Wheelock, 1992).

Table 1 reports coefficients from regressions of Federal Reserve credit (i.e., rediscounts and purchases of bankers' acceptances) on 12 monthly dummy variables, a national index of industrial production (measured relative to January 1915) to capture business cycle effects on the demand for Federal Reserve credit, and a time trend. We estimate the regression for each Reserve Bank separately and also for the System as a whole. The coefficients from the System regression are reported in the first column and those of the individual Reserve Banks in the other columns. We estimate the regressions using monthly data for January 1922 to August 1931. We omit the early years of the System through 1921 to avoid the war years, when the Fed offered preferential discount rates on loans secured by U.S. government securities and

Figure 11

Seasonal Patterns in Federal Reserve Credit by Federal Reserve District, 1922-31



NOTE: Each panel plots for the indicated Federal Reserve district the demeaned coefficients on monthly dummy variables from a regression of Federal Reserve credit (the sum of discount window loans and bankers' acceptance holdings) on the dummy variables, an index of industrial production (relative to its level in January 1915), and a time trend.

SOURCE: Federal Reserve Credit: Board of Governors of the Federal Reserve System (1943). Industrial Production: Miron and Romer (1989).

those loans dominated the Fed's lending. We end the sample in August 1931 to avoid the effects of a major banking panic that began in September 1931.

As shown in Table 1, the coefficients on the monthly dummies indicate notable seasonal patterns in Fed lending and tend to be larger toward the end of the year. The coefficient on the December dummy variable is the largest of the monthly dummy coefficients for the System as a whole as well as for the Boston, New York, Philadelphia, Cleveland, and Chicago Reserve Banks. For the other Reserve Banks, the largest monthly coefficient is usually for an autumn month—September, October, or November.

Figure 10 plots the difference of each monthly dummy coefficient from the mean of the 12 monthly coefficients for the Federal Reserve System as a whole. At the System level, Federal Reserve credit typically exceeded the mean level during September to December and was below the mean during January to August. Seasonal peaks occurred in September and December.

Figure 11 presents similar plots for each Reserve Bank, again illustrating that Federal Reserve loans tended to peak in December in the Boston, New York, Philadelphia, Cleveland, and Chicago districts, but in earlier months in all other districts. The four districts in the Northeast had large manufacturing and financial sectors and relatively small agricultural sectors. In those districts, strong demands for currency and reserves were likely dominated by the December holiday shopping season and end-of-year payments for settlement of business and financial transactions. The Chicago District was more diverse, having large manufacturing and agricultural sectors as well as a large financial center. By contrast, the remaining districts had relatively larger farming sectors—and greater demand for harvest-related funds. Hence, the differences in the seasonal lending patterns of the Reserve Banks are broadly consistent with the differences in the timing of local demands for money and bank loans observed by Kemmerer (1910) and are evidence that the Federal Reserve was likely responsible for the substantial reduction in seasonal liquidity pressures on bank balance sheets and money markets.

IMPACT ON BANK LIQUIDITY AND THE SIZE OF THE INTERBANK SYSTEM

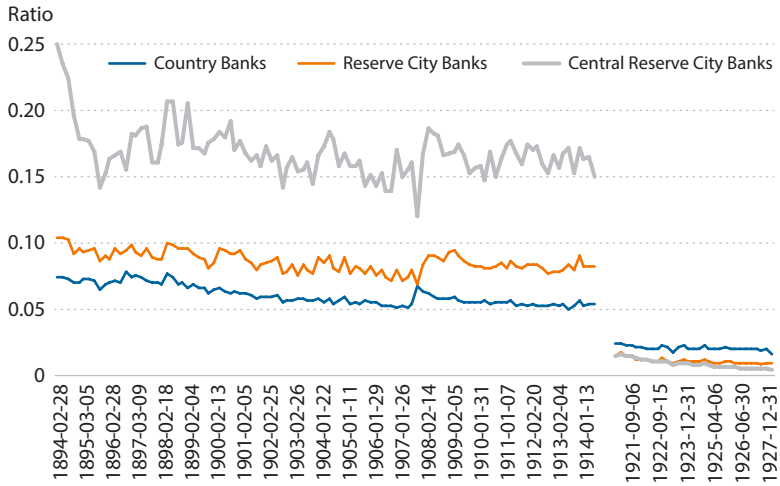
The reductions in the seasonal variation of flows of interbank deposits and bank reserves/assets ratios after the founding of the Fed, as well as the size and seasonal timing of the Fed's lending, is strong evidence that the Fed accomplished the founders' goal of providing seasonal liquidity to the banking system. The Fed's founders also intended the System to largely supplant the existing interbank network of correspondent relationships. All national banks were required to join the Federal Reserve System (membership was optional for banks with state charters). Member banks were required to purchase stock in their local Federal Reserve Bank and to maintain a deposit with the Reserve Bank to satisfy their statutory reserve requirement. After a three-year transition period, deposits held with national banks in reserve cities or central reserve cities no longer counted toward a member bank's reserve requirement.

In addition to providing a ready source of liquidity to the banking system, the Federal Reserve was also intended to make the U.S. payments system more efficient. The Reserve Banks offered check clearing and other payments services to their member banks. The Fed quickly acquired a large share of the interregional check clearing market and seems to have made the clearing system more efficient (Gilbert 1998, 2000).

By providing a ready source of currency and making the payments system more efficient, the Federal Reserve likely enabled banks to hold a smaller share of their total assets in the form of vault cash. Further, an amendment to the Federal Reserve Act in 1917 specified that a member bank's full reserve requirement must be held as a deposit with a Federal Reserve Bank; member banks no longer had the option of using vault cash to meet their statutory requirement. Not surprisingly, national banks maintained substantially lower cash-to-assets ratios during the 1920s than they had during the 20 years before the Fed's founding in 1914. Figure 12 plots aggregated cash/assets for country national banks, national banks in 18 reserve cities, and national banks in the three central reserve cities on each reporting date during 1894-1914 and 1921-28.[13] Whereas national banks held vault cash equal to between 5 and 20

Figure 12

Vault Cash as a Share of Total Bank Assets, 1894-1914 and 1921-28



NOTE: The figure plots the ratio of vault cash to total assets for country national banks, reserve city banks, and central reserve city banks. Data for country banks are aggregated across all U.S. states; data for reserve city banks are aggregated across 18 long-time reserve cities; and data for central reserve cities are aggregated across the three central reserve cities (New York City, Chicago, and St. Louis). St. Louis is treated as a central reserve city throughout the period even though its designation was changed to reserve city in 1922.

SOURCE: National bank data through 1910: Weber (2000). National bank data after 1910: U.S. Office of the Comptroller of the Currency (1917-28).

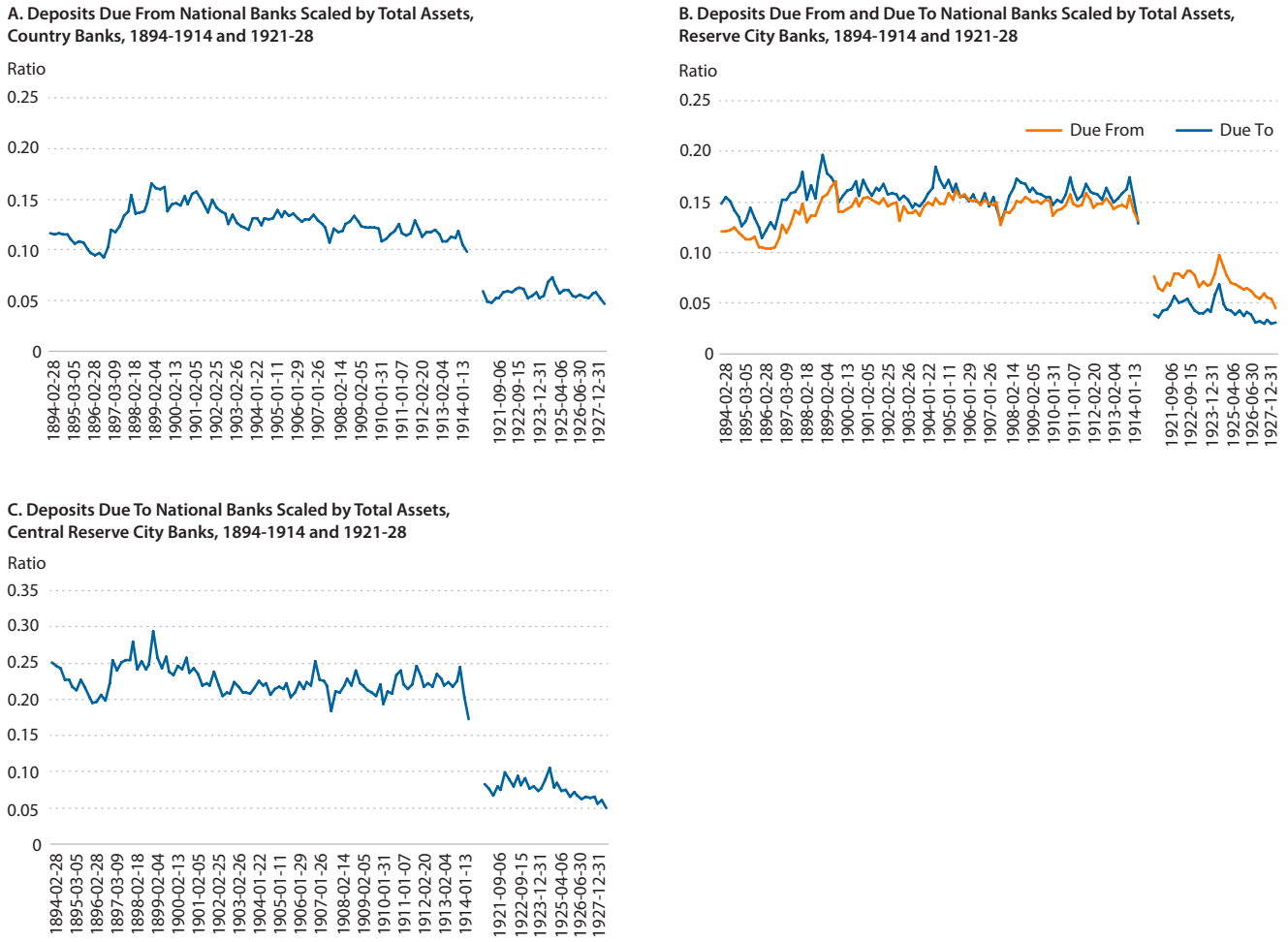
percent of their total assets during the 20 years before the founding of the Fed, during the 1920s, vault cash comprised 2.5 percent or less of their total assets.

National banks also held much lower levels of interbank deposits during the 1920s, as shown in Figure 13. The share of total country bank assets held in the form of deposits due from other national banks, shown in Panel A, declined from an average of 12.6 percent during 1894-1914 to just 5.7 percent during 1921-28. For national banks in 18 reserve cities, the percentage of assets held as deposits *due from* other national banks, shown in Panel B, fell from an average of 15.6 percent during 1894-1914 to just 4.3 percent during 1921-28. At the same time, reserve city banks saw a decline in the volume of their deposits *due to* other national banks, from 14.2 percent to 6.9 percent (relative to their total assets).[14] Similarly, for central reserve city banks, deposits due to other national banks, shown in Panel C, declined from 22.5 percent to 7.5 percent (relative to total central reserve city bank assets). Thus, the Fed's founders appear to have accomplished their goal of shrinking the private interbank market.[15]

Figure 14 presents additional information indicating that national banks were generally less liquid in the 1920s than they had been before the founding of the Fed. For each class of national bank, the figure plots a measure of reserves consisting of vault cash, cash items in the process of collection, and deposits with reserve agents. As Figure 14 shows, national banks held substantially lower reserves/assets ratios during the 1920s than they had during the 20

Figure 13

Interbank Deposits, 1894-1914 and 1921-28



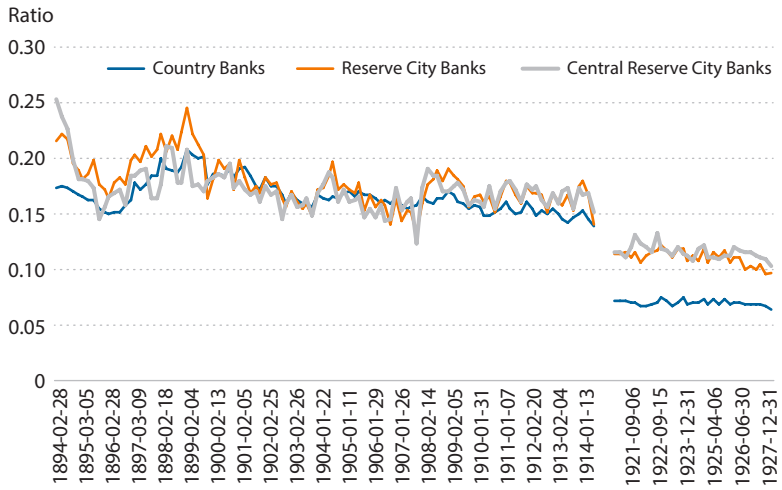
NOTE: Data for country banks are aggregated across all U.S. states; data for reserve city banks are aggregated across 18 long-time reserve cities; and data for central reserve cities are aggregated across the three central reserve cities (New York City, Chicago, and St. Louis). St. Louis is treated as a central reserve city throughout the period even though its designation was changed to reserve city in 1922.

SOURCE: National bank data through 1910: Weber (2000). National bank data after 1910: U.S. Office of the Comptroller of the Currency (1917-28).

years prior to the Fed's establishment. A portion of the decline was undoubtedly due to reductions in the required reserve ratios applied to national banks in 1913 and 1917. However, Carlson and Wheelock (2016b) show that, as a share of total assets, national banks held somewhat lower levels of reserves in excess of legal requirements in the 1920s than they had prior to 1914.[16] Conceivably, banks responded to the presence of the Fed—specifically to its promise to supply liquidity as needed—by holding less liquidity themselves. In effect, by enabling or encouraging banks to devote larger shares of their balance sheets to relatively illiquid but higher-earning assets, the presence of the Fed may have increased the need for a lender of

Figure 14

Reserve Ratios, 1894-1914 and 1921-28



NOTE: The figure plots the ratio of reserves to total assets for banks in each tier group, where reserves are the sum of vault cash, cash items in the process of collection, and deposits with reserve agents. Data for country banks are aggregated across all U.S. states; data for reserve city banks are aggregated across 18 long-time reserve cities; and data for central reserve cities are aggregated across the three central reserve cities (New York City, Chicago, and St. Louis). St. Louis is treated as a central reserve city throughout the period even though its designation was changed to reserve city in 1922.

SOURCE: National bank data through 1910: Weber (2000). National bank data after 1910: U.S. Office of the Comptroller of the Currency (1917-28).

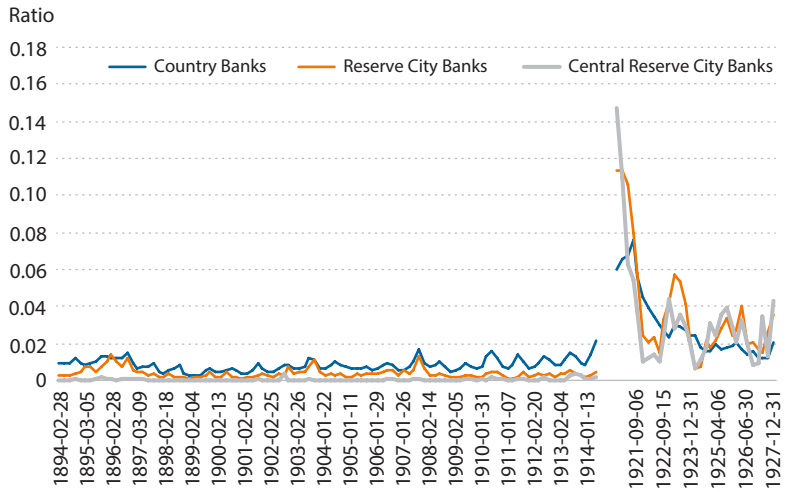
last resort. Ultimately, therefore, whether the Fed would be successful in preventing banking panics, and thus achieving the main objective of the System's founders, would depend on how well the Fed performed as a lender of last resort.

Impacts on Bank Liabilities, Capital, and Lending

As noted previously, before the founding of the Fed, country banks often borrowed for short periods from their correspondents, particularly at times of the year when local demands for currency and loans were at their highest. Reserve city banks generally borrowed less frequently than country banks, and central reserve city banks hardly borrowed at all.[17] Figure 15 plots the short-term borrowing of banks in each class of national banks (scaled by their total assets). Short-term borrowing was never large in the aggregate, even for country banks. Over the 20 years prior to the establishment of the Fed, borrowing approached 2 percent of country bank assets only during banking panics in 1907 and 1914. Banks tended to borrow more during the 1920s, mostly at the Fed's discount window.[18] During World War I and for several months afterward, the Fed offered a preferential discount rate on loans secured by U.S. government bonds, and banks borrowed heavily from the Fed. The preferential rate was eliminated, and Reserve Banks hiked their discount rates sharply in 1920-21. Although the level of borrowing then fell, relative to total bank assets, borrowing was considerably higher

Figure 15

Short-Term Borrowing Scaled by Total Assets, 1894-1914 and 1921-28



NOTE: The figure plots the ratio of short-term borrowing to total assets for banks in each tier group. Data for country banks are aggregated across all U.S. states; data for reserve city banks are aggregated across 18 long-time reserve cities; and data for central reserve cities are aggregated across the three central reserve cities (New York City, Chicago, and St. Louis). St. Louis is treated as a central reserve city throughout the period even though its designation was changed to reserve city in 1922.

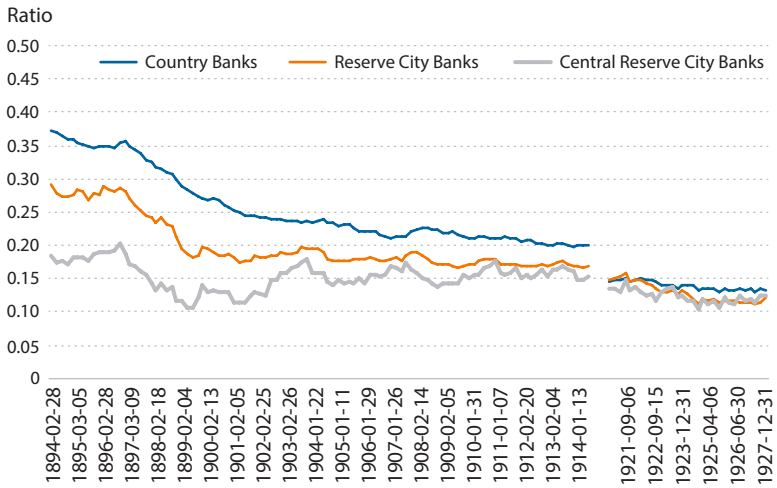
SOURCE: National bank data through 1910: Weber (2000). National bank data after 1910: U.S. Office of the Comptroller of the Currency (1917-28).

during the remainder of the 1920s, especially for reserve city and central reserve city banks, than it had been before 1914.

While the Federal Reserve Act changed the level and structure of reserve requirements imposed on banks that became members of the Federal Reserve System, the act did not change their capital requirements (except by requiring state-chartered banks that joined the System to conform with national bank requirements). At the time, federal statutes specified minimum capital levels for national banks based on the size of the city in which a bank was located, but did not specify minimum capital-to-assets ratios (or, equivalently, maximum leverage ratios).[19] Many states set even lower minimums for their state-chartered banks, which apparently was one reason why few state-chartered banks elected initially to join the Federal Reserve System (White, 1983, pp. 14-23). As shown in Figure 16, the aggregate capital/assets ratio of country national banks declined from over 0.35 (35 percent) in 1896 to 0.20 in 1914. By 1921, the ratio had fallen to 0.15, or less than half the level of 1896. The aggregate capital/assets ratios for reserve city and central reserve city banks also declined over time and ranged between 0.10 and 0.15 during the 1920s. The declines for reserve city and central reserve city banks were less dramatic than those for country banks, however, with less of a discrete decline between the pre-Fed period and the 1920s. Hence, it is unclear whether the Fed's founding explains the lower national bank capital/assets ratios in the 1920s.

