Money and Banking

Jón Steinsson*
University of California, Berkeley

January 17, 2025

The past few centuries have seen a remarkable transformation of our monetary system. For many centuries, the primary form of money in much of the world was gold, silver, and copper coins. Other objects such as cowrie shells, coca beans, and cattle were also used as money. Monetary systems based on these objects faced various challenges as we discussed in some detail in the last chapter. Perhaps the most basic challenge was simply their bulk. Another was their vulnerability to theft. A third was their uneven size, weight, and quality (fineness in the case of coins). These characteristics resulted in substantial transactions costs, which gave traders, merchants, and the population in general a strong incentive to invent better forms of money.

For much of history monetary innovation centered on finding ways to make it less costly for people to make payments. In other words, monetary economics was largely about the development of a more efficient payment system. Innovations that made money more uniform and its value more easily verifiable were an important strand of this process. The invention of coins in antiquity was a major step forward, as were subsequent innovations such as milled edges on coins and token coins.

In this chapter, we focus on monetary innovations that have allowed people to shift away from using coins as a medium of exchange and towards using paper documents and ledger entries (either physical or electronic) as media of exchange. Over the past millenium, many different types of paper documents have been developed

*I would like to thank Julian Alcazar, Michael Bordo, John Cochrane, Colin Drumm, Barry Eichengreen, Thomas Eisenbach, Fumiko Hayashi, Kenneth Isaacson, Stephen Luck, Emi Nakamura, Larry Neal, Ricardo Reis, Gary Richardson, Jared Rubin, Veronica Santarosa, Nathan Sussman, Alan Taylor, Richard von Glahn, Chenzi Xu, and Yao Zeng for valuable comments and discussions. I thank Paul Moloney and Roger Tufts for sharing data on bank capital. First posted in January 2025.

for this purpose. Some of the most important are bills of exchange, bank notes, and checks. Ledger-based payment systems also have a long history, as we will see. But the rise of electronic payment systems over the past half-century – e.g., credit cards, debit cards, ACH transactions, and wire transfers – have dramatically increased the importance of ledger-based payment systems at the expense of bank notes and other paper documents. In the not too distant future, our monetary system may be fully electronic and paper money (not to mention coins) may have become a historical relic.

The development of a paper and ledger-based monetary system has been fundamentally connected with the development of banking. Paper monetary instruments are typically issued by or drawn on banks. Payments in ledger-based payment systems are typically settled with transfers of funds between accounts at banks.

It is sometimes underappreciated that one of the principle roles of banks in the economy, both historically and today, is to facilitate payments. The quantity of payments in modern economies is astronomically large and the smooth functioning of the economy relies heavily on the payment system working efficiently without fail. The highly interconnected nature of the economy implies that even minor glitches in the payment system can quickly snowball into major problems (if A can't pay B, then B can't pay C, and C can't pay D, etc.). In this chapter we discuss how the modern payment system works, and how it developed over the last several centuries.

The shift from a coin-based payment system to a paper and ledger-based payment system has lowered transactions costs enormously and thus (arguably) contributed a great deal to economic growth. But this shift has also exposed the economy to new risks. The most important such risk arises because the vast majority of paper and ledger-based forms of payment are bank liabilities.

While banks are central to the functioning of our payment system, they are fundamentally fragile institutions. They are prone to "runs" that can turn into "panics", and if these panics are not stopped they can lead the financial system and payment system to collapse with colossal consequences for output and employment. An important aim of this chapter is to explain why banks are fragile, how this fragility results in the risk of "runs" and "panics", and what economic policies are needed to avoid the occurrence of such financial crises.

For several hundred years after the introduction of paper- and ledger-based forms of payment, all such forms of payment were directly or indirectly "backed" by specie. In other words, they were directly or indirectly promises by banks to pay a certain quantity of gold or silver. Over the past one hundred years, a further

transformation has occurred: the link between money and specie has been gradually severed. Today our monetary system is a purely *fiat* system. This means that the monetary instruments we use today are not claims on any object of intrinsic value (such as gold or silver).

The notion of fiat money introduces an added level of conceptual complexity into monetary economics. To keep things as simple as possible in this chapter, we postpone our main discussion of fiat money until chapter XX [Monetary policy chapter]. In this chapter we (for the most part) maintain the assumption that all forms of money are convertible into specie. Since this was (for the most part) true historically until the 20th century, describing the system in this way also allows us to easily discuss various historical episodes.

1 Credit as Money

The most fundamental change in our monetary system over the past millenium has been the increased use of various types of credit instruments as money. Today, most forms of money are a credit instrument – i.e., someones liability. When you pay for something, you don't hand over an object of intrinsic value (such as a cow or a piece of gold). Rather you swipe or tap a card, pay with your phone, make an online payment, or (increasingly rarely) hand over a special piece of paper – a bank note or a check. In all of these cases, you are using a credit instrument as a form of payment.

In most cases, the credit instrument you use to make payment is a bank liability. For example, when you make a card payment, you are paying with bank deposits (a bank liability): the seller receives the payment in the form of an increased balance in their bank account and the balance in one of your bank accounts (checking or credit card account) decreases by the amount of the payment. Our monetary system is designed to make it easy for you to make payments using bank deposits.

Prior to the rise of electronic payments, it was more common for payments to be made by the transfer of a bank note. Bank notes are another form of bank liability. Early in the development of paper money, private banks issued bank notes. These notes were their liabilities: the banks in question promised to redeem their notes in specie. In the 19th and 20th centuries, governments granted central banks a monopoly on the issue of bank notes. This means that bank notes today are liabilities of the central bank.

The development of credit instruments as media of exchange was motivated by

a desire to lower the cost of transferring money from person to person. Credit instruments have obvious advantages in this regard. In particular, they weigh very little and take up very little space. (In the case of electronic ledger entries, they weigh nothing and take up no space.) This contrasts with gold and silver. Also, the use of credit instruments that have little intrinsic value as money allows society to economize on the use of costly objects (such as cattle or gold) as money. Furthermore, credit instruments can be designed to guard against theft: bills of exchange and checks can be made payable to a particular person, while electronic payment systems are typically designed to be highly (if not completely) secure.

But credit instruments also raise new challenges and exacerbate some old challenges. Two such challenges are counterfeiting and risk of default. Paper money is very cheap to produce. This implies that the incentive to counterfeit paper money is very large. Ever since the first issue of paper money in China nearly a thousand years ago, governments have tried hard to prevent counterfeiting (with varying success). Early paper money in China was printed using multiple colors on high quality paper and included stern warnings about counterfeiting. Modern bank notes include a number of security features such as a serial number, watermarks, a security thread, a 3D security ribbon, microprinting, and color shifting ink.

Risk of default is another potentially serious problem when credit instruments are used as money. Risk of default makes the value of the instrument uncertain, which raises transactions costs when the instrument is being used for payment. For a credit instrument to be a low cost medium of exchange, its value must be known and agreed upon by all potential users. In other words, money should ideally have a "no questions asked" property; it should be accepted immediately without the seller needing to research its value.

In the United States today, a dollar bill is worth a dollar. This may seem obvious and trivial. But this has not always been the case and need not be the case. In the free banking era in the United States (1837-1863), bank notes circulated at a discount that fluctuated and was uncertain. This created transaction costs as sellers needed to research the value of the money a buyers offered as payment. Imagine standing in a checkout line at the supermarket and having a minute or two added to each customer's checkout experience because the employee working the register needs to look up the value of the money offered by the customer. This was the state of affairs in the United States (with some exaggeration) prior to the creation of a "uniform national currency" in 1864.

There are two basic types of instruments that form the backbone of monetary

exchange with credit instruments. The first of these is the "I owe you" (IOU) instrument. Bank notes and bank deposits are IOUs. In both cases, it is a bank that is promising to pay on demand. Bank notes were traditionally promises to pay specie, while bank deposits have typically been a promise to pay bank notes.

The other type of instrument critical to monetary exchange is the "please pay them" instrument. The check is a "please pay them" instrument. The writer of a check requests that their bank pay whoever the check is made out to the amount written on the check. An earlier example of a "please pay them" instrument is the bill of exchange. As we note above, "please pay them" instruments have long been critical to trade because they are less subject to theft (can be made out to a specific person). This implies that they can be used to make large payments over long distances at low cost.

An important difference between an IOU and a "please pay them" is that an IOU is a binding contract, while a "please pay them" is a request. In the case of a check, the bank may refuse the request. The bank does this if the writer of the check does not have a large enough balance in their account. In this case, we say that the check "bounces". The same is true of a bill of exchange. If a check bounces, the seller must go back to the buyer and ask for an alternative form of payment. This can be difficult as the seller may have a hard time locating the buyer. This problem is a principle source of fraud associated with checks and other "please pay them" instruments.

Electronic transactions such as debit and credit card transactions involve "please pay them" requests. The main difference between these and checks or bills of exchange is that they do not involve paper documents and (increasingly) they are accepted and the transfer is made virtually instantaneously. Our modern ways of making payments may seem fundamentally different from something as antiquated as a check. But the fundamentals of the transaction are actually the same. Only form and speed have changed.

2 The Modern Payment System

Perhaps the most underappreciated marvel of the modern economy is the payment system. The quantity of payments made in the modern economy is staggering. The payment system processes all these payments so seamlessly and efficiently day in and day out, year in and year out, for years on end that we tend to take its functioning for granted. In fact, it works so well, that the problems it is designed to solve

have become largely invisible and therefore hard to notice and discuss in a modern context.

To better grasp the problems that our modern hyper-efficient payment system is designed to solve, later in this chapter we will chart the evolution of payment methods over the last millenium. But first it is useful to describe the principle pillars of the payment system as it exists today. For concreteness, we focus mostly on the U.S. payment system. Payment systems in other advanced economies are similar if not identical.

Recent decades have seen a great deal of innovation in the payment system. Today households and firms have access to a range of different payment methods. These payment methods vary in their characteristics and therefore serve different parts of the economy. Most everyday transactions by consumers are made using retail payment methods. In the Unted States, these include credit cards, debit cards, Automated Clearinghouse (ACH) transactions, checks, and cash. More recent additions include Apple Pay, Google Pay, Venmo, PayPal, Zelle, and FedNow. Businesses use a combination of retail and wholesale payment methods including wire transfers, checks, credit cards, and ACH transactions.

The most important retail payment systems judged by value of transactions (as of this writing) are the ACH network, the card networks (principally VISA and Mastercard), and checks. Card transactions are the predominant method of payment for point-of-sale (POS) transactions such as paying at the grocery store or other retail outlets. They are also the predominant method of payment for online consumer transactions. The ACH network is used for payroll, utility bills, mortgage payments, social security payments, rent payments, and a great variety of business-to-business payments. In 2021, the total value of card payments in the United States was $9.4 trillion (40% of GDP), the total value of ACH payments was $91.9 trillion (almost 400% of GDP), and the total value of payments by check was $27.2 trillion (115% of GDP). For comparison, $730 billion (3% of GDP) of cash was withdrawn from ATMs. (These estimates are from the Federal Reserve's Payment Study.)

While the various payment methods mentioned above differ in important ways, they mostly share the same common core structure. The execution of a payment typically involves six parties. First, there are the payer and the payee. Then, there are the payer's bank and the payee's bank. Finally, there is a settlement bank (typically the central bank) and the network operator. The execution of a payment can be broken into two stages. The first stage is the clearing stage. This stage involves a number of payment messages being sent between the parties involved. The second

stage is the settlement stage. This stage involves funds being transferred between different parties.

2.1 ACH Transactions

The easiest way to explain how payments are cleared and settled is to consider an example. Let's start by considering ACH transactions. There are actually two distinct types of ACH transactions. The first type is one where the payer initiates the payment ("push" transactions). These are called ACH credits. An example of such a payment is a payroll transaction initiated by an employer (the payer) whose employee has set up direct deposit. The second type is one where the payee initiates the payment ("pull" transactions). These are called ACH debits. An example of such a payment is a utility bill payment. Many households have set up autopay for their utility bills. In this case, it is the utility company (payee) that initiates the transaction by requesting that it be paid.

Figure 1 depicts the flow of payment messages and payments for an ACH credit transaction. The dashed lines represent payment messages, while the solid lines represent payments. The transaction begins by the payer contacting their bank and requesting that the payment be made (arrow 1). The payer's bank will typically debit this amount from the payer's account immediately (arrow 1'). Since the ACH network is not a real-time payment system but rather processes payments in batches several times a day, the payer's bank will typically collect many such requests from its customers and then send a batch of them to their ACH operator (arrow 2). (There are two ACH operators in the U.S.: the Federal Reserve and The Clearing House Payments Company.)

The ACH operator collects such payment requests from various banks and batch processes them. This means that it sorts the requests and sends them on to the banks of the payees (arrow 3). If certain criteria are met (e.g., routing and account information is accurate), the payee's bank accepts the payment, notifies the payee, and credits them the amount in question (arrows 4 and 4'). Once this process is complete, the payment is said to have "cleared". Clearing of ACH credits takes one business day (although the system also offers "same day" ACH processing for an extra fee).

ACH debits are a little more complex than ACH credits in that the payee (who initiates the transaction in that case) usually does not receive the funds until several days after the transaction clears. The reason for this is that the payer may have

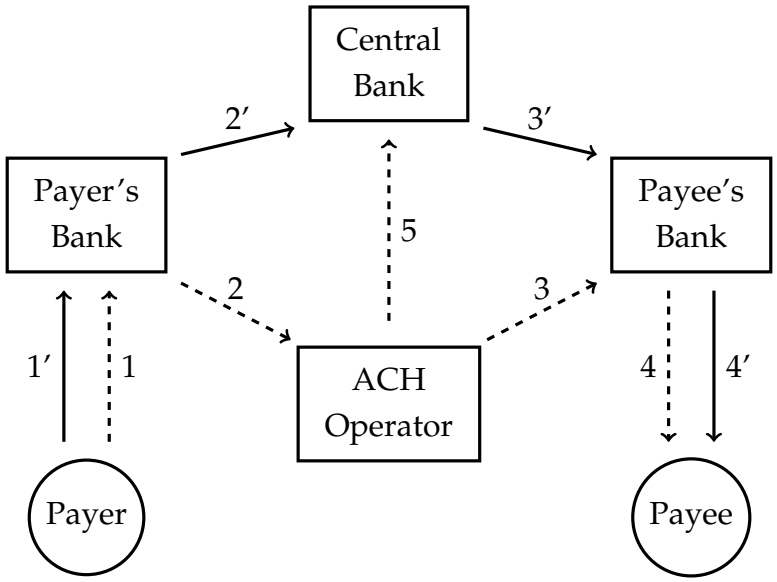

Figure 1: Clearing and Settlement of ACH Credit Payment

Note: The dashed arrows represent payment messages. Messages 1 through 4 are part of the clearing stage, while message 5 is part of the settlement stage. The solid arrows represent the flow of funds.

insufficient funds or the account in question may no longer exist. If this is the case, the payer's bank will initiate a "return" transaction to undo the initial transaction. This is analogous to a check "bouncing". The payer's bank has two days to send a return transaction. The payee's bank usually does not deposit funds into the payee's account until these two days have passed. This means that funds usually take three days to arrive in the case of ACH debits.

The second phase of an ACH transaction is settlement. Each ACH transaction triggers three financial obligations. First, by accepting the payment, the payee's bank is obliged to pay the payee. Second, the payer's bank is obliged to pay the payee's bank. Finally, the payee is obliged to pay their bank. Settlement is the discharge of these obligations. Notice that in the case of an ACH credit two of these three financial obligations are settled at the time of clearing: the payer's account is debited at their bank and the payee's account is credited at their bank. In the case of an ACH debit, only one financial obligation is settled at the time of clearing: the payer's account is debited (while the payee's account is not credited until two days later).

The final step of the settlement stage is for the financial obligation between the payer's bank and the payee's bank to settle. Interbank settlement typically involves a settlement bank at which both the payee and payer banks have accounts. For

domestic interbank transactions, this settlement bank is often the central bank. In the United States, banks have "master" accounts at the Federal Reserve. Interbank settlement of ACH transactions are settled with credits and debits in these accounts. For the ACH credit transaction depicted in Figure 1, the master account of the payer's bank at the Federal Reserve is debited, while the master account of the payee's bank is credited (arrows 2′ and 3′ in Figure 1).

2.2 Card Transactions

Card transactions work in a similar manner to ACH transactions. When a customer pays for a product with a card (credit card or debit card), the merchant sends information about the transaction to it's bank. The merchant's bank forwards this information to the card network (e.g., VISA or Mastercard) which forwards it on to the customer's bank. The customer's bank either approves or denies the payment request. This information is relayed back to the merchant in real time and if the payment request is approved, the customer's bank places a hold on the funds in the customer's account. The merchant does not receive funds immediately. Rather they collect transactions over some period of time (usually a day) and then sends them to the card network for batch settlement.

The card networks collect such batches over a day and tally up how much each bank owes or is owed on net due to all card transactions on that network. Some banks primarily issue cards to consumers. These banks will typically owe money on behalf of their customers. Other banks primarily service merchants. These banks are typically owed money due to sales made by their merchant customers. Many larger banks engage in both lines of business. A fraction of the transactions involving their customers will net out.

To settle the balances owed by and owed to the banks at the end of the day, each card network works with a (large) bank that acts as a settlement bank for that card network. For banks that are owed funds, the card network requests that its settlement bank transfer funds to that bank. For banks that owe funds, the card network requests that the bank transfer fund to the settlement bank. In most cases, these interbank transfers are "wire transfers" on the Fedwire system (see discussion below) but they may in some cases be ACH transactions.

Merchants typically get paid once the interbank transaction has settled. Some merchants work with specialized payment processing companies to handle card transactions. In these cases, the payment processor must transfer the funds it re-

ceives on behalf of that merchant to the merchant's bank. This can add an additional delay of up to a few days. (In some cases, these payments are ACH payments.) The accounts of card users are typically credited at the end of the day.

In the United States, the fees associated with card transactions are very large. The typical fee is on the order of 2.25% (i.e., merchants get paid 2.25% less than the nominal value of the transaction). The largest part of this fee is the so-called interchange fee, which is a payment from the bank of the merchant to the bank of the card user. A typical value of this fee in the United States is 1.75%. Part of this fee stays with the issuing bank. But a large portion of it funds rewards programs (e.g., 1% cash back). These high fees raise the prices of goods and services in the United States. In addition, interchange fees and rewards programs for cards available to high income households tend to be larger than those for cards available to low income households. This fee/reward structure therefore benefits high income households at the expense of low income households. (See Wang (2024) for a more detailed discussion.) In much of Europe, interchange fees are capped by regulators at much lower levels.

Today, many consumer payments are made using ApplePay and Google Pay. But as of this writing, these services are simply digital wallets: they just store one's card information. The transaction that results is a card transaction using the traditional card networks (e.g., VISA and Mastercard). Venmo and Paypal are other newcomers to the payment space. They offer both digital wallet services and peer-to-peer payment networks. For example, Venmo offers users three ways to pay: using a card, using their bank account (i.e., an ACH transaction), or using their Venmo balance. When Venmo users receive funds, the funds get added to their Venmo balance. Importantly, since Venmo is not an actual bank (at the time of this writing) Venmo balances are not covered by deposit insurance and Venmo does not have direct access to emergency borrowing from the Federal Reserve in case of a run on its balances. Venmo users can transfer balances to their bank account (by ACH transaction).

2.3 Large-Value Payment Systems

The payment systems discussed above are retail payment systems designed to handle relatively low value transactions. The average value of a card transaction in the United States in 2021 was $60, while for ACH the average value was about $2,500 and the average check was about $2,400. Large-value payment systems are designed – as the name suggests – to handle larger value transactions. In the United

States, the two principle general-purpose large-value payment systems are Fedwire and CHIPS. In 2021, the total value of Fedwire transactions was $991 trillion (about 4,200% of GDP). The average value of a Fedwire transaction was $4.85 million. The total value of CHIPS transactions in 2021 was $449 trillion (about 1,900% of GDP), with an average transactions value of $3.51 million.

The basic functioning of Fedwire is similar to that of other payment systems: a transaction involves a payer, a payee, the payer's bank, the payee's bank, and the Federal Reserve, which acts both as the settlement bank and the network operator. Suppose a payer would like to "wire" funds to a payee. They ask their bank to initiate the wire transfer. The payer's bank debits the payer's account and sends a wire transfer request to the Fedwire system. In most cases, as long as the request is correctly formatted with correct account information, the request is automatically accepted. Funds are then "transferred" from the payer bank's master account to the payee bank's master account at the Federal Reserve. More precisely, the Fedwire system debits the payer bank's master account and credits the payee bank's master account. The payee bank is notified of the credit and what account the funds are meant for. It then credit's the payee's account.

An important feature of the Fedwire system is that it is a real-time payment system. This means that settlement of payments occurs virtually instantaneously. For this reason, Fedwire is used for highly time-sensitive payments. Think of a large business deal where lawyers of two parties are sitting in a conference room waiting for funds to be transferred from the one party to the another before other parts of the deal can be finalized or undertaken. Fedwire is also used for all manner of interbank transfers, which tend to involve very large amounts. The counterparts of Fedwire in other countries include TARGET and EURO1 in the Euro Area, CHAPS in the United Kingdom, BoJ-NET in Japan, and LVTS in Canada.