Figure 16

Bank Equity Scaled by Total Assets, 1894-1914 and 1921-28



NOTE: The figure plots the ratio of equity (paid in capital, surplus, and undistributed profits) to total assets for banks in each tier group. Data for country banks are aggregated across all U.S. states; data for reserve city banks are aggregated across 18 long-time reserve cities; and data for central reserve cities are aggregated across the three central reserve cities (New York City, Chicago, and St. Louis). St. Louis is treated as a central reserve city throughout the period even though its designation was changed to reserve city in 1922.

SOURCE: National bank data through 1910: Weber (2000). National bank data after 1910: U.S. Office of the Comptroller of the Currency (1917-28).

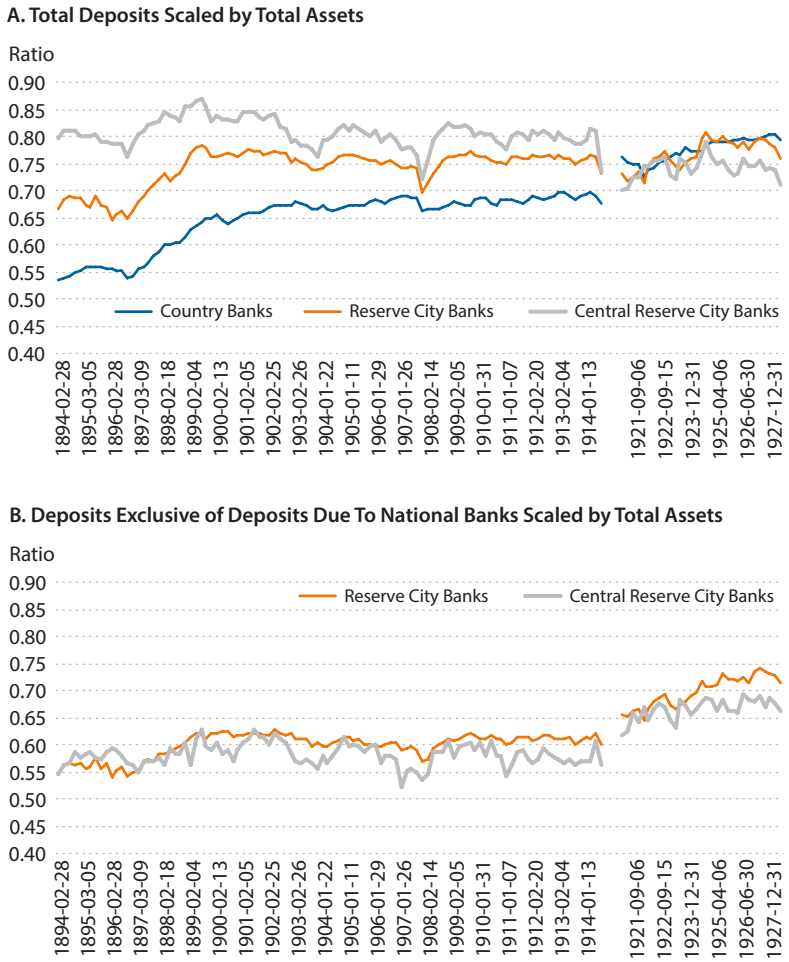
Whereas bank capital/assets ratios declined between the pre-Fed period and the 1920s, the deposits/assets ratios rose, at least among country national banks, as shown in Panel A of Figure 17.[20] Reserve city banks experienced a much smaller increase in deposits/assets ratios, while central reserve city banks saw a decline. However, both reserve city and central reserve city banks saw substantial increases in deposits/assets ratios exclusive of their deposits due to national banks, as shown in Panel B. The Fed's promise to provide the liquidity needed to meet their customers' demand for cash, as well as improved efficiency of check collection (Gilbert 1998, 2000), probably made banks more willing to offer demand deposits and led to higher deposits/assets ratios.

Finally, Figure 18 plots data on loans/assets ratios for each of the three classes of national banks. There are no obvious patterns. Loans generally ranged between 50 and 60 percent of total assets for each class during both the pre-Fed period and 1920s. The Fed's founding did not produce an obvious change in the shares of national bank assets composed of loans.[21] Instead, national banks devoted larger shares of their portfolios to various types of securities. During the 20 years before the Fed was established, securities typically comprised about 15 percent of total national bank assets. However, in the 1920s, they usually exceeded 20 percent of assets.[22]

In sum, the presence of the Fed had its largest impact on banks' demand for liquid assets and interbank balances. Relative to their total assets, national banks held substantially lower

Figure 17

Deposits Scaled by Total Assets, 1894-1914 and 1921-28



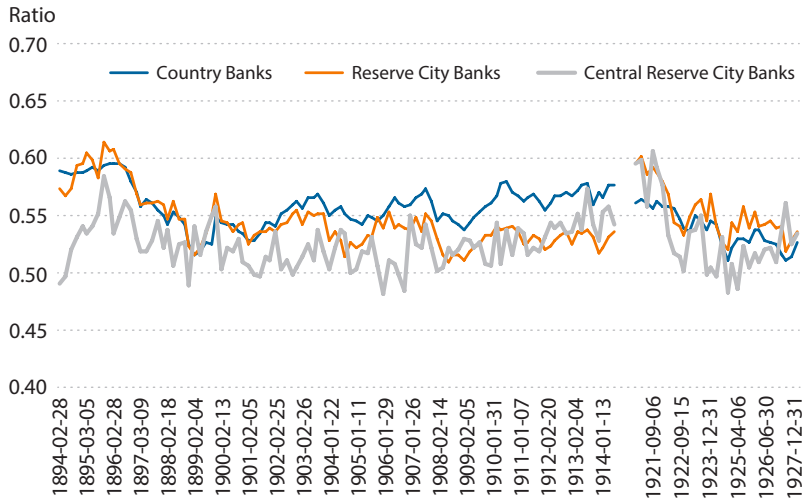
NOTE: The figure plots the ratio of total deposits to total assets for banks in each tier group (Panel A) and the ratio of deposits exclusive of deposits due to other national banks to total assets for reserve city and central reserve city banks (Panel B). Data for country banks are aggregated across all U.S. states; data for reserve city banks are aggregated across 18 long-time reserve cities; and data for central reserve cities are aggregated across the three central reserve cities (New York City, Chicago, and St. Louis). St. Louis is treated as a central reserve city throughout the period even though its designation was changed to reserve city in 1922.

SOURCE: National bank data through 1910: Weber (2000). National bank data after 1910: U.S. Office of the Comptroller of the Currency (1917-28).

levels of reserves and interbank deposits in the 1920s than they had before the Fed's founding. National banks also had higher deposits/assets ratios (excluding interbank deposits) and lower capital/assets ratios during the 1920s. Although we have not formally tested whether the presence of the Fed caused or contributed to these changes in national bank portfolios, they are consistent with the idea that the Fed's presence gave banks confidence to assume greater liquidity risk and leverage.

Figure 18

Total Loans Scaled by Total Assets, 1894-1914 and 1921-28



NOTE: The figure plots the ratio of loans to total assets for banks in each tier group. Data for country banks are aggregated across all U.S. states; data for reserve city banks are aggregated across 18 long-time reserve cities; and data for central reserve cities are aggregated across the three central reserve cities (New York City, Chicago, and St. Louis). St. Louis is treated as a central reserve city throughout the period even though its designation was changed to reserve city in 1922.

SOURCE: National bank data through 1910: Weber (2000). National bank data after 1910: U.S. Office of the Comptroller of the Currency (1917-28).

CONCLUSION

The establishment of the Federal Reserve System greatly altered the operating environment for U.S. banks. The Federal Reserve Act changed the structure of reserve requirements imposed on banks that joined the Federal Reserve System, created a new currency and payments provider, and introduced a lender of last resort. The System's founders sought to protect the banking system from banking panics, which they attributed largely to the nation's "inelastic currency" and the fragility of the interbank system. Accordingly, the Fed was designed to hold the reserves of its member banks, supply currency and reserves to meet recurring and extraordinary demands, and provide check clearing and other payments services.

Our analysis of national bank balance sheet data provides new evidence of reduced seasonal liquidity pressures on the banking system after the founding of the Fed. Seasonal movements in interbank deposits and country national bank borrowing and reserve/asset ratios were considerably smaller during the 1920s than they had been over the two decades preceding the Fed's establishment. At the same time, Federal Reserve lending was markedly seasonal, and seasonal peaks in lending by the individual Reserve Banks coincided with seasonal peaks in liquidity demands in their districts. Thus, the evidence from national bank data and Federal Reserve lending patterns, coupled with evidence of reduced seasonal pressures in money mar-

kets, suggests strongly that the System's founders accomplished their goal of accommodating seasonal liquidity demands.

In addition to exhibiting reduced seasonality, the correspondent deposits and liquid reserves of national banks were also much smaller relative to total bank assets during the 1920s than they had been before the founding of the Fed. Banks also relied more heavily on deposits for funding and had somewhat lower average equity/asset ratios in the 1920s. By encouraging banks to be less liquid themselves, it was especially important that the Fed perform as the System's founders had intended and provide a liquidity backstop in the event of a panic. That is, the presence of a lender of last resort encouraged banks to act in ways that increased the likelihood that the support of a lender of last resort would be needed.

The absence of significant banking panics throughout the 1920s, despite large numbers of bank failures, was viewed as important evidence that the problem of banking panics had indeed been solved. Congress awarded this apparent success in 1927 by making the charters of the 12 Federal Reserve Banks, which the Federal Reserve Act had set at 20 years, permanent. Moreover, quick action by the Federal Reserve Bank of Atlanta to stem a local banking panic in Florida in 1929 suggested that the Fed had the tools to prevent major panics of the sort that had plagued the banking system before the Fed's founding (Carlson, Mitchener, and Richardson, 2011).

Banking panics returned with a vengeance in the early 1930s, however, and the Federal Reserve proved incapable or unwilling to perform as lender of last resort on a national scale. The banking panics of the Great Depression showed that elimination of seasonal liquidity pressures was not sufficient to prevent banking panics, and the variety of responses by the individual Reserve Banks suggest that the Fed's structure inhibited a response to a nationwide crisis.[23] Thus, while the Fed's founders achieved their proximate goals, the banking panics of the Great Depression showed that further reforms were needed to achieve the founders' ultimate goal of ending banking panics. ■

NOTES

[1] See, for instance, Yellen (2016) and the Basel Committee on Banking Supervision (2016).

[2] For example, Bianchi (2016) examines the trade-offs associated with financial bailouts in a quantitative equilibrium model of financial crises. By easing balance sheet constraints, lender of last resort interventions can mitigate the severity of a crisis-induced recession, but bailouts generally lead to greater risk-taking and thus make economies more vulnerable to crises. However, Bianchi finds that economies are significantly less exposed to crises when lender of last resort actions are systemic and broad-based than when they are idiosyncratic and targeted.

[3] Bernstein, Hughson, and Weidenmier (2010) find that the volatilities of both interest rates and stock prices during the months of September and October—the two months of peak volatility during the late-nineteenth and early-twentieth centuries—were significantly lower after the Fed's founding. However, if years of banking panics are omitted from the sample, then volatilities were similar between the pre-Fed and post-1914 years. They argue that the Fed's primary effect was to reduce liquidity risk in years when there was a business cycle turning point and financial crisis. Further, there is some debate about whether the Fed was responsible for a reduction in seasonal interest rate volatility that occurred in many countries around 1914 (see Wheelock, 1992, and the references therein).

[4] A "correspondent" bank is one with whom another bank, known as the "respondent," maintains a deposit or receives services.

[5] Calomiris and Gorton (1991) contend that the Fed's founders misunderstood (or mischaracterized) the causes of banking panics. Rather than being caused by random, unexpected, or seasonal withdrawals of cash from banks, Calomiris and Gorton (1991) argue that panics occurred when adverse news about the macroeconomy caused bank depositors to revise their perceptions about the risk of bank debt. Because depositors are unable to distinguish among individual bank risks, they withdrew a large volume of deposits from all banks. Calomiris and Gorton (1991) find little association between seasonal liquidity demands and the incidence of panics.

[6] Banks chartered by state governments were also connected to the interbank system. State banks were subject to state-specific reserve requirements. Many state banks maintained deposits with national banks to satisfy reserve requirements or to obtain services, and some national banks held deposits with state banks. Consistent balance sheet data are not available for state-chartered banks, however, and this article is based solely on data for national banks.

[7] During World War I and for several months afterward, the Federal Reserve Banks offered low discount rates on loans secured by U.S. government bonds, and member banks borrowed heavily from the Fed. The Reserve Banks increased their discount rates sharply in 1920 and borrowing dropped precipitously. By 1922, discount window lending and bank balance sheets displayed regular seasonal patterns and levels. Hence, to assess the long-run impact of the Fed, we focus on the period 1922-28 (we end our comparison period in 1928 both to avoid distortions caused by the Great Depression and because of changes in how national bank balance sheet data were reported). This article uses national bank balance sheet data for country banks aggregated at the level of states and for individual reserve and central reserve cities, which are available for 1880-1910 from the Federal Reserve Bank of Minneapolis digital archive (http://cdm16030.contentdm.oclc.org/cdm/singleitem/collection/p16030coll4/id/6/rec/2). For other years, the article uses data from annual reports of the Comptroller of the Currency, which are available from the Federal Reserve Bank of St. Louis (https://fraser.stlouisfed.org/title/56).

[8] For each state, we calculate the net deposits due from reserve agents and other national banks as the difference between deposits *due from* banks and deposits *due to* banks. For the 1920s, deposits due from agents represent deposits that banks held with the Federal Reserve.

[9] Changes in the frequency and timing of reporting dates over the sample period make it impossible to draw firm conclusions about seasonal patterns in bank balance sheets. See Carlson and Wheelock (2016a,b) for more discussion and evidence about changes in seasonal pressures on the interbank system with the establishment of the Fed.

[10] Short-term borrowing consisted of "bills rediscounted," that is, loans sold with recourse, and "bills payable," that is, promissory notes of the borrowing bank, and includes borrowing from Federal Reserve Banks in the 1920s.

[11] See Odell and Weiman (1998) on the selection of Atlanta and Dallas for Reserve Banks, and Jaremski and Wheelock (2017) and Wheelock (2015) for discussion and analysis of the selection of Reserve Bank cities and branches and district boundaries.

[12] At the time, short-term commercial and agricultural loans were typically made on a discount basis. A member bank could obtain cash or reserve deposits from the Fed by rediscounting a loan—in effect selling the loan to the Federal Reserve Bank at the Reserve Bank's discount rate (the member bank remained liable for payment of the loan if the borrower defaulted). An amendment to the Federal Reserve Act in 1916 permitted the Fed to offer "advances," or loans of up to 15 days, to member banks on their own promissory notes, with loans or securities that were eligible for rediscounting serving as collateral. Advances have been the main form of discount window loans since the Great Depression. See Hackley (1973) for a history of the Federal Reserve lending function and Carlson and Duygan-Bump (2016) for information about the mechanics and volumes of Federal Reserve discount window lending and purchases of bankers' acceptances during the 1920s.

[13] The number of designated reserve cities rose over time. For consistency, the information and figures in this article are based on data for 18 cities that were reserve cities throughout the sample period. Similarly, for consistency, we treat St. Louis as a central reserve city throughout the 1920s, even though it was downgraded to reserve city status in 1922.

[14] Recall that deposits "due to" banks are deposits that other banks hold with a correspondent bank and are thus liabilities of the correspondent bank. By contrast, deposits "due from" banks are deposits that a given bank holds with correspondents and, thus, are assets of that bank. For consistency in assessing changes in the importance of interbank deposits, we scale deposits due to banks by total assets, even though such deposits are liabilities of the bank in which they are held.

[15] Although the relative size of the interbank system fell after the Fed was established, it did not disappear altogether. Watkins (1929) studied how the Fed's establishment affected the demand for interbank balances and concluded that national banks continued to hold some deposits with correspondents for interest income (unlike the Fed, correspondent banks typically paid interest on interbank deposits) and because correspondent banks would invest surplus funds in securities markets on behalf of their respondents.

[16] During 1894-1914, the excess reserves/assets ratios of country, reserve city, and central reserve city banks averaged 0.11, 0.12, and 0.08, respectively. For 1921-28, the corresponding averages were 0.09, 0.08, and 0.06.

[17] Although central reserve city banks rarely borrowed, during the major banking panics, some of those banks borrowed heavily from their local clearinghouses. See Gorton (1985). The data shown in Figure 15 do not include loans from clearinghouses.

[18] National bank reports do not indicate the source of loans. However, Federal Reserve data indicate that loans from the Fed comprised a large share of total borrowing by national banks (and other Fed member banks) during the 1920s.

[19] The Comptroller of the Currency recommended legislation to prohibit banks from having total deposits in excess of 10 times their capital (U.S. Department of the Treasury, 1914, p. 21). However, Congress did not mandate maximum leverage ratios (or, equivalently, minimum capital/assets or capital-to-deposit ratios) until the 1980s (Federal Deposit Insurance Corporation, 2003).

[20] The total deposits data underlying Figure 17 include deposits of individuals, firms, financial institutions, and state and local governments, but do not include federal government deposits.

[21] One exception might be for country national banks in the South, where the average loans/assets ratio among the eight southern states substantially exceeded the U.S. average in the 1920s. However, the average among the southern states also exceeded the U.S. average during 1910-14 and exhibited considerable intra-year volatility, making it difficult to attribute changes in the loans/assets ratio after 1914 to the presence of the Fed.

[22] Securities issued by the federal government comprised about 8 percent of total national bank assets before 1914 and 10 percent during the 1920s; hence, private securities also comprised larger shares of bank assets in the 1920s.

[23] Richardson and Troost (2009) show that aggressive lender of last resort action by the Federal Reserve Bank of Atlanta enabled the banks and economy of the southern half of Mississippi to fare better during the crisis than did the banks and economy of the northern half of Mississippi, which was in the district of the less aggressive Federal Reserve Bank of St. Louis. Bordo and Wheelock (2013) attribute the Fed's failure to respond aggressively to the banking panics of the Great Depression to restrictions on Fed lending imposed by the Federal Reserve Act as well as to the Fed's decentralized structure.

REFERENCES

Basel Committee on Banking Supervision. *Eleventh Progress Report on Adoption of the Basel Regulatory Framework. Bank for International Settlements*, October 2016; http://www.bis.org/bcbs/publ/d388.pdf.

Bernstein, Asaf; Hughson, Eric and Weidenmier, Marc D. "Identifying the Effects of a Lender of Last Resort on Financial Markets: Lessons from the Founding of the Fed." *Journal of Financial Economics*, October 2010, *98*(1), pp. 40-53; https://doi.org/10.1016/j.jfineco.2010.04.001.

Bianchi, Javier. "Efficient Bailouts?" *American Economic Review*, December 2016, *106*(12), pp. 3607-59; https://doi.org/10.1257/aer.20121524.

Board of Governors of the Federal Reserve System. *Banking and Monetary Statistics, 1914-1941*. 1943; https://fraser.stlouisfed.org/title/38, accessed July 12, 2017.

Bordo, Michael D. and Wheelock, David C. "The Promise and Performance of the Federal Reserve as Lender of Last Resort, 1914-1933," in Michael D. Bordo and William Roberds, eds., *A Return to Jekyll Island: The Origins, History, and Future of the Federal Reserve*. Cambridge and New York: Cambridge University Press, 2013, pp. 59-98; https://doi.org/10.1017/CBO9781139005166.004.

Calomiris, Charles W. and Carlson, Mark A. "Interbank Networks in the National Banking Era: Their Purpose and Their Role in the Panic of 1893." *Journal of Financial Economics*, September 2017, *125*(3), pp. 434-53; https://doi.org/10.1016/j.jfineco.2017.06.007.

Calomiris, Charles W. and Gorton, Gary. "The Origin of Banking Panics: Models, Facts, and Bank Regulation," in R. Glenn Hubbard, ed., *Financial Markets and Financial Crises*. Chicago: University of Chicago Press, 1991, pp. 109-173.