While large-value payment systems are typically "fast" they are not always real-time systems. CHIPS in the United States is an example of a 'deferred settlement' large-value payment system. Deferred settlement systems have the important advantage that they can take advantage of 'netting' of different payments. Consider, for example, an end-of-day settlement system. Over the course of a day, the customers of a particular bank that is a part of the system make large numbers of payments, but they also receive large numbers of payments. Each time a customer makes a payment the position of their bank in the system falls, while each time a customer receives a payment the position of their bank in the system rises. In a deferred settlement system, the bank does not need to settle each of these payments

one-by-one. Rather many of them net out over the day and the bank only needs to settle its net position at the end of the day.

The fact that many payments net out in a deferred settlement system implies that the banks in the system need not hold as large balances in the system for the purpose of settling payments as they would in a real-time "gross" settlement system. Deferred settlement systems thus economize on 'liquidity.' A drawback of such a system is that it introduces settlement risk. Something may happen within the settlement period that prevents a bank from settling its net position at the end of the settlement period. If this happens, other banks in the system will not receive the funds they expects and may run into difficulties of their own.

If this risk is large enough, the banks in the system may not release funds to their customers until the end of the settlement period, which may seriously reduce the usefulness of the system. CHIPS and other deferred net settlement systems try to balance these advantages and drawbacks, i.e., they try to take advantage of as much netting as possible without introducing too much settlement risk. Prior to 2000, CHIPS was an end-of-day settlement system. In 2000, it introduced an algorithm for within-day settlement of some payments to reduce settlement risk.

In recent years, a number of real-time retail payment networks have been developed in the United States. These include FedNow developed by the Federal Reserve and RTP developed by the Clearing House Payments Company. Over time, it is likely that these services will grow in importance at the expense of ACH and checks. This will mean that it will become easier for individuals and businesses to make instant payments and transfer funds between accounts at different institutions in real time at dramatically lower cost than in the past.

2.4 International Payments

Since most payment systems are confined to a particular country, international payments pose particular issues. We can most easily illustrate this with an example. Consider a Norwegian importing firm that would like to pay a U.S. supplier for goods that it is importing to Norway. This payment is depicted in Figure 2. The Norwegian importer initiates the transaction by asking its bank in Norway to handle the payment. The importer's bank then debits the importer's account (arrow 1 in Figure 2).

The key complication is that the importer's bank in Norway does not have direct access to the U.S. payment system. It, therefore, cannot directly handle the

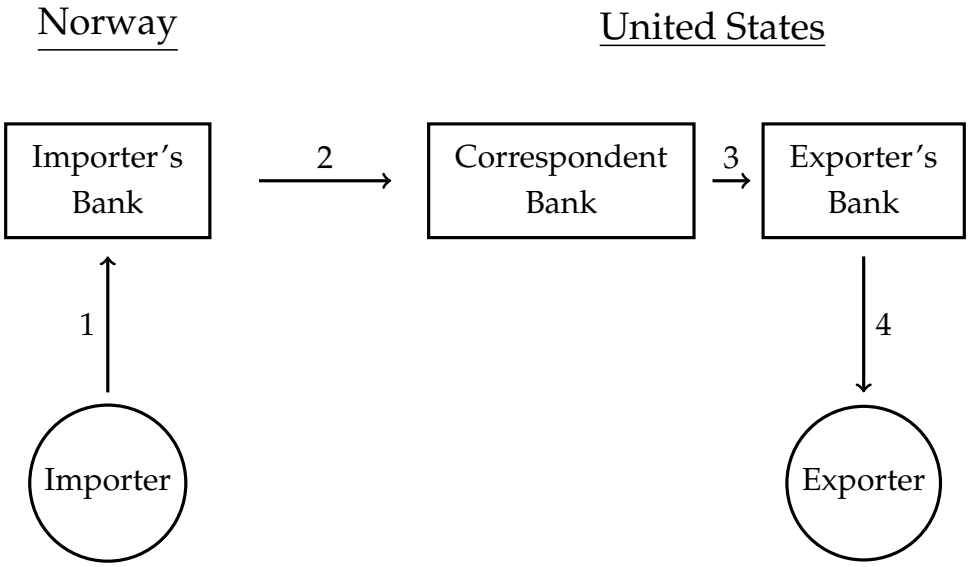

Figure 2: An International Payment

payment. Rather, it must have a correspondent bank in the United States, which it asks to make the payment on its behalf. This request is typically sent through the SWIFT messaging system. The correspondent bank debits the account of the importer's bank in its books by the amount of the payment (arrow 2 in Figure 2). The correspondent bank then initiates a payment in the U.S. payment system (for example, a Fedwire payment) to the exporter (arrows 3 and 4 in Figure 2).

Notice that in this transaction money does not "move" from Norway to the U.S. to pay for the goods. Rather, the balance in the account of the importer's bank at its correspondent bank in the U.S. falls by the amount of the payment. This is true quite generally for international transactions. In an earlier era, gold might be shipped between countries to pay for goods (more on that below). Today the physical shipment of "money" across borders is insignificant. Virtually all international payments are made through accounts at correspondent banks.

In many cases, international payments are more complex than the one depicted in Figure 2 because the payment needs to go through a chain of correspondent banks. For example, a payment to a smaller country may involve a correspondent bank in London or New York which itself has a correspondent bank in the small country.

Foreign exchange transactions raise additional complications. While banks may offer their customers foreign currency accounts and may even have some foreign currency bills available for tourists, larger value foreign currency transactions in-

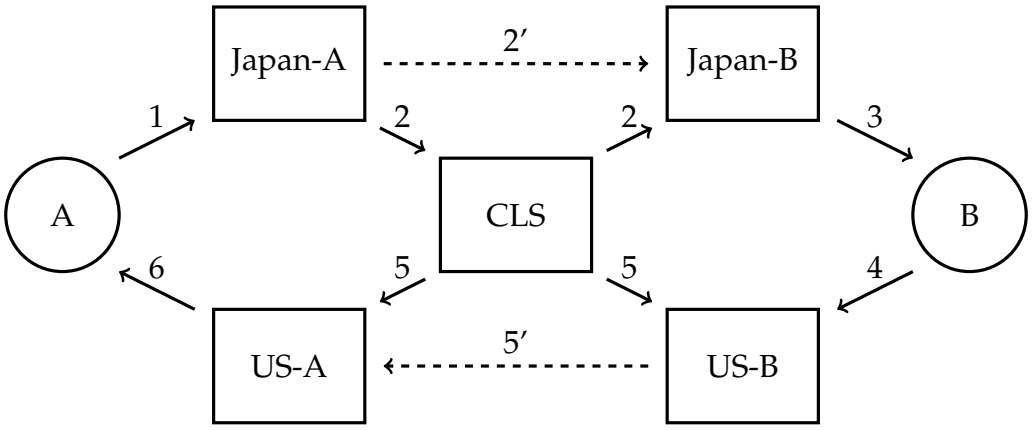

Figure 3: Foreign Exchange Transaction

volve interbank settlement in two different payment systems. Consider, for example, a foreign exchange transaction where two banks in London – let's call them bank A and bank B – exchange U.S. dollars for Japanese yen. For concreteness, let's suppose bank A is buying dollars and selling yen. Settlement of this transaction is depicted in Figure 3.

Let's first consider the traditional way such a transaction was settled. On the settlement day (which might be a day or two after the trading day) bank A would instruct its correspondent bank in Japan (Japan-A in Figure 3) to send yen to bank B's correspondent bank in Japan (Japan-B). This would typically be handled as a transaction on Japan's large-value payment system (BoJ-NET) where Japan-A's account was debited and Japan-B's account was credited (arrow 2′ in Figure 3). Once Japan-B receives the funds, it debits bank B's account and informs bank B that the funds have arrived.

On that same day, bank B would instruct its correspondent bank in the U.S. (US-B) to send dollars to bank A's correspondent bank in the U.S. (US-A). This transaction would typically occur through FedWire or CHIPS (arrow 5′). Once US-A receives the funds, it would debit bank A's account and inform bank A that the funds have arrived.

The risk associated with this traditional settlement method is that one leg of the trade may be settled before the other. Suppose bank A sends the yen early in the day, while bank B plans to send the dollars later in the day. (Recall that Japan is in a time zone that is 13 or 14 hours ahead of New York depending on the time of year.) This discrepancy in timing leaves bank A exposed to settlement risk until bank B settles its leg of the transaction. This particular type of settlement risk is often referred to as

Herstatt risk on account of a German bank that was closed on June 26 1974. At the time it was closed, some of its foreign exchange counterparties had sent it Deutsche marks but had not yet received the U.S. dollars Herstatt had agreed to supply them with in return.

While settlement risk for any given foreign exchange transactions is small, the overall risk to the financial system is potentially very large because of the enormous volume of foreign exchange trading (several trillion dollars per day). To mitigate this risk, a specialized bank named CLS bank (which stands for Continuous Linked Settlement) was set up to handle the settlment of foreign exchange transactions. With the advent of CLS bank, Japan-A will instruct CLS bank to pay yen to Japan-B and US-B will instruct CLS bank to pay dollars to US-A. CLS bank will perform these two payments simultaneously and thereby implement "payment versus payment" for foreign exchange transactions (Lindley, 2008). If one side of the transaction fails, CLS bank will ask its liquidity provider in the relevant currency to pick up that side of the trade and otherwise return the funds on the other side of the trade. As of this writing, CLS bank handles transactions in 18 currencies and settles about 6.5 trillion dollars of transactions per day.

3 The Bill of Exchange

The transition from a system where coins served as the principle form of money to our modern system of largely electronic ledger-based payments took hundreds of years to occur. The transition involved many innovations that helped solve different problems. The next few sections discuss some of this history with an eye towards the key innovations.

A particularly important early innovation that allowed for more efficient long-distance trade was the *bill of exchange*. The use of coins as a means of payment was particularly cumbersome in the case of long-distance trade since transporting large amounts of bullion long distances involved substantial hassle and risk. To avoid this, merchants might sell goods in a town for local currency and then buy other goods in the same town. However, the need for a double coincidence of wants of sorts – an attractive opportunity to sell and an attractive opportunity to buy in the same town – was detrimental to efficient trade. The bill of exchange allowed long-distance trade to occur with vastly less transportation of specie.

The bill of exchange is often said to have been invented by Italian merchants

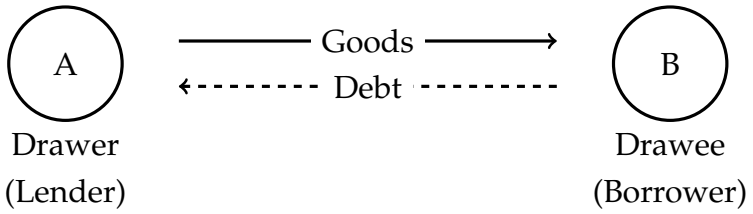

Figure 4: A Simplified Bill of Exchange

Note: The arrows depict flows at the time the bill of exchange is signed. When the bill matures, the flows reverse: cash flows from B to A and B's debt is canceled.

in or around the 13th century. However, similar instruments have been used in various parts of the world for much longer. These include the hawala and safatij in the Middle East, the hundi on the Indian sub-continent, and so-called flying cash in Tang Dynasty China (618-907 CE). There are even passages in the Bible (Exodus) that have been interpreted as describing the use of instruments that might be called bills of exchange. The European bill of exchange that emerged in Northern Italy around the 13th century did, however, have some distinctive characteristics relative to these earlier instruments. Given the importance of the bill of exchange and the fact that the terminology associated with it is quite involved, it is worth describing in some detail.

At its core, the bill of exchange is simply a document requesting that one party pay another party a specific sum of money. Before describing the typical ways in which bills of exchange worked, let us consider a simplified setting where a merchant sells goods to another merchant and accepts payment in the form of a promissory note (i.e., accepts that the buyer will owe them money for the goods to be paid later). We refer to this promissory note as a simplified bill of exchange. In this case, the seller would draw up a bill and presents it to the buyer. If the buyer agrees with the content of the bill, they would sign it. This act is often described as the buyer accepting the bill. Once accepted, the bill becomes a binding legal obligation.

Notice, that it is the seller who draws up the bill – and is referred to as the "drawer" – and the buyer (the person who owes money) who accepts the bill – and is referred to as the "accepter" or "drawee." The seller is said to draw a bill on the buyer. And the buyer is said to buy goods against a bill of exchange. This situation is depicted in Figure 4: A is the seller and thus also the drawer and lender, while B is the buyer, the drawee, and the borrower. A bill of exchange is sometimes also referred to as a draft or acceptance bill. (The word for bill of exchange in Latin is cambium.)

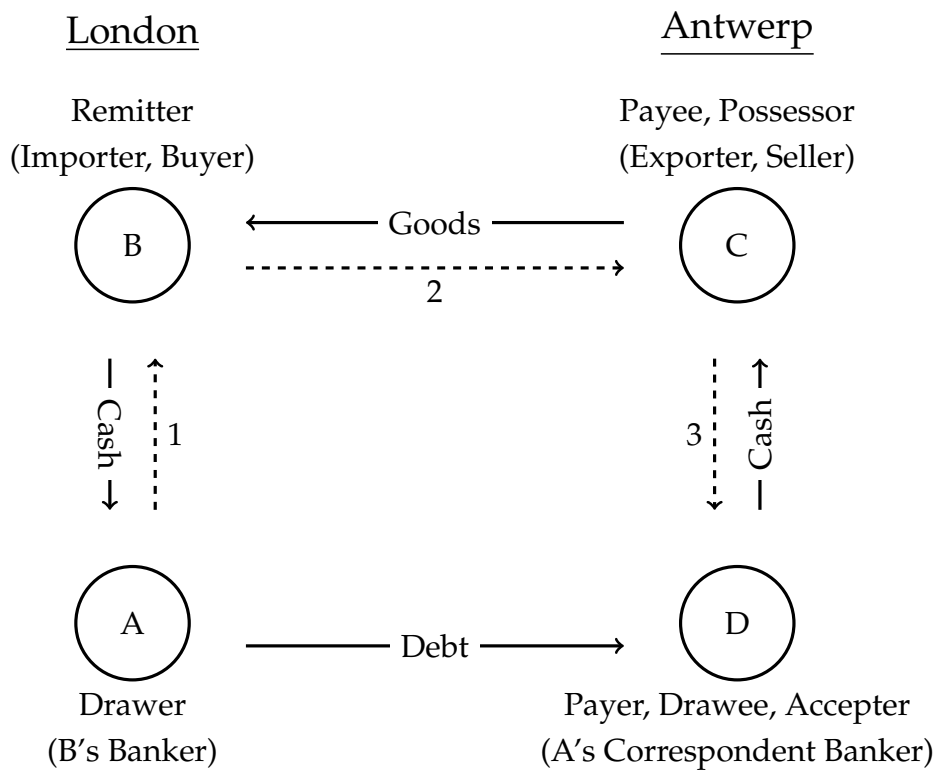

Figure 5: A Foreign Bill of Exchange Used to Make Payment

Note: The dashed arrows depict the movements of the bill of exchange. The bill is drawn up by A on D at B's request. B purchases the bill from A in exchange for local money. B remits the bill to C in exchange for goods. C presents the bill to D, who accepts it and pays C in local money. C might alternatively sell the bill to someone else; the bill might then circulate as a medium of exchange until it is eventually presented to D. Typically, A has an account with D which is debited when D accepts the bill. Alternatively, A and D may net this bill against another that D draws on A for one of his clients.

3.1 Economizing on the Need to Transport Coins

A core function of bills of exchange in the late middle ages and early modern period was to minimize the need for merchants to transfer coins from one location to another. A simple use-case along these lines involving four parties is depicted in Figure 5. In this case, B is an importer in London who is purchasing goods from an exporter C in Antwerp. Rather than send coins as payment, B approaches A – a banker in London who has a relationship with a banker D in Antwerp – and asks A to draw up a bill of exchange on D. B then purchases the bill from A and remits it to C as payment for the goods. C takes the bill to D, who accepts it and pays C when the bill matures.

A few things are noteworthy about this situation. First, notice that the bill is not

an IOU, rather it is a "please pay them" – A is requesting that D pay C. The bill is a request that only becomes binding once D accepts it. D may "protest" the bill – i.e., refuse to accept it – in which case C would send it back to B (or A) and demand alternate payment.

A bill of exchange typically specified payment either on sight (sometimes called a sight draft) or at a specific future date (sometimes called a time draft). In the late medieval and early modern period, most bills were time drafts and the maturity of these time drafts was standardized for many city pairs. Such bills were said to be payable at "usance." The usance for a city pair depended on the travel time between the cities. For example, usance between London and Antwerp, Amsterdam, or Hamburg was one month, while usance between London and Northern Italy was three months (de Roover, 1944).

As we note above, the primary purpose of the bill of exchange in Figure 5 is to allow the importer B to avoid shipping coins from London to Antwerp as a payment for the goods they are importing. Instead, payment is made by drawing down the balance of A with D. This may involve D extending credit to A, if A does not have a sufficient balance with D to begin with. A and D will typically be bankers that engage in many such transactions. Some of these transactions will involve bills being sent from London to Antwerp as in Figure 5, while others will involve bills being sent the other way (and good sent from London to Antwerp).

Importantly, bills traveling in one direction will net out against bills traveling the other direction. Bills sent from London to Antwerp will reduce the balance of A with D, while those sent from Antwerp to London will increase this balance. If the quantity of trade between London and Antwerp is balanced, the value of bills going each way will perfectly net out and no coins will need to flow between the cities. If trade is not balanced, the net amount may need to flow between the cities (although these bankers and cities are part of a broader trading network with other bankers and cities where further netting out can occur). In this way, bills of exchange dramatically reduced the need to ship coins from city to city.

3.2 Hiding Interest Payments

A second important use of bills of exchange was as a means to extend credit. In the late middle ages, usury laws placed severe restrictions on lending at interest. This meant that merchants and bankers needed to resort to various sophisticated mechanisms to hide the interest involved in extending credit. Bills of exchange were

one such mechanism. The use of bills of exchange as credit usually involved a pair of bills going back and forth. This is often referred to as exchange and re-exchange (cambium and recambium in Latin). I will also refer to these two bills as the initial bill and the return bill.

Consider the example in Figure 6. A is an exporter in London looking to export goods to Antwerp. A needs financing to be able to purchase the goods for export. A therefore draws up a bill of exchange on D (their agent in Antwerp) and sells this bill to their London banker B, who by purchasing the bill is in effect lending money to A. In this type of trade, A is often referred to as the "taker" as they take the money, and B is often referred to as the "deliverer" as they deliver the money. A uses the money to buy goods in Britain and sends these goods to Antwerp. In Antwerp, D arranges for the sale of the goods for A. Meanwhile, B sends the bill of exchange to C (their agent in Antwerp). C presents the bill to D, who accepts the bill. At usance, D pays C.

C now needs to find a way to transfer the funds they have received back to B. (Remember, C is acting as B's agent in Antwerp.) In order to get the money back to B, C finds an exporter in Antwerp who is looking to export goods to London and is in need of financing. Let's call this exporter F. C purchases a bill of exchange that F has drawn on E (their agent in London). F uses these funds to purchase goods in Flanders and ships them to London. In London, E arranges for the sale of the goods for F. Meanwhile, C sends the return bill to B. B presents it to E, who accepts it. At usance, E pays B.

Since usance for bills of exchange between London and Antwerp is one month, this whole sequence of transactions takes two months to unfold. B has therefore, in effect, lent out funds for two months. B must be compensated both for the opportunity cost of the money over these two months and also for various risks involved in this sequence of trades (e.g., default by one of the parties). Since lending at interest was not allowed, B's compensation for lending the funds was hidden in the exchange rates used in the transactions. Typically, domestic currency was worth more locally than abroad and this resulted in B earning a return from the exchange / re-exchange transaction.

This is best illustrated by an example. Following de Roover (1944), I use the exchange rates between pound sterling (London money) and Flemish pounds (Antwerp money) discussed in the report prepared by the British royal commission on the exchanges of 1564. At this time, the exchange rate in London was 22s. 6d. Flemish per pound sterling, while the exchange rate in Antwerp was 22s. 2d.

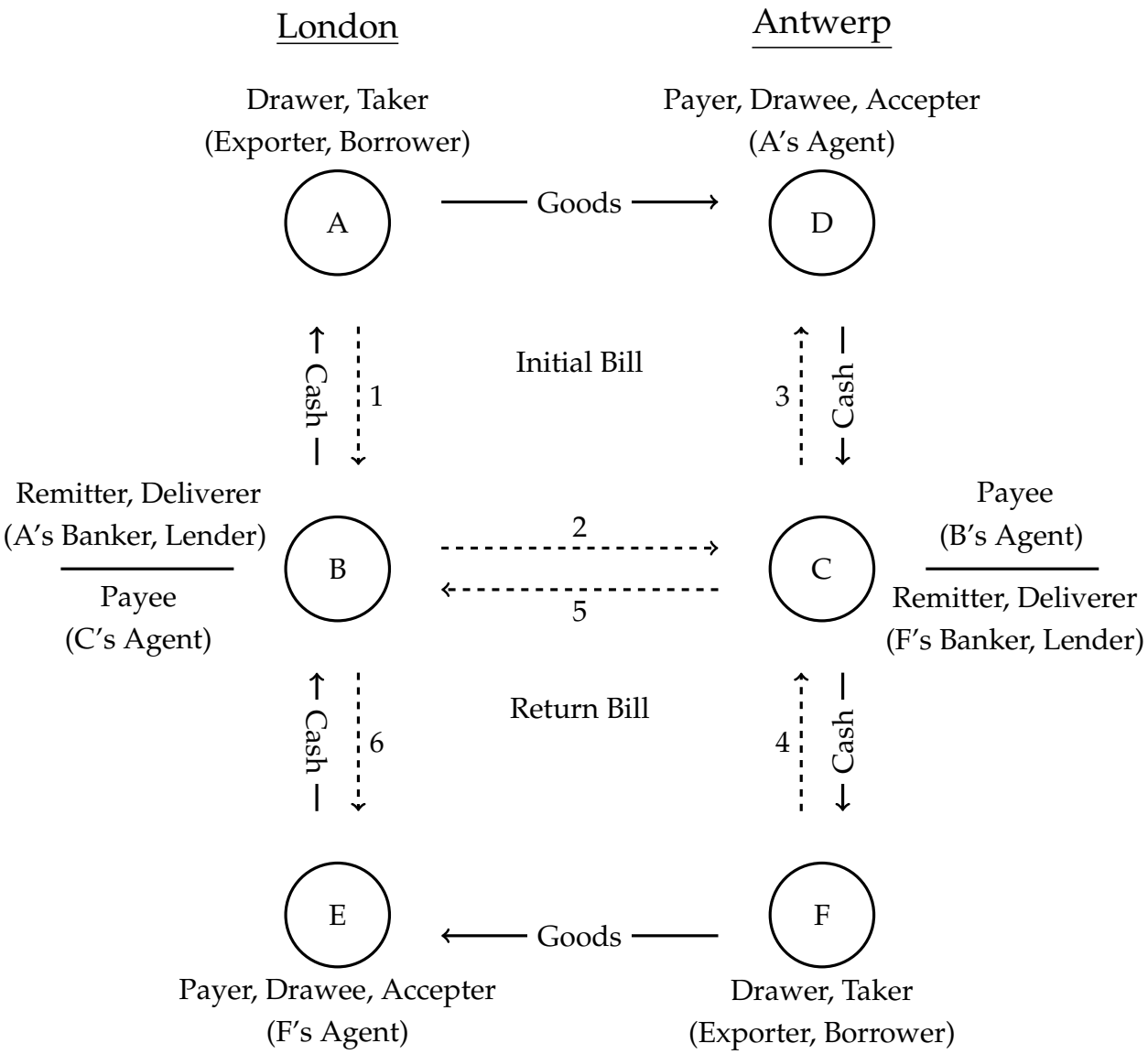

Figure 6: A Pair of Foreign Bills of Exchange Used as a Credit Instrument

Note: The dashed arrows depict the movements of the two bills of exchange. The initial bill is drawn up by A on D. B lends local money to A by purchasing the bill from A. A uses the money to buy goods and sends these goods to D who – acting as A's agent – sells them. B sends the bill to C. C presents the bill to D, who accepts it and pays C in local money at the stated maturity date (usance). C lends this money to F in exchange for a bill of exchange drawn on E (the return bill). F buys goods and sends them to E, who sells them for F. C send the return bill to B who present it to E. E accepts it and pays B in local money at the stated maturity date. The exchange rate between pounds sterling (London money) and Flemish pounds (Antwerp money) stipulated in each bill is raised artificially in favor of B (the original lender). This allows B to make a profit from lending without formally charging interest (which might violate usury laws).

Flemish per pound sterling. Recall that the monetary systems of both England and Flanders were of the traditional form where there were 20 shillings in a pound, and

12 pence in a shilling: £1 = 20s. = 240d. (a penny is abbreviated by d. because in Latin a penny is denarius). It is the 4d. difference between the exchange rate in London and the exchange rate in Antwerp that results in it being profitable to lend through an exchange / re-exchange transaction.

Suppose B in Figure 6 lends £100 to A by buying a bill of exchange A has drawn on D in Antwerp at the prevailing exchange rate of 22s. 6d. Flemish per pound sterling. This means that the bill stipulates that D will pay C £112 10s. 6d. Flemish at usance. At that point, B has £112 10s. 6d. Flemish on deposit with C in Antwerp. Not having direct use for these funds, B instructs C to remit them back to London via a re-exchange. C does this by purchasing a return bill for £112 10s. 6d. drawn by F on E in London at the prevailing exchange rate in Antwerp of 22s. 2d. Flemish per pound sterling. At usance, B collects £101 10s. 1d. sterling from E.

This set of transactions nets B a profit of £1 10s. 1d. sterling on £100 invested for two months. (We are ignoring any fees B pays to C for their assistance.) This amounts to an annualized return of roughly 9.4% since $(1 + 361/24000)^6 \approx 1.0937$. Importantly, the exchange rate difference between London and Antwerp yields a similar profit for a lender in Antwerp who does these transactions in reverse. The exchange rate difference should therefore be thought of as arising so as to incentivize trade finance. An increase in the exchange rate difference would imply that the implicit interest rate in this type of finance would increase.

B's return from these transactions is uncertain at the outset. The uncertainty arises not only because of default risk, but also because the exchange rate might move between the time the initial bill is drawn and the return bill is drawn. If the exchange rate in Antwerp has risen to a level above 22s. 6d. by the time the return bill is drawn, B will make a loss. One reason the exchange rate might move is a trade imbalance between London and Antwerp. Another reason is that one of the two currencies might be debased.

This uncertainty and, in particular, the possibility of loss was crucially important since this helped merchants argue that dealing in bills of exchange was not usury. Cannon law was typically interpreted such that only certain gain from a loan was considered usury. Somewhat disingenuously, bankers claimed that dealing in bills of exchange was not lending at all, but rather money exchange. Most authorities accepted this dubious argument. Thus, as de Roover (1966, p. 14) put it "the odium attached to usury swerved around the merchant-bankers to fall with all its impact on petty money-lenders and pawnbrokers." The sequence of transactions depicted in Figure 6 is, of course, a very roundabout way to extend credit. But such was the

power of religious dogma regarding usury that merchants and merchant-bankers went to all this trouble to avoid being labelled as usurers.

3.3 Discounting of Bills

The heavy use of bills of exchange in international trade in the late middle ages and early modern period meant that international trade was in effect conducted largely with "paper money" (the bills). However, an important limitation of early bills of exchange was that they did not circulate in the way that later paper money did. The reason for this is that they could not be "discounted."

Discounting refers to the sale of a bill (or any other debt instrument) for less than the face value of the bill. Consider again the initial bill of exchange discussed above. Once D has accepted the bill, it is effectively an "I owe you" on D for £112 10s. 6d. Flemish to be paid at usance. If C needs funds before usance, they may want to sell this bill. The buyer would then be giving up funds today in exchange for funds later and would want to be compensated for this. A simple form for this compensation to take is for the price of the bill to be below the bill's face value – i.e. for the bill to be discounted. For example, if the bill is sold 15 days before usance at a price of £112 2s., this would imply compensation of 8s. 6d., which amounts to an annualized rate of return of about 9.5%.

In the late middle ages, discounting was considered usury. It was not until the sixteenth and seventeenth centuries with the weakening of usury laws especially in Protestant areas of Western Europe that bills of exchange became "negotiable" – i.e., could be transferred (sold) by the payee to someone else (typically at a discount) – and could therefore circulate.

When the payee discounted a bill, they would "endorse" the bill by writing on the back of the bill to whom it should be paid and then signing it. Interestingly, endorsements made bills more secure since all those that had endorsed the bill were liable for the payment in the event that the payer defaulted. In some cases, bills had "bearer clauses" – i.e., they stated that they should be paid to a specific person *or to the bearer*. Such bills could circulate from hand to hand without any documentation of the transfer, further reducing the cost of transactions.

Usury laws may seem like a barbarous relic. But as with many barbarous relics, they may have served a useful purpose in an earlier era. Economic historian Charles Kindelberger provides the following historical explanation for usury laws in his financial history of Western Europe:

> The basis for the prohibition against charging interest is found in the ethical prescription in a primitive society, close to the subsistence level, against taking advantage of the misfortunes of others. When a crop fails and a family goes hungry, brotherhood exacts a charitable response, not an exploitative one. As capital starts to become productive, however, there is no ethical requirement for the owner to share its fruit and to lend to others for their positive advantage. (Kindleberger, 1993, p. 43)

4 Early Ledger-Based Payment Systems

Bills of exchange allowed merchants engaged in long-distance trade to avoid transferring coins from city to city. They, however, did not eliminate the need to use coins in local transactions. (Consider, for example, the transactions between A and B, on the one hand, and C and D, on the other, in Figure 5.) Such transactions might be carried out using high denomination coins. But even the use of high denomination coins involved non-trivial transactions costs. Such coins must be counted, weighed, and assayed, and the handling and storage of large quantities of coins, even locally, involved risk of robbery.

An additional problem, especially for small states, was that many different types of coins circulated their simultaneously. In 1606, for example, the Dutch Republic officially recognized 25 gold and 14 silver high-denomination coins (Quinn and Roberds, 2014). This along with the varying quality of coins raised the issue of what type and quality of coin was considered valid to discharge a debt. Cities would issue ordinances regarding this matter, but enforcement was costly and imperfect. In practice, debtors and private bankers had an incentive to discharge debts in light (debased) coins (an instance of Gresham's Law). This incentive contributed to a persistent trend of coin debasement and much discontent among creditors. In essence, the *settlement asset* for bills of exchange was not clearly defined which raised the cost of trade. In his famous digression on banks of deposit in the *Wealth of Nations*, Adam Smith argued that it was to solve this problem that many smaller states established public deposit banks: "this bank being always obliged to pay, in good and true money, exactly according to the standard of the state" (Smith, 1776/2000, p. 511). We will discuss these public banks in some detail below.

4.1 Local Trade: Sophisticated Barter and Private Credit

To avoid the hassle, risk, and cost associated with using coins, many local transactions were either "sophisticated barter" or involved private ledger-based credit. Most people lived on farms and rarely if ever handled coins. A typical farmer would engage in sophisticated barter: they would take their goods to a trading post, sell them to a merchant operating a general store in exchange for goods that they needed. For a literary reference of this type of trade, consider Pa's trips to town in *Little House on a Prairy*. The farmer might on occasion need goods before harvest. In this case, the merchant might extend them credit to be settled at harvest time.

The local merchants themselves might also purchase goods on credit from wholesale merchants and engage in sophisticated barter on a larger scale selling the products they bought from local farmers in exchange for the goods they intended to sell to the local farmers. Some of this trade might be carried out with bills of exchange. But some was carried out with ledge-based credit. Each merchant would keep an extensive set of books where they recorded debts they were owed and credits they owed other merchants and customers. Most debits and credits would be canceled through later trades in goods rather than with coins. This system of private credit and sophisticated barter trade dramatically reduced the need to use coins in local transactions.

As time went on, some merchants came to specialize in financial dealings. They would purchase bills of exchange, extend credit, and engage in money changing. They might also accept deposits. In effect, they became bankers for the local merchant community. Since many people in the local community had accounts with these merchant bankers, payments could be made simply by transferring balances from one account to another on the ledger of the merchant banker. The merchant bankers, therefore, contributed to making the local payment system more efficient. This was an important innovation. However, private merchant bankers had an important weakness in that they were prone to "runs" that could cause them to fail for reasons that we will explore in section 8.

4.2 Public Deposit Banks

The central hubs of the trading network needed larger institutions to serve as central clearing institutions for large value payments. In many of the main commercial centers of Western Europe, public deposit banks were formed to serve this purpose.

The first of these public deposit banks seem to have been established at the beginning of the 15th century, e.g., Taula de Canvi in Barcelone (1401) and Banco di San Giorgio in Genoa (1407). These public banks took in deposits of coins and allowed payments to be made by bank transfer (giro). Depositors could withdraw their deposits on demand at any time. Since the banks paid out depositors in good coin, they were unwilling to accept bad coins at their nominal value. Overdrafts were forbidden and the banks were not allowed to make loans. All deposits were backed 100% by coin reserves (in theory at least). The public banks were thus what would today be called "narrow banks." For this reason, they were not subject to failure due to runs (in theory).

In practice, the public banks came to take on other roles than solely facilitating payments. In particular, they came under pressure to make loans to their sponsoring city and other "systemically important institutions," especially in times of emergency. Many public banks eventually failed as a consequence of fiscal exploitation by the state. The merchant community recognized this risk and was in many cases reluctant to deposit funds at the public banks, preferring to rely on a voluntary private payment system despite the associated risks and costs. The sponsoring city would at times issue ordinances that compelled merchants to make use of the public bank for large-value payments. Overall, the success of these public deposit banks varied wildly.

4.3 The Bank of Amsterdam

The history of the most important early public deposit bank – the Bank of Amsterdam (Amsterdamsche Wisselbank in Dutch) – illustrates well both the success and difficulties of these institutions. The Bank of Amsterdam was established by the city of Amsterdam in 1609. Its core function was to standardize the settlement of bills of exchange. The bank took deposits and offered customers the service of making payments by transferring funds between accounts at no fee. The city ordinance that initially regulated the bank induced merchants to use it by requiring that all bills of exchange over 600 guilders (100 pounds Flemish) must be settled through the bank and also by outlawing private bankers (called cashiers) who had previously engaged in settlement of bills of exchange.

As an exchange bank (i.e., large-value payment system), the Bank of Amsterdam was phenomenally successful for almost two hundred years. Over this period, Amsterdam flourished as a center of trade. Huge quantities of bills of exchange were

drawn on merchants and bankers in Amsterdam and settled at the Bank of Amsterdam. Merchants in Amsterdam were "the envy of Europe for their access to deposit, transfer, and payment services that were trustworthy, safe, efficient, and virtually costless" (de Vries and van der Woude, 1997, p. 131). The Bank of Amesterdam became "the clearinghouse of world trade" (ibid) and the bank's money became "the dominant international currency of the late seventeenth and early eighteenth centuries" (Quinn and Roberds, 2014).

One important reason for the success of the Bank of Amsterdam as an exchange bank was that the bank guaranteed the quality of coins upon withdrawal. In other words, the bank created a settlement asset for bills of exchange of know quality. This reduced uncertainty about the value of the money a merchant would receive when settling a bill of exchange. The bank only accepted certain types of coins at their official ordinance value (i.e., by tale). All other coins were accepted at their underlying metallic value (i.e., by weight). The bank would send these inferior coins to the Mint to be minted into full bodied commercial coins.

For much of the seventeenth century, the bank charged a fee of 1.5% for withdrawals. One reason for this fee was that minting coins was costly. Had the bank not charged a withdrawal fee, owners of inferior coins who wanted to convert these coins into full bodied commercial coins could have used the Bank of Amsterdam to avoid mint charges. They could have deposited the inferior coins at the bank (at their metallic value) and then asked to withdraw these funds and received full bodied coins equal to the metallic value of their previous coins. This would have resulted in a loss for the bank equal to the mint charges. The bank thus needed to charge a withdrawal fee that was at least equal to the mint charges.

The existence of the withdrawal fee meant that merchants had a significant incentive to avoid withdrawals. This led to a market arising where "bank money" was exchanged for "current money" – i.e., coins outside the bank. A merchant who needed to make a payment in bank money but didn't have enough bank money would use current money to purchase bank money from another merchant who desired to withdraw bank money. (The cashiers had reinvented themselves as intermediaries in this trade.) This transaction allowed the second merchant to exchange bank money for current money without incurring the withdrawal fee. The exchange rate between bank money and current money in the market would usually fluctuate in such as way that the two merchants shared the gain from having avoided the withdrawal fee. In other words, the exchange rate fluctuated between 0.985 and 1 current florins for each bank florin (Quinn and Roberds, 2014). (The Dutch unit of

account was the florin, which was also called the guilder.)

The incentive to avoid withdrawals also indirectly meant that merchants had an incentive not to deposit more money in the bank than they needed to. In fact, the amount of money deposited in the Bank of Amsterdam was modest relative to the overall money supply in the Netherlands and did not grow much in the second half of the seventeenth century. This fact is sometimes interpreted as a sign of failure. But that interpretation is questionable. Quite to the contrary, the success of a payment system should be judged by the volume of transactions it can handle per florin (or dollar) deposited. The ideal payment system can handle massive volumes of transactions while tying up minimal funds in the process. Judged by this metric, the Bank of Amsterdam was far ahead of its time.

4.4 Problems Associated with Debasement of Coins

The Dutch Republic suffered from persistent debasement of coins during the sixteenth and seventeenth centuries. This debasement caused substantial difficulty for the Bank of Amsterdam. The debasement resulted principally from competition among the many mints in the Republic and from mints in the Spanish Netherlands: each mint had an incentive to mint coins with slightly lower silver content than the prevailing coin at the time and try to get these debased coins to circulate at the same nominal value as the prevailing heavier coin.

The rixdollar was one of the main commercial coins accepted at its official ordinance value by the Bank of Amsterdam. Mints in the Spanish Netherlands minted a similar coin called the patagon, which had slightly lower silver content (about 5% lower). In the early seventeenth century, the patagon was not officially recognized by mint ordinances in the Dutch Republic. Nevertheless, due to its similarity with the rixdollar, it circulated at a value equal to the mint ordinance value of the rixdollar. This eventually lead the market price of rixdollars to rise above its official ordinance value.

Dutch authorities struggled throughout this period with the following policy trade-off. They wanted to keep the official ordinance value of coins consistent with their market value. Doing this usually entailed raising the official ordinance value of coins up to their market value. But raising the ordinance value of coins was unpopular with creditors since it devalued debts (debtors were obliged to pay debts in coins valued at their ordinance value). In addition, officials worried that raising the official ordinance value of coins could kick off a new round of debasement. Since the

authorities also wanted to limit the debasement of the unit of account (the florin), this led them to be reluctant to adjust ordinance values. As a consequence, authorities adjusted ordinance values relatively seldom and sometimes incompletely.

In 1619, the Dutch authorities increased the official value of the rixdollar from 2.4 florins to 2.5 florins. This devalued deposits at the Bank of Amsterdam by about 5% when viewed in terms of rixdollars since a set number of florins deposited now fetched fewer rixdollars. In 1622, they officially recognized the patagons and gave it a mint ordinance value of 2.35 florins even though it was by then circulating at a value of 2.5 florins. In 1638, they sought to realign the market and ordinance value of the patagon by raising its ordinance value to 2.5 florins. This new mint ordinance did not however raise the ordinance value of the rixdollar. As a consequence, both the rixdollar and the patagon had the same ordinance value of 2.5 florins although the rixdollar had roughly 5% higher silver content (Van Dillen, 1934, p. 87-92).

The 1638 mint ordinance made it profitable for anyone who could obtain a rixdollar at its ordinance value to melt that coin and mint patagons. As Quinn and Roberds (2007) explain, this was a serious problem for the Bank of Amsterdam since it was bound to deliver rixdollars at their mint ordinance value. In effect, the mint ordinance created an arbitrage opportunity: deposit a patagon at the bank, withdraw an equal number of rixdollars less the 1.5% withdrawal fee, melt the rixdollars and mint new patagons. Repeat. Unsurprisingly, huge quantities of rixdollars were withdrawn from the bank. By 1641, the bank was forced to stop giving out rixdollars and switch to patagons. This was officially in violation of the ordinances governing the bank, but the city was forced to permit it.

Curiously, the bank did not value patagons at their market price of 2.5 florins. Rather, it applied a "haircut" to them, valuing them as 2.4 florins. This practice was accepted by the city and in 1659 a mint ordinance was passed that officially recognized this new state of affairs. As a consequence, a system of two officially sanctioned units of account arose in Amsterdam: bank money and current money. An owner of a patagon could use it to purchase something that cost 2.5 florins. Alternatively, they could deposit the coin at the Bank of Amsterdam and receive a deposit of 2.4 bank florins. Due to the withdrawal fee discussed above, most owners or patagons desiring deposits would, in practice, purchase deposits from other depositors desiring patagon coins (this would avoid the withdrawal fee).

The exchange rate between bank money and current money in these transactions was called the agio. It was quoted as a percentage premium on bank money, i.e., an agio of 4% indicated that 1.04 florins of current money could purchase 1 florin of

bank money. The agio tended to hover around 4% most of the time. It was usually slightly lower than the legal difference in the value of bank money and current money which meant that the two parties shared the gains from avoiding the withdrawal fee. An important advantage of this new system of two units of account was that further debasement of coins was decoupled from the metallic value of deposits at the Bank of Amsterdam. Were the market value of patagon's to rise further, this would simply increase the agio. A patagon would remain worth 2.4 bank florins no matter how many current florins it fetched.

4.5 The Bank's Golden Age and Decline

Throughout most of its long history, a key reason for the success of the Bank of Amsterdam was that it managed to largely avoid fiscal exploitation. The ordinance regulating the bank banned lending to customers by overdraft. Despite this ban, the bank did lend to the Dutch East India Company, the city of Amsterdam, and the province of Holland. But these loans were usually relatively modest in size. Quinn and Roberds (2014) report that in 1669 the bank had 6.1 million florins of deposits and 4.5 million florins of coins in its vaults (a 74% reserve ratio).