Carlson, Mark and Duygan-Bump, Burcu. "The Tools and Transmission of Federal Reserve Monetary Policy in the 1920s." Board of Governors of the Federal Reserve System *FEDS Notes*, November 22, 2016; https://www.federalreserve.gov/econresdata/notes/feds-notes/2016/tools-and-transmission-of-federal-reserve-monetary-policy-in-the-1920s-20161122.html.

Carlson, Mark; Mitchener, Kris J. and Richardson, Gary. "Arresting Banking Panics: Federal Reserve Liquidity Provision and the Forgotten Panic of 1929." *Journal of Political Economy*, October 2011, *119*(5), pp. 889-924; https://doi.org/10.1086/662961.

Carlson, Mark and Wheelock, David C. "Interbank Markets and Banking Crises: New Evidence on the Establishment and Impact of the Federal Reserve." *American Economic Review: Papers & Proceedings*, May 2016a, *106*(5), pp. 533-37; https://doi.org/10.1257/aer.p20161044.

Carlson, Mark A. and Wheelock, David C. "Did the Founding of the Federal Reserve Affect the Vulnerability of the Interbank System to Systemic Risk?" Working Paper No. 2016-012B, Federal Reserve Bank of St. Louis, July 2016b.

Federal Deposit Insurance Corporation. "Basel and the Evolution of Capital Regulation: Moving Forward, Looking Back." January 14, 2003; https://www.fdic.gov/bank/analytical/fyi/2003/011403fyi.html.

Friedman, Milton and Schwartz, Anna J. *A Monetary History of the United States, 1867-1960*. Princeton: Princeton University Press, 1963.

Gilbert, R. Alton. "Did the Fed's Founding Improve the Efficiency of the U.S. Payments System?" Federal Reserve Bank of St. Louis *Review*, May/June 1998, *80*(3), pp. 121-42; https://files.stlouisfed.org/files/htdocs/publications/review/98/05/9805ag.pdf.

Gilbert, R. Alton. "The Advent of the Federal Reserve and the Efficiency of the Payments System: The Collection of Checks, 1915-1930." *Explorations in Economic History*, April 2000, *37*(2), pp. 121-48; https://doi.org/10.1006/exeh.2000.0736.

Gorton, Gary. "Clearinghouses and the Origin of Central Banking in the United States." *Journal of Economic History*, June 1985, *45*(2), pp. 277-83; https://doi.org/10.1017/S0022050700033957.

Hackley, Howard H. *Lending Functions of the Federal Reserve Banks: A History*. Washington D.C.: Board of Governors of the Federal Reserve System, May 1973.

Jaremski, Matthew and Wheelock, David C. "Banker Preferences, Interbank Connections, and the Enduring Structure of the Federal Reserve System." *Explorations in Economic History*, October 2017, *66*, pp. 21-43; https://doi.org/10.1016/j.eeh.2016.08.002.

Kemmerer, Edwin W. *Seasonal Variations in the Relative Demand for Money and Capital in the United States*. National Monetary Commission. Washington DC: Government Printing Office, 1910.

Miron, Jeffrey A. "Financial Panics, the Seasonality of the Nominal Interest Rate, and the Founding of the Fed." *American Economic Review*, March 1986, *76*(1), pp. 125-40.

Miron, Jeffrey A. and Romer, Christina D. "A New Monthly Index of Industrial Production, 1884-1940." NBER working paper No. 3172, National Bureau of Economic Research, November 1989.

Odell, Kerry A. and Weiman, David F. "Metropolitan Development, Regional Financial Centers, and the Founding of the Fed in the Lower South." *Journal of Economic History*, March 1998, *58*(1), pp. 103-25; https://doi.org/10.1017/S0022050700019902.

Redenius, Scott A. and Weiman, David F. "Banking on the Periphery: The Cotton South, Systemic Seasonality, and the Limits of National Banking Reform," in Paul W. Rhode, Joshua L. Rosenbloom, and David F. Weiman, eds., *Economic Evolution and Revolution in Historical Time*. Redwood City, CA: Stanford University Press, 2011, pp. 214-42; https://doi.org/10.11126/stanford/9780804771856.003.0009.

Reserve Bank Organization Committee. *Report to the Reserve Bank Organization Committee by the Preliminary Committee on Organization*. 1914; https://fraser.stlouisfed.org/scribd/?title_id=609&filepath=/files/docs/historical/federal%20reserve%20history/Reserve%20Bank%20Organization.pdf.

Richardson, Gary and Troost, William. "Monetary Intervention Mitigated Banking Panics during the Great Depression: Quasi-experimental Evidence from a Federal Reserve District Border, 1929-1933." *Journal of Political Economy*, December 2009, *117*(6), pp. 1031-73; https://doi.org/10.1086/649603.

Sprague, O.M.W. *History of Crises under the National Banking System*. National Monetary Commission. Senate Document No. 538, 61st Congress, 2d Session. Washington DC: Government Printing Office, 1910.

U.S. Department of the Treasury. *Annual Report of the Comptroller of the Currency*. Washington DC: Government Printing Office, 1915.

U.S. Office of the Comptroller of the Currency. *Annual Report of the Comptroller of the Currency*. 1917-1928; https://fraser.stlouisfed.org/title/56, accessed July 12, 2017.

Watkins, Leonard L. *Bankers Balances: A Study of the Effects of the Federal Reserve System on Banking Relationships*. Chicago: A.W. Shaw Company, 1929.

Weber, Warren E. "Disaggregated Call Reports for U.S. National Banks, 1880-1910." Research Division Digital Archives, Federal Reserve Bank of Minneapolis, 2000; http://research.mpls.frb.fed.us/research/economists/wewproj.html.

Wheelock, David C. "Seasonal Accommodation and the Financial Crises of the Great Depression: Did the Fed 'Furnish an Elastic Currency?'" Federal Reserve Bank of St. Louis *Review*, November/December 1992, *74*(6), pp. 3-18; https://files.stlouisfed.org/files/htdocs/publications/review/92/11/Seasonal_Nov_Dec1992.pdf.

Wheelock, David C. "Economics and Politics in Selecting Federal Reserve Cities: Why Missouri Has Two Reserve Banks." Federal Reserve Bank of St. Louis *Review*, Fourth Quarter 2015, *97*(4), pp. 269-88; https://doi.org/10.20955/r.2015.269-88.

White, Eugene N. *The Regulation and Reform of the American Banking System, 1900-1929*. Princeton: Princeton University Press, 1983; https://doi.org/10.1515/9781400857449.

Yellen, Janet L. "Supervision and Regulation." Testimony before the Committee on Financial Services, U.S. House of Representatives, Washington DC, September 28, 2016; https://www.federalreserve.gov/newsevents/testimony/yellen20160928a.htm.

Credit Cycles and Business Cycles

Costas Azariadis

Unsecured firm credit moves procyclically in the United States and tends to lead gross domestic product, while secured firm credit is acyclical. Shocks to unsecured firm credit explain a far larger fraction of output fluctuations than shocks to secured credit. This article surveys a tractable dynamic general equilibrium model in which constraints on unsecured firm credit preclude an efficient capital allocation among heterogeneous firms. Unsecured credit rests on the value that borrowers attach to a good credit reputation, which is a forward-looking variable. Self-fulfilling beliefs over future credit conditions naturally generate endogenously persistent business cycle dynamics. A dynamic complementarity between current and future borrowing limits permits uncorrelated belief shocks to unsecured debt to trigger persistent aggregate fluctuations in both secured and unsecured debt, factor productivity, and output. The author shows that these sunspot shocks are quantitatively important, accounting for around half of output volatility. (JEL D92, E32)

Federal Reserve Bank of St. Louis *Review*, First Quarter 2018, *100*(1), pp. 45-71.
https://doi.org/10.20955/r.2018.45-71

1 OVERVIEW

Two prominent characteristics of the business cycle are the high autocorrelations of credit and output time series and the strong cross-correlation between those two statistics. Understanding these correlations, without the help of large and persistent shocks to the productivity of financial intermediaries and to the technical efficiency of final goods producers, has been a long-standing goal of macroeconomic research and the motivation for the seminal contributions mentioned in Section 2. Is it possible that cycles in credit, factor productivity, and output are not the work of large and persistent productivity shocks that afflict all sectors of the economy simultaneously? Could these cycles instead come from shocks to people's confidence in the credit market?

This article gives an affirmative answer to both questions within an economy in which part of the credit firms require to finance investment is secured by collateral and the remainder

Costas Azariadis is a research fellow at the Federal Reserve Bank of St. Louis and a professor at Washington University in St. Louis. This essay is based on the longer and more technical article Azariadis, Kaas, and Yi (2016). The author thanks Steve Williamson and two anonymous referees for constructive comments.

is based on reputation. The main contribution is to emphasize and quantify the role of reputational credit. Unsecured firm credit in the U.S. economy from 1981 to 2012 was quite important; it is strongly correlated with gross domestic product (GDP) and leads it by about a year. In our model, unsecured credit improves debt limits, facilitates capital reallocation, and helps aggregate productivity, provided that borrowers expect plentiful unsecured credit in the future. Favorable expectations of future debt limits increase the value of remaining solvent and on good terms with one's lenders. Widespread doubts, on the other hand, about future credit will lead to long-lasting credit tightening with severe macroeconomic consequences: Productive firms are unable to purchase or rent capital from unproductive ones, which end up owning too much capital. Aggregate productivity drops and GDP follows suit.

Total factor productivity (TFP) in this setting is sensitive to credit availability, and credit shocks drive TFP shocks. In this way, dynamic complementarity between current and future lending connects macroeconomic performance over time and endows one-time expectational impulses with long-lasting responses. A calibrated version of our economy matches well with the observed autocorrelations and cross-correlations of output, firm credit, and investment. Using the model to identify structural shocks to collateral credit, unsecured credit, and aggregate technology, I find that sunspot shocks to unsecured credit account for around half the variance in all major time series, while collateral shocks explain roughly one-third and technology shocks play a rather minor role. On the other hand, if the endogenous influence of sunspots on credit conditions is excluded a priori, the results show that too much output volatility would be incorrectly attributed to exogenous movements in aggregate technology—a standard result in the literature. I conclude that self-fulfilling and endogenously propagated credit shocks are quite important in U.S. business cycles.

2 CONNECTIONS WITH THE LITERATURE

Most citizens believe that credit availability is important for housing and other forms of investment as well as for job creation. If they are correct in this belief, there must be broad-based shocks to credit availability that real-world financial markets cannot easily overcome, as they would in the canonical Arrow-Debreu model of general equilibrium.

Starting with contributions of Bernanke and Gertler (1989) and Kiyotaki and Moore (1997), much of recent research is devoted to exploring how difficulties, or "frictions," in financial markets amplify and propagate disruptions to macroeconomic fundamentals, such as shocks to TFP or to monetary policy. More recently, and to some extent motivated by the events of the last financial crisis, several contributions argue that shocks to the financial sector itself may not only lead to severe macroeconomic consequences but can also contribute significantly to business cycle movements. For example, Jermann and Quadrini (2012) develop a model with collateral constraints, which they identify as residuals from aggregate time series of firm debt and collateral capital. Estimating a joint stochastic process for TFP and borrowing constraints, they find that both variables are highly autocorrelated and that financial shocks play an important role in business cycle fluctuations.[1] But what drives these shocks to financial conditions and to aggregate productivity? And what makes their responses so highly persistent?

This paper focuses on unsecured firm credit, a variable that is overlooked in the literature but seems to be of key importance for answering these questions. I first document new facts about secured versus unsecured firm credit. Most strikingly, for the U.S. economy over the period 1981-2012, I find that unsecured debt is strongly procyclical, with some tendency to lead GDP, while secured debt is at best acyclical, thus not contributing to the well-documented procyclicality of total debt. This finding provides some challenge for business cycle theories based on the conventional view of Kiyotaki and Moore (1997) that collateralized debt amplifies and even generates the business cycle. When credit is secured by collateral, a credit boom is associated with not only a higher leverage ratio but also a higher value of the collateralized assets. Conversely, an economic slump is associated with deleveraging and a decrease in the value of collateral. This suggests that secured debt, such as mortgage debt, should be strongly correlated with GDP. But this is not what I find; to the contrary, based on firm-level data from Compustat and on aggregate data from the flow of funds accounts of the Federal Reserve Board, I show that it is the *unsecured part of firm credit* that strongly comoves with output.

The model is a standard stochastic growth model that comprises a large number of firms facing idiosyncratic productivity shocks. In each period, productive firms wish to borrow from their less-productive counterparts. Besides possibly borrowing against collateral, the firms exchange unsecured credit, which rests on reputation. Building upon Bulow and Rogoff (1989) and Kehoe and Levine (1993), I assume that a defaulting borrower is excluded from future credit for a stochastic number of periods. As in Alvarez and Jermann (2000), endogenous forward-looking credit limits prevent default. These credit limits depend on the value that a borrower attaches to a good reputation, which itself depends on future credit market conditions.

One contribution of this article is the tractability of this framework, which permits me to derive a number of insightful analytical results in Section 3. With standard and convenient specifications of preferences and technology, I characterize any equilibrium by one backward-looking and one forward-looking equation (Result 1). With this characterization, I prove that unsecured credit cannot support first-best allocations, thereby extending related findings of Bulow and Rogoff (1989) and Hellwig and Lorenzoni (2009) to a growth model with idiosyncratic productivity (Result 2). I then show the existence of multiple stationary equilibria for a range of parameter configurations (Result 3). While there is always an equilibrium without unsecured credit, there can also exist one or two stationary equilibria with a positive volume of unsecured credit. One of these equilibria supports an efficient allocation of capital between firms, and another one features a misallocation of capital. The latter equilibrium is the one that provides the most interesting insights, since unsecured credit is traded and yet factor productivity falls short of the technology frontier.[2] I show that this equilibrium is always locally indeterminate and hence permits the existence of sunspot cycles fluctuating around the stationary equilibrium (Result 4). Moreover, output and credit respond persistently to a one-time sunspot shock.

One way to understand the role of expectations is to view unsecured credit like a bubble sustained by self-fulfilling beliefs, as has been argued by Hellwig and Lorenzoni (2009). Transitions from a "good" macroeconomic outcome with plenty of unsecured credit to a "bad"

outcome with low volumes of unsecured credit can be triggered by widespread skepticism about the ability of financial markets to continue the provision of unsecured credit at the volume needed to support socially desirable outcomes, which is similar to the collapse of a speculative bubble. The emergence and the bursting of rational bubbles in financially constrained economies has received attention in a number of recent contributions, for example, Caballero and Krishnamurthy (2006); Kocherlakota (2009); Farhi and Tirole (2012); and Miao and Wang (2015).

This work is also related to a literature on sunspot cycles arising from financial frictions. In an early contribution, Woodford (1986) shows that a simple borrowing constraint makes infinitely lived agents behave like two-period-lived overlapping generations, so that endogenous cycles can occur with sufficiently strong income effects or with increasing returns to production (see, e.g., Benhabib and Farmer, 1999, for a survey).[3] Harrison and Weder (2013) introduce a production externality in a Kiyotake-Moore (1997) model and show that sunspots emerge for reasonable values of returns to scale. Benhabib and Wang (2013) show how the interaction between collateral constraints and endogenous markups can lead to indeterminacy for plausibly calibrated parameters. Liu and Wang (2014) find that the financial multiplier arising from credit constraints gives rise to increasing returns at the aggregate level, which facilitates indeterminacy.

Other recent contributions find equilibrium multiplicity and indeterminacy in endowment economies with limited credit enforcement under specific assumptions about trading arrangements (Gu et al., 2013) and about the enforcement technology (Azariadis and Kaas, 2013). Azariadis and Kaas (2016) study a related model with limited enforcement, also documenting equilibrium multiplicity. That article builds on a stylized model with linear production technologies, which is not suited for a quantitative analysis; it does not consider sunspot shocks and focuses on a multi-sector economy without firm-specific risk.

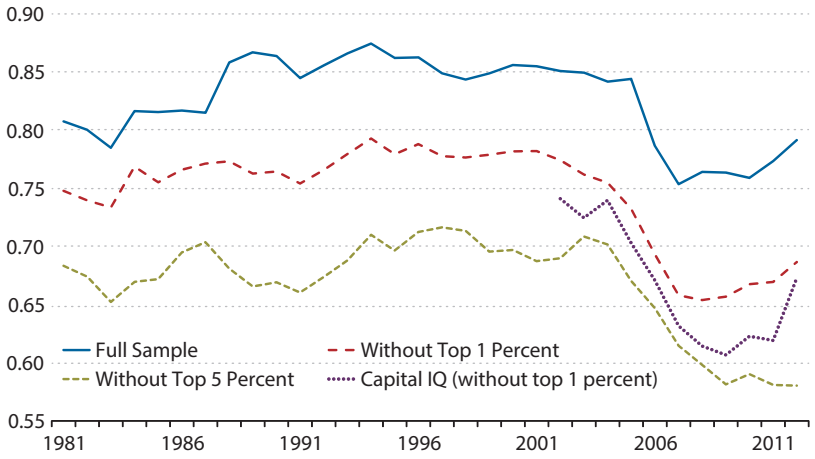
The rest of this article is organized as follows. The next section documents empirical evidence about secured and unsecured firm credit in the U.S. economy. In Section 4, I lay out the model framework; all equilibria by a forward-looking equation in the reputation values of borrowers; and derive my main results on equilibrium multiplicity, indeterminacy, and sunspot cycles. In Section 5, I extend the model in a few dimensions and conduct a quantitative analysis to explore the impacts of sunspot shocks and fundamental shocks on business cycle dynamics. Section 6 concludes.

3 UNSECURED VERSUS SECURED FIRM DEBT

This section summarizes evidence about firms' debt structure and its cyclical properties. I explore different firm-level data sets, covering distinct firm types, and relate the findings to evidence obtained from the flow of funds accounts. In line with previous literature,[4] I show that unsecured debt constitutes a substantial part of firms' total debt and is typically lower for samples including smaller firms. Time-series variation, whenever available, further indicates that unsecured debt plays a much stronger role in aggregate output dynamics than debt secured by collateral. I first describe the data and the variables measuring unsecured and secured debt and then report business cycle features.

Figure 1

Share of Unsecured Debt in Total Debt for Firms in Compustat and Capital IQ



SOURCE: Compustat and Capital IQ.

3.1 The Share of Unsecured Debt

I start with the publicly traded U.S. firms covered by Compustat for the period 1981-2012 for which Compustat provides the item "dm: debt mortgages and other secured debt." In line with Giambona and Golec (2012), I use this item to measure secured debt, and I attribute the residual to unsecured debt.[5] The *unsecured debt share* is then defined as the ratio between unsecured debt and total debt. To clean the data, I remove financial firms and utilities, and I also remove those firm-year observations where total debt is negative, where item "dm" is missing, or where "dm" exceeds total debt. Since Compustat aggregates can easily be biased by the effect of the largest firms in the sample, I also consider subsamples where I remove the largest 1 percent or 5 percent of the firms by their asset size. To see the impact of the largest firms for unsecured borrowing, Figure 1 shows the series of the unsecured debt shares for the three samples obtained from Compustat. The role of the largest firms is quite important for the level of the unsecured debt share, although much less for the time-series variation.[6] The very biggest firms are likely to have better access to bond markets and hence borrow substantially more unsecured debt. Removing the largest 1 percent (5 percent) of firms, however, cuts out 45 percent (75 percent) of the aggregate firm debt in the sample. Interestingly, in the years prior to the financial crisis of 2007-08, the unsecured debt share fell substantially, as firms expanded their mortgage borrowing relatively faster than other types of debt, with some reversal after 2008.

While Compustat covers public firms, the vast majority of U.S. firms are privately owned. To complement the above evidence, I also explore two data sets to obtain debt information for private firms. I first look at firms included in the database of Capital IQ, which is an affili-

ate of Standard and Poor's that produces the Compustat database but covers a broader set of firms. Since coverage by Capital IQ is comprehensive only from 2002 onward, I report these statistics for the period 2002 to 2012. I clean the data in the same way as above and consider aggregates for the full sample (without financial firms and utilities) and for the sample without the 1 percent (5 percent) of the largest firms. Similar to the Compustat definition, I use the Capital IQ item "SEC: Secured Debt" and the residual "DLC + DLTT – SEC" to measure unsecured debt. The resulting unsecured debt shares show a similar cyclical pattern as those from Compustat during the same period. For visual clarity, Figure 1 includes only the Capital IQ series with the largest 1 percent of firms removed. Note, however, that including larger firms or removing the top 5 percent of firms has similar effects as for Compustat, though it does not affect the U-shaped cyclical pattern in the graph. Relative to the corresponding series in Compustat, firms in Capital IQ borrowed more secured debt in all years, which is possibly explained by the fact that these firms have lower market transparency and hence less access to bond markets.[7]

It is worth emphasizing that even the private firms included in the Capital IQ database are relatively large firms with some access to capital markets, so they are also not fully representative of the U.S. business sector. To obtain evidence on the debt structure of small firms, I use the data collected in the Survey of Small Business Finances (SSBF) conducted by the Board of Governors of the Federal Reserve System in 2003. Earlier surveys, conducted in 1987, 1993, and 1998, do not contain comparably comprehensive information on collateral requirements, so I cannot obtain evidence across time. Firms in this survey report their balances in different debt categories (and within each category for up to three financial institutions). For each loan, they report whether collateral is required and which type of collateral is used (real estate, equipment, or other). I aggregate across firms for each debt category and measure as secured debt all the loans for which collateral is required, while unsecured debt comprises credit card balances and all loans without reported collateral requirements. I minimally clean the data by only removing observations with zero or negative assets or equity. Table 1 shows the results of this analysis. While mortgages and credit lines constitute the largest debt categories of small firms, accounting for almost three-quarters of the total, significant fractions of the other three loan categories are unsecured. This results in an unsecured debt share of 19.3 percent for firms in the SSBF.[8]

The evidence presented in Figure 1 and in Table 1 suggests that the unsecured debt share varies between about 20 percent (for the smallest firms) and 75 percent (for Compustat firms excluding the largest 1 percent).[9] To obtain a rough estimate for the average share of unsecured debt, I can further use the information in the flow of funds accounts, in which firm debt is categorized into several broad categories. About 95 percent of all credit market liabilities of non-financial firms are either attributed to mortgages (31 percent), loans (31 percent), or corporate bonds (33 percent). While mortgages are clearly secured and bonds are unsecured types of debt, the security status classification is ambiguous for loans. Among the non-mortgage loans in Table 1, around 30 percent are unsecured; this is a similar fraction as found in other studies.[10] Taken together, this suggests that around 45 percent ($\approx (0.33 + 0.31 \cdot 0.30)/(0.95)$) of the credit liabilities of non-financial firms is unsecured. In Section 4, I use an unsecured

Table 1

Secured and Unsecured Debt in the SSBF (2003, percent)

Debt category	Share of debt	Secured by real estate/equipment	Secured by other collateral	Unsecured
Credit cards	0.6	0.0	0.0	100.0
Lines of credit	36.5	39.4	38.5	22.1
Mortgages	38.0	98.0	0.4	1.7
Motor vehicle loans	4.8	52.1	2.1	45.8
Equipment loans	6.5	62.0	1.7	36.4
Other loans	13.6	53.6	6.3	40.1
Total	100.0	65.4	15.2	19.3

SOURCE: SSBF.

debt share of 50 percent as a calibration target.

Table 2

Relative Volatility and Comovement with Output (Compustat)

	Volatility relative to GDP			Correlation with GDP		
	Full sample	Without top 1%	Without top 5%	Full sample	Without top 1%	Without top 5%
Secured debt	3.61	3.39	2.76	–0.15	–0.05	0.15
Unsecured debt	4.19	3.73	4.43	0.70	0.70	0.75

SOURCE: Compustat.

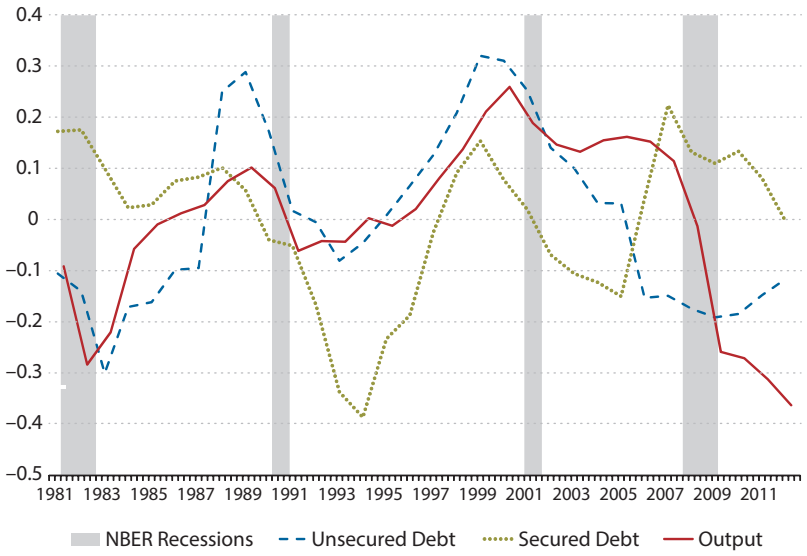
3.2 Business Cycle Features

I consider the time series from Compustat, deflate them by the price index for business value added, and linearly detrend the real series.[11] Table 2 reports the volatility of secured and unsecured debt (relative to output) as well as the contemporaneous correlations with output. Secured debt is weakly negatively correlated with GDP in the full sample; it becomes zero and weakly positive once I exclude the top 1 percent or 5 percent firms. In sharp contrast, unsecured debt is always strongly positively correlated with GDP. Thus, the well-known pro-cyclicality of total firm credit is driven by the independent role of unsecured debt. Both secured and unsecured debt are about three to four times as volatile as output.

Figure 2 shows the detrended time series of unsecured and secured debt for the full sample over the observation period, together with GDP. While unsecured debt comoves strongly with output, secured debt is only weakly related. Between the mid-1990s and the mid-2000s, both debt series move together, but they exhibit quite different patterns before and after this period. Unsecured debt falls much more sharply than secured debt during all recessions except the one in 2007-09.

Figure 2

Unsecured and Secured Debt for Compustat Firms, and GDP Multiplied by Factor Four (1981-2012)



NOTE: Annual linearly detrended series. Gray bars indicate recessions as determined by the National Bureau of Economic Research (NBER).

SOURCE: Compustat.

4 A MODEL OF UNSECURED FIRM CREDIT

To capture the prominent role of unsecured firm credit, I develop in this section a macroeconomic model in which heterogeneous firms face idiosyncratic productivity shocks and borrow up to endogenous credit limits that preclude default in equilibrium. For expositional reasons, I present first a benchmark model featuring only unsecured credit, along with a fixed labor supply and i.i.d. firm-specific productivity shocks. I also do not consider aggregate shocks to economic fundamentals. All these assumptions will be relaxed in the next section. Tractability and the main theoretical findings are preserved in these extensions.

4.1 The Setup

The model has a continuum $i \in [0,1]$ of firms, each owned by a representative owner, and a unit mass of workers. At any time t, all individuals maximize expected discounted utility,

$$\mathbb{E}_t (1-\beta) \sum_{\tau \geq t}^{\infty} \beta^{\tau - t} \ln(c_\tau),$$

over future consumption streams. Workers are perfectly mobile across firms; they supply 1 unit of labor per period, have no capital endowment, and do not participate in credit markets. Firm owners hold capital and have no labor endowment.[12] They produce a consumption and

investment good y_t using capital k_t' and labor ℓ_t with a common constant returns technology $y_t = (k_t')^{\alpha}(A\ell_t)^{1-\alpha}$. Aggregate labor efficiency A is constant for now, which will be relaxed in Section 5.

Firms differ in their ability to operate capital investment k_t. Some firms are able to enhance their invested capital according to $k_t' = a^p k_t$; they are labeled "productive." The remaining firms are labelled "unproductive"; they deplete some of their capital investment such that $k_t' = a^u k_t$. I assume that $a^p > 1 > a^u$ and write $\gamma \equiv a^u/a^p (<1)$ for the relative productivity gap. Productivity realizations are independent across agents and uncorrelated across time; firms are productive with probability π and unproductive with probability $1 - \pi$. Thus, a fraction π of the aggregate capital stock K_t is owned by productive firms in any period. Uncorrelated productivity simplifies the model; it also implies that the dynamics of borrowers' net worth do not propagate shocks as in Kiyotaki and Moore (1997) and Bernanke and Gertler (1989). At the end of a period, all capital depreciates at common rate δ.

Timing within each period is as follows: First, firm owners observe the productivity of their business, then they borrow and lend in a centralized credit market at a gross interest rate R_t and hire labor in a centralized labor market at wage w_t. Second, production takes place. Third, firm owners redeem their debt; they consume and save for the next period. All prices and credit constraints (as defined below) possibly depend on the realization of sunspot shocks.

In the credit market, productive firms borrow from unproductive firms, drawing on a short-term credit line that depends on their equity position and on a small array of macroeconomic variables. I rule out long-term loans and pre-determined interest rates that do not depend on current economic conditions. See Garriga et al. (2016) on the role of long-term nominal contracts. All credit is unsecured and is available only to borrowing firms with a clean credit history. If a firm decides to default in some period, the credit reputation deteriorates and the firm is banned from unsecured credit.[13] Defaulting firms can continue to operate their business; hence, they are able to produce or to lend their assets to other firms.[14] Each period after default, the firm recovers its credit reputation with probability ψ ($\geq$0), in which case it regains full access to credit markets.[15]

Since no shocks arrive during a credit contract (that is, debt is redeemed at the end of the period before the next productivity shock is realized), there exist default-deterring credit limits, defined similarly as in the pure exchange model of Alvarez and Jermann (2000). These limits are the highest values of credit that prevent default. Unsecured borrowing is founded on a borrower's desire to maintain a good credit reputation and continued access to future credit. Below I prove that credit constraints are necessarily binding in equilibrium (see Result 2).

Workers do not participate in the credit market and hence consume their labor income w_t in every period. This assumption is not as strong as it may seem; in the steady-state equilibrium it only requires that workers are not permitted to borrow. This is because the steady-state gross interest rate R satisfies $R < 1/\beta$ (see Corollary 1), which means that workers are borrowing constrained and do not desire to save.[16]

At the beginning of the initial period $t = 0$, a firm owner i is endowed with capital (equity) e_0^i; hence, the initial equity distribution $(e_0^i)_{i \in [0,1]}$ is given. In any period $t \geq 0$, let θ_t denote the constraint on a borrower's debt-to-equity ratio in period t. This value is common for all bor-

rowing firms because the value of solvency and default are proportional to the equity position of borrowers with a homothetic utility function, as shown below. It is endogenously determined to prevent default (cf. property (iii) of the following equilibrium definition). A productive firm i entering the period with equity (capital) e_t^i can borrow up to $b_t^i = \theta_t e_t^i$ and invest $k_t^i = e_t^i + b_t^i$. An unproductive firm lends out capital, so $b_t^i = \theta_t e_t^i$ and investment is $k_t^i = e_t^i + b_t^i \leq e_t^i$. Although the constraints $b_t^i \leq \theta_t e_t^i$ seem to resemble the collateral limits in the literature emanating from Kiyotaki and Moore (1997), I emphasize that θ_t here has very different features: It describes the size of an unsecured credit line, not the value of collateralizable equity. It is also a forward-looking variable that reacts to changes in credit market expectations.

The budget constraint for firm i with capital productivity $a^i \in \{a^p, a^u\}$ reads as

$$c_t^i + e_{t+1}^i = \left(a^i k_t^i\right)^{\alpha} \left(A\ell_t^i\right)^{1-\alpha} + (1-\delta) a^i k_t^i - w_t \ell_t^i - R_t b_t^i. \tag{1}$$

I am now ready to define equilibrium.

Definition 1 *A competitive equilibrium is a list of consumption, savings, and production plans for all firm owners,* $(c_t^i, e_t^i, b_t^i, k_t^i, \ell_t^i)_{i \in [0,1], t \geq 0}$, *conditional on realizations of idiosyncratic productivities and sunspot shock; consumption of workers,* $c_t^w = w_t$; *factor prices for labor and capital* (w_t, R_t); *and debt-equity constraints* θ_t, *such that the following occur:*

(i) $(c_t^i, e_t^i, b_t^i, k_t^i, l_t^i)$ *maximizes firm-owner i's expected discounted utility* $\mathbb{E}_t \sum_{t \geq 0}^{\infty} \beta^t \ln\left(c_t^i\right)$, *subject to the budget constrain (1) and credit constraints* $b_t^i \leq \theta_t e_t^i$.

(ii) The labor market and the credit market clear in all periods $t \geq 0$:

$$\int_0^1 \ell_t^i \, di = 1, \quad \int_0^1 b_t^i \, di = 0.$$

(iii) If $b_t^i \leq \theta_t e_t^i$ *is binding in problem (i), the firm-owner i is exactly indifferent between debt redemption and default in period t, where default entails exclusion from credit for a stochastic number of periods with readmission probability* ψ *in each period following default.*

A typical equilibrium in this economy will lead all high-productivity firms to borrow their entire credit line and low-productivity firms to be indifferent between producing and lending. TFP in this situation is a weighted average of all individual TFP's, which reflects the misallocation of capital. Small credit lines will correspond to significant misallocation and low aggregate TFP. Generous credit lines will improve both TFP and GDP.

4.2 Equilibrium

The model permits a tractable description of individual choices. This is because individual firm's policies (i.e., borrowing/lending, saving, and employment) are all linear in the firm's equity position and independent of the firm's history, which in turn implies that these decisions can be easily aggregated. Furthermore, default incentives are also independent of the current size of the firm, which implies that all borrowing firms face the same constraint on their debt-to-equity ratio. Uncorrelated idiosyncratic productivities simplify the model further

because all firms have the same chance to become productive in each period, so that the distribution of wealth is irrelevant.[17]

Since firms hire labor to equate the marginal product to the real wage, all productive (unproductive) firms have identical capital-labor ratios; these are linked by a no-arbitrage condition implied by perfect labor mobility:

$$\frac{k_t^p}{\ell_t^p} = \gamma \frac{k_t^u}{\ell_t^u}. \tag{2}$$

With binding credit constraints, a fraction $z_t \equiv \min[1, \pi(1 + \theta_t)]$ of the aggregate capital stock K_t is operated by productive firms. It follows from (2) and labor market clearing that

$$\frac{k_t^p}{\ell_t^p} = \frac{a_t K_t}{a^p} \leq K_t < \frac{a_t K_t}{a^u} = \frac{k_t^u}{\ell_t^u},$$

where $a_t \equiv a^p z_t + a^u(1 - z_t)$ is the average capital productivity. The gross return on capital for a firm with capital productivity $a^s \in \{a^u, a^p\}$ is then $a^s R_t^*$ with $R_t^* \equiv [1 - \delta + \alpha A^{1-\alpha}(a_t K_t)^{\alpha-1}]$.

In any equilibrium, the gross interest rate cannot exceed the capital return of productive firms $a^p R_t^*$, and it cannot fall below the capital return of unproductive firms $a^u R_t^*$. Thus it is convenient to write $R_t = \rho_t a^p R_t^*$, with $\rho_t \in [\gamma, 1]$. When $\rho_t < 1$, borrowers are credit constrained. In this case, the leveraged equity return $[1 + \theta_t(1 - \rho_t)]a^p R_t^*$ exceeds the capital return $a^s R_t^*$. Unproductive firms, on the other hand, lend out all their capital when $\rho_t > \gamma$; they only invest in their own inferior technology if $\rho_t = \gamma$. Therefore, credit market equilibrium is equivalent to the complementary-slackness conditions:

$$\rho_t \geq \gamma, \quad \pi(1+\theta_t) \leq 1. \tag{3}$$

With this notation, the firm owner's budget constraint (1) simplifies to $e_{t+1} + c_t = R_t e_t$ when the firm is unproductive in t and to $e_{t+1} + c_t = [1 + \theta_t(1 - \rho_t)]a^p R_t^* e_t$ when the firm is productive. It follows from logarithmic utility that every firm owner consumes a fraction $(1 - \beta)$ of wealth and saves the rest.

To derive the endogenous credit limits, let $V_t(W)$ denote the continuation value of a firm owner with a clean credit reputation who has wealth W at the end of period t, prior to deciding consumption and saving. These values satisfy the recursive equation[18]:

$$\begin{aligned} V_t(W) &= (1-\beta)\ln\left[(1-\beta)W\right] \\ &\quad + \beta \mathbb{E}_t\left[\pi V_{t+1}\left(\left[1+\theta_{t+1}(1-\rho_{t+1})\right]a^p R_{t+1}^* \beta W\right) + (1-\pi)V_{t+1}\left(R_{t+1}\beta W\right)\right]. \end{aligned}$$

The first term in this equation represents utility from consuming $(1 - \beta)W$ in the current period. For the next period, $t + 1$, the firm owner saves equity βW, which earns leveraged return $[1 + \theta_{t+1}(1 - \rho_{t+1})]a^p R_{t+1}^*$ with probability π and return R_{t+1} with probability $1 - \pi$. It follows that continuation $t + 1$ values take the form $V_t(W) = \ln(W) + V_t$, where V_t is independent of wealth, satisfying the recursive relation:

$$V_t = (1-\beta)\ln(1-\beta)+\beta\ln\beta + \beta\mathbb{E}_t\left[\pi\ln\left(\left[1+\theta_{t+1}(1-\rho_{t+1})\right]a^p R^*_{t+1}\right)+(1-\pi)\ln(R_{t+1})+V_{t+1}\right]. \tag{4}$$

For a firm owner with a default flag and no access to credit, the continuation value is $V^d(W) = \ln(W) + V^d$, where V^d satisfies, analogously to equation (4), the recursion

$$V^d_t = (1-\beta)\ln(1-\beta)+\beta\ln\beta + \beta\mathbb{E}_t\left[\pi\ln\left(a^p R^*_{t+1}\right)+(1-\pi)\ln(R_{t+1})+V^d_{t+1}+\psi\left(V_{t+1}-V^d_{t+1}\right)\right]. \tag{5}$$

This firm owner cannot borrow in period $t + 1$, so the equity return is $a^p R^*_{t+1}$ with probability π and R_{t+1} with probability $1 - \pi$. At the end of period $t + 1$, the credit reputation recovers with probability ψ, in which case the continuation utility increases from V^d to V_{t+1}.

If a borrower has a clean credit reputation and enters period t with equity e_t, the debt-equity constraint θ_t makes him exactly indifferent between default and debt redemption if

$$\ln\left(\left[1+\theta_t(1-\rho_t)\right]a^p R^*_t e_t\right)+V_t = \ln\left(a^p R^*_t(1+\theta_t)e_t\right)+V^d_t.$$

Here the right-hand side is the continuation value after default: The firm owner invests $(1 + \theta_t)e_t$, earns return $a^p R^*_t$, and does not redeem debt. The left-hand side is the continuation value under solvency, where the borrower earns the leveraged equity return $[1 + \theta_t(1 - \rho_t)]a^p R^*_t$.

Defining $v_t \equiv V_t - V^d \geq 0$ as the value of reputation, this equation can be solved for the default-deterring constraint on the debt-to-equity ratio:

$$\theta_t = \frac{e^{v_t}-1}{1-e^{v_t}(1-\rho_t)}. \tag{6}$$

This constraint is increasing in the reputation value v_t: A greater expected payoff from access to unsecured credit makes debt redemption more valuable, which relaxes the self-enforcing debt limit. In the extreme case when the reputation value is zero, unsecured credit cannot be sustained so that $\theta_t = 0$.