Having managed to resist fiscal exploitation turned out to be crucial in 1672 when France attacked the Dutch Republic. The attack lead to heavy withdrawals of coin from the bank. The bank's ample reserves meant that it withstood this panic. The commissioners of the bank are said to have opened the vaults of the bank to dramatically demonstrate their ample reserves some of which were blackened due to a fire 20 years earlier (Frost, Shin, and Wierts, 2020). Several other exchange banks who had strayed further away from full backing of deposits were forced to suspend payments to depositors in this episode (e.g., the exchange banks in Middleburg, Rotterdam, and Hamburg). Surviving this episode substantially boosted the reputation of the Bank of Amsterdam.

The bank's extremely strong reputation allowed it to do something that no other public bank had previously done: it eliminated the right of depositors to withdraw their funds. No formal announcement to this effect was ever made. Withdrawals were rare due to the 1.5% withdrawal fee. Furthermore, in 1683, the bank instituted a new system where depositors received a receipt that allowed them to withdraw the specific type of coin they had deposited within 6 months at a much smaller fee (this change circumvented the minting arbitrage discussed above). These receipts were tradeable, and with this change, withdrawals of prior deposits seem to have

fallen into disuse. By the early eighteenth century the right to withdraw seems to have been eliminated completely (Van Dillen, 1934, p. 101-102).

During the eighteenth century, therefore, bank money at the Bank of Amsterdam was an inconvertible currency. Some authors have argued that this makes bank money at the Bank of Amsterdam an early precursor to our modern system of fiat money. This analogy is rather imperfect for several reasons. Most importantly, for much of the eighteenth century the Bank of Amsterdam maintained very substantial coin reserves. In sharp contrast, the asset side of the balance sheet of today's U.S. Federal Reserve (as of 2022) consists almost entirely of assets denominated in U.S. dollars (i.e., denominated in its own liability). A run on the U.S. dollar might well result in these assets losing value.

Writing during the period in question, Adam Smith argues that in the event of a crisis the Bank of Amsterdam would reestablish the right of depositors to withdraw – i.e., reestablish "convertibility" into coins (Smith, 1776/2000, p. 517). Furthermore, the Bank of Amsterdam did not allow the exchange rate of bank florins to fluctuate much versus current florins. It maintained a policy of intervening in the market for bank florins, buying bank florins when the agio fell to about 4.25% and selling bank florins when the agio rose to about 4.875%. The Bank of Amsterdam in the eighteenth century, therefore, seems more analogous to a central bank in the modern era that maintains a fixed exchange rate to a foreign currency and has enormous foreign currency reserves.

The period from 1683 to 1780 was a golden era for the Bank of Amsterdam. In addition to serving its role as a large-value payment system, the bank helped facilitate the international bullion trade in which Amsterdam played a key role, and it provided substantial working capital to the Dutch East India Company (Vereenigde Oost Indische Compagnie or VOC). In most periods, the advances the bank made to the VOC were quickly paid off using proceeds from auctions of colonial products the company held annually. The bank was therefore in most cases only helping the VOC manage seasonal fluctuations in its capital needs. However, there are several periods when the VOC's debt to the bank persisted for a number of years and grew to considerable amounts. Up until 1780, the VOC always managed to clear its position during profitable periods (such as during the Seven Years War of 1756-1763).

All this changed with the Fourth Anglo-Dutch War of 1780 to 1784. This war was a shock of sufficient magnitude for the Dutch Republic and the city of Amsterdam that prior restraint on fiscal exploitation of the bank by the state was no longer tenable. During the war, the state exploited the bank massively and the bank never

recovered from this. At the beginning of the war, the bank had 23.2 million florins in deposits, 20.0 million florins of metal in its vaults (a reserve ratio of 86%), and loans of 3.1 million florins to the VOC and the city of Amsterdam. By the end of the war, it has 19.6 million in deposits, only 6.6 million of metal in its vaults (a reserve ratio of 34%), and it had been forced to lend roughly 9 million to the VOC and the city (Van Dillen, 1934, p. 122).

Confidence in the bank began to sag and the agio with it. When the bank did not purchase bank florins to maintain the agio people feared its weak position prevented it from doing so. In 1790, the agio fell below zero for the first time. The city and the bank made efforts to shore up confidence in the bank over the next few years and the agio did rise above zero for a time. But then the final blow came in the fall of 1794 with the French invasion of the Dutch Republic. Deposits at the bank fell from 22.2 million florins in 1793 to only 8.1 million florins in 1797. The new government in Amsterdam made further efforts to revive the bank. But confidence was permanently lost and the bank shriveled up and was finally dissolved in 1819.

5 Paper Money

The Bank of Amsterdam was primarily a clearing center for large value transactions. It allowed merchants to settle bills of exchange in a standardized unit of account (the bank florin). This avoided the need to handle large quantities of gold and silver coins, and to assess their weight and fineness. In a nutshell, the Bank of Amsterdam was what today is called a large-value payment system. This makes the Bank of Amsterdam an early example of a monetary institution based on a ledger system (i.e., a spreadsheet) where payments are made not by handing over a physical object – such as a coin or a piece of paper – but by adjusting the balance of an account in a bank or some other ledger.

Another major response to the inconvenience of using coins as money was the development of paper money in the form of bank notes. Paper and printing were invented in China long before they were brought to Europe (or reinvented in the case of printing). This is one reason why paper money appeared at an earlier date in China than in Europe. Another reason is that China suffered from a sever shortage of silver and gold prior to the rise of inter-continental trade after 1500. This shortage of silver and gold led to the use of baser metals as coins (such as bronze and iron) in China. Coins made of base metals have low value per unit weight. This makes

them highly inconvenient media of exchange.

5.1 Paper Money in China

The basic monetary unit in China was a coin (*wen*) that typically had a hole through the middle so that it could be strung together into units of 100 (*mo*) and 1,000 (*guan*) for use by merchants. The Tang Dynasty (618-907 CE) and particularly the Song dynasty (960-1276 CE) saw a huge increase in the demand for money as the economy underwent a 'commercial revolution'. The government struggled to supply enough coin to meet this increase in demand. One consequence of this was that various "short-string" *guan* standards emerged, with fewer than 1,000 coins. The official Song dynasty short-string standard had 770 coins in a *guan* (Von Glahn, 2005).

Von Glahn (2005, 2018) describe how paper money in China originated in the province of Sichuan around the year 1,000 CE. Due to its relative isolation at the time and a shortage of bronze coins, Sichuan had developed an iron-coin-based monetary system after the fall of the Tang dynasty in the 10th century. The iron coins had extremely low value relative to their weight: "A housewife would have to bring a pound and a half of iron coin to the marketplace to buy a pound or salt." (von Glahn, 2005, p. 67) When a rebellion against the Song dynasty broke out in Sichuan in 993, the region's mint was forced to close. The resulting shortage of coins put further strain on the already ill-functioning monetary system and prompted merchants to start issuing paper "exchange bills" (*jiaozi*) which began to circulate as money.

This privately issued paper money in Sichuan suffered from some of the problems we describe in section 9. It was soon heavily regulated and then completely taken over by the state. Initially, the paper money was redeemable in coin and maintained its value. But in times of fiscal stress, the state stopped redeeming the *jiaozi* and massively expanded the quantity it issued. This led the value of the currency to collapse. In 1107, the Sichuan government issued a new form of paper money called the *qianyin* which was not redeemable in coin. Later in the 12th century, paper money spread to the Southern Song capital of Hangzhou and the southeast part of the Song empire with the issue of the *huizi*. In the 13th century, the Song issued separate paper currencies in the regions of Huainan (*jiaozi*) and Hubei (*huhui*).

These four paper currencies were not the units of account in the areas in which they circulated. The units of account remained the bronze and iron coins that circulated alongside them. Paper money became the dominant medium of exchange in merchant transactions, while bronze and iron coins remained the the currency of

petty transactions. Periods of monetary stability were punctuated by periods of instability when the exchange value of the paper currencies fell below their face value and fluctuated in some cases by large amounts. Monetary instability often coincided with times of fiscal stress such as wars. During these times the government tended to massively expand the issuance of paper money. At other times, the government would seek to bolster confidence in the paper money by intervening to purchase paper money in exchange for silver (von Glahn, 2018).

When the Mongols conquered China in the 13th century, they gradually unified the monetary system. Khubilai Khan (r. 1260-1294) issued a new paper currency (the *Zhongtong chao*) denominated in bronze coin. He banned the use of coin, gold, and silver in exchange and largely succeeded in his effort to get all trade to be conducted using paper currency. Importantly, Khubilai issued small denomination paper notes in contrast to the Song emperors. Marco Polo reported to his European readers that the great khan's subjects used nothing but paper money in trade. This notion was considers so outlandish in Europe that many concluded that Polo's writings were pure fabrication. As with earlier paper money in China, periods of stability were punctuated with periods of instability when the paper money lost substantial value. When Mongol rule crumbled in the 1350s, their paper money became worthless. Ming emperors tried to revive the use of paper money but did not succeed. China reverted to a silver money standard that lasted until the 20th century.

5.2 Paper Money in Europe

In Europe, paper money in the form of bank notes first appeared in the 17th century. But, as we have discussed in section 3, various forms of paper documents – most notably bills of exchange – had been used to facilitate trade for centuries before this. The path from medieval bills of exchange to bank notes involves several steps. One problem with the medieval bills of exchange was that they were not "negotiable" – i.e., they could not be transferred (sold) by the payee to someone else. This, of course, meant that they could not circulate as money. It was not until the 16th and 17th centuries that bills of exchange became negotiable – first in Protestant regions of Europe.

Initially, bills of exchange were transferred by endorsement. The payee would write a note on the back of the bill explaining to whom the bill should be payed and sign this note. While this was a major step forward in terms of the liquidity of bills of exchange, it had significant limits relating to the space on the back of the bill and

the fact that all endorsers were liable for the payment in the event that the payer defaulted. At some point, however, bills with "bearer clauses" arose. This allowed bills of exchange to circulate without endorsement or any other documentation of transfer, further enhancing their liquidity.

Bills of exchange typically had a particular maturity date. This meant that the payer was not obliged to pay the bill until that date (often referred to as "usance"). Bank notes, in contrast, are redeemable on demand at any time. The fixed maturity of bill of exchange implied that they were typically sold at a discount from face value ("discounted"). The discount represented interest on the bill from the date of transfer to the maturity date. In contrast, modern bank notes are not discounted, they are typically used as a means of payment at their face value.

Finally, bills of exchange are "bespoke" instruments – i.e., each one is tailor-made to a particular transaction. This means that they did not have standardized denominations that were round numbers. Rather each bill was drawn for a different amount relating to the specific transaction it related to. In some regions bills of exchange could *only* be created to facilitate payment in commercial trade. Such bills were referred to as "real" bills. Furthermore, early bills of exchange were handwritten and varied in terms of the text they contained, rather than being printed and standardized. All of this made them more cumbersome to use as a means of payment than modern bank notes.

5.2.1 Stockholms Banco

The seventeenth and eighteenth centuries saw the emergence of various types of bills and notes that gradually introduced the features that make modern bank notes have low transaction costs. The earliest instance of a strikingly modern form of bank notes appeared for a brief period (3 years) in Sweden. Just as in Sichuan China, Sweden's precociousness regarding paper money is likely in part due to the particularly cumbersome nature of its coins. Sweden has ample copper mines and likely for this reason used copper coins. As Heckscher (1934) puts it, these coins were "almost inconceivably cumbrous." The most common denomination was a two dollar "plate" which weighed 3.2kg. "Even the payment of small sums made the use of carriers and horses necessary." (ibid) Obviously, transactions costs were very high in this environment.

In 1656, the king of Sweden granted Johan Palmstruch – an immigrant from the Netherlands with a colorful past – a charter for a bank (*Stockholms Banco*). The idea

(at least officially speaking) was to copy the Bank of Amsterdam. But Stockholms Banco ended up being quite distinct. Rather than being mainly an exchange bank, it mostly engaged in lending. As with all lending banks, this led it to be vulnerable to deposit withdrawals (see section 7). In 1660, the government debased the currency, which led to large withdrawals of copper deposits at the bank (Wetterberg, 2009, p. 37).

In response to this crisis, Palmstruch took the highly innovative step of issuing bank notes. These were printed notes in standard denominations that gave the bearer a claim on Stockholms Banco. That is, the bank promised to exchange the notes for copper coin on demand. Palmstruch seems to have found inspiration for his bank notes in receipts issued by a Swedish copper concern to its miners in exchange for the copper they had mined. These receipts came to circulate as money in the local community of the mine.

Despite being issued as an emergency measure, the bank notes of Stockholms Banco quickly became a huge success. People sought to exchange their copper coins for paper notes at such a rate that the bank had trouble meeting demand. Small wonder, given the extremely clunky nature of the domestic coinage.

As is often the case with financial innovation, the issuance of these bank notes led to a boom and a bust. The bank notes implied that the amount of lending done by Stockholms Banco was not directly dependent on the amount of deposits it had. The bank could simply print more bank notes. At first, the bank treaded carefully. But in 1663 it expanded its lending by a large amount and started opening branches all over Sweden. But by the fall of that year it was starting to have trouble honoring its promise to exchange notes for coin on demand. Once this became known, its troubles intensified and sellers began demanding a premium for taking notes as payment. The bank never recovered from this but hobbled along for another few years. In 1668 it was finally liquidated by the government and then resurrected as a government bank. Bank notes were banned in Sweden until the early 18th century. But people found innovative ways to use deposit receipts as money to the chagrin of authorities. The demand for a convenient means of payment was clearly very strong.

5.2.2 Goldsmith Banking in England

In England, paper money developed gradually over the course of the 17th and 18th centuries along with the emergence of banking. Economic historians Richard D.

Richards and Albert Feavearyear describe how members of four professions developed into bankers in 17th century England: merchants, brokers, scriveners, and goldsmiths (Richards, 1929, ch. 1; Feavearyear, 1963, ch. 5). The best known pioneers of banking in England were the goldsmiths. Goldsmiths were originally metal workers who produced jewelry and other items made of gold and silver. Some also did business exchanging coins. Merchants frequently needed to exchange coins from one country for coins from another. Goldsmiths would engage in this trade and keep reserves of many types of coins for this purpose. This exchange business also conveniently provided goldsmiths with a cover for a different line of business: sifting through coins and (illegally) melting down the heavy coins while passing on the lighter ones. This was at times a very profitable business (e.g., in the 1630s and 1640s). Some goldsmiths may also have engaged in counterfeiting which was a rampant problem.

Goldsmiths had a comparative advantage when it came to safe storage as they maintained strong-rooms for their traditional business. They thus also came to function as custodians of money. Merchants and others needed a place to safely store their gold. Goldsmiths were a natural choice. This line of business was enhanced by the general insecurity of the English Civil War and in particular the 'Tower incident' of 1640, when King Charles I forced merchants to lend him a large amount of gold that they had sent to be coined at the Mint in the Tower of London.

The growth of safe storage of money by goldsmiths paved the way towards banking. At first, goldsmiths held money in trust and were not allowed to lend it out. But as Richards explains "from being a bailee of money was but a short step to the accepting of demand deposits with full authority to make use of such deposits as loans to customers" (Richards, 1929, p. 37). Goldsmiths would issue notes as receipts to depositors (sometimes in convenient denominations) and over time these goldsmith notes came to circulate, i.e., they came to be used for payment (as money). Finally, in the words of Hartley Withers, "some ingenious goldsmith conceived of the epoch-making notion of giving notes, not only to those who had deposited metal, but to those who came to borrow it, and so founded modern banking" (Withers, 1916, p. 24).

The goldsmiths banks in 17th century England were, therefore, a very early instance of "banks of issue", i.e., banks that issued notes that circulated as money. Initially, the goldsmith notes needed to be endorsed when they circulated and often a witness was needed. Initially, they were also handwritten one-by-one by the goldsmith (counterfeiting was a major concern). Over time the notes evolved to

become payable to 'the bearer' and endorsement ceased to be necessary. Also, the notes came to be printed and became more standardized. This process took over a century.

While goldsmith notes did circulate in the second half of the 17th century, their circulation was imperfect. Goldsmith bankers were prone to failure. Those holding their notes knew this and thus knew that holding their notes was not risk free. Furthermore, the legal status of goldsmith notes was not as firmly established as that of bills of exchange, which had become fully negotiable (transferable) in 1666 when the English courts decided that the customs of merchants were part of the law of the land but did not consider goldsmith notes to be customary.

Goldsmith notes seemed to have circulated widely among merchants in London but were only reluctantly taken as payment by various branches of the government. Horsefield (1977) concludes from this that goldsmith notes were not true paper money, which he defines as "anything which is generally acceptable in final settlement of a debt." This may be true. But this is a high bar for the 17th century. The goldsmith notes were a form of proto paper money, but a big step towards the paper money that dominated most economies in the 19th and 20th centuries.

5.2.3 Exchequer Orders and the "Stop of the Exchequer"

Another early form of paper money in England were Exchequer orders first issued by the English Treasury (the Exchequer) between 1667 and 1672. The Exchequer had long issued tallies – pieces of notched wood – to those that it owed money on a short term basis. Someone who made an advance to the King would receive a tally with notches indicating the amount. The holder of the tally would then have the right to intercept money from revenue officials on their way to the Exchequer. Some tallies were issued on security of specific taxes, such as excise taxes on particular products. A disadvantage of tallies was that they did not circulate freely as it was not easy to write legible endorsements on them.

In 1667, the Exchequer introduced paper "orders" that similarly acknowledged a debt of the king. These were paid (i.e., exchanged for coin) by the Exchequer in the order that they had been issued as general tax revenue arrived. A major advantage of these "Exchequer orders" was that they were easily negotiable (transferable) by endorsement. Exchequer orders quickly came to circulate as money and were issued in convenient denominations such as £1, £2, and £5. They were, thus, the first form of government paper money in England.

When England went to war with Holland in 1672, the king's expenses rose dramatically. The king's financial position was weak to begin with as he had already issued a large quantity of Exchequer orders in the preceding years. It was not clear how the king could both honor the existing debt and pay for military expenditures. The solution the king adopted was to stop payment on a large portion of existing Exchequer orders – those issued to bankers and others that had made direct advances to the king. (Contractors, suppliers, and government employees were exempted.) The king then redirected the tax revenue towards military expenditures.

This event is referred to as the *Stop of the Exchequer*. The Stop promptly led to a run on many goldsmith bankers who were known to have lent heavily to the Exchequer (including many of the largest goldsmith banks). These banks immediately "suspended convertibility" – i.e., seized to pay out deposits on demand – and many went bankrupt. The king eventually agreed to make partial repayments on the affected orders and some banker reopened. But losses and bankruptcies were widespread both among bankers and their depositors (Feavearyear, 1963, ch. 5). This episode is an early example of a banking crisis.

Despite the substantial advancement of banking in England over the course of the 17th century, there was a widespread sentiment that commerce in England suffered because the English did not have a bank like the Bank of Amsterdam or the Bank of Venice. Many banking schemes were proposed. Richards argues that one reason why none of these schemes came to fruition before the Glorious Revolution was "fear of Royal confiscation, a fear intensified by the Tower incident and the "Stop" of the Exchequer" (Richards, 1929, p. 105).

In 1694, a major banking scheme was finally adopted and the Bank of England founded. The immediate impetus was the dire revenue need of the crown as it fought yet another war. The stock subscribers to the Bank agreed to lend £1,200,000 to the king at 8% interest in exchange for the right to form a joint-stock bank. From the start, the Bank of England became a major bank of issue. It issued several different types of bills and notes that circulated as money. Early on, some of these bills and notes paid interest. But as time wore on, this seized to be the case. Early on, the notes were hand-written and circulated by endorsement. But as time wore on, the notes were printed, engraved, and water-marked to make counterfeiting more difficult, and they circulated without endorsement.

5.2.4 Country Banking in England

The banking developments discussed above were largely confined to London. Elsewhere in England banking did not develop until the middle of the 18th century as capitalist commerce spread through England and especially as the Industrial Revolution gathered steam. The banks that sprung up outside of London in the 18th century were called country banks. They often had their origins in the local shopkeeper who retailed goods from London. In agricultural areas, farmers would receive inland bills of exchange as payment when they sold their goods to London. These bills were drawn on the buyer's bank in London. (Think of them as checks.) To avoid having to travel to London to receive payment for a bill (i.e., cash a check) and then travel back with the silver or gold coins (risking robbery), the farmer needed to find someone in their local community that was willing to purchase the bill (discount it).

Local shopkeepers were natural candidates to discount bills on London since they purchased goods in London and could therefore use funds in London. These shopkeepers in many cases had accounts at banks in London. They could sell the bills to their London bankers or ask these bankers to act as their correspondents in receiving payment when the bills came due. The shopkeepers would keep deposits at their London banks and draw on these deposits when they purchased goods in London.

In turn, the farmers in the local community would find it convenient to leave their funds with the shopkeepers and draw on these funds when they purchased goods from them. In this way, the shopkeepers came to take deposits from the local community. Some shopkeepers would also issue notes as receipts and these would circulate as money in the local community. These communities, thus, used a combination of paper money (the shopkeeper notes) and ledger-based money (transfers between accounts at the shopkeeper) in addition to coins. Over time, some shopkeepers developed into country bankers that specialized in brokering bills (i.e., buying and selling bills of exchange), taking deposits, and making loans. After some time they might exit their original line of business to focus on banking.

The Industrial Revolution led to large flows of funds between regions in England. The industrial regions in the north borrowed heavily to fund investment, while the agricultural regions in the south and west supplied funds (in addition to food). Feavearyear (1963, ch. 7) explains how a network of banks enabled this flow of funds. In the agricultural regions, farmers would sell bills of exchange they received as payment when selling their wares in London to their local country bank

and receive deposits or country bank notes. The country bank in the agricultural region would sell this same bill to its correspondent bank in London in exchange for deposits at the London bank. The London bank could then lend these funds to country banks in the industrial regions and these country banks could lend them on to industrialists.

The loans from the country bank in the industrial regions to the industrialists often took the form of a bill of exchange drawn on a London bank. The country bank in the industrial region would debit an industrialist with a loan and – rather than giving them cash – would give them a bill of exchange drawn on their London bank. This bill of exchange would then circulate as money by endorsement. Feareayear argues that such bills were the only form of paper money that circulated in Lancashire and West Riding in the 18th century.

The monetary system in England was thus quite heterogeneous in the 18th century. In London, goldsmith bankers gradually exited the business of issuing notes and Bank of England notes were the primary form of paper money. In the agricultural regions, country bank notes were the primary form of paper money, while in the industrial regions bills of exchange drawn on London banks where the primary form of paper money.

5.2.5 Small Notes

Paper money was at first primarily used in wholesale trade and other high value transactions as most bank notes were high denomination. Petty transactions were mostly carried on by sophisticated barter, proprietor credit, and coins (see section 4.1). However, as time passed, bank notes of small denomination became more and more common. These small denomination notes were quite controversial. Feavearyear (1963) explains how the "bankruptcy, at periods of strain, of many of the issuing parties [of these small denomination notes], brought great distress to the poorer classes." (p. 174) In an apparent early act of consumer protection, Parliament outlawed notes under £1 in 1775 and regulated notes under £5 out of existence in 1777.

These policies were reversed in 1797 when fear of a French invasion caused a panic and the Bank of England was forced to suspend convertibility of its notes to gold. With gold coins no longer available, the government quickly passed an act allowing the Bank of England to issue small denomination notes (i.e., £1 and £5). This led to a large expansion of the use of paper money in England.

When Britain went back on the gold standard after the end of the Napoleonic Wars, it was the intention of the government and the Bank of England to redeem small denomination notes (less then £5) and replace them with gold coins. However, the public had grown accustomed to using paper bills and found them more convenient than gold coins. The plan to withdraw small denomination bills was unpopular and was initially rescinded. However, a banking crisis in 1825 was partly blamed on excessive note issue by country banks and an act of Parliament prohibited the issue of small notes in 1826. The Bank of England had already by this time withdrawn much of its small denomination notes due to large numbers of forgeries (Feavearyear, 1963, ch. 10).

5.3 Paper Money in America

The British colonies in North America suffered from a chronic "shortage" of gold and silver coins in the 17th and 18th centuries, according to contemporaries and later scholars (Grubb, 2023). Throughout the colonial period, exports of English coin to the colonies were prohibited. The main coins used in the colonies were Spanish silver dollars – "pieces of eight" – and Portuguese Johanneses – "joes". The colonists acquired these coins when they exported goods. But by the same token, imports of goods were a constant drain on specie.

The fundamental difficulty was that the use of gold and silver coins as a medium of exchange called for substantial imports of specie in excess of exports of specie. But to accomplish this, the colonies needed to run a persistent current account surplus. In this sense, it was quite expensive for the colonies to use gold and silver as money. Every pound sterling of specie used as a medium of exchange meant a pound sterling less of imports from abroad, imports that were sorely needed in the fast growing colonies.

One response to the shortage of coins was to adjust their price. The colonial assemblies did this to varying degrees by raising the nominal value of coins above the British Mint (sterling) value. While the British Mint value of a Spanish piece of eight coin was 4s 6d, its value in 1700 was 6s in Boston, 6s 9d in New York, 7s 8d in Jersey and Pennsylvania (Brock, 1975, p. 8). By doing this, the colonies, in effect, created their own local units of account that differed from the British pound sterling. As with the debasement of coins, this had temporary effects. In the short run, it encouraged the importation of coins. But in the longer run, prices adjusted.

The high cost of carrying out transactions with specie coins and the strong need

for imports in fast growing colonies, meant that the colonists sought to find substitutes for gold and silver money. Foreign trade and merchant commerce could often be conducted using bills of exchange on London. Much small-scale trade was conducted by sophisticated barter or local credit. But certain transactions needed to be conducted using "legal tender" – which meant coins. These included the payment of taxes and often the payment of debts. Also, credit transactions were limited to parties that had established some degree of mutual trust sharply limiting its scope.

For these reasons, there was persistent demand for non-coin legal tender. One form of non-coin legal tender was *commodity money*. The most important example of commodity money was tobacco, which was used as money, for example, in Virginia. The problem with using tobacco as money was that people would pay with low quality tobacco (an instance of Gresham's Law). Virginia reacted to this by enacting a law that required that tobacco meant to be used for payment be brought to warehouses for inspection. Inspectors notes were then issued in exchange for the tobacco and these were legal tender in the county in question and adjacent counties. These inspectors notes were a form of paper currency backed by tobacco. This system prevailed in Virginia with some interruptions from 1730 to 1775 (Brock, 1975, ch. 1).

Another – more innovative – form of non-coin legal tender issued by the colonies was "bills of credit". These were a form of paper currency and were first issued in Massachusetts in 1690 and then by a number of colonies in the 18th century. Formally, the bills of credit were either backed by future taxes of the colony or issued as loans backed by collateral (mostly real estate) posted by the borrowers. Bills of credit backed by future taxes were meant to be retired as the taxes were collected, but the earmarked tax revenue was sometimes diverted to other uses or the bills reemitted to pay additional expenses of the colony. Bills issued as loans were meant to be retired as the loans matured, but again, were often reissued.

Over time, a substantial amount of bills came to circulate in the colonies. The bills were substitutes for coins. Their presence resulted in a substantial fraction of coins in the colonies being exported in the first half of the 18th century – especially in New England. In this way, the issuance of paper money allowed to colonies to run a persistent current account deficit.

Issuing bills of credit yielded seigniorage revenue for the colonies. This was particularly valuable in times of war when government expenses were high. The vast majority of bills were issued at times of war to pay for military expenditures. Since issuing bills increased the money supply, it also tended to boost output – at

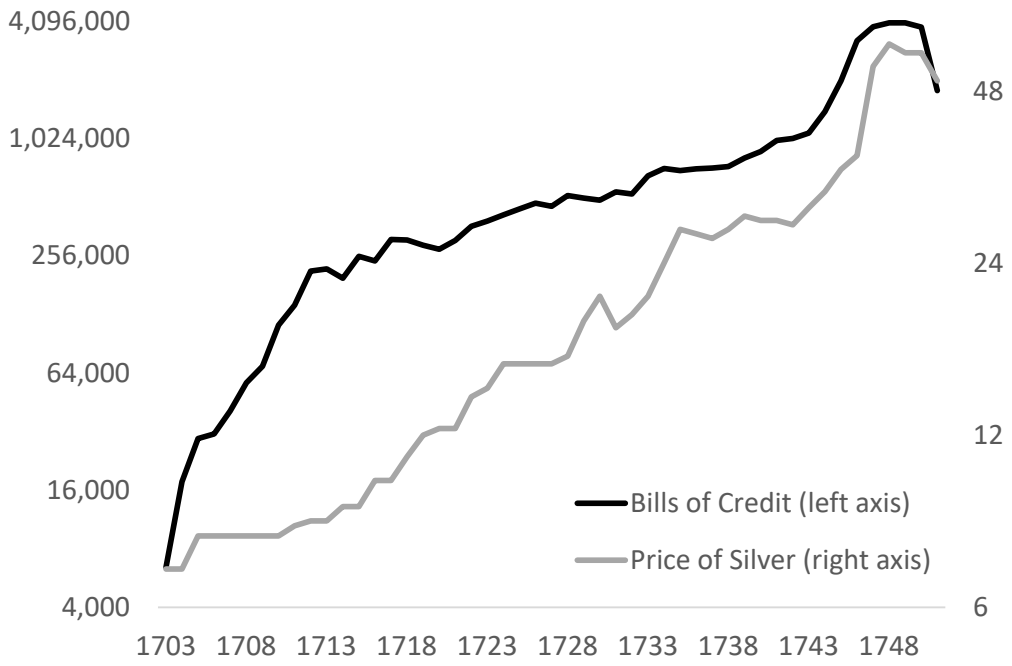

Figure 7: Bills of Credit Outstanding and Price of Silver in New England, 1703-1751

Note: This figure plots the quantity of bills of credit outstanding in New England (black line) as well as the price of silver in Boston (gray line) between 1703 and 1751. The quantity of bills and price of silver is listed in old tenor terms. The data are from Brock (1975).

least in the short run – for the reasons discussed in chapter XX [quantity theory chapter]. A downside of issuing bills was that it could result in inflation, which creditors disliked. Some also argued against bills because they viewed the bills as debt and didn't approve of public debt.

The quantity of bills issued varied greatly over time and across colonies. In New England, the bills of all four colonies – Massachusetts, Connecticut, New Hampshire, and Rhode Island – circulated freely across the entire region. This gave the smaller colonies a particularly strong incentive to issue bills. The seigniorage benefits accrued entirely to the issuing colony, while the inflationary consequences were spread across the entire region. Rhode Island took advantage of this and issued roughly 10 times as many bills per capita as its much larger neighbor Massachusetts (Brock, 1975, ch. 2).

The colonies in New England together issued sufficient quantities of bills that they soon began to depreciate relative to silver coins. Figure 7 plots the quantity of bills of credit outstanding in New England from 1703 to 1751 along with the price of silver in terms of bills. The degree to which the bills depreciated relative to silver

was moderate prior to 1740 (at least by 20th century standards). In the mid-1730s, the average annual rate of depreciation was about 5%. However, with the outbreak of King George's War in 1744 (known as the War of the Austrian Succession in Europe), government spending in the colonies of New England rose dramatically. This was financed with large issues of bills of credit and led to a substantial inflation in the region. The inflation resulted a backlash against paper money. In 1750, Massachusetts returned to a silver standard.

The experience of the middle colonies with bills of credit was different than that of New England. Prior to 1754, these colonies generally issued smaller amounts relative to the size of their economies and experienced only very modest amounts of depreciation relative to silver. This is likely due to their more modest involvement in wars with the French prior to 1754. With the outbreak of the French and Indian War (known as the Seven Years War in Europe) in 1754, the middle colonies issued large quantities of bills to finance the war. For example, bills outstanding in New York rose from about £200,000 to £500,000 over the course of a few years, and bills outstanding in Pennsylvania rose from about £80,000 to £500,000. Yet, these colonies experienced very little inflation and very little depreciation of their paper currency with respect to silver (Brock, 1975, ch. 3 and 7).

With the outbreak of the American Revolution, the Second Continental Congress issued bills of credit – the continental dollar – to pay for the war. These depreciated rapidly over the course of the war and became virtually worthless by the end of the war. Hence the phrase "not worth a continental." The high level of inflation in continentals, likely played a role in causing the framers of the US constitution to ban the issue of bills of credit both at the state and national level. Another factor that may have played a role in this regard is the interests of the nascent banking industry that wanted to be able to issue paper bills of credit and didn't want to have to compete with the state in this regard. After the ratification of the US constitution, the issuance of paper money in America shifted to private banks for over half a century, but was then gradually reclaimed by the government. We discuss these developments in some detail in section 9.

6 Defining the Money Supply

In an economy where gold and silver coins are the only form of money, defining the money supply is a relatively simple matter: it is the sum of all of these gold

and silver coins in circulation. Once paper documents and ledger entries come to be used as media of exchange, defining the money supply becomes more complicated. At a conceptual level, the goal is typically to sum up all assets that are used as media of exchange. However, this simple idea is tricky to operationalize.

The most basic reason why this is tricky is that it is hard to draw a clear line between the class of assets that are used as media of exchange and those that are not. Consider the situation in the 17th and 18th centuries discussed above. There were various forms of bills, notes, and deposits that were used to make payments to a varying degree. Which of these should be considered money and which not is not easy to determine.

In addition to this, there have long been important classes of assets that are not themselves used as media of exchange but can easily be transformed into assets that are used as media of exchange. Deposits in a savings accounts are perhaps the most obvious example of such an asset class in modern times. One doesn't directly pay for things with funds that are in one's savings account. Rather one transfers those funds into ones checking account to make a payment (for example, with a debit card or check). But such transfers are sufficiently straightforward to perform (in the U.S. in the 21st century) that the funds that one has in one's savings account are, for all practical purposes, available to make payments.

Financial innovation has meant that the set of assets that are used as media of exchange has varied over time. We have discussed the increased "negotiability" of various forms of bills and notes above. In more modern times, the ease with which funds in various types of accounts at financial institutions can be used to make payment has evolved considerably (checking accounts, savings accounts, money market accounts, brokerage accounts, etc.). This financial innovation implies that the definition of the money supply must change over time. In addition to this, the fact that different forms of payment are used in different countries implies that the definition of the money supply must differ across space.

Finally, at a practical level, data limitations can affect how the money supply is measured since data on some asset classes are not readily available. For example, data on the notes issued by country banks in England in the 18th century do not exist (to my knowledge). In other cases, it may be important to have a particular breakdown, for example, between demand deposits (checking accounts) and time deposits (savings accounts), and this may not exist.

6.1 M1, M2, M3, ...

These complications have led central banks and monetary scholars to use several different definitions of the money supply (M1, M2, M3, etc.) and to change these definitions frequently in response to financial innovation. To keep things as simple as possible in this section (section 6) and the next section (section 7), we will focus on three types of assets: specie (gold and silver coins), paper currency (bank notes), and bank deposits (we ignore the difference between checking and savings accounts).

In an economy where only these three asset types are used as media of exchange, the money supply is the sum of specie, bank notes, and bank deposits in the hands of the public. This definition corresponds relatively closely to the definition of M1. We will therefore sometimes refer to this definition as M1. (Before 2022, the definition of M1 used by the Federal Reserve in the United States did not include savings accounts. For that period, our definition is closer to M2, which has always included savings accounts.)

The clause "in the hands of the public" in the definition above is important. It means that specie and bank notes in bank vaults or government coffers do not count as part of the money supply. Consider the case of a person who deposits $100 of bank notes into a bank. This does not affect the money supply. Before the deposit is made, the person has the bank notes, i.e., the bank notes are "in the hands of the public." After depositing the bank notes, the person has $100 in bank deposits, but the bank notes are no long in the hands of the public. They are now in a bank vault. Making the deposit changes the breakdown of the money supply between currency and deposits ($100 more deposits, $100 less currency), but does not change the overall money supply.

Similarly, deposits of one bank at another bank also do not count as part of the money supply. Only deposits of "the public" (non-bank entities) count. Consider an instance where the central bank purchases Treasury bills for $100 million from a commercial bank and pays by crediting the reserve (master) account of the commercial bank in question $100 million. (Deposits of banks at the central bank are often called reserves.) In this case, the aggregate amount of reserves banks hold at the central bank has risen by $100 million. But the money supply has not changed at all since the amount of specie, currency, and bank deposits held by the public is unchanged.

6.2 The Monetary Base

It is useful to introduce one additional monetary aggregate at this stage: the *monetary base* (sometimes also referred to as high-powered money or outside money). The monetary base is a completely different concept from the money supply in the hands of the public and it is important not to confuse the two. To understand what the monetary base is, we start by considering bank deposits. Bank deposits are promises to pay something. That something is the monetary base. In other words, the monetary base is the asset that banks hold as reserves for deposits. Importantly, the monetary base is the total quantity of that asset in the economy including both the amount held by the public and the amount held by banks.

From the 17th century to the early 19th century, bank deposits and bank notes were typically promises to pay gold and silver coins on demand. During this era, the quantity of gold and silver coins (both in the hands of the public and in bank vaults) were the monetary base. As central banks became established, the nature of bank deposits shifted. They became promises to pay paper currency (i.e., bank notes issued by the central bank). Once this transition had fully transpired and gold and silver coins were no longer in circulation, the monetary base was the quantity of paper currency in the hands of the public and on reserve at banks (other than the central bank).

In between these two eras, there was typically a transitional period during which the definition of the monetary base was more complicated. At this time, some deposits were promises to pay gold or silver coins while others were promises to pay paper currency (issued by a central bank). Also, both paper currency and specie circulated. During this era, both paper currency and specie in the hands of the public and in bank vaults where part of the monetary base. In some cases – such as in the United States during the National Banking Era – bank notes issued by private banks were sufficiently securely backed by government securities (which were in turn backed by gold) that they have been counted as part of the monetary base. This transitional period can be complicated to comprehend.

7 Bank Lending and Fractional Reserve Banking

Our discussion up until this point has mostly focused on the liability side of banks' balance sheets. It is bank liabilities of various kinds that we use to make payment: bank notes are bank liabilities; bank deposits are bank liabilities; bills of exchange

and checks are bank liabilities or requests to transfer bank liabilities. In a modern economy, even coins are a bank liability. (This has been true since the advent of token coins (see chapter XX [quantity theory chapter]).) These bank liabilities – along with full-bodied gold and silver coins to the extent they are held by the public – are the "money" in the economy, the medium of exchange.

But to fully understand the transition to paper and ledger-based money, we must also consider the asset side of banks' balance sheets. This is what we now turn to.

7.1 Bank Lending and Money Creation

A simplistic view of banking is that people deposit their money in a bank and the bank simply stores the money for them. This is how children typically think of banks. "Their" money is at the bank and they can get that specific money back if they go to the bank. This is, of course, not how things work with actual banks. But as we saw in section 5.2.2, this was in some cases how things started when banking was developing: someone gave a goldsmith or a merchant their gold and silver coins for safe keeping and received a receipt in return. The depositor was free to request their money back at any time. All the goldsmith or merchant was doing was providing safe storage for the money.

Suppose this "proto-banker" took such deposits from many people. As time passed, they realized several things. First, it was unlikely that all the depositors would ask for their money back at the same time. Most of the time, a lot of the money was sitting idle in the proto-banker's strong room. Second, providing safe storage for money was perhaps a fine business, but the proto-banker could earn more if they put the money to use rather than letting it sit idle.

Now, suppose the proto-banker decides at some point to lend some of the money out to merchants, entrepreneurs, home buyers, etc. at interest. If these loans turn out to be well placed and the borrowers pay the loans back with interest, this lending will be profitable. The profitability of the lending will make deposits more valuable. Bankers will then compete for depositors by paying them interest rather than charging the depositors for keeping their money safe.

This bank lending potentially creates a great deal of value. It is one way to channel funds from those with more funds than they know what to do with, to those with more projects than they have funds to finance on their own. But our interest is in a different consequence of bank lending, namely its effect on the money supply.

Borrower's Bank

Assets		Liabilities	
Loan to Borrower	$1,000	Borrower's deposits	$1,000

Borrower

Assets		Liabilities	
Deposits at Borrower's Bank	$1,000	Loan from Borrower's Bank	$1,000

(a) Loan Amount Deposited in Borrower's Account

Borrower's Bank

Assets		Liabilities
Loan to Borrower	$1,000	
Currency	-$1,000	

Borrower

Assets		Liabilities	
Currency	$1,000	Loan from Borrower's Bank	$1,000

(b) Loan Paid Out to Borrower in Central Bank Notes

Figure 8: A Bank Makes a Loan

7.1.1 How Banks Create Money

When a bank makes a loan, the money supply increases by the size of the loan. To see this, consider Figure 8. It depicts two different ways in which a bank can pay out a loan it has made. Either the bank credits the borrower's checking account with the loan amount, or the bank pays the loan amount out to the borrower in the form of bank notes, a wire transfer, a check, or in some other similar manner.

Panel A of Figure 8 considers the case where the bank issues deposits equal to the loan amount to the borrower, while panel B considers the case where the loan is paid out in the form of currency (bank notes issued by the central bank). Each panel shows T-accounts for the borrower's bank and the borrower. These T-accounts are simplified representations of the balance sheets of the parties (we only list changes to the balance sheet for each party). Changes to assets are listed on the left hand side and changes to liabilities on the right hand side.

Consider first panel A in Figure 8. The loan shows up as a new asset for Borrower's Bank, while the new deposit of the borrower shows up as a liability. Con-

versely, the loan is a new liability for the borrower, while the deposit is a new asset for them. In panel B, Borrower's Bank acquires a new asset (the loan) in exchange for another asset it gives up (the currency), while the borrower's situation is more similar to panel A. Other variants of this transaction are very similar to either panel A or panel B.

Notice that in both of these cases, the act by Borrower's Bank of making the new loan increases the money supply (M1) by the amount of the loan. In panel A, it is deposits held by the public that increase by the loan amount, while in panel B it is currency in the hands of the public that increase by the loan amount. Either way, the money supply increases by the loan amount.