Using (4) and (5), reputation values satisfy the recursive identity:

$$\begin{aligned} v_t &= \beta\mathbb{E}_t\left[\pi\ln\left(1+\theta_{t+1}(1-\rho_{t+1})\right)+(1-\psi)v_{t+1}\right] \\ &= \beta\mathbb{E}_t\left[\pi\ln\left(\frac{\rho_{t+1}}{1-e^{v_{t+1}}(1-\rho_{t+1})}\right)+(1-\psi)v_{t+1}\right]. \end{aligned} \tag{7}$$

I summarize this equilibrium characterization as follows:

Result 1 *Any solution $(\rho_t,\theta_t,v_t)_{t\geq 0}$ to the system of equations (3), (6), and (7) gives rise to a competitive equilibrium with interest rates $R_t = \rho_t a^p R^*_t$, capital returns $R^*_t = 1 - \delta + \alpha A^{1-\alpha}(a_t k_t)^{\alpha-1}$, and average capital productivities $a_t = a^u + (a^p - a^u) \cdot \min[1,\pi(1 + \theta_t)]$. The capital stock evolves according to*

$$K_{t+1} = \beta\left[\left(1-\delta\right)+\alpha A^{1-\alpha}\left(a_t K_t\right)^{\alpha-1}\right]a_t K_t. \tag{8}$$

An implication of this result is that any equilibrium follows two dynamic equations: the backward-looking dynamics of aggregate capital, equation (8), and the forward-looking dynamics of reputation values, equation (7), or, equivalently, equation (9) below. The latter identity is independent of the aggregate state K_t and hence permits a particularly simple equilibrium analysis.

Using Result 1, I obtain two immediate results. First, an equilibrium with no unsecured credit always exists ($v_t = 0$, $\theta_t = 0$, and $\rho_t = \gamma$ in all periods). Intuitively, if there is no value to reputation, any borrower prefers to default on unsecured credit so that debt limits must be zero. Second, I show that constraints on unsecured credit are necessarily binding. This is in line with earlier results by Bulow and Rogoff (1989) and Hellwig and Lorenzoni (2009), who show that the first-best[19] cannot be implemented by limited enforcement mechanisms that ban defaulting agents from future borrowing but not from future lending. It differs decisively from environments with two-sided exclusion, as in Kehoe and Levine (1993) and Alvarez and Jermann (2000), where first-best allocations can be sustained with unsecured credit under certain circumstances.[20] The intuition for this result is as follows: If borrowers were unconstrained, the interest rate would coincide with the borrowers' capital return. Hence, there would be no leverage gain and access to credit would have no value. In turn, every borrower would default on an unsecured loan, no matter how small. I summarize this finding as follows:

Result 2 *Any equilibrium features binding borrowing constraints. Specifically, given any time and history, there exists some future time and continuation history in which the borrowing constraint is binding.*

It follows immediately that the equilibrium interest rate is smaller than the rate of time preference.

Corollary 1 *In any steady-state equilibrium, $R < 1/\beta$.*

4.3 Multiplicity and Cycles

Although borrowers must be constrained, the credit market may nonetheless be able to allocate capital efficiently. In particular, when the reputation value v_t is sufficiently large, credit constraints relax and the interest rate exceeds the capital return of unproductive firms who then lend out all their capital. Formally, when v_t exceeds the threshold value,

$$\overline{v} = \ln\frac{1}{1-\gamma\left(1-\pi\right)} > 0,$$

the equilibrium conditions (3) and (6) are solved by $\theta_t = (1 - \pi)/\pi$ and $\rho_t = [1 - e^{-v_t}]/(1 - \pi) > \gamma$. Conversely, when v_t falls short of v, credit constraints tighten, the interest rate equals the capital return of unproductive firms ($\rho_t = \gamma$), who are then indifferent between lending out capital or investing in their own technology, so that some capital is inefficiently allocated. I can use this insight to rewrite the forward-looking equation (7) as

$$v_t = \mathbb{E}_t f\left(v_{t+1}\right) \tag{9}$$

with

$$f(v) = \begin{cases} \beta(1-\psi)v + \beta\pi \ln\left[\dfrac{\gamma}{1-e^{v}(1-\gamma)}\right], \text{if } v \in [0,\bar{v}], \\ \beta(1-\pi-\psi)v + \beta\pi\ln(1/\pi), \qquad \text{if } v \in [\bar{v}, v_{\max}]. \end{cases}$$

Here $v = v_{max} = \ln(1/\pi)$ is the reputation value, where the interest rate reaches $\rho = 1$ and borrowers are unconstrained. It is straightforward to verify that f is strictly increasing if $\pi + \psi < 1$, convex in $v < \bar{v}$, and it satisfies $f(0) = 0$, $f(\bar{v}) > \bar{v}$ if γ is small enough, and $f(v_{max}) < v_{max}$. This reconfirms that the absence of unsecured credit ($v = 0$) is a stationary equilibrium. Depending on economic fundamentals, there can also exist one or two steady states exhibiting positive trading of unsecured credit. Panel A of Figure 3 shows a situation in which function f has three intersections with the 45 degree line: $v = 0$, $v^* \in (0,\bar{v})$, and $v^{**} \in (\bar{v}, v_{max})$ The steady states at $v = 0$ and at v^* have inefficient capital allocation, whereas capital is efficiently allocated at $v^{**} > \bar{v}$. Panel B of Figure 3 shows a possibility with only two steady states, at $v = 0$ and at $v^{**} > \bar{v}$. A third possibility (not shown in the figure) is that $v = 0$ is the unique steady state, so that unsecured credit is not enforceable. The following result describes how the set of stationary equilibria changes as the productivity ratio $\gamma = a^u/a^p$ varies:

Result 3 *For all parameter values* (β,π,ψ,γ), *there exists a stationary equilibrium without unsecured credit and with inefficient capital allocation. In addition, there are threshold values* $\gamma_0, \gamma_1 \in (0,1)$ with $\gamma_0 < \gamma_1$ for the productivity ratio γ such that

(i) for $\gamma \in (\gamma_0,\gamma_1)$, *there are two stationary equilibria with unsecured credit: one at* $v^* \in (0,\bar{v})$ *with inefficient capital allocation and one at* $v^{**} \in (\bar{v}, v_{max})$ *with efficient capital allocation;*

(ii) for $\gamma \le \gamma_0$, *there exists a unique stationary equilibrium with unsecured credit and efficient capital allocation at the reputation value* $v^{**} \in (\bar{v}, v_{max})$*; and*

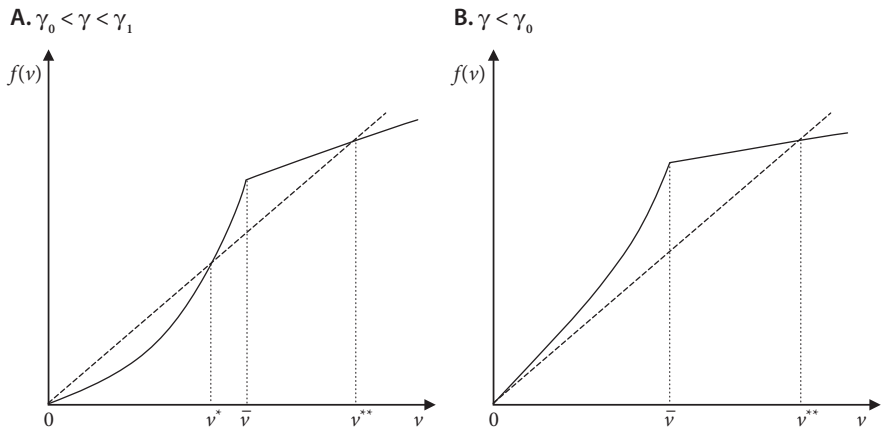
(iii) for $\gamma > \gamma_1$, *there is no stationary equilibrium with unsecured credit.*

For small enough idiosyncratic productivity fluctuations $\gamma > \gamma_1$, unsecured credit is not enforceable, because firm owners value participation in credit markets too little. Conversely, for larger idiosyncratic shocks, exclusion from future credit is a sufficiently strong threat so that unsecured credit is enforceable without commitment. When idiosyncratic productivity shocks are sufficiently dispersed, the unique steady state with unsecured credit has an efficient factor allocation, while for intermediate values of γ, a third equilibrium emerges with unsecured credit and some misallocation of capital.

The explanation for equilibrium multiplicity is a *dynamic complementarity* between endogenous credit constraints, which are directly linked to reputation values. Borrowers' expectations of future credit market conditions affect their incentives to default now, which in turn determine current credit constraints. If future constraints are tight, the payoff of a clean credit reputation is modest so that access to unsecured credit has low value. In turn,

Figure 3

Steady States at $v = 0, v^{*} v^{**}$



current default-deterring credit limits must be small. Conversely, if borrowers expect future credit markets to work well, a good credit reputation has high value, which relaxes current constraints.

As Figure 3 shows, multiplicity follows from a specific nonlinearity between expected and current reputation values. To understand this nonlinearity, it is important to highlight the different impacts of market expectations on borrowing constraints and on interest rates. In the inefficient regime $v \leq \bar{v}$, improvements in credit market expectations relax credit constraints without changes in the interest rate, which leads to particularly large gains from participation and hence to a strong impact on the current value of reputation. Conversely, if $v > \bar{v}$, beliefs in better credit conditions also raise the interest rate, which dampens the positive effect and hence mitigates the increase in the current reputation value.

Even when unsecured credit is available and possibly supports efficient allocations of capital, that efficiency rests on the confidence of market participants in future credit market conditions. When market participants expect credit constraints to tighten rapidly, the value of reputation shrinks over time, which triggers a self-fulfilling collapse of the market for unsecured credit. For instance, if $\gamma < \gamma_0$, the steady state at v^{**} is determinate and the one at $v = 0$ is indeterminate; see Panel B of Figure 3. That is, there exists an infinity of nonstationary equilibria $v_t = f(v_{t+1}) \rightarrow 0$ where the value of reputation vanishes asymptotically. These equilibria are mathematically similar to the bubble-bursting equilibria in overlapping-generation models or in Kocherlakota (2009). If $\gamma \in (\gamma_0, \gamma_1)$, the two steady states at $v = 0$ and at v^{**} are determinate, whereas the one at v^{*} is indeterminate. In that situation, a self-fulfilling collapse of the credit market would be described by an equilibrium with $v_t \rightarrow v^{*}$, where a positive level of unsecured credit is still sustained in the limit. In both of these events, a one-time belief shock can lead to a permanent collapse of the credit market. But in the latter case, indeterminacy also permits stochastic business cycle dynamics driven by self-fulfilling beliefs (sunspots). Sunspot fluctu-

ations vanish asymptotically if $\gamma < \gamma_0$, but they give rise to permanent volatility around the indeterminate steady state v^* if $\gamma \in (\gamma_0, \gamma_1)$.

Result 4 *Suppose that $\gamma \in (\gamma_0, \gamma_1)$, as defined in Result 3. Then there exist sunspot cycles featuring permanent fluctuations in credit, output, and TFP.*

The dynamic complementarity between current and future endogenous credit constraints not only creates expectations-driven business cycles, it also generates an endogenous propagation mechanism: Because of $f'(v^*) > 1$, a one-time belief shock in period t triggers a persistent adjustment dynamic of reputation values v_{t+k} (and thus of credit, investment, and output) in subsequent periods. Intuitively, a self-fulfilling boom (slump) in unsecured credit in period t can emerge only if the boom (slump) is expected to last for several periods.

Corollary 2 *A one-time sunspot shock $\varepsilon_t > 0 (\varepsilon_t < 0)$ in period t induces a persistent positive (negative) response of firm credit and output.*

Although an endogenous propagation mechanism is not a necessary feature of any sunspot model, it tends to be associated with a large class of neoclassical models with local indeterminacy, such as the one in Benhabib and Farmer (1994). Local indeterminacy introduces additional state variables that tend to generate endogenous propagation mechanisms. The model differs from other sunspot models in that it uses borrower reputation as an additional state variable. The difference this makes is that sunspots are tied specifically to confidence in credit markets. I show in the next section that self-fulfilling beliefs in future credit conditions can indeed generate output fluctuations broadly similar to the data.

5 QUANTITATIVE ANALYSIS

The previous section demonstrates how self-fulfilling belief shocks can generate procyclical responses of unsecured credit, with potentially sluggish adjustment dynamics. In this section, I introduce some additional features to this model and calibrate it to the U.S. economy in order to examine the business cycle features of sunspot shocks as well as of fundamental shocks.

5.1 Model Extension

I extend the model in three directions. First, I include a variable labor supply. Second, I allow firms to issue debt secured by collateral. Third, I introduce aggregate fundamental shocks to technology and to firms' collateral capacity.

Specifically, I modify workers' period utility to $\ln\left(C_t - \frac{\varphi}{1+\varphi} L_t^{(1+\varphi)/\varphi}\right)$, where L_t is the labor supply and φ is the Frisch elasticity. Regarding secured borrowing, I assume that a fraction $\lambda_t < 1$ of a firm's end-of-period assets can now be recovered by creditors in a default event, instead of no recovery at all as in Section 4. Since all firms can pledge collateral to their creditors, the relevant outside option of a defaulter is the exclusion from unsecured credit while retaining access to collateralized credit. As before, all credit is within the period and no default occurs in equilibrium, which implies that secured and unsecured credit carry the same interest rate R_t. Besides sunspot shocks, I allow for shocks to λ_t and to aggregate labor efficiency A_t.

The first type of shock directly affects the tightness of borrowing constraints, much like the financial shocks considered by Jermann and Quadrini (2012). Shocks to labor efficiency account for those movements in aggregate output that are not generated by the endogenous response of aggregate productivity to changes in the allocation of capital.

All productive firms in period t can borrow secured credit up to the debt-equity limit θ_t^s, which is determined from $R_t\theta_t^s = \lambda_t a^p R_t^*(1 + \theta_t^s)$. For each unit of equity, the firm borrows θ_t^s so that a fraction λ_t of the end-of-period assets $a^p R_t^*(1 + \theta_t^s)$ fully protect the lenders who provide secured credit. On top of that, firms can borrow unsecured credit up to the endogenous debt-equity limit θ_t^u. This extended model leads to the following constraint on the total debt-to-equity ratio $\theta_t = \theta_t^s + \theta_t^u$, which precludes default:

$$\theta_t = \frac{e^{v_t} - 1 + \lambda_t}{1 - \lambda_t - e^{v_t}\left(1 - \rho_t\right)}, \tag{10}$$

where v_t is again the value of reputation, that is, the utility benefit of a clean credit reputation, and $\rho_t = R_t/(a^p R_t^*)$. This relation extends equation (6) to the case where some assets can be collateralized. One implication of equation (10) is that an exogenous reduction in collateral borrowing, of the type that may have triggered the 2007-09 Great Recession when housing prices dropped, will lead to a decline in total borrowing because unsecured borrowing does not expand fast enough to counteract the contraction in collateral. Observe that all borrowing must be secured; that is, $\theta_t = \theta_t^s$ if reputation has no value ($v_t = 0$). If $v_t > 0$, borrowing in excess of θ_t^s is unsecured. Note, however, that the share λ_t of the unsecured debt obligation $R_t\theta_t^u$ could be recovered if a firm opted for default. This is certainly a realistic feature since bond holders, for example, can recover a substantial fraction of their assets after a default. I also generalize equation (9) to a forward-looking equation:

$$v_t = \mathbb{E}_t f\left(v_{t+1}, \lambda_{t+1}\right). \tag{11}$$

Therefore, I obtain a similar dichotomy as before: The dynamics of reputation values is independent of the capital stock, labor market variables, and technology shocks. I also confirm that, for specific parameter constellations, a steady state with unsecured credit and inefficient capital allocations exists; I choose this equilibrium for the calibration of model parameters.[21] This steady state is again indeterminate, so self-fulfilling belief shocks impact the dynamics of unsecured credit.

5.2 Calibration

I calibrate this model to the U.S. economy, choosing parameters so that the indeterminate steady-state equilibrium matches suitable long-run properties. The calibration targets correspond to statistics obtained for the U.S. business sector in the period 1981-2012. As the best available data source on unsecured versus secured credit is available at annual frequency, I calibrate the model annually and set δ, α, and β in a standard fashion to match plausible values of capital depreciation, factor income shares, and the capital-output ratio, respectively.[22] The Frisch elasticity is set to $\varphi = 1$. I normalize average capital productivity in steady state to $a = 1$

Table 3

Parameter Choices

Parameter	Value	Explanation/Target
δ	0.078	Depreciation rate
α	0.3	Capital income share
β	0.89	Capital-output ratio
φ	1	Frisch elasticity
ψ	0.1	10-year default flag
π	0.18	Share of productive firms (credit volume)
λ	0.43	Recovery parameter (unsecured debt share)
a^u	0.779	Lowest productivity (debt-to-equity ratio $\theta = 3$)
a^p	1.080	Highest productivity (normalization $a = 1$)

and steady-state labor efficiency to $A = 1$. I set the exclusion parameter $\psi = 0.1$ so that a defaulting firm owner has difficulty obtaining unsecured credit for 10 years after default.[23] I choose the remaining parameters π, λ, and a^u to match the following three targets[24]: (i) credit to nonfinancial firms is 82 percent of annual GDP; (ii) the debt-to-equity ratio of constrained firms is $\theta = 3$; and (iii) unsecured credit is 50 percent of total firm credit.[25] Given that this model has a two-point distribution of firm productivity (and hence of debt-to-equity ratios), the choice of target (ii) is somewhat arbitrary. I also calibrate the model with $\theta = 2$ and obtain very similar results. All parameters are listed in Table 3.

Despite the simplicity of this model, it is worth noting that this calibration has a reasonably low share of credit-constrained firms ($\pi = 18$ percent) and that the mean debt-to-capital ratio ($\theta\pi = 54$ percent) is in line with empirical findings (cf. Rajan and Zingales, 1995). Further, the parameterization produces a plausible cross-firm dispersion of TFP. With firm-level output equal to $y^i = (a^i - 1)k^i + (A\ell^i)^{1-\alpha}(a^i k^i)^\alpha$, I calculate a standard deviation of log TFP equal to 0.33, which is close to the within-industry average 0.39 reported in Bartelsman et al. (2013).

5.3 Persistence of Sunspot Shocks

For illustrative purposes, I first suppose that fundamental shocks are absent; that is, λ_t and A_t are at their steady-state values, while sunspot shocks are the only source of business cycle dynamics. In this case, the log-linearized dynamics of the credit-to-capital ratio[26] follows

$$\hat{\theta}_{t+1} = \frac{1}{\varphi_2}\hat{\theta}_t + d_1\varepsilon^s_{t+1},$$

where coefficients d_1 and φ_2 are constant terms and ε^s_{t+1} is a sunspot shock. In particular, I find that the autocorrelation coefficient is

$$\frac{1}{\varphi_2} = \frac{1}{\beta(1-\psi) + \beta\pi(1+\theta)\dfrac{a^p - a^u}{a^u}},$$

Table 4

Model Statistics with Uncorrelated Sunspot Shocks

	Output	Credit	Investment	Consumption	Employment
U.S. data (1981-2012)					
Relative volatility	1	2.73	2.43	0.80	0.69
Autocorrelation	0.848	0.832	0.618	0.899	0.893
Correlation with output	1	0.620	0.715	0.969	0.910
Model					
Relative volatility	1	2.59	3.28	0.84	0.35
Autocorrelation	0.925	0.903	0.791	0.978	0.978
Correlation with output	1	0.993	0.771	0.923	0.923

SOURCE: Compustat.

NOTE: Output and investment are for the U.S. business sector. Credit is for the Compustat firm sample considered in Section 2 without the largest 1 percent of firms. All variables are deflated, logged, and linearly detrended. Model statistics are based on 100,000 simulations of 32 periods. The volatility of sunspot shocks is set so that the model-generated output volatility matches the one in the data.

which equals 0.949 for the calibrated model parameters. That is, when I feed the model with uncorrelated sunspot shocks, the endogenous dynamics of credit are highly persistent, actually more so than in the data.[27] Table 4 confirms this finding and reports business cycle statistics under sunspot shocks. Most importantly, uncorrelated sunspot shocks generate persistent business cycle dynamics with autocorrelation coefficients that are somewhat above their data counterparts. Volatilities and comovement of consumption and investment are plausible, whereas credit is too strongly correlated with output, which comes as no surprise since all output dynamics are induced by the sunspot-driven dynamics of credit.[28]

5.4 Multiple Shocks

To evaluate the relative importance of sunspot shocks for the overall business cycle dynamics, I include fundamental shocks to the financial sector (collateral parameter λ_t) as well as to the real sector (labor efficiency parameter A_t). I identify sunspot shocks as well as fundamental shocks as follows: I use the Compustat series for secured credit to compute the secured credit-to-capital ratio whose cyclical component measures $\hat{\theta}_t^s$. Similarly, all Compustat credit (secured and unsecured) identifies the series $\hat{\theta}_t$. I then use those two series to back out the (log deviations of) reputation values $\hat{v}_t$ and collateral parameters $\hat{\lambda}_t$. Labor efficiency $\hat{A}_t$ is identified so as to match the cyclical component of output. Hence, it picks up all output dynamics left unexplained by financial shocks (shocks to collateral $\hat{\lambda}$ and to unsecured credit $\hat{v}$). Therefore, all three shocks together generate by construction the output dynamics of the data. I can therefore measure how each one contributes to the total volatility and how it accounts for output movements in specific episodes.