This example shows that banks create money when they make loans. Many people find this simple fact shocking. Some worry that it confers dangerous powers on banks. Others worry that it limits society's ability to set up a monetary system in which the money supply and the price level are stable. If money can simply "flow from the fountain pens" of commercial bankers, how can the monetary system be managed so as to bring about monetary and financial stability? We consider these worries in detail below. But first it is useful to develop our example a bit more fully.

7.1.2 Most Transactions Don't Affect Money Supply

While new loans create new money, it is important to recognize that most other transactions in the economy leave the money supply in the economy unchanged. It is worth working through one transaction in detail to see this. Consider the borrower discussed above. Once they have the funds from the loan, they may use these funds to, for example, purchase a machine. Let's suppose the price of the machine is $1,000. If the borrower pays with a debit card or with a check, this simply transfers $1,000 from the borrower's checking account to the checking account of the machine's seller at their bank. Such a transaction does not affect the aggregate amount of deposits in the banking system as a whole and therefore does not affect the money supply.

To see this point clearly, Figure 9 works through a debit card transaction made by the borrower to pay for the machine in detail. As we discussed in section 2, when the borrower uses their debit card to pay for the machine, they are requesting that their bank transfer funds from their checking account to the checking account of the merchant selling the machine. The Borrower's bank will debit the borrower's account. It will request that the merchant's bank credit the merchant's account.

Central Bank

Assets		Liabilities	
		Borrower's Bank reserves	-$1,000
		Merchant's Bank reserves	$1,000

Borrower's Bank

Assets		Liabilities	
Reserves at Central Bank	-$1,000	Borrower's deposits	-$1,000

Merchant's Bank

Assets		Liabilities	
Reserves at Central Bank	$1,000	Merchant's deposits	$1,000

Borrower

Assets		Liabilities	
Machine	$1,000		
Checking account	-$1,000		

Merchant

Assets		Liabilities	
Machine	-$1,000		
Checking account	$1,000		

Figure 9: How a Typical Transaction Moves Through the Banking System

Interbank settlement between the borrower's bank and the merchant bank then involves two transactions. Recall that each card network (e.g., Visa and Mastercard) has a settlement bank. The borrower's bank will transfer reserves from its master account at the central bank to the settlement bank, while the settlement bank will transfer funds from its master account at the central bank to the merchant's bank. The net position of the settlement bank is unchanged. We therefore omit this bank from Figure 9. The net results is that Borrower's bank has $1,000 less reserves at the central bank, while merchant bank has $1,000 more reserves at the central bank.

Notice that the aggregate quantity of deposits held by the public does not change when this transaction occurs. The borrower's deposits at Borrower's Bank fall by

Bank (before loan)

Assets		Liabilities	
Treasury bills	$1,000	Deposits	$1,000
Currency	$1,000	Net worth	$1,000

Bank (before loan)

Assets		Liabilities	
Treasury bills	$1,000	Deposits	$2,000
Currency	$1,000	Net worth	$1,000
Loan	$1,000		

Figure 10: Balance Sheet Consequences of a Bank Making a Loan

$1,000, while the merchant's deposits at Merchant Bank rise by $1,000. The same is true of most other transactions in the economy. There are only a few exceptions to this. The first is when banks make new loans (or when borrowers pay back their bank loans). The second is when the economy experiences an inflow or outflow of specie. The third is when the central government – U.S. Treasury in the case of the U.S. – makes or receives a payment.

7.2 Fractional Reserve Banking

Most banks have more deposits than they have cash reserves backing these deposits. Banks in this position are practicing *fractional reserve banking*. Figure 10 illustrates this with an example. Consider a bank that a group of owners initially start with $1,000 that they invest in Treasury bills. The bank then receives $1,000 in deposits from new customers in the form of currency. The balance sheet of the bank at this point is depicted on the left in Figure 10. The bank has two assets: $1,000 of Treasury bills and $1,000 of currency. The bank's liabilities are $1,000 of deposits. The bank's net worth (assets less liabilities) is $1,000.

Now suppose the bank decides to make a $1,000 loan. Suppose specifically, that the bank credits the borrower with $1,000 of deposits when the loan is made. The balance sheet of the bank after the loan is made is depicted on the right in Figure 10. Now the bank has a third asset (the loan) and the bank's deposits have risen to $2,000.

Before the loan was made, the bank had a dollar of currency in reserve for every dollar of deposits. The bank's reserve ratio was therefore 100%. After the loan is made, this is no long the case. At this point, the bank has $2,000 of deposits, but

only $1,000 of currency on reserve. The bank's reserve ratio has, therefore, fallen to 50%.

If the borrower withdraws some of the funds they have on deposit, this will further reduce the bank's reserve ratio. Suppose for example, that the borrower withdraws $500 from their account. This reduces the bank's cash reserves to $500. It also reduces the bank's deposits to $1,500. However, since the reserves were smaller to begin with, the proportional reduction of reserves is larger than the proportional reduction in deposits and the reserve ratio falls from 50% to 33% ($500/$1,500 = 1/3).

Most transactions by bank customers are not cash withdrawals, but rather card transactions, ACH trancations, check payments, wire transfers, or the like. To settle such transactions for its customers, the bank must transfer funds into its reserve account at the central bank. Suppose, starting again from the situation depicted on the right in Figure 10, the bank deposits $500 of its currency into its reserve account at the central bank. Making this transfer does not affect the total amount of reserves the bank has since reserves at the central bank and currency in the bank's vault both count as reserves. Now suppose one of its customers makes a $500 purchase and pays with their debit card. If the merchant receiving the payment is a customer of a different bank, this will result in this bank's reserves at the central bank falling by $500. (This transaction works like the one depicted in Figure 9.) So, in this case, just as with a cash withdrawal, the bank's total reserves fall from $1,000 to $500 and its reserve ratio falls from 50% to 33%.

Bank customers are constantly making payments and receiving payments. Banks are therefore constantly seeing their reserves rise and fall. Many of these transactions net out as time passes: at one moment, a customer makes a payment and the bank's reserves fall; the next moment, a different customer receives a payment and the bank's reserves rise. For a large bank with many customers, the net movement of reserves over a day – payments made by customers less payments received by customers – is usually vastly smaller than the gross movement – payments made plus payments received. Most days, therefore, the bank's reserve ratio doesn't change very much. On occasion, however, it might happen that a large customer makes a large payment or an unusually large number of customers make an unusually large quantity of payments. In such cases, the bank's reserves may fall noticeably. The risk that an event like this may occur is one reason why banks maintain substantial reserves.

Figure 11 plots the aggregate reserve ratio of all banks in the United States – i.e.,

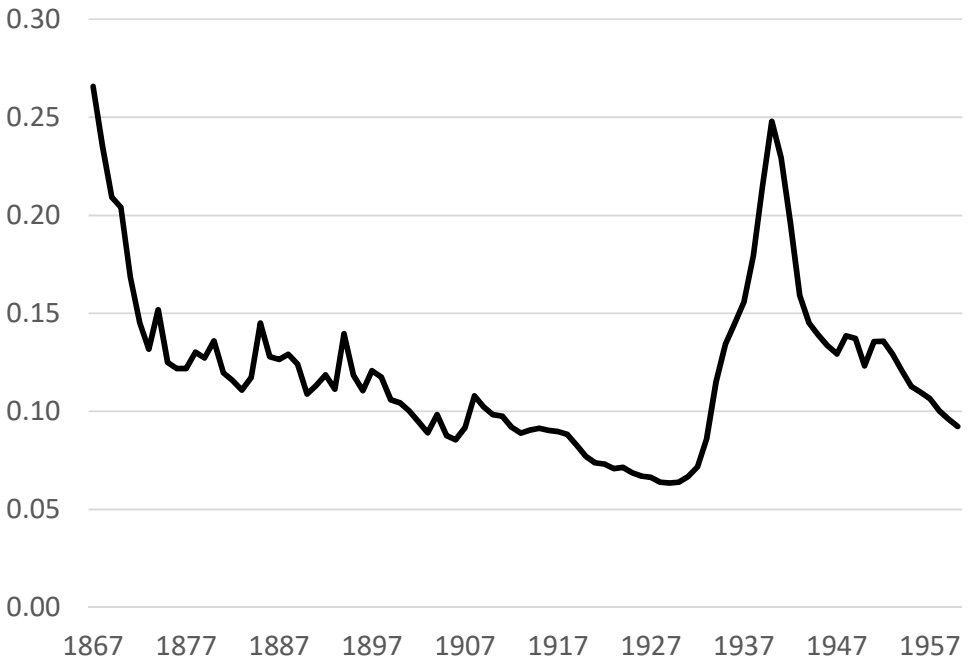

Figure 11: Reserve Ratio of Banks in the United States

Note: This figure plots the ratio of reserves to deposits for U.S. banks between 1867 and 1960. The data are from Friedman and Schwartz (1963).

total bank reserves divided by total bank deposits – between 1867 and 1960. These data were constructed by Milton Friedman and Anna Jacobson Schwartz for their monumental book *A Monetary History of the United States, 1867-1960*. The reserve ratio was above 25% in the immediate aftermath of the Civil War, partly due to high reserve requirements that had been imposed by the government during the war. The reserve ratio fell rapidly in the late 1860s and early 1870s as reserve requirements were relaxed and the quantity of deposits rose. The reserve ratio continued to fall more gradually until the onset of the Great Depression in 1930. It reached a low of only 6.3% in 1929, but then rose very sharply during the Great Depression, peaking at 25% in 1940, before falling back to about 10% in the late 1950s.

We see from Figure 11 that around the turn of the 20th century, a typical bank had a reserve ratio of only about 10%. Figure 12 depicts an example of what the balance sheet of such a bank might have looked like. The bank has assets of $11,000. The vast majority of these assets are loans ($9,000). The bank has $1,000 in Treasury bills and $1,000 in cash reserves. These are listed as currency, but some of them may have been deposited with another bank. (Recall that the U.S. did not have a central bank at the time.) The liabilities of the bank are $10,000 in deposits. Since the bank's

Assets		Liabilities	
Treasury bills	$1,000	Deposits	$10,000
Currency	$1,000	Net worth	$1,000
Loan	$9,000		

Figure 12: Balance Sheet of a Typical Bank

deposits are ten times larger than the bank's cash reserves, the reserve ratio is 10%.

The amount of reserves a bank decides to hold is partly determined by regulatory reserve requirements imposed by the central bank. We will discuss these more below. But let's suppose for now that there are no reserve requirements. In this case, banks hold reserves purely for precautionary reasons: to meet unusually large withdrawals and to bolster confidence in their ability to meet withdrawals.

Holding reserves is costly. Currency earns no interest and until 2008 reserves at the central bank earned no interest in the United States. By holding reserves, banks are therefore foregoing the interest they could earn by lending out these reserves or using them to purchase interest bearing assets such as Treasury bills. Banks would like to hold as little reserves as they can get away with subject to not running out of funds on days when outflows are unusually large.

The quantity of reserves banks choose to hold depends on their perception of the risk that they may face unusual net withdrawals (i.e., how volatile are the payments made and received by their customers) and it also depends on the bank's confidence that their customers will not loose confidence in them and start withdrawing large amounts. In addition to these factors, the reserve ratio a bank chooses also depends on how quickly the bank can raise additional cash were it to need to.

The bank depicted in Figure 12, can sell the $1,000 of Treasury bills it has. This can be done relatively quickly. It may also be able to borrow money from other banks or from the central bank (using the loans it has made as collateral). This may take more time. Finally, it will take some time to distribute any cash it acquires through these means to its various branches.

We see in Figure 11 that the reserve ratio of banks in the U.S. fell gradually from the 1870s until the 1920s. This fall, partly reflected a fall in precautionary demand for reserves because perceived risk of unusual withdrawals fell, confidence in banks

Assets		Liabilities	
Treasury bills	$1,000	Deposits	$10,000
Currency	$1,000	Net worth	$890
Loan	$8,890		

Figure 13: Balance Sheet After a Loss of 1% of Assets

rose, and the ability of banks to raise cash on short notice improved. (It also partly reflects falling regulatory reserve requirements.) As we discussed in section 4, the efficiency of a payment system should be judged by the volume of transactions it can handle per dollar held in reserves. The fact that the reserve ratio fell over this period is therefore evidence of improved efficiency of the payment system.

7.3 Bank Leverage and Risk

We have seen that bank lending creates money and that it leads the bank's reserve ratio to fall. Another consequence of bank lending is that it expands a bank's balance sheet and leads the bank's *leverage* to grow. We define leverage as total assets divided by equity. The bank depicted in Figure 12 is levered 11 to 1. (Equivalently, it has a leverage ratio of 11.) A related metric is the debt-to-equity ratio. The bank depicted in Figure 12 has a debt-to-equity ratio of 10.

Suppose the bank depicted in Figure 12 makes a $4,000 loan. This increases its loans from $9,000 to $13,000 and it increases its deposits from $10,000 to $14,000 (at least until the borrower starts spending down their new deposits). The total size of the bank's balance sheet increases from $11,000 to $15,000, while the bank's equity remains unchanged. The bank's leverage ratio therefore increases from 11 to 15.

Leverage is intimately connected to risk. Other things equal, a bank that is more leveraged is riskier. To see this, consider a case where the bank depicted in Figure Figure 12 suffers a loss on its assets that is equal to 1% of the value of its assets. An example of such a loss would be a loan loss equal to $110. Figure 13 depicts the balance sheet of this bank after such a loan loss.

Since the bank's liabilities have not changed, the $110 loan loss translates into a $110 fall in the bank's net worth. But notice that $110 is a much larger proportion

of equity than it is of assets. The bank's assets have fallen by 1%, but this translates to an 11% fall in the bank's equity. Since the bank has a leverage ratio of 11, a 1% asset loss leads to an 11% drop in the bank's equity. If the bank suffers a 3% drop in assets, its equity will fall by 33%, and a roughly 9% drop in assets will completely wipe out this bank's equity.

Suppose instead that the bank was leveraged 20 to 1. In this case, a 1% drop in assets would reduce the bank's equity by 20% and only a 5% drop in assets would wipe out the bank's equity completely. On the eve of the 2008-2009 financial crisis, the leverage ratio of Goldman Sachs – at the time the largest investment bank in the United States – was roughly 26. At this time, Goldman Sachs could, therefore, only withstand losses equal to about 4% of assets. Investment banks were much more leveraged in 2008 than commercial banks. For example, Citibank's leverage ratio was "only" about 6.5. It could therefore withstand losses equal to about 15% of its assets.

Clearly, other things equal, a bank's equity is riskier the more leveraged the bank is. Given their extremely high leverage, banks tend to invest in relatively safe assets. They attempt to diversify their risk of loan losses by making many modestly sized loans, rather than fewer larger loans. They also purchase relative safe bonds, such as Treasury bills and bonds. But their high leverage implies that even a modest mistake can land a bank in big trouble.

7.4 The Money Multiplier

Let's now return to the question of what determines the money supply in an economy with paper and credit money. We saw above that banks can create money by making loans. But how much money do they end up creating in this manner? To gain a better understanding of this, it is useful to derive a formula for the so called *money multiplier*. In section 6, we discussed the definition of the money supply and the definition of the monetary base. Recall that a somewhat simplified definition of the money supply is currency in circulation plus bank deposits:

$$M = C + D,$$

where M denotes the money supply, C denotes currency in circulation (i.e., in the hands of the public), and D denotes deposits (we are again not making a distinction between demand deposits and time deposits, for simplicity). The definition of the

monetary base is currency in circulation plus bank reserves:

$$M_b = C + R,$$

where M_b denotes the monetary base and R denotes bank reserves (vault cash plus reserves at the central bank).

Let's now take a ratio of these two equations:

$$\frac{M}{M_b} = \frac{C + D}{C + R}.$$

Next we divide both the numerator and the denominator on the right-hand side by D to get

$$\frac{M}{M_b} = \frac{(C/D) + 1}{(C/D) + (R/D)}. \tag{1}$$

This equation can be rewritten

$$M = \left[\frac{(C/D) + 1}{(C/D) + (R/D)}\right] M_b = B_m M_b, \tag{2}$$

where we refer to B_m as the money multiplier. The logic for this name is the notion that the monetary system starts off with the monetary base – think of gold coins in the 17th century or paper currency in 20th century – and then the interactions of banks and their customers "multiply" the amount of money in the economy (the money supply) by a factor equal to the money multiplier.

Notice that the money multiplier is a function of two ratios: 1) the ratio of currency in circulation to deposits C/D, and 2) the ratio of bank reserves to deposits R/D. C/D is determined by the behavior of households and firms. In contrast, R/D is determined by the behavior of banks. Both C/D and R/D can in principle be larger than one, but were typically quite a bit smaller than one during the 20th century and the latter part of the 19th century. C/D is currently trending towards zero in large parts of the world as electronic payment methods lead people to hold less and less currency. However, R/D has risen above one in recent years. Why this has occurred is a topic for chapter XX [Monetary Policy chapter].

Figure 14 plots C/D and R/D for the period 1873 to 1913 in the United States. Over this period, both ratios were substantially below one and were generally falling. There are a few modestly sized upward spikes in these series, which generally coincide with financial panics. Reductions in C/D and R/D increase the money multiplier. The combined downward trend in these series over this period led the money multiplier to increase from about three in 1873 to about 5.5 in 1913.

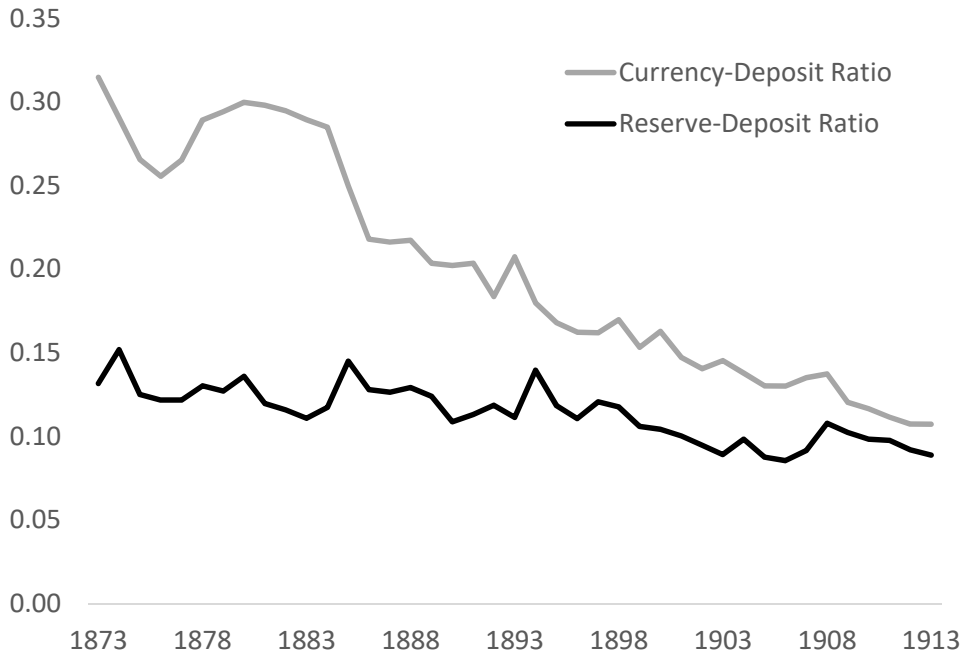

Figure 14: Currency and Reserve to Deposit Ratios in the United States, 1873-1913

Note: This figure plots the ratios of currency held by the public to deposits (gray line) and reserves to deposits (black line) for U.S. banks between 1873 and 1913. The data are from Friedman and Schwartz (1963).

7.5 Determinants of the Currency-Deposit Ratio

As we mention above, it is households and firms that determine the C/D ratio. Several factors determine what ratio of currency to deposits households and firms choose to hold. One factor is convenience. Prior to the widespread adoption of electronic forms of retail payment (debit cards, credit cards, tap-and-pay, etc.) currency was demanded for its convenience in making small transactions. People would pay with cash at the grocery store, restaurant, hardware store, etc. Some establishments would allow customers to pay with checks. However, this was less convenient than cash for small transactions and there was a risk that the establishment would refuse to accept a check. This demand for cash has been eroding rapidly over the past few decades with the rise of electronic payment methods.

For larger transactions, checks, bills of exchange and more recently card transactions, ACH transactions and wire transfers have been the more common means of payment. Transporting large amounts of cash is inconvenient and risky. But to write a check, draw up a bill of exchange, pay with a card, etc., one needs to hold deposits (or have credit) at a bank. Deposits therefore yield convenience for those

engaged in large transactions.

A second factor determining the currency-deposit ratio is safety. As with convenience, safety can favor either currency or deposits depending on the circumstance. Holding cash carries the risk of being robbed. Having depisits in a bank carries the risk that the bank may fail. Both of these risks vary over time and space. Banking crises are times when the risk of bank failure rises dramatically leading to potentially large shifts away from deposits and towards cash as people withdraw money from banks they fear may fail.

The currency-deposit ratio fell quite substantially over the period plotted in Figure 14. This large drop was likely caused by increased confidence of the public in banks. Over time, people became more comfortable with holding their money at banks. More people deposited money into banks and the ratio of deposits to currency rose (the ratio of currency to deposits fell). The ability to pay with a check may also have improved making deposits more convenient.

A third factor determining the currency-deposit ratio is the relative return of these two assets. Currency, of course, earns no interest. Deposits can earn interest. This favors deposits over currency. When interest rates rise, this should reduce the currency-deposit ratio as households and firms face a stronger incentive to economize on cash relative to deposits. In many cases, deposits earn rather modest interest in comparison to other asset classes. An increase in interest rates, therefore, typically reduces the overall demand for money (currency plus deposits) as we discuss in chapter XX [IS-LM chapter].

A fourth factor is the absence of a paper trail. Transactions involving deposits leave a paper trail, while cash transactions can be conducted in such a way that they do not leave a paper trail. This implies that cash is favored for transactions that people want to keep secret. Illegal transactions are an important category of such transactions. Cash transactions are quite prevalent in the illegal drug trade, prostitution, and transactions that involve tax evasion. A surprisingly large fraction of the population is strongly opposed to the elimination of cash despite very efficient and safe electronic alternatives. It seems that many people value very strongly the ability to engage in some number of transactions that they can keep secret.

7.6 Reserve Requirements

In the United States, legal reserve requirements for banks originated with the National Banking Act of 1863. The original rationale for these requirements was to

ensure that banks had sufficient liquidity to meet withdrawals. While high required reserves may have helped build confidence in the new National Banking System, the liquidity rationale for reserve requirements is flawed. Reserve requirements, in fact, do not result in banks having additional liquidity in times of stress (unless the requirements are soft constraints). The reason for this is simply that a dollar of reserves cannot both be used to meet a customer's demand for cash and to satisfy a reserve requirement.

If a bank has $1,500 in reserves, but is required to hold a minimum of $1,000 in reserves, it only has $500 to meet withdrawals. Once these are exhausted, the bank will need to borrow additional reserves or sell assets. Economists Armen Alchain and William Allen drew a colorful analogy: "To rely upon a reserve requirement for the meeting of cash-withdrawal demands of banks' customers is analogous to trying to protect a community from fire by requiring that a large tank of water be kept full at all times: the water is useless in case of emergency if it cannot be drawn from the tank." (Alchian and Allen, 1967)

A different rationale for reserve requirements is that they help the central bank regulate the money supply. We saw in section 7.4 that the money multiplier is a function of the reserve-deposit ratio (equation (1)). Higher required reserves will result in a higher reserve-deposit ratio and a smaller money multiplier. A smaller money multiplier implies that a dollar deposited into a bank will have a smaller affect on the money supply. The extra deposit will induce the bank to make more loans. But this will be more constrained by the need to hold extra reserves. Higher reserve requirements may thus help reduce volatility in the money supply (other things equal).

This logic, however, depends on how monetary policy is conducted. If the central bank fixes the quantity of the monetary base, a dollar deposited into a bank will increase the money supply as discussed above. But this will lower interest rates in the economy since households and firms will need to be induced to hold more money by lower returns on other assets (more on this in chapter XX [IS-LM chapter]). If instead the central bank targets an interest rate, a dollar being deposited into a bank will induce the central bank to reduce the monetary base so as to avoid interest rates falling. In this second case, the dollar being deposited into the bank does not affect the overall money supply. More generally, it is not clear that reserve requirements play a useful role for monetary policy if central banks target interest rates rather than monetary aggregates.

In the 19th and 20th centuries, reserves (even those held at the central bank) did

not pay interest. A cost to banks of holding reserves was therefore the foregone interest they could have earned if they instead held a different asset. To the extent reserve requirements affected the quantity of reserves banks held, they were thus a tax on banks. The banks were forced to hold zero-interest liabilities of the central bank as reserves. This expanded the balance sheet of the central bank. The interest on the additional assets the central bank held enhanced its profits which were typically handed over to Treasury as seignorage revenue. The reserve requirement tax raised banks' costs of making loans and therefore reduced lending in the economy. This effect was modest in size when interest rates were low but more substantial when interest rates were high (such as in the 1970s).

In recent years, reserve requirements have been dramatically reduced or completely eliminated in many countries. In addition to this, it has become more common that central banks pay interest on reserves. In the United States, the Federal Reserve began paying interest on reserves in 2008 and eliminated required reserves in 2020. As of this writing, the European Central Bank still maintains a positive but very low required reserve ratio of 1%. It pays interest on reserves. In Japan, required reserves are also very low at 0.8%. The central banks of the United Kingdom, Canada, Australia, Sweden, New Zealand, among others have eliminated required reserves. Feinman (1993) provides a more detailed account of the history of reserve requirements in the United States.

7.7 The Widow's Curse

Consider a situation where a bank makes a loan and pays the loan out to the borrower as cash. Suppose the borrower deposits this cash at their bank. That bank now has more deposits and decides to make an additional loan. The second borrower deposits the loan at their bank. That bank now having additional deposits decides to make an additional loan. One might imagine this process continuing forever in which case commercial banking would possess a "widow's curse" of unlimited supply of loans. Since each loan increases the money supply, a particular worry is that this widow's curse might imply uncontrolled monetary expansion.

We have seen in the preceding sections several mechanisms through which this widow's curse is limited. One such limit is reserve requirements. If reserve requirements are positive, the banks must in each round put aside some funds as reserves. But we have also seen that reserve requirements have been dramatically reduced or completely eliminated in many countries.

A second limit is the fact that even absent requirements banks may choose to hold some fraction of cash as reserves. Additionally, households may choose to hold some fraction of their money in cash. But one might worry that even if this is true on average it might not be true on the margin. Perhaps demand for cash and reserves has at some point been satiated. Beyond that point, perhaps there is no limit on the widow's curse.

Nobel laureate James Tobin argued that this is not the case because additional lending affects the equilibrium interest rate on loans and on deposits in a way that limits lending (Tobin, 1963). The more banks lend, the less profitable will be the marginal potential project. Also, as the money supply expands, the public will need to be induced to hold the extra deposits with higher rates of return on deposits. At some point, these changes in rates of return on loans and deposits will imply that it is no longer profitable for banks to continue lending. Thus, the ultimate limit on the widow's curse of bank lending is the same as for any other economic activity: downward-sloping demand curves. As Tobin puts it: "Evidently, the fountain pens of commercial bankers are essentially different from the printing presses of the government."

8 Bank Runs and Banking Panics

All banks – even the best capitalized and best run – are fragile. The reason for this is that the liabilities of banks tend to be short-term debt obligations (e.g., demand deposits), while their assets tend to be long-term loans. This implies that the balance sheets of banks feature a *maturity mismatch*. This maturity mismatch is quite fundamental to banking since the most important roles of banks are: 1) to assist their customers in making payments, which involves supplying customers with demand deposits, and 2) financing investment projects through loans, many of which are long-term in nature. Given this dual role of banks, it is fundamental to banking that they engage is maturity transformation by holding long-term assets but issuing short-term liabilities.

The maturity transformation that banks perform is an extremely valuable activity for society since it creates a vast amount of very "liquid" assets (demand deposits) which are useful for making the payment system in the economy work well. Alternatively, we can say that the maturity transformation of banks allows for a vast amount of long-term funding of investment projects despite the demand of retail in-

vestors for highly liquid assets.

But the fact that banks engage in this maturity transformation also has an important downside: it makes banks fragile. In particular, banks are vulnerable to loss of confidence by their depositors. If depositors lose confidence in a bank, the bank can get in trouble very quickly. The bank's assets (long-term loans) are illiquid in the sense that they are hard to sell or call in quickly. If enough depositors lose confidence and withdraw their deposits, the bank will run out of money (i.e., funds that can be used as a means of payments) and fail unless other banks, large investors, or the government are willing to extend it potentially huge amounts of credit. This is true of all banks, even the best capitalized and best run banks.

What is more, loss of confidence by depositors can become a *self-fulfilling prophesy*. Each depositor understands that if other depositors lose confidence the bank will fail. Each depositor may thus fear loss of confidence by other depositors. If enough depositors act on such fear, the bank will in fact fail. In such cases, the bank is failing solely as a result of self-fulfilling fear (in the sense that absent the fear the bank was well capitalized and well run).

8.1 A Little Bit of Game Theory

To understand how loss of confidence by depositors can be a self-fulfilling prophesy, it is useful to introduce a few concepts from game theory. Game theory is the branch of economics that studies strategic interactions between different actors in the economy. The most famous game studied in game theory is the *prisoner's dilemma*. I will use this game as a vehicle to introduce the concepts we need to study bank fragility.

Consider a situation where the police have arrested two people on suspicion of having committed a crime. The two suspects are held in separate cells and cannot communicate with one another. The evidence the police has is not sufficient to convict the prisoners without at least one of them confessing. They each face a choice between confessing and not confessing. They each know that the other prisoner faces this same choice. And the police make sure they both understand that the consequences of their choices will be the following. If both confess, they will both be sentenced to 5 years in prison. If neither confesses, they will both be convicted of a minor offense and sentenced to 1 year in prison. If one confesses but the other does not, the one that confesses will go free, while the one that does not will be sentenced to 10 years in prison.

Figure 15 depicts this situation using a matrix. As noted above, each of the

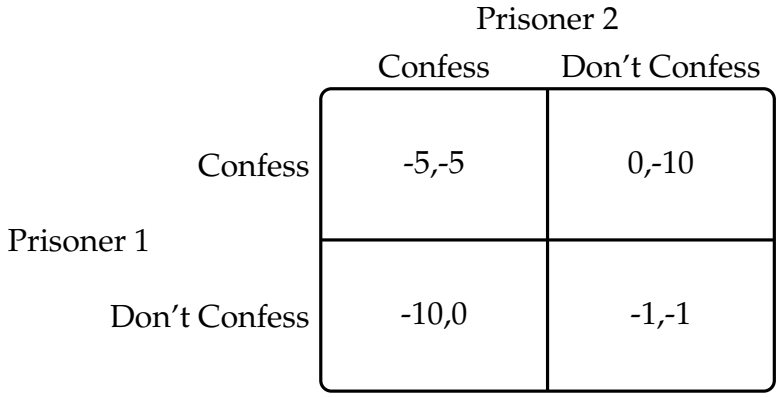

Figure 15: The Prisoner's Dilemma Game

players has two actions (strategies) they must choose between: confess and don't confess. The rows in the matrix correspond to the actions of prisoner 1, while the columns in the matrix correspond to the actions of prisoner 2. The numbers in each cell are the payoffs for the players if that set of actions is chosen. The first number in each cell is the payoff for prisoner 1, while the second number is the payoff for prisoner 2. For example, the south-west cell represents the outcome that prisoner 1 does not confess, while prisoner 2 confesses. In this case, prisoner 1 goes to prison for 10 years, which is listed as a payoff of -10; while prisoner 2 goes free, which is listed as a payoff of 0.

Let's consider the situation from prisoner 1's perspective. He thinks to himself: Suppose prisoner 2 confesses. In this case, my payoff will be -5 if I confess, while it will be -10 if I don't confess. So, in this case, I should confess. But what about if prisoner 2 doesn't confess. In this case, my payoff is 0 if I confess, while it is -1 if I don't confess. So, in this case as well, I should confess.

This logic implies that no matter what prisoner 2 does, prisoner 1's *best response* is to confess. But notice that the game is symmetric: prisoner 2's problem is identical to prisoner 1's problem. It is therefore also the case that no matter what prisoner 1 does, prisoner 2's best response is to confess. If both players play their best responses, they both confess and both land in prison for 5 years.

Game theorists refer to the (confess, confess) outcomes as a *Nash equilibrium*. Formally, a Nash equilibrium is an outcomes where each player of a game is playing a strategy that is a best response to the strategies that all other players in the game are playing. The prisoner's dilemma game is a bit special in that the confess strategy is a best response no matter what strategy the other player is playing. A strategy that satisfies this condition is called a dominant strategy. More generally, the strategy

that each player plays need only be a best response to the strategies other players *actually play* for the outcome to be a Nash equilibrium. Since in a Nash equilibrium each player plays a best response to the strategies other players actually play, it embeds the notion that each player has *rational expectations* about what other players will do.

The prisoner's dilemma game has a unique Nash equilibrium. (Confess, confess) is the only outcome where both players are playing a best response. The other three outcomes all involve one or both of the players not playing a best response strategy. But (confess, confess) is a much worse outcome for the prisoners than (don't confess, don't confess). If neither confessed, they would both shave 4 years off their prison sentences. This is what makes the prisoner's dilemma game such an interesting game. Mutual cooperation would benefit both players, but (unfortunately for them) it is not in their individual self interest, which makes it hard to achieve.

8.2 The Diamond-Dybvig Model of Bank Runs

The economists Douglas Diamond and Philip Dybvig used game theory to develop a model that explains how bank runs can be self-fulfilling prophesies (Diamond and Dybvig, 1983). In 2022, they were awarded the Nobel Prize in Economics for these insights. We now consider a variant of this model. The model considers a bank that takes deposits from a large number of depositors. The bank lends out most of these funds to companies that are seeking to finance investment projects. The companies need funds for two periods to be able to complete their projects. The bank therefore extends them two-period loans, i.e., the bank commits to fund the companies for two periods and cannot call in the loans after one period. The interest rate the bank charges on these loans is such that the companies on average end up paying the bank R_ℓ dollars per dollar invested. We assume that $R_\ell > 1$.

Next consider the bank's depositors. They are constantly needing to make and receive payments. On average, they make and receive the same quantity of payments each period. Some make more payments than they receive (and see their deposits decline temporarily), while others receive more payments than they make (and see their deposits increase temporarily). At the level of the bank, all this churn of normal payments averages out and does not affect the funds the bank has at its disposal. But the fact that the depositors need to be able to make payments at all times implies that they demand that their deposits be completely liquid (i.e., demand deposits). In particular, the depositors are free to withdraw their deposits

after one period.

The bank competes with other banks to attract depositors. For simplicity, we assume that the market for depositors is perfectly competitive. This implies that the interest rate on deposits is bid up to the point where the bank is breaking even. We suppose that the deposit rate over two period is R_d and that $R_d > 1$. (R_ℓ is larger than R_d, the difference reflecting other costs the bank must bear to administer the loans and deposits.) If we consider for simplicity a depositor that happens not to make or receive any payments over these two periods and doesn't withdraw their funds from the bank, this depositor will have R_d dollars in deposits at the bank per dollar of deposits they had at the beginning of the two periods.

With this setup, we get to the heart of the model, which is the situation of the depositors after one period has passed. At this point, the bulk of the bank's assets are illiquid, as the bank has lent out (for two periods) most of the funds it received. In contrast, the bank's liabilities are liquid, as each depositor is allowed to withdraw their funds at will.

Let's consider first what happens if a single depositor gets cold feet about the bank and decides to withdraw all their deposits. Suppose that this one depositor accounts for a very small fraction of the bank's total deposits. The bank is therefore able to absorb this withdrawal without trouble. However, for the depositor, switching banks (or stuffing their money under their mattress) entails some cost. We denote this cost by ϵ (the Greet letter epsilon). The payoff this depositor receives is therefore $R_d - \epsilon$ over the two periods.

Contrast this with what happens if all (or almost all) depositors get cold feet about the bank and decide to withdraw all their deposits. In this case, the bank gets into trouble. Its cash reserves are not sufficient to handle such large withdrawals. It may be able to borrow funds from other banks, sell some of its loan portfolio to other banks, or raise fresh equity capital from large investors. But if the the depositor withdrawals occur quickly enough and in sufficient numbers, these efforts will not be sufficient and the bank will fail.

If the bank fails, not all depositors receive their money back. Some are lucky enough to arrive at the bank (or at their mobile banking app) before others manage to. These depositors are able to withdraw their funds and lose only ϵ. Others are less fortunate and are not able to withdraw before the bank suspends withdrawals. These depositors lose much more. For simplicity, we assume that these late depositors receive nothing.

Before the bank run occurs, each depositor doesn't know whether they will be

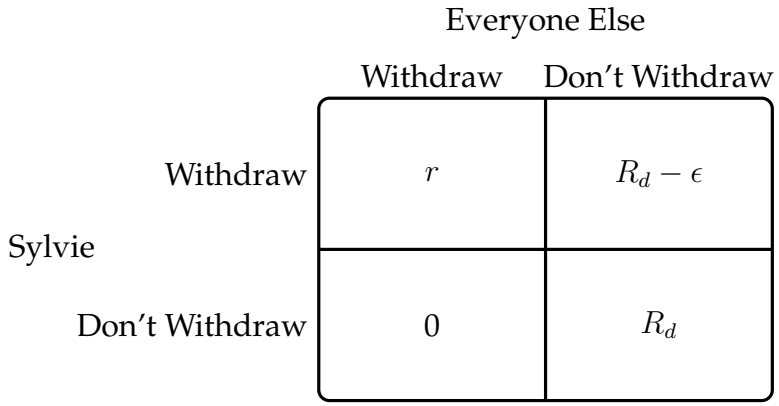

Figure 16: The Diamond Dybvig Game

one of the lucky ones that is quick to react or one of the unlucky ones that doesn't react fast enough. Let's suppose that on average the depositors that attempt to withdraw receive r dollars per dollar of deposits they had with the bank, where $r < 1$. (Here, r is a weighted average of $R_d - \epsilon$ and 0.) Any depositor who doesn't attempt to withdraw when all or most others do, receives nothing.

Figure 16 depicts the situation the depositors face in period one. Since the Diamond-Dybvig model is not a two-player game like the prisoner's dilemma, the depiction in Figure 16 differs from that in Figure 15. In particular, Figure 16 depicts the situation from the perspective of one particular depositor, which we refer to as Sylvie. The rows correspond to the actions that Sylvie must choose between: withdraw or don't withdraw. The columns represent what all other depositors are doing. The payoffs listed in each cell of the matrix are Sylvie's payoffs. For example, the south-west cell gives Sylvie's payoff when she does not withdraw, while all other depositors do withdraw. In this case, Sylvie receives a payoff of zero. Since all depositors face the same situation, each one faces the problem depicted in Figure 16.

Let's now consider Sylvie's best responses to the actions of other depositors. If other depositors do not withdraw, Sylvie's payoffs are given by the two cells on the right in Figure 16. Comparing these two payoff reveals that Sylvie's best response is to not withdraw in this case. If she withdraws, she receives a payoff of $R_d - \epsilon$, while if she does not withdraw she receives a payoff of R_d. Since the other depositors don't withdraw, the bank doesn't fail. In this case, not withdrawing simply saves Sylvie the transactions cost ϵ.

Recall that each depositor faces the same problem as Sylvie. This implies that if no one else withdraws, each depositor's best response is to not withdraw (just like

Sylvie). This implies that it is a Nash equilibrium for no depositor to withdraw.

Next consider Sylvie's situation if other depositors do withdraw. We do this by comparing Sylvie's payoffs in the two cells on the left in Figure 16. In this case, Sylvie's best response is to withdraw. If she withdraws, she receives a payoff of r (on average), while if she does not withdraw she receives a payoff of 0. The crucial difference versus the case considered above is that in this case the withdrawals of other depositors lead the bank to fail. Given this, attempting to withdraw is better than not attempting to withdraw.

Just as in the earlier case, all depositors face the same problem as Sylvie. This implies that if all other depositors are withdrawing, each depositor's best response is to withdraw. As a consequence, it is a Nash equilibrium for all depositors to withdraw.

This analysis shows that the Diamond-Dybvig model has *multiple equilibria*. Everyone not withdrawing is a (Nash) equilibrium. But everyone withdrawing is also a (Nash) equilibrium. So, the model has two (symmetric) Nash equilibria. Let's focus on the run equilibrium (the one in which everyone withdraws). In this case, it is each depositor's belief that other depositors will run that makes it a best response for that depositor to run. There is absolutely nothing "fundamental" that is wrong with the bank, i.e., we have not assumed that there are any indications of loan losses, fraud, or mismanagement at the bank. The run occurs *purely* due to beliefs by each depositor that *others* will withdraw. In this sense, the run is a self-fulfilling prophesy.

Importantly, the no-run equilibrium of the Diamond-Dybvig model is a much better outcome than the run equilibrium. In the no-run equilibrium, the bank does not fail and the depositors receive $R_d > 1$. In the run equilibrium, the bank fails and the depositors receive on average $r < 1$. This feature of the Diamond-Dybvig model captures the fact that bank runs destroy value. In many cases, they destroy a huge amount of value.

8.3 Banking Panics

A bank run can destroy a great deal of value. But worse still, a run on one bank may trigger a run on other banks, and in some cases may trigger a generalized panic with people running on large numbers of banks at the same time. Banking panics have occurred time and again. In England, panics occurred in 1672, 1763, 1772, 1793, 1796, 1811, 1825, 1847, 1857, and 1866 (Feavearyear, 1963). In the United States, major banking panics occurred in 1814, 1833, 1837, 1857, 1873, 1893, 1907, 1930-33

and smaller panics occurred on more than 20 other occasions (Hammond, 1957; Jalil, 2015).

Banking panics often contribute to causing deep recessions. They are thus among the most severe economic calamities that can occur in capitalist economies. Most notably, the Great Depression of 1929-1933 was arguably transformed from a relatively normal downturn into the most serious economic downturn in recorded history by a series of banking panics. Chapter XX is devoted to analyzing the Great Depression in detail.