I consider the following SVAR:

$$(12)\qquad \begin{pmatrix} \hat{A}_t \\ \hat{\lambda}_t \\ \hat{\nu}_t \end{pmatrix} = \mathbf{B}\begin{pmatrix} \hat{A}_{t-1} \\ \hat{\lambda}_{t-1} \\ \hat{\nu}_{t-1} \end{pmatrix} + \begin{pmatrix} e_{1t} \\ e_{2t} \\ e_{3t} \end{pmatrix}$$

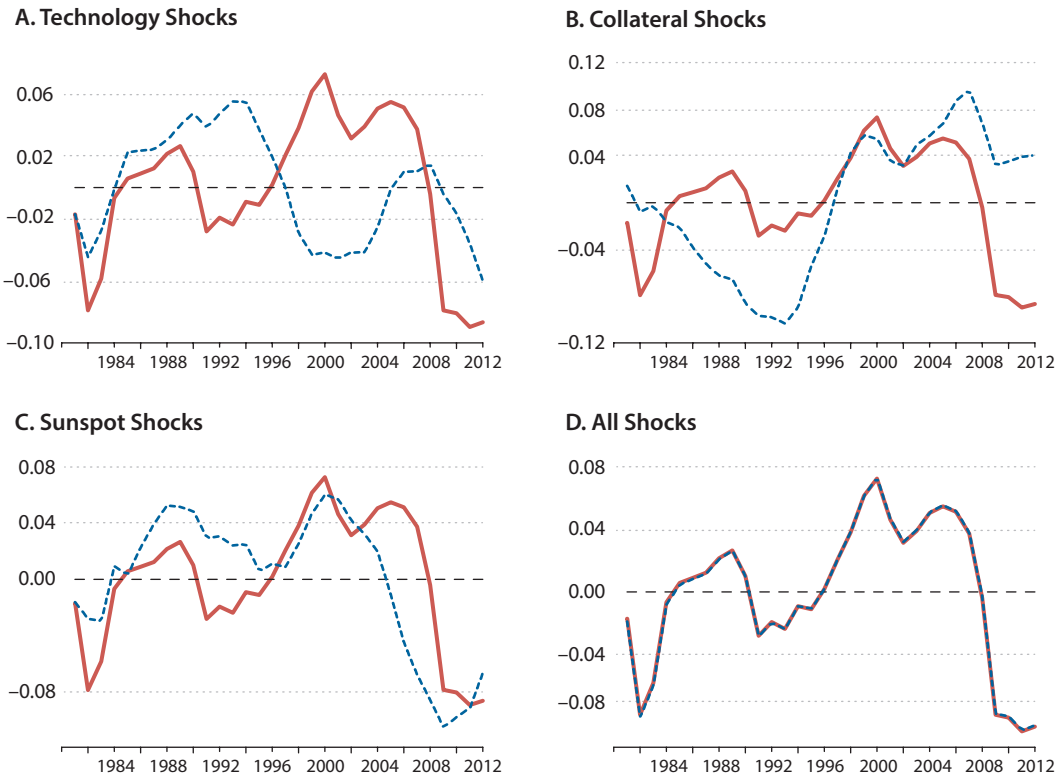
with coefficient matrix **B**, and apply the Choleski decomposition such that $\mathbf{e} = (e_{1t},e_{2t},e_{3t}) = \mathbf{C}(\varepsilon_{1t},\varepsilon_{2t},\varepsilon_{3t})$ with lower triangular matrix **C**. I call ε_{1t} the technology shock, ε_{2t} the collateral shock, and ε_{3t} the sunspot shock. By ordering the sunspot shock as the last variable in the SVAR, I assume that those shocks can impact only credit market expectations contemporaneously, while all correlations in the innovations to $(\hat{A}_t,\hat{\lambda}_t,\hat{\nu}_t)$ are attributed to technology shocks and to collateral shocks. In other words, I may be attributing too much influence to technology and collateral shocks, thus providing a lower bound on the contribution of sunspot shocks. I take into account that the forward-looking equation for reputation values (11) imposes a restriction on the last row in equation (12). I therefore estimate only the first two equations and impose the model restriction on the last row in equation (12).[29]

Figure 4 shows the implied time-series decomposition of output into the three components associated with the three identified structural shocks $(\varepsilon_{1t},\varepsilon_{2t},\varepsilon_{3t})$, where the solid line in each window represents the data output and the dashed line represents the predicted output when only one of the structural shocks is active. Panel D puts all three shocks together, which by construction explains all output variation. Sunspot shocks, ε_{3t}, account for the broad business cycle features of output quite well (Panel C); this is despite the fact that I have attributed all the contemporaneous correlations of the three innovations to technology and collateral shocks. Collateral shocks seem to matter for the credit-expansion periods in the late 1990s and mid-2000s, while they only account for a moderate portion of the decline in 2007-09. Technology shocks do not appear to matter much for output movements since the 1990s, although they are responsible for a substantial fraction of the output drop after the Great Recession.[30]

I can also decompose the total variance of output (more specifically, the power spectrum) into the three structural components, with each coming separately from one of the three identified shocks. I find that sunspot shocks account for 51 percent of the total output variance, collateral shocks account for 44 percent, and technology shocks account for only the remaining 5 percent. This result is quite striking: Even though the $\hat{A}_t$ series is constructed to match all output dynamics not explained by financial shocks, shocks to $\hat{A}_t$ play a rather minor role in the total output variance. This result—that the two financial shocks account for the vast majority of output dynamics—differs markedly from Jermann and Quadrini (2012), who find that productivity shocks and financial shocks both explain around half of output fluctuations. But the present model generates a similar result when I shut down sunspot shocks. Precisely, when I set $\hat{\nu}_t = 0$ and identify $\hat{\lambda}_t$ and $\hat{A}_t$ to account for the dynamics of total firm credit and output, I find that structural shocks to collateral and to technology each account for around half of output volatility. Put differently, technology shocks pick up a large fraction of the output dynamics that come from the self-fulfilling belief shocks that drive unsecured credit in

Figure 4

Decomposition of Output for the Three Shocks



NOTE: In panel A to C, the solid curves are data output and the dashed curves are model-generated output dynamics with only the designated shock active. In the bottom-right panel, all three shocks are active.

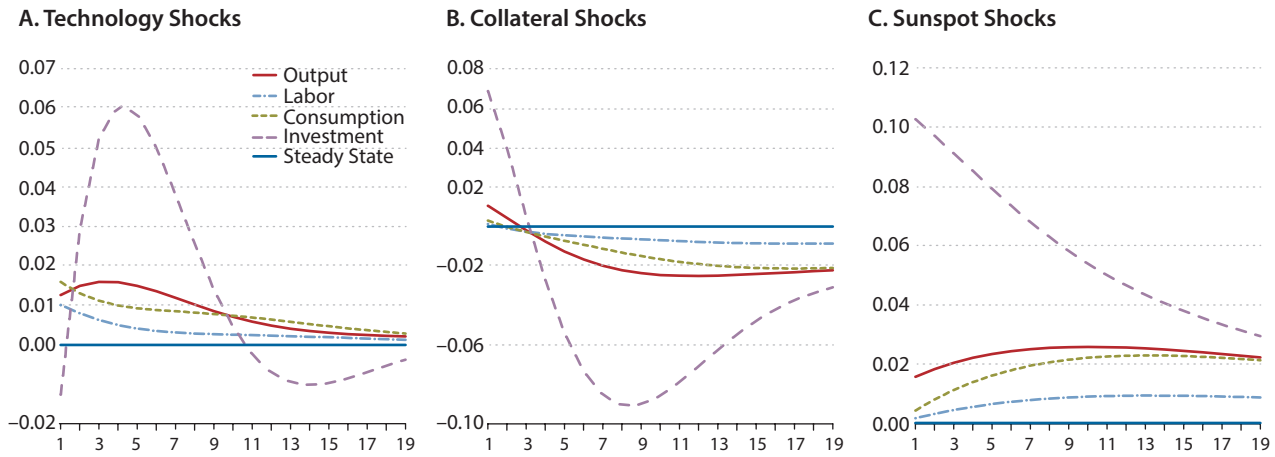
the model.[31] Sunspot shocks not only matter for output but also for the dynamics of other macroeconomic variables: In a variance decomposition, I find that sunspot shocks account for around 50 percent of the variance of employment, consumption, investment, and firm credit, whereas technology shocks account for less than 10 percent.[32]

By construction, the three structural shocks not only describe output fluctuations but also the secured and unsecured credit dynamics. It is interesting to examine to what extent the three shocks contribute to these separate credit cycles and how these components relate to output dynamics. Sunspot shocks turn out to account for the major movements in unsecured credit, while technology shocks and collateral shocks do not capture the pattern of unsecured credit well. Given this finding and that unsecured credit is strongly procyclical, I can infer that sunspot shocks are the major driving force of the credit cycle.

Lastly, in Figure 5 I show the impulse responses of output, investment, consumption, and employment to the three orthogonal shocks (one standard deviation). Sunspot shocks generate a stronger and more persistent response than the other two shocks. In particular, a

Figure 5

Impulse Responses to the Three Shocks



NOTE: In panel A to C, the solid curves are data output and the dashed curves are model-generated output dynamics with only the designated shock active. In the bottom-right panel, all three shocks are active.

collateral shock implies that the positive output response turns negative only two years after the shock, which is at odds with the vector autoregression evidence on the real effects of credit market shocks (e.g., Lown and Morgan, 2006, and Gilchrist et al., 2009). This suggests that sunspot shocks (on top of or independent of collateral shocks) are an important contributing factor.

5.5 Autocorrelated Productivity

A strong simplifying assumption in the benchmark model is that firm productivity is drawn each period independently from a two-point distribution. The main benefit is that this makes the model very tractable, permitting a complete analytical characterization of the global dynamics in Section 3. This framework can be readily extended to account for an autocorrelated idiosyncratic productivity process (still on a two-point distribution). All of the main results survive. In particular, there are multiple steady-state equilibria; further, a steady-state equilibrium with unsecured credit and some misallocation of capital is typically indeterminate and hence gives rise to sunspot-driven dynamics.

6 CONCLUSIONS

To study the role of unsecured firm debt, I develop and analyze a small dynamic general equilibrium model with heterogeneous firms and limited credit enforcement. In the model, credit constraints and aggregate productivity are endogenous variables. Constraints on unsecured credit depend on the value that borrowers attach to future credit market conditions,

which is a forward-looking variable. Aggregate productivity depends on the reallocation of existing capital among heterogeneous firms which, among other things, depends on current credit constraints. When these constraints bind, they slow down capital reallocation between firms and push aggregate factor productivity below its frontier. I show that this model exhibits a very natural equilibrium indeterminacy that gives rise to endogenous cycles driven by self-fulfilling beliefs in credit market conditions (sunspot shocks). In particular, a one-time sunspot shock triggers an endogenous and persistent response of credit, productivity, and output.

Cycles persist because of a dynamic complementarity in endogenous constraints on unsecured credit. Borrowers' incentives to default depend on their expectations of future credit market conditions, which in turn influence current credit constraints. If borrowers expect a credit tightening over the next few periods, their current default incentives become larger, which triggers a tightening of current credit. This insight also explains why a one-time sunspot shock *must be followed* by a long-lasting response of credit market conditions (and thus of macroeconomic outcomes): If market participants expect that a credit boom (or a credit slump) will die out quickly, these expectations could not be powerful enough to generate a sizable current credit boom (or slump). ■

NOTES

1 Other examples of financial shocks include the work of Kiyotaki and Moore (2012), who introduce shocks to asset resaleability, and Gertler and Karadi (2011), who consider shocks to the asset quality of financial intermediaries. These papers also impose or estimate highly persistent shock processes.

2 The other determinate steady states of this model either do not sustain unsecured credit (and hence resemble similar dynamics as in a Kiyotaki-Moore-type model with binding collateral constraints) or they have an efficient allocation of capital (and hence exhibit the same business cycle properties as a frictionless real business cycle model).

3 Although earlier work on indeterminacy has shown that sunspot shocks can induce persistent macroeconomic responses (e.g., Farmer and Guo, 1994, and Wen, 1998), the adjustment dynamics are typically sensitive to the particular specifications of technologies and preferences. In the model, persistent responses arise necessarily due to the dynamic complementarity in unsecured credit conditions.

4 See, in particular, recent corporate finance contributions examining heterogeneity in the debt structure across firms (e.g., Rauh and Sufi, 2010; Giambona and Golec, 2012; and Colla et al., 2013). These do not address business cycles, however.

5 This classification means that unsecured debt is not explicitly backed by collateral; it does not mean that it has zero (or little) recovery value in the case of default.

6 While the effect of the largest firms is also important for total debt growth, it is not important for its cyclicality. Total debt over this period grew much faster than GDP.

7 Firms in the Capital IQ sample are actually bigger than those in the Compustat samples. In the period 2002-12, the average asset size of firms in the full (bottom 99 percent, bottom 95 percent) Compustat sample is $2,602 million ($1,230 million, $550 million), whereas in the Capital IQ sample it is $3,391 million ($2,028 million, $1,142 million). In total, there are about twice as many observations in Compustat than in Capital IQ for each year.

8 Because the collateral requirement is a dummy variable, only a fraction of these loans might actually be secured by collateral. This measure of unsecured credit should therefore be regarded as a lower bound.

9 Note that the latter number is consistent with those found in two other studies about the debt structure of Compustat firms. Rauh and Sufi (2010) examine the financial footnotes of 305 randomly sampled non-financial firms in Compustat. Based on different measures, their unsecured debt share (defined as senior unsecured plus

subordinated debt relative to total debt) is 70.3 percent. Giambona and Golec (2012) look at the distribution of unsecured debt shares for Compustat firms, reporting mean (median) values of 0.63 (0.75).

[10] Using bank survey data, Berger and Udell (1990) find that around 70 percent of all commercial and industrial loans in the United States are secured. Booth and Booth (2006) find that 75 percent of their sample of syndicated loans are secured.

[11] I use a linear trend to capture the low-frequency movements in credit and output that are quite significant over the period 1981-2012.

[12] The assumption of a representative owner by no means restricts this model to single-owner businesses. All it requires is that the firm's owners desire a smooth dividend stream, for which there is ample evidence (e.g., Leary and Michaely, 2011).

[13] That is, lenders receive no payment in a default event. In the next section, I relax this assumption by introducing collateral assets and secured credit. In this extension, a fraction of unsecured borrowing can also be recovered.

[14] We can think of such default events as either a liquidation, in which case the firm owners can start a new firm that needs to build up a reputation, or as a reorganization, in which case the firm continues operation.

[15] With permanent exclusion of defaulters ($\psi = 0$), this enforcement technology corresponds to the one discussed by Bulow and Rogoff (1989) and Hellwig and Lorenzoni (2009), who assume that defaulters are excluded from future credit but are still allowed to save.

[16] Outside the steady state, the workers' first-order condition $\mathbb{E}_t[\beta R_t w_t / w_{t+1}] < 1$ is satisfied in the log-linear approximation of the model for the calibrated parameters and for shocks of reasonable magnitude.

[17] If productivity shocks are autocorrelated, the wealth distribution becomes a state variable, but the model remains tractable since only a single variable, the wealth share of borrowing firms, matters for aggregate dynamics. This follows again because linear policy functions permit aggregation.

[18] In the absence of sunspot shocks, the expectations operator could be dropped from this and from subsequent equations because I abstract from aggregate shocks to economic fundamentals in this section.

[19] In the first-best equilibrium of this economy, there are no credit constraints; the interest rate equals the capital return of productive firms, $R_t = a^p R_t^*$, so all firms (productive and unproductive) earn the same return. All capital is employed at productive firms, and the model is thus isomorphic to a standard growth model with a representative firm.

[20] In endowment economies with permanent exclusion of defaulters, it is well-known that perfect risk sharing can be implemented if the discount factor is sufficiently large, if risk aversion is sufficiently strong, or if the endowment gap between agents is large enough (see, e.g., Kehoe and Levine, 2001). Azariadis and Kaas (2013) show that the role of the discount factor changes decisively if market exclusion is temporary. Note that the multiplicity results discussed in this article do not change under permanent exclusion of defaulters.

[21] As in the simpler model of the previous section, the other (determinate) steady states either feature efficient factor allocations or do not sustain unsecured credit. Hence, their business cycle properties either resemble those of a standard frictionless model or those of an economy with collateral-based credit constraints.

[22] Output is real value added in the business sector, and the capital stock is obtained from the perpetual inventory method based on total capital expenditures in the business sector. This yields 1.49 as the target for the capital-output ratio.

[23] This 10-year default flag corresponds to the bankruptcy regulation for individual firm owners who file for bankruptcy under Chapter 7 of the U.S. Bankruptcy Code. Generally, business firms in the United States can file for bankruptcy under either Chapter 7 (which leads to liquidation) or Chapter 11 (which allows continued operation after reorganization). In either case, it is plausible to assume that the reputation loss from default inhibits full access to credit for an extended period.

[24] The normalization $a = a^u + \pi(1 + \theta)(a^p - a^u) = 1$ then yields parameters a^p and $\gamma = a^u / a^p$.

[25] Regarding (i), credit market liabilities of non-financial business are 0.82 of annual output (averaged over 1981-2012, flow-of-funds accounts of the Board of Governors of the Federal Reserve System, Z.1, Table L.101). Regarding (ii), debt-to-equity ratios below 3 are usually required to qualify for commercial loans (see Herranz et al., 2017). Further, in the SSBF (Capital IQ, Compustat) samples, the mean debt-to-equity ratios are 3.04 (3.15, 2.43).

[26] The hat symbol over a variable indicates the log deviation of that variable from the steady state. The credit-to-capital ratio in the model is $\theta_t \pi$, for which the log deviation equals that of the borrowers' credit-equity ratio θ_t, because π is constant.

[27] To obtain data analogs for the (linearly detrended) credit-to-capital ratio, I can either use firm credit from the flow-of-funds accounts or from Compustat, which yield annual autocorrelation coefficients of 0.883 (flow of funds) and 0.817 (Compustat).

[28] When I decompose total credit into secured and unsecured components, I find that both are strongly procyclical when sunspots are the only source of shocks; correlations with output are 0.83 (0.95) for secured (unsecured) credit. Secured credit is however much less volatile; relative standard deviations are 1.36 (4.13) for secured (unsecured) credit.

[29] I also perform a similar analysis in which I estimate equation (12) without any restrictions on matrix **B**. The main findings are similar and attribute an even larger role to sunspot shocks. Particularly, I find that sunspot shocks account for around 70 percent of the variance of output, employment, consumption, and investment. They also induce similarly persistent impulse responses.

[30] Somewhat surprisingly, Panel A of Figure 4 shows a negative correlation between detrended GDP and technology shocks over the period of the "Great Moderation," roughly 1988-2008. One possible explanation is that credit markets improved markedly over this period, reducing capital misallocation and boosting the effective rate of capital utilization by too big a margin relative to observed GDP growth. A drop in TFP would then exactly match GDP with the Solow residual of a model with full capital utilization. Cette et al. (2016) list some evidence favoring slower TFP growth since the millennium.

[31] The model further differs from Jermann and Quadrini (2012) in that aggregate productivity is partly endogenous and hence correlates positively with financial conditions.

[32] The standard errors for these point estimates are small. For example, the one-standard-error bands for the sunspot contributions of any of these variables are between 1- and 11-percentage-points wide.

REFERENCES

Alvarez, Fernando and Jermann, Urban. "Efficiency, Equilibrium, and Asset Pricing with Risk of Default." *Econometrica*, July 2000, *68*(4), pp. 775-97; https://doi.org/10.1111/1468-0262.00137.

Azariadis, Costas and Kaas, Leo. "Endogenous Credit Limits with Small Default Costs." *Journal of Economic Theory*, March 2013, *148*(2), pp. 806-24; https://doi.org/10.1016/j.jet.2012.08.004.

Azariadis, Costas and Kaas, Leo. "Capital Misallocation and Aggregate Factor Productivity." *Macroeconomic Dynamics*, March 2016, *20*(2), pp. 525-43; https://doi.org/10.1017/S1365100514000236.

Azariadis, Costas; Kaas, Leo and Wen, Yi. "Self-Fulfilling Credit Cycles." *Review of Economic Studies*, October 2016, *83*(4), pp. 1364-405; https://doi.org/10.1093/restud/rdv056.

Bartelsman, Eric; Haltiwanger, John and Scarpetta, Stefano. "Cross-Country Difference in Productivity: The Role of Allocation and Selection." *American Economic Review*, February 2013, *103*(1), pp. 305-34; https://doi.org/10.1257/aer.103.1.305.

Benhabib, Jess and Farmer, Roger. "Indeterminacy and Increasing Returns." *Journal of Economic Theory*, June 1994, *63*(1), pp. 19-41; https://doi.org/10.1006/jeth.1994.1031.

Benhabib, Jess and Farmer, Roger. "Indeterminacy and Sunspots in Macroeconomics," in J.B. Taylor and Michael Woodford, eds., *Handbook of Macroeconomics*. Volume 1A. Amsterdam: North-Holland, 1999, pp. 387-448; https://doi.org/10.1016/S1574-0048(99)01009-5.

Benhabib, Jess and Wang, Pengfei. "Financial Constraints, Endogenous Markups, and Self-fulfilling Equilibria." *Journal of Monetary Economics*, October 2013, *60*(7), pp. 789-805; https://doi.org/10.1016/j.jmoneco.2013.06.004.

Berger, Allen and Udell, Gregory. "Collateral, Loan Quality and Bank Risk." *Journal of Monetary Economics*, 1990, *25*(1), pp. 21-42; https://doi.org/10.1016/0304-3932(90)90042-3.

Bernanke, Ben S. and Gertler, Mark. "Agency Costs, Net Worth, and Business Fluctuations." *American Economic Review*, March 1989, *79*(1), pp. 14-31; https://www.jstor.org/stable/1804770?seq=1#page_scan_tab_contents.

Booth, James R. and Booth, Lena Chua. "Loan Collateral Decisions and Corporate Borrowing Costs." *Journal of Money, Credit and Banking*, February 2006, *38*(1), pp. 67-90; https://doi.org/10.1353/mcb.2006.0011.

Bulow, Jeremy and Rogoff, Kenneth. "Sovereign Debt: Is to Forgive to Forget?" *American Economic Review*, March 1989, *79*(1), pp. 43-50; https://www.jstor.org/stable/1804772?seq=1#page_scan_tab_contents.

Caballero, Ricardo, J. and Krishnamurthy, Arvind. "Bubbles and Capital Flow Volatility: Causes and Risk Management." *Journal of Monetary Economics*, 2006, *53*(1), pp. 35-53; https://doi.org/10.1016/j.jmoneco.2005.10.005.

Cette, Gilbert; Fernald, John G. and Mojon, Benoit. "The Pre-Great Recession Slowdown in Productivity." *European Economic Review*, September 2016, *88*, pp. 3-20; https://doi.org/10.1016/j.euroecorev.2016.03.012.

Colla, Paolo; Ippolito, Filippo and Li, Kai. "Debt Specialization." *Journal of Finance*, October 2013, *68*(5), pp. 2117-41; https://doi.org/10.1111/jofi.12052.

Farhi, Emmanuel and Tirole, Jean. "Bubbly Liquidity." *Review of Economic Studies*, 2012, *79*(2), pp. 678-706; https://doi.org/10.1093/restud/rdr039.

Farmer, Roger and Guo, Jang-Ting. "Real Business Cycles and the Animal Spirits Hypothesis." *Journal of Economic Theory*, June 1994, *63*(1), pp. 42-72; https://doi.org/10.1006/jeth.1994.1032.

Garriga, Carlos; Kydland, Finn E. and Sustek, Roman. "Nominal Rigidities in Debt and Product Markets." Working Paper 2016-017A, Federal Reserve Bank of Saint Louis, 2016; https://research.stlouisfed.org/wp/2016/2016-017.pdf.

Gertler, Mark and Karadi, Peter. "A Model of Unconventional Monetary Policy." *Journal of Monetary Economics*, January 2011, *58*(1), pp. 17-34; https://doi.org/10.1016/j.jmoneco.2010.10.004.

Giambona, Erasmo and Golec, Joseph. "The Growth Opportunity Channel of Debt Structure." Working paper, Amsterdam Business School Research Institute, 2013; http://hdl.handle.net/11245/1.408704.

Gilchrist, Simon; Yankov, Vladimir and Zakrajsek, Egon. "Credit Market Shocks and Economic Fluctuations: Evidence from Corporate Bond and Stock Markets." *Journal of Monetary Economics*, April 2009, *56*(4), pp. 471-93; https://doi.org/10.1016/j.jmoneco.2009.03.017.

Gu, Chao; Mattesini, Fabrizio; Monnet, Cyril and Wright, Randall. "Endogenous Credit Cycles." *Journal of Political Economy*, October 2013, *121*(5), pp. 940-65; https://doi.org/10.1086/673472.

Harrison, Sharon G. and Weder, Mark. "Sunspots and Credit Frictions." Macroeconomic Dynamics, July 2013, 17(5), pp. 1055-69; https://doi.org/10.1017/S1365100511000836.

Hellwig, Christian and Lorenzoni, Guido. "Bubbles and Self-Enforcing Debt." *Econometrica*, July 2009, *77*(4), pp. 1137-64; https://doi.org/10.3982/ECTA6754.

Herranz, Neus; Krasa, Stefan and Villamil, Anne. "Entrepreneurs, Legal Institutions and Firm Dynamics." *Economic Theory*, January 2017, *63*(1), pp. 263-85; https://doi.org/10.1007/s00199-016-1026-8.

Jermann, Urban and Quadrini, Vincenzo. "Macroeconomic Effects of Financial Shocks." *American Economic Review*, February 2012, *102*(1), pp. 238-71; https://doi.org/10.1257/aer.102.1.238.

Kehoe, Timothy and Levine, David. "Debt-Constrained Asset Markets." *Review of Economic Studies*, October 1993, *60*(4), pp. 865-88; https://doi.org/10.2307/2298103.

Kehoe, Timothy and Levine, David. "Liquidity Constrained Markets Versus Debt Constrained Markets." *Econometrica*, May 2001, *69*(3), pp. 575-98; https://doi.org/10.1111/1468-0262.00206.

Kiyotaki, Nobuhiro and Moore, John. "Credit Cycles." *Journal of Political Economy*, April 1997, *105*(2), pp. 211-48; https://doi.org/10.1086/262072.

Kiyotaki, Nobuhiro and Moore, John. "Liquidity, Business Cycles, and Monetary Policy." NBER Working Paper No. 17934, National Bureau of Economic Research, March 2012; http://www.nber.org/papers/w17934.

Kocherlakota, Narayana. "Bursting Bubbles: Consequences and Cures." Paper presented at the Macroeconomic and Policy Challenges Following Financial Meltdowns Conference, Federal Reserve Bank of Minneapolis, April 2009; https://www.imf.org/external/np/seminars/eng/2009/macro/pdf/nk.pdf.

Leary, Mark T. and Michaely, Roni. "Determinants of Dividend Smoothing: Empirical Evidence." *Review of Financial Studies*, October 2011, *24*(10), pp. 3197-249; https://doi.org/10.1093/rfs/hhr072.

Liu, Zheng and Wang, Pengfei. "Credit Constraints and Self-Fulfilling Business Cycles." *American Economic Journal: Macroeconomics*, January 2014, *6*(1), pp. 32-69; https://doi.org/10.1257/mac.6.1.32.

Lown, Cara and Morgan, Donald. "The Credit Cycle and the Business Cycle: New Findings Using the Loan Officer Opinion Survey." *Journal of Money, Credit and Banking*, September 2006, *38*(6), pp. 1575-97; https://doi.org/10.1353/mcb.2006.0086.

Miao, Jianjun and Wang, Pengfei. "Banking Bubbles and Financial Crises." *Journal of Economic Theory*, May 2015, *157*(5), pp. 763-92; https://doi.org/10.1016/j.jet.2015.02.004.

Rajan, Raghuram G. and Zingales, Luigi. "What Do We Know About Capital Structure? Some Evidence from International Data." *Journal of Finance*, December 1995, *50*(5), pp. 1421-60; https://doi.org/10.1111/j.1540-6261.1995.tb05184.x.

Rauh, Joshua D. and Sufi, Amir. "Capital Structure and Debt Structure." *Review of Financial Studies*, October 2010, *23*(12), pp. 4242-80; https://doi.org/10.1093/rfs/hhq095.

Wen, Yi. "Capacity Utilization Under Increasing Return to Scale." *Journal of Economic Theory*, July 1998, *81*(1), pp. 7-36; https://doi.org/10.1006/jeth.1998.2412.

Woodford, Michael. "Stationary Sunspot Equilibria in a Finance-Constrained Economy." *Journal of Economic Theory*, October 1986, *40*(1), pp. 128-37; https://doi.org/10.1016/0022-0531(86)90011-6.

The Aggregate Implications of Size-Dependent Distortions

Nicolas Roys

This article examines the aggregate implications of size-dependent distortions. These regulations misallocate labor across firms and hence reduce aggregate productivity. The author then considers a case study of labor laws in France, where firms with 50 employees or more face substantially more regulation than firms with fewer than 50. The size distribution of firms is visibly distorted by these regulations: There are many firms with exactly 49 employees. A quantitative model is developed with a payroll tax of 0.15 percent that applies only to firms with more than 50 employees. Removing the regulation while holding total employment constant leads to an increase in output of around 0.3 percent. (JEL E23, O1, O40)

Federal Reserve Bank of St. Louis *Review*, First Quarter 2018, *100*(1), pp. 73-85.
https://doi.org/10.20955/r.2018.73-85

1 INTRODUCTION

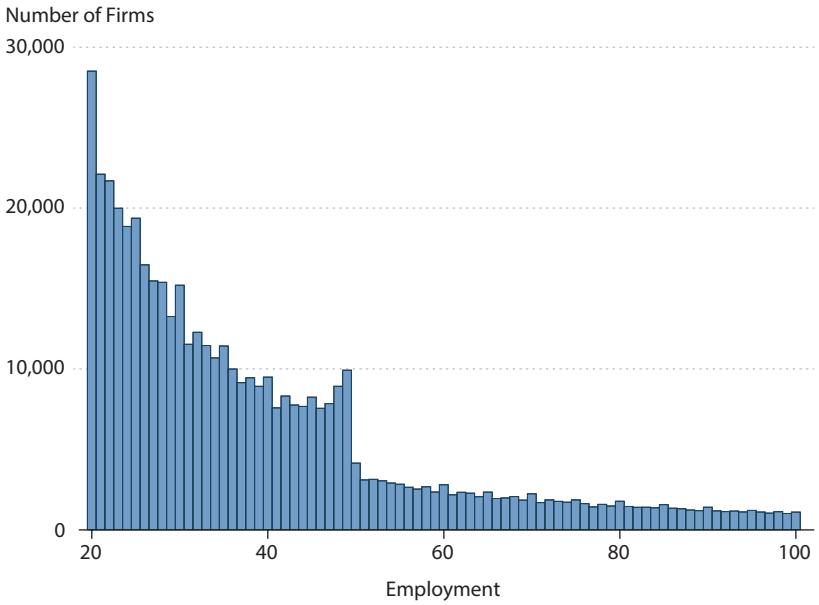
In the United States, new firms created 2.9 million jobs per year on average over the period 1980-2010.[1] While new firms clearly play an important role in job creation, many fail after a short period of time or do not grow. Are regulations preventing young businesses from expanding? Many regulators seem to think so: In many countries small firms face lighter regulation than large firms. The rationale for exempting small firms from some regulations is that the compliance cost is too high relative to their sales. A necessary consequence, however, is that regulations are phased in as the firm grows, generating an implicit marginal tax. Because regulations are typically phased in at a few finite points, they are sometimes referred to as "threshold effects."

Regulation, broadly defined, takes many forms—from hygiene and safety rules, to mandatory elections of employee representatives, to larger taxes. Under the Affordable Care Act, firms with 50 or more full-time equivalent employees are required to offer health insurance to their full-time employees. This requirement raises concerns that firms cut employment to

Nicolas Roys was an economist at the Federal Reserve Bank of St. Louis and is a senior lecturer at Royal Holloway, University of London. This *Review* article relates to Gourio and Roys (2014). The author thanks Joseph McGillicuddy for excellent research assistance.

Figure 1

Distribution of French Firms with 20 to 100 Employees



SOURCE: INSEE, BRN.

stay below the threshold or substitute some of their full-time workers with part-time workers. Similarly, regulations that alter the incentives to expand explain the large number of small community banks in the United States.

These distortions have attracted attention in public policy circles. The common wisdom, as reflected in numerous reports by blue-ribbon panels, is that these regulations significantly impede the growth of small firms and should be suppressed or smoothed out. However, there is little work formally modeling these policies to understand and evaluate their effects. This article proposes a simple model and gives a quantitative evaluation of this common wisdom. What are the potential benefits of removing, or smoothing, the regulation thresholds? To answer this question, this article considers a case study of regulations that apply only to firms in France with more than 50 employees. The firm-size distribution is distorted: There are few firms with exactly 50 employees and a large number of firms with 49 employees. Figure 1 plots the firm-size distribution in our French data, illustrating this well-known pattern. The visibly distorted firm distribution suggests that productivity could be increased if firms close to the threshold grow, as labor would be reallocated toward more-productive firms. Because these regulations depend on a precise threshold, the behavior of firms around the threshold is particularly informative on the effects of distortions.

The rest of the article proceeds as follows. Section 2 presents a model to study regulations that limit firm scale. Section 3 presents a case study of labor laws in France that differ depend-

ing on the side of the employment threshold firms stand on. Section 4 applies the model of Section 2 to study these distortions. Section 5 proposes a quantitative analysis. Section 6 concludes.

2 THE MODEL

This section introduces a simple model of production and employment, based on Lucas (1978), to evaluate the impact of size-dependent distortions.

2.1 Environment

There is a continuum of firms with production function

$$y = e^z n^\alpha,$$

where n is employment and e^z is a firm's productivity level (e denotes the exponential function). The distribution of productivity in the population is characterized by the density f. Production displays decreasing returns $\alpha \in (0,1)$.[2] Aggregate output, Y, is defined as the integral of the production of each firm $y(z)$,

$$Y = \int e^z n(z)^\alpha f(z) dz, \tag{1}$$

where $n(z)$ is the employment of firms with productivity z.

Firms hire labor in a competitive labor market where workers supply labor inelastically. Let total employment be denoted by N. The wage rate, w, taken as given by each firm, is such that the labor market in equilibrium is

$$\int n(z) f(z) dz \leq N. \tag{2}$$

Labor costs for the firm are equal to the wage bill multiplied by a size-dependent tax $T(n)$.

2.2 Labor Demand

A firm with productivity z solves the optimization problem,

$$\max_n \left\{ e^z n^\alpha - wn\left(1 + T(n)\right) \right\}.$$

If T is differentiable, labor demand satisfies the first-order condition,

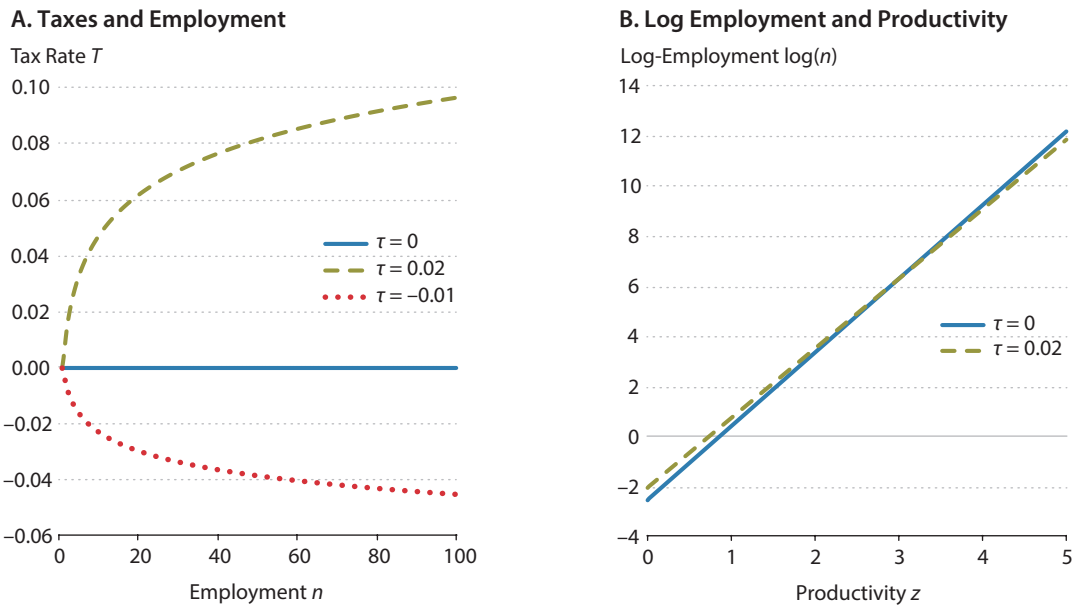
$$\alpha e^z n^{\alpha-1} = w\left(1 + T(n) + nT'(n)\right).$$

If $T(n) = \tau$, distortions are size independent and the first-order condition simplifies to $\alpha e^z n^{\alpha-1} = w(1+\tau)$. The following functional form will be used for the remainder of this section:

$$T(n) = n^\tau - 1.$$

Figure 2

Taxes, Employment, and Productivity



NOTE: The relations plotted in the right panel were calculated by assuming the distribution of productivity is exponential with rate parameter 3.2, setting the curvature parameter to 0.66, and normalizing the labor supply to 1.

Panel A of Figure 2 displays this tax function for different values of τ. If $\tau = 0$, there are no distortions. If $\tau > 0$, distortions are size dependent, and larger establishments face higher distortions than smaller ones. For instance, with $\tau = 0.02$, the tax rate for firms with fewer than 20 employees is at most 6 percent, while the tax rate for firms with more than 100 employees is close to 10 percent.

Labor demand can be solved in closed form:

$$n = \left(\frac{\alpha e^{z}}{w(1+\tau)} \right)^{\frac{1}{1-\alpha+\tau}}. \tag{3}$$

With distortions, the link between employment and productivity becomes weaker. More-productive firms are relatively smaller and less-productive firms are relatively larger compared with the baseline scenario with no distortions, as Panel B of Figure 2 shows.