The Diamond-Dybvig model suggests that bank runs – and by extension banking panics – can occur for no good reason. If depositors get spooked, this can spark a run, which can spark a panic. Presumably something leads the investors to get spooked. But this something may be nothing more than a false rumor. In that sense, the run can occur totally out of the blue.

The Diamond-Dybvig model is deliberately stylized. It is meant to make the point – in as stark terms as possible – that bank runs *can* occur for no good reason. In reality, however, the likelihood that a bank faces a run is affected by the behavior of the bank. Correia, Luck, and Verner (2023) show that bank failures in the United States in the second half of the 19th century and early part of the 20th century were highly predictable. Banks that failed typically experienced a boom-bust pattern. They first experienced unusually rapid lending growth financed by expensive non-core sources of funding. This was followed by a period of decline with rising loan losses leading up to the time of failure. Furthermore, waves of bank failure (i.e., panics) tended to occur when many banks were at risk of failure due to weak fundamentals. Others have also argued that banking panics tended to occur when bad shocks hit a particular region (e.g., Gorton, 1988; Calomiris and Mason, 2003b).

An example of a banking panic initiated by banks with weak fundamentals is the panic of 1907. This began in mid October of 1907 when eight banks "controlled through stock ownership on margin by a few men of no great financial standing, who used the banks to further speculation in the stocks of copper-mining companies", required assistance (Friedman and Schwartz, 1963). It then spread to the Knickerbocker Trust Company – a large New York bank – because of that bank's connections to the troubled banks.

While the banks that initially faced trouble in 1907 had clearly taken undue risks, a general panic developed in the following weeks with depositors withdrawing funds from all manner of banks. The panic therefore (arguably) morphed from crisis driven by fundamentals into a crisis driven by the self-fulfilling nature of bank runs

along the lines of the Diamond-Dybvig model. Further bank failures were eventually curtailed by a general suspension of convertibility of deposits into currency (Sprague, 1910; Friedman and Schwartz, 1963). This was common in the 19th and early 20th centuries. While panics did cause some bank failures, their primary negative consequence was more general disruption in financial markets and the payment system.

Discussions of bank runs often center on the question of whether the run is a "liquidity crisis" or a "solvency crisis." This language is meant to distinguish between cases where the bank is fundamentally solvent and is only in trouble because of the run (as in the run equilibrium of the Diamond-Dybvig model) and cases where the bank is fundamentally insolvent. It is often argued that banks facing liquidity crises should be saved, while those facing a solvency crisis should be let fail.

It is not clear that this liquidity vs. solvency crisis view is a good way to approach the policy response to a banking panic. An important challenge to this view is that it is often difficult to assess whether a bank is solvent in the midst of a crisis. One reason for this is that many of the bank's assets are highly illiquid and therefore do not have easily measurable market prices (think of loans the bank has made to firms and households).

Another reason is that banking panics can result in *fire sales* of assets. Fire sales can occur when large quantities of assets are offered for sale over a short period of time (as can happen when banks facing runs seek to sell assets quickly to raise funds so that they can meet deposit redemptions). In such cases, there may not be enough buyers that are sufficiently knowledgable about the assets to be willing to purchase them at their true fundamental value. This can drive the price of these assets below their fundamental value. Exacerbating this, potential buyers may worry that the banks will continue to sell and that this will continue to drive prices down. This may lead potential buyers to wait on the sidelines for the crisis to "bottom out" before they start buying (or even to amplify the downward spiral by themselves selling).

These problems can lead prices of assets to temporarily fall far below "fundamental value" during a crisis. If the assets of banks are *marked to market* – i.e., their value is updated from whatever the banks value them at in their books to the then-current market value – the banks may seem insolvent. But this may be an artifact of the ongoing fire sale in asset markets. The banks may be solvent if the fire sale were to subside and prices were to return to their fundamental value. In addition to this, the banks may have substantial franchise value – value from future operations. Of

course, these factors are hard to judge in the midst of a crisis since both the fundamental value of many assets and the bank's franchise value are highly uncertain.

In addition to this, it is unclear that banks should always be allowed to fail even if they are insolvent. The failure of large banks or large numbers of banks can cause macroeconomic distress by contracting credit in the economy. A large empirical literature has shown that negative shifts in the supply of bank credit has important consequences for employment, sales, and investment of the firms that are customers of the affected banks and the regions in which these firms operate (see, e.g. Peek and Rosengren, 2000; Calomiris and Mason, 2003a; Chodorow-Reich, 2014; Huber, 2018). The reason for this is that much economic activity is highly reliant on bank credit. When banks fail, the ability of the financial system to intermediate credit is temporarily impaired and this has negative consequences for firms and households. In this sense, large banks are arguably *systemically important* for the economy and may be *too big to fail*.

Given their potential severity, the prevention of banking panics is arguably among the most important public policy problems we face as a society. The long list of banking panics in England and the United States discussed above indicates that banking panics were quite frequent in the 19th century. But then at some point, they became much less frequent. Banking panics ceased to occur in England after 1866 and in the United States they ceased to occur after 1933. We discuss below how this was arguably a consequence of improved economic policy.

8.4 Suspension of Convertibility

The simplest and crudest "tool" available to arrest a bank run is for the bank to restrict or fully suspend convertibility of deposits. A mild version of such a policy is for a bank to offer extremely slow service to depositors seeking withdrawal during a run (e.g., counting and recounting currency slowly). This allows the bank more time to sell assets or raise funds in other ways. Also, the fears of depositors may fade over time as they see the bank honor withdrawal requests.

A bank that is under more severe pressure may place restrictions on deposit withdrawals. For example, it may place an upper limit on the size of withdrawals allowed by each depositor on each day. This will again slow the outflow of deposits.

Finally, if withdrawals become severe enough, a bank may simply close its doors to depositors seeking withdrawal. This is commonly referred to as suspension of convertibility. As with milder restrictions, the aim of suspending convertibility is to

buy time. The bank can attempt to raise funds. Those considering buying the bank's assets, lending the bank funds, or injecting equity into the bank have more time to gather information about the bank and its assets. If a fire sale has lowered the prices of assets the bank holds, suspending convertibility allows the bank to wait and sell assets when market conditions return to normal. If the bank finds it hard to raise funds during the crisis (perhaps because other banks are also facing runs), the bank can also wait to raise funds until the crisis subsides. It may also be that the fears of depositors subside.

However, suspension of convertibility has a serious downside. It effectively creates a dual monetary system. To the extent that currency is needed to make certain payments, deposits become an inferior form of money. They will thus trade at a discount relative to currency. This discount will fluctuate depending on the perceived health of the bank and the perceived length of the suspension. As a result, much of the money in the economy will lose its "not questions asked" property and transactions costs in the economy will rise substantially.

In some cases, various other forms of private liabilities may start circulating as money further increasing the complexity of monetary exchange. For example, banks may issue cashiers checks, banking associations may issue loan certificates, and employers may issue payment certificates in lieu of cash pay. These private liabilities may then circulate as money.

Widespread suspension of convertibility occurred in the United States in the severe financial crises of 1893, 1907, and 1933. In the first two of these cases, convertibility was suspended quickly after the onset of the crises and was coordinated by clearinghouses (Gorton, 1985). In sharp contrast, convertibility was suspended in 1933 after three years of ongoing crisis. Friedman and Schwartz (1963, ch. 4.3) argue that the suspension in 1907 avoided a much more serious collapse of the banking system than actually occurred. Yet, suspension resulted in many households and firms having difficulty making needed payments. Especially in the case of the suspension in 1893, many firms had difficulty making payroll payments which resulted in hardship for employees and their families.

8.5 Lender of Last Resort

A central bank with unlimited resources can stop a banking panic by acting as a *lender of last resort*. Suppose the central bank is willing to lend enough to a bank facing a run that the bank can honor the withdrawal requests of depositors even if

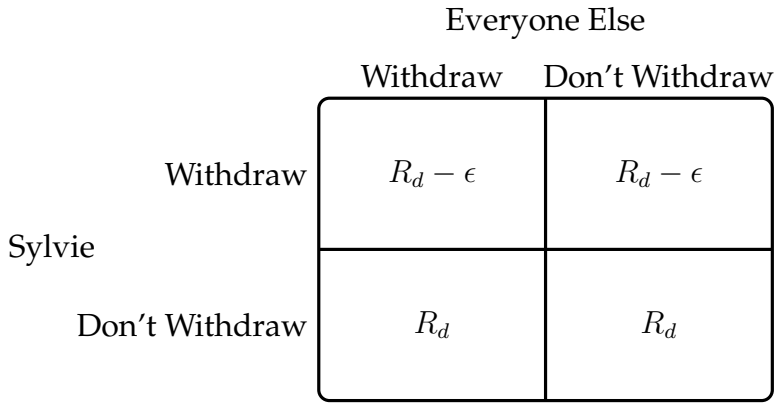

Figure 17: The Diamond Dybvig Game with Lender of Last Resort

all depositors request withdrawal. This completely changes the incentives facing the bank's depositors. Since the bank doesn't fail even if there is a run, the depositors don't have an incentive to run.

Figure 17 depicts the payoffs our representative depositor Sylvie faces when her bank is backed by a lender of last resort. In this case, Sylvie's payoffs do not depend on what other depositors do. Whether or not other depositors withdraw, she can rest assured that she will receive R_d if she does not withdraw. If other depositors withdraw, the central bank lends the bank enough funds that it can pay those depositors. The depositors that don't withdraw are therefore unaffected by the withdrawals of other depositors. In this case, therefore, Sylvie's best response is to not withdraw whether or not other depositors withdraw.

As before, all depositors face the same incentives as Sylvie. This means that it is a best response for all depositors to not withdraw irrespective of what other depositors do. As a consequence, the unique Nash equilibrium of the Diamond Dybvig game in the presence of a lender of last resort is for no depositor to withdraw. In other words, the presence of a lender of last resort eliminates the run equilibrium in the Diamond Dybvig game.

Interestingly, the mere public announcement (if it is credible) by the central bank that it will lend as much as is needed to a bank that is in danger of a run can prevent the run from happening. In this case, the central bank will not actually have to do any lending. This may seem like magic. But it is magic that is backed up by the immense resources of the central bank.

8.6 Bagehot's Principles

Early in the history of banking, central banks did not exist. Furthermore, the idea that a lender of last resort was an effective means of preventing banking panics was not well understood. It was in the 19th century that these ideas were developed. A classic early exposition of the logic of last resort lending by a central bank was provided by Henry Thornton in his 1802 book *The Paper Credit of Great Britain* (see, in particular, chapter 7 of that book). But the most famous exposition of these ideas is that of Walter Bagehot (pronounced "badge-et") in his 1873 book *Lombard Street*.

Bagehot's policy prescriptions for a central bank in a crisis are commonly summarized by the following principles:

1. Lend freely
2. At a penalty rate
3. Against good collateral.

To understand the logic of these principles and the essential role that central banks play during a banking crisis, it is useful to consider the backdrop against which Bagehot originally made his arguments, i.e., that of England in the late 18th and early 19th century.

At that time, the English financial system consisted of hundreds of country banks outside of London – many of which issued their own paper notes – dozens of London banks (and bill brokers) – these provided financial services in London and also serviced the country banks – and the Bank of England. The Bank of England was a private, for-profit bank. However, it enjoyed certain special privileges by law: a monopoly on joint stock banking in England until 1826 and a monopoly on the issue of bank notes within a sixty-five mile radius of London after that date. In return, it was expected to provide certain public services. But exactly what was required of it in this regard was vague and at times controversial.

Importantly, the essential character of a banking crisis was the same 200 years ago as it is today: bank customers lose confidence in banks and begin withdrawing their funds. In early 19th century England, this might manifest itself in several ways. For example, customers of country banks might begin exchanging their country bank notes for gold, or customers of London banks might withdraw deposits in exchange for Bank of England notes. Often, the loss of confidence was somewhat gradual to begin with, but could spiral out of control quickly.

Let's consider a period of moderate alarm about banks when only their most attentive and risk averse customers are withdrawing funds. These withdrawals cause a drain of the reserve assets held at the banks in question. For concreteness, consider a case where customers of country banks are demanding gold in exchange for country bank notes. This leads to a drain of gold reserves at these banks.

In response to this type of drain, the country banks would seek to shore up their reserves of gold by drawing down their deposits at their correspondent banks in London or by discounting (i.e., selling) some of their financial assets (typically bills of exchange) to their correspondent banks. This would transfer the gold drain from the country banks to the London banks. The London banks would then seek to shore up their gold reserves by drawing down their deposits at the Bank of England or discounting bills of exchange at the Bank of England. This would transfer the drain to the Bank of England.

In addition to this, a natural reaction of banks during a time of alarm is to hoard gold reserves. Each bank is worried that the crisis may intensify. Each bank therefore has an incentive to increase its precautionary holdings of gold reserves. This leads each bank to become more reluctant to buy bills at discount and make loans. This precautionary reaction makes the overall situation more difficult, since it makes it more difficult for those banks facing a gold drain from customers to acquire the needed gold.

Were the Bank of England to view itself as just another bank, it would restrict its lending in an effort to act prudently and hoard gold just as other banks were restricting their lending. But if all banks hoard gold, then there is no source of gold that can accommodate the increased demand for gold from the public.

The fundamental problem in such a circumstance is that the alarm has led to an increase in the *aggregate* demand for gold. The public is demanding more gold, and this is leading the banks to demand more gold as well. Unless someone is willing to supply more gold to the market (or a substitute for gold) the banks facing a drain from customers will be at risk of depleting their gold reserves and "failing" (i.e., suspending convertibility of their notes and deposits). Such failures are likely to intensify the alarm and may turn it into a serious panic.

This is what happened, for example, in England in 1825. A speculative boom had crested and some market participants were getting worried about a possible bust and the consequences of such a bust for banks. This alarm led to withdrawals from country banks. By November that year, the alarm had turned into a crisis that intensified week by week. The Bank of England refused to grant banks assistance.

By mid December, some 61 country banks and 6 important London banks had suspended convertibility and every day brought news of new failures (Feavearyear, 1963, p. 237). Finally, after a week of full blown panic, the Bank of England was compelled to act. It began lending freely both gold and its own bank notes. This ended the panic and restored confidence.

Bagehot argued that the Bank of England should as a matter of policy stand ready to lend freely in a crisis. It should do this from the very beginning of a crisis, and it should make it known that this was its policy. By doing this, the Bank would supply the financial system with the additional gold and bank notes needed to accommodate the increased demand for these forms of money. Furthermore, by making it know that the Bank would act in this way – i.e., act as a lender of last resort in a crisis – the Bank would calm markets and dissuade many depositors from withdrawing their funds. The Bank could in this way eliminate self-fulfilling runs, as discussed above. In Bagehot's words:

> In wild periods of alarm, one failure makes many, and the best way to prevent the derivative failures is to arrest the primary failure which causes them. (Bagehot, 1873/1999, p. 51)

8.6.1 Last Resort Lending and Inflation

An objection to Bagehot's proposal which was voiced at the time and has been voiced frequently since is that lending freely will increase the monetary base (in some cases massively) and thus lead to inflation. Those voicing this objection fail to appreciate that banking crises involve large increases in *demand* for high-powered money (gold and Bank of England notes in Bagehot's time). If the central bank refuses to accommodate this increase in demand for high-powered money, the result will be severe deflationary pressure.

We can see this most easily using the quantity theory model discussed in chapter XX [quantity theory chapter] and adopting a narrow definition of money (as only high-powered money). Recall that the quantity equation says that $M_t V_t = P_t Y_t$. In chapter XX, we assumed that velocity V_t was constant. But the increased demand for high-powered money in a banking crisis is a negative shock to velocity V_t. If this is not offset by an equiproportionate increase in the monetary base M_t, the banking crisis will result in large negative pressure on $P_t Y_t$.

If one alternatively adopts a broader definition of money (as M1 or M2), it is crucial to appreciate that banking crises involve a decrease in the demand for bank de-

posits (and bank notes of country banks in Bagehot's time). People are withdrawing money from banks and thus converting bank deposits to currency. This increases the currency-deposit ratio (C/D) and decreases the money multiplier (see section 7.4). Again, if the central bank refuses to offset the decrease in the money multiplier with an equiproportionate increase in the monetary base, the overall money supply will decrease resulting in potentially large deflationary pressure.

Rather than causing inflation, lender of last resort lending during a crisis will allow the economy to avoid deflation. The US Federal Reserve failed to react to severe banking panics during the Great Depression for three years. One result of this was in fact a huge deflation.

8.6.2 Fiscal Risks of Last Resort Lending

A different objection to Bagehot's proposal is that lending freely is imprudent from the perspective of the central bank's own finances. In the early 19th century, the Bank of England still thought of itself largely as a private for-profit bank. Holding large reserves in good times to be able to lend freely in a crisis meant not lending out those reserves and thus reduced the Bank's profits. Also, lender of last resort actions could result in massive increases in the size of the Bank's balance sheet and, therefore, a large increase in the Bank's leverage ratio. This would make the Bank's capital much more vulnerable to modest proportional loan losses, as explained in section 7.3.

Today, these issues is viewed differently. Central banks are public (or quasi-public) institutions. They are generally viewed as being backed by the resources of the government. They are generally not run with an aim to maximize profits and also not liable to become insolvent in the way a private bank might. Large amounts of last resort lending nonetheless carry risks of large losses that ultimately would be borne by taxpayers.

One rationale for Bagehot's prescription that last resort lending should be done at a penalty rate (he used the term 'high rate') is that this compensates the central bank for the risk it is taking and the service it is providing banks. Last resort lending is a form of insurance provided by the central bank against the risk of a panic. Providing such insurance is costly. The central bank must in good times maintain ample reserves and ample financial capacity (capacity to expand its balance sheet) so that it is able to provide assistance when a crisis erupts. Maintaining reserves and financial capacity is costly. The central bank should be compensated for this.

In addition, last resort lending is risky lending. By its very nature, last resort lending is done in very uncertain times (during a crisis). The banks that are seeking central bank financing at penalty rates are doing so because they find it hard to finance themselves by other means. While purely self-fulfilling bank runs are a possibility, we have seen the banks are more likely to face runs if their fundamentals are weak. The central bank typically has limited time to conduct due diligence on the balance sheet of the banks it lends to during a crisis. It is therefore taking risks when it engages in robust amounts of last resort lending. Lending at a penalty rate compensates the central bank for taking these risks.

Bagehot's third principle – that central banks should lend against good collateral – is meant to limit the risks the central bank takes when it engages in last resort lending. In theory, if the central bank only lends against good collateral, its risk will be minimal, since it can seize and sell the collateral in the event that the borrowing bank fails. A related practical benefit of lending against good collateral is that the central bank does not need to evaluate the entire portfolio of the bank it is lending to. As noted above, last resort lending must often be done very quickly. This makes it hard to evaluate the bank's entire portfolio to assess whether the bank is solvent. It is much simpler to evaluate a particular set of securities offered by the bank as collateral.

While the idea that a central bank should lend freely as long as it is lending against good collateral is a good rule of thumb, it is not a panacea. As we observed above, many bank assets are not traded on markets (e.g., loans to households and firms) and, therefore, don't have easily referenced market prices. Also, panics can result in fire sales of various assets that drive the prices of these assets far below their "fundamental value." These complications imply that a good deal of judgment is inevitable when the central bank engages in last resort lending even when such lending is collateralized.

A key question that arises is whether the central bank should value collateral at current market prices or the prices the assets would fetch in the absence of the panic. Proponents of robust last resort lending typically argue for the latter. Critics sometimes decry such lending as bank bailouts for reasons discuss in more detail in section 8.7.

A related question is how large a "haircut" the central bank should set in its collateralized last resort lending. The haircut in collateralized lending refers to the fraction of the value of a piece of collateral that the central bank allows a bank to borrow when posting that collateral. For example, if the central bank sets a 20%

haircut on a particular class of collateral, this means that a bank can borrow 80% of the value of the collateral that it posts in that class. The lower is the haircut, the more the central bank allows banks to borrow. Typically, the central bank will set a relatively modest haircut for lending against very safe collateral (e.g., government bonds) but a higher haircut for lending against less safe collateral (e.g., corporate bonds or mortgage banked securities).

8.6.3 Last Resort Lending and a Run on the Currency

Another (related) concern in Bagehot's time was that lender of last resort actions resulted in potentially substantial decreases in the Bank of England's gold cover ratio (the ratio of gold reserves to bank notes outstanding). Lending freely during a crisis could involve large increases in notes outstanding. Crises would also often be times of substantial decreases in gold reserves both due to domestic demands for gold and foreign outflows. For both of these reasons the gold cover ratio of the bank would decrease. This posed the risk that the Bank might face a run on its gold reserves. Whether this was more likely when the Bank acted aggressively as a lender of last resort (potentially averting a panic) or when the bank did not act (potentially allowing a crisis to spiral out of control) is not clear. Thornton and Bagehot argued that it was less likely when the bank acted aggressively as a lender of last resort.

A similar concern arises today for central banks that fix their exchange rate to another currency. These central banks must maintain a reserve of that currency (a foreign exchange reserve) to be able to honor their commitment to fix the exchange rate. This is analogous to being on the gold