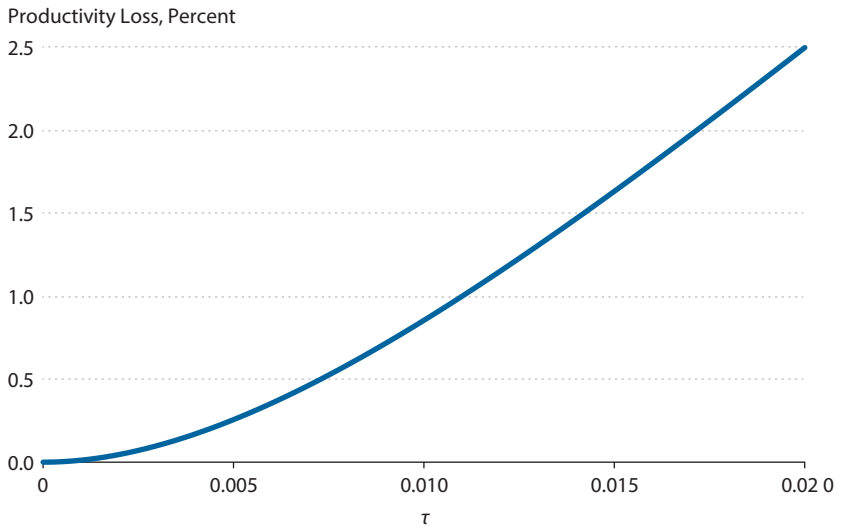
2.3 Aggregates

Using labor demand (equation (3)) and the resource constraint (equation (2)), the equilibrium wage rate w can be expressed as

$$(1+\tau)w = \alpha \left(N^{-1} \int e^{\frac{z}{1-\alpha+\tau}} f(z)dz \right)^{1-\alpha+\tau}.$$

Figure 3

Productivity Losses and Distortions



Using the equilibrium wage rate and inserting labor demand in equation (1), aggregate output, Y, can be characterized in closed form:

$$Y = \left(\int e^{\frac{z}{1-\alpha+\tau}} f(z)dz \right)^{-\alpha} \left(\int e^{\frac{z(1+\tau)}{1-\alpha+\tau}} f(z)dz \right) N^{\alpha}.$$

It is a Cobb-Douglas function in aggregate employment and a productivity index. The productivity index is a weighted average of the productivity level of each firm in the economy.

How should the planner allocate labor across firms to maximize aggregate output, Y? It corresponds to the competitive equilibrium when $\tau = 0$. Then, aggregate output simplifies to

$$Y = \left(\int e^{\frac{z}{1-\alpha}} f(z)dz \right)^{1-\alpha} N^{\alpha}.$$

Further, when $\tau = 0$, the first-order condition of each firm is

$$\alpha e^{z} n^{\alpha-1} = w.$$

The efficient allocation equates marginal products $\alpha e^{z} n^{\alpha-1}$ across all firms. In other words, without distortions, high-productivity firms and low-productivity firms have the same marginal productivity of labor. In the distorted economy, there is dispersion across firms in average labor productivity. Formally, average labor productivity is

$$\frac{y}{n} = \left(\frac{w(1+\tau)}{\alpha} \right)^{\frac{1-\alpha}{1-\alpha+\tau}} e^{\frac{\tau z}{1-\alpha+\tau}}.$$

How large are the output losses due to these distortions? Assume the distribution of productivity is exponential: $f(z) = 3.2e^{-3.2z}$, $\forall z \in [0,\infty)$. The curvature parameter α is set to 0.66. And labor supply is normalized to 1. Figure 3 reports the output loss (in percent) for different values of τ. The upper bound of $\tau = 0.02$ corresponds to a tax rate of about 10 percent for a firm with 100 employees. One can see that distortions can lead to GDP losses of up to 2.5 percent.

3 SIZE-DEPENDENT REGULATIONS IN FRANCE

The previous section shows that the output losses due to distortions can potentially be large. It also raises the question of how large these distortions are in the real world. The rest of the article is devoted to quantifying the effect of a particular distortion. It is a case study of the impact of distortions on productivity that looks at size-dependent regulations in France. Because these regulations depend on a precise threshold, the behavior of firms around the threshold is particularly informative on the effects of distortions.

3.1 Institutional Background

Labor laws in France as well as various accounting and legal rules make special provisions for firms with more than 10, 11, 20, or 50 employees. These regulations, however, are not all based on the same definition of "employee." Labor laws, which are likely the most important, are based on the full-time equivalent workforce. The full-time equivalent workforce is computed as an average of a firm's workforce over the past 12 months, including part-time workers and temporary workers but not trainees or *contrats aidés* (a class of government-subsidized, limited-duration contract workers, which may include people that face "special difficulties" in finding employment, such as the very long-term unemployed or unskilled youth). Hence, it seems fairly difficult for firms to work around the regulations. The main additional regulations as the firm reaches 50 employees are

- possible mandatory designation of an employee representative;
- formation and training of a committee for hygiene, safety, and work conditions;
- formation of a *comité d'entreprise* (works council) that must meet at least every other month, have some office space, and receive a subsidy equal to 0.2 percent of the total payroll and that has both social objectives (e.g., organizing cultural or sports activities for employees) and an economic role (mostly on an advisory basis);
- a higher payroll tax rate, which increases from 0.9 percent to 1.5 percent, to subsidize training (*formation professionelle*); and
- if more than nine workers are fired for "economic reasons," a special legal process must be followed (*plan social*), which increases dismissal costs and creates legal uncertainty for the firm.

This list is not exhaustive, but clearly one would expect the costs of these regulations to be significant. Some of these costs are also difficult to model in a tractable manner. In some cases—in particular, the *comité d'entreprise*—the firm is required to fund additional worker

Table 1

Distribution of French Firms with 40 to 59 Employees

Employees	Fraction	S.E.	No. of Firms	Employees	Fraction	S.E.	No. of Firms
40	8.42	0.28	9,486	50	3.67	0.29	4,140
41	6.72	0.29	7,575	51	2.75	0.29	3,097
42	7.38	0.29	8,311	52	2.78	0.29	3,130
43	6.88	0.29	7,752	53	2.70	0.29	3,040
44	6.81	0.29	7,666	54	2.57	0.29	2,901
45	7.31	0.29	8,239	55	2.51	0.29	2,826
46	6.70	0.29	7,548	56	2.34	0.29	2,638
47	6.96	0.29	7,841	57	2.24	0.29	2,526
48	7.92	0.29	8,916	58	2.37	0.29	2,670
49	8.80	0.28	9,916	59	2.08	0.29	2,344

SOURCE: INSEE, BRN.

NOTE: Fraction is the number of firms for each employment size (40 to 59 employees) divided by the total number of firms with 40 to 59 employees; S.E. is the associated standard error; and No. of Firms is the raw number of firms in each bin.

benefits. To the extent that the process is reasonably efficient, these rules might simply amount to a substitute form of compensation and have limited effects: The higher benefits may allow firms to attract better workers or to pay them less.

3.2 Data

The data come from a panel of firms assembled by the French National Institute of Statistics and Economic Studies (INSEE) that covers the 1994-2000 period. This panel, known as BRN (Bénéfices Réels Normaux), contains employment and standard accounting information on total compensation costs, value added, current operating surplus, gross productive assets, etc. The BRN data include all private companies in France with a sales turnover of more than 3.5 million francs (around 530,000 euros) and liable to corporate taxes under the standard regime. These data also include some other smaller firms. The 3.5 million threshold implies that all firms with more than 30 employees or so are included. Hence, I focus on the threshold at 50 employees, for which the data are essentially exhaustive. When I estimate the model, I remove from the sample firms with strictly fewer than 20 employees. This generates a sample of 44,189 firms that we follow for 7 years, or 309,323 firm-year observations.

Figure 1 plots the distribution of employment for the entire period (1994-2000), truncating at 100 employees. There is clearly a large discontinuity around the threshold of 50 employees. Many surveys reveal "rounding" of employment, but this figure shows the opposite pattern.

Table 1 reports the distribution of firms with 40 to 59 employees. There is a clear drop in the number of firms after 49 employees. For example, there are more than three times as many firms with 49 employees as firms with 51 employees.

4 MODEL APPLICATION

I apply the model of Section 2 to the case of size distortions in France. I replace the smooth function T with a step function to mimic the regulations described in the previous section. Firms face a regulation that requires them to pay a higher proportional tax on wages τ if they currently have more than $\underline{n}$ employees. Formally, if n is greater than $\underline{n}$, a proportional payroll tax τ applies. The proportional tax applies to all employment, including that below $\underline{n}$. For simplicity, there is only one threshold and $\underline{n} = 50$.

4.1 Labor Demand

Take a firm that operates below the threshold. The firm solves the following problem:

$$\pi(z) = \max_{0 \le n < \underline{n}} \left\{ e^z n^\alpha - wn \right\}.$$

There is some value of z, say $\underline{z}$, such that the firm's optimal labor demand equals $\underline{n}$ when the restriction $n < \underline{n}$ is ignored. Note that $n(z)$ is strictly increasing in z absent any restrictions on n. Thus, if z is below $\underline{z}$, the firm will hire $n(z) = \left(\frac{\alpha}{w} \right)^{\frac{1}{1-\alpha}} e^{\frac{z}{1-\alpha}} < \underline{n}$ employees and receive profit $\pi(z) = e^{\frac{z}{1-\alpha}} \left(\frac{\alpha}{w} \right)^{\frac{\alpha}{1-\alpha}} (1-\alpha)$. If z is greater than or equal to $\underline{z}$, the best the firm can do is hire right at (or just below) the threshold due to the restriction $n < \underline{n}$. In this case, optimal labor demand is $n(z) = \underline{n}^-$ (where $\underline{n}^-$ indicates a value just below $\underline{n}$) and the firm receives profit $\pi(z) = e^z \underline{n}^\alpha - w\underline{n}$.

There is some point, though, where a firm's productivity is great enough such that it would be more profitable to operate above the threshold and pay the higher proportional tax on wages τ than to operate just below the threshold and avoid the additional tax. A firm that operates above the threshold solves the following problem:

$$\pi(z) = \max_{n \ge \underline{n}} \left\{ e^z n^\alpha - w(1+\tau)n \right\}.$$

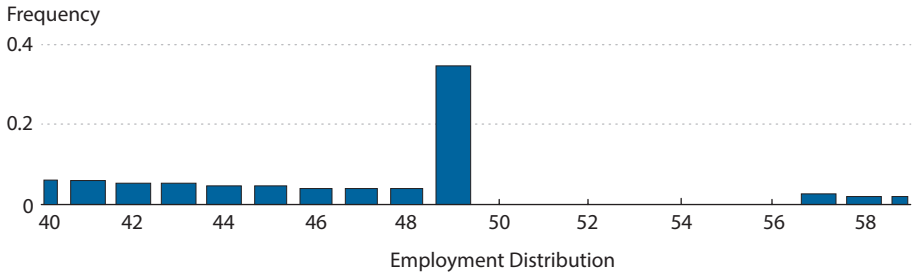
Call the value of z for which a firm is indifferent between operating above and below the threshold as $\bar{z}$. $\bar{z}$ is defined as the solution to

$$e^{\frac{z}{1-\alpha}} \left(\frac{\alpha}{w(1+\tau)} \right)^{\frac{\alpha}{1-\alpha}} (1-\alpha) = e^z \underline{n}^\alpha - w\underline{n}.$$

If a firm has productivity $z > \bar{z}$, it operates above the threshold, selecting optimal employment $n(z) = \left(\frac{\alpha}{w(1+\tau)} \right)^{\frac{1}{1-\alpha}} e^{\frac{z}{1-\alpha}} > \underline{n}$ and receiving profit $\pi(z) = e^{\frac{z}{1-\alpha}} \left(\frac{\alpha}{w(1+\tau)} \right)^{\frac{\alpha}{1-\alpha}} (1-\alpha)$. If z is less than or equal to $\bar{z}$, then the firm operates below the threshold, making decisions as described in the preceding paragraph. It is easy to see that $\bar{z} > \underline{z}$, provided that there is a cost of operating above the threshold $\tau w \underline{n} > 0$.

Figure 4

Distribution of French Firms with 40 to 59 Employees, Model Without Measurement Error



To obtain the formula for the profit of a firm with productivity z, note the following summary of the above results: (i) if $z < \underline{z}$, the firm will earn more profit below the threshold than above since the firm pays lower wages; (ii) if $z > \bar{z}$, the firm will decide to operate above the threshold; and (iii) if $z \in [\underline{z}, \bar{z}]$, it is optimal for the firm to remain just below the threshold. Hence,

$$\pi(z) = \begin{cases} e^{\frac{z}{1-\alpha}} \left(\frac{\alpha}{w}\right)^{\frac{\alpha}{1-\alpha}} (1-\alpha) \text{ for } z < \underline{z}, \\ e^{z} \underline{n}^{\alpha} - w\underline{n} \text{ for } \underline{z} \le z \le \bar{z}, \\ e^{\frac{z}{1-\alpha}} \left(\frac{\alpha}{w(1+\tau)}\right)^{\frac{\alpha}{1-\alpha}} (1-\alpha) \text{ for } z > \bar{z}. \end{cases}$$

For completeness, I also state the employment demand:

$$n(z) = \begin{cases} \left(\frac{\alpha}{w}\right)^{\frac{1}{1-\alpha}} e^{\frac{z}{1-\alpha}} \text{ for } z < \underline{z}, \\ \underline{n}^{-} \text{ for } \underline{z} \le z \le \bar{z}, \\ \left(\frac{\alpha}{w(1+\tau)}\right)^{\frac{1}{1-\alpha}} e^{\frac{z}{1-\alpha}} \text{ for } z > \bar{z}. \end{cases}$$

Overall, firms are distributed above and below the threshold and bunched just below the threshold.

4.2 Firm Distribution

Firm productivity, z, has an exponential distribution with parameter λ. Since log employment is proportional to z, employment follows a Pareto distribution with parameter

Table 2

Economic Parameters

Parameters	Values	Definition
α	0.66	Curvature profit function
σ_{mrn}	0.0324	Measurement error
τ	0.0015	Payroll tax above $\underline{n}$
λ	3.6829	Exponential distribution

Table 3

Moments

	Data	Model
Standard deviation $\Delta \log n$	0.1561	0.1561
Power law coefficent	2.2522	2.2522
Density of firms in each bin		
40-46	0.0718	0.0666
47-49	0.0790	0.0783
50-52	0.0307	0.0341
53-59	0.0240	0.0281

$\beta = \lambda(1-\alpha) + 1$.[3] Figure 4 displays the firm-size distribution implied by the model around the threshold. There is a substantial "hole" in the distribution, with no firms whatsoever between 50 and 55 employees. This is an empirical challenge, because in the data there are many firms with an employment level slightly greater than 49. I attribute the presence of all these firms to measurement error.

5 QUANTITATIVE ANALYSIS

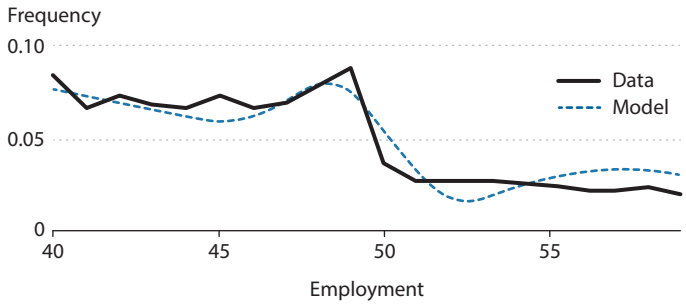
This section proposes a simple calibration of the model and evaluates the aggregate effect of the distortions described in the previous section.

5.1 Calibration

Table 2 lists the calibrated values of the parameters. I incorporate measurement error in (log) employment. Formally, measured employment is equal to the product of the true value and a lognormal error term, with standard deviation σ_{mrn} and a mean equal to unity. Measurement error also helps capture model misspecification, which can take several forms. First, the measure of employment is the arithmetic average of the number of employees at the end of each quarter. This is the relevant measure of employment for some but not all of the regulations. Some regulations apply based on employment measured as the full-time equivalent

Figure 5

Distribution of French Firms with 40 to 59 Employees



NOTE: The distributions are normalized by the total number of firms with 40 to 59 employees.

workforce, and others apply if there are more than 50 employees in the firm for more than 12 months. Second, measurement error also captures adjustment costs or search frictions, which lead to an imperfect control of the size of the workforce.

The wage rate is normalized to 1. The curvature parameter α is set to 0.66. This last parameter is a reduced form for the labor share, decreasing returns to scale, and the elasticity of demand.

Table 3 lists the target moments: (i) the volatility of growth in employment and the slope of the power law and (ii) the distribution of employment around the threshold, as approximated by the density of firms with 40 to 46 employees, 47 to 49 employees, 50 to 52 employees, and 53 to 59 employees.[4] The rational for the first group of moments (i) is that I want the model to be consistent with key features of firm dynamics. The rationale for the distribution of employment around the threshold (ii) is that I want to reproduce well the discontinuity in the firm-size distribution, which is the *prima facie* evidence that the regulation matters.

Table 3 evaluates the fit of the model. Overall, the data are consistent with a small but significant proportional payroll tax of 0.15 percent. This value is lower than the taxes that are actually set in the law, which presents an apparent puzzle. One possible interpretation is that some of these regulations are indeed not as costly as they appear and represent benefits that are valued by workers. The model requires a measurement error of around 3 percent, or on average two workers around the threshold. In spite of its parsimony, the model is able to reproduce reasonably well all the targeted moments and, in particular, the discontinuities in the distributions. A graphical illustration is provided in Figure 5.

5.2 Policy Experiments

I use the calibrated model to infer the aggregate effect of the regulation on productivity. The results are reported in Table 4. From the point of view of a social planner, the regulation misallocates labor across firms and hence reduces aggregate productivity. I now perform the same calculation as in Section 2. Precisely, I ask how much of an increase in output can be

Table 4

Policy Experiments

Experiment	Gains (%)
Benchmark	0.30
Apply regulation to all firms	–2.50
Apply regulation to firms above 75 employees	0.06

NOTE: Gains are relative to the scenario when the regulation is applied to firms with more than 50 employees.

obtained, holding total employment constant, by reallocating labor across firms.

The gain in total output, holding total labor constant, is 0.30 percent, which is significant. Second, one might ask how much of the efficiency gain can be achieved by extending the threshold to 75 employees rather than 50. The answer is, not much: The gains are reduced to 0.06 percent. Third, the motivation for the phase-in of the regulation at 50 employees is that it is too costly to impose the compliance cost on small firms. I evaluate this argument by considering the counterfactual: What would happen if all firms were subject to the regulation? It would reduce output by 2.5 percent. It is safe to say, then, that applying the regulation to all firms would be quite costly, which suggests that the phase-in is perhaps not such a bad policy.

6 CONCLUDING REMARKS

This article studies a particular regulation that clearly distorts the firm-size distribution, leading to an obvious misallocation of labor—a channel that has been emphasized in the recent literature. The model fits the size-distribution discontinuity around the threshold well. Removing the regulation leads to an increase of output close to 0.3 percent, holding employment fixed.

These results suggest that size distortions have a fairly moderate aggregate impact. What can explain the small benefits in Section 5 with the potentially large benefits in Section 2? Further research is needed to conclusively address the issue. There are at least three reasons to believe the effects could be bigger. First, this is just one example of distortions among many others. France is characterized by, for instance, stringent employment-protection legislation, and more than 15 percent of workers are affected by minimum wage increases. Second, the model abstracts from the notion of match quality and assortative matching. It might be missing some of the negative effects of the regulation. For instance, some talented workers might be stuck in small unproductive firms because of the regulation but would contribute more to aggregate output by working in larger firms. Last, the proposed framework is static and there may be dynamic effects of these policies that are missed by the current analysis. These distortions reduce the value of investment and the value of entry. ■

NOTES

[1] See Decker et al. (2014).

[2] This formulation is equivalent to a linear production technology, where firms have some market power. In this case, $\alpha < 1$ is equal to the inverse of the demand elasticity.

[3] Many studies have found that this is a good approximation of the firm-size distribution. See Gabaix (2016) for a review of power law in economics.

[4] The distribution is the number of firms in each bin, divided by the length of the bin (7 or 3), and further divided by the total number of firms with 40 to 59 employees.

REFERENCES

Decker, Ryan; Haltiwanger, John; Jarmin, Ron and Miranda, Javier. "The Role of Entrepreneurship in US Job Creation and Economic Dynamism." *Journal of Economic Perspectives*, Summer 2014, *28*(3), pp. 3-24; https://doi.org/10.1257/jep.28.3.3.

Gabaix, Xavier. "Power Laws in Economics: An Introduction." *Journal of Economic Perspectives*, Winter 2016, *30*(1), pp. 185-206; https://doi.org/10.1257/jep.30.1.185.

Gourio, François and Roys, Nicolas. "Size-Dependent Regulations, Firm Size Distribution, and Reallocation." *Quantitative Economics*, July 2014, *5*(2), pp. 377-416; https://doi.org/10.3982/QE338.

Lucas, Robert E. Jr. "On the Size Distribution of Business Firms." *Bell Journal of Economics*, Autumn 1978, *9*(2), pp. 508-523; https://doi.org/10.2307/3003596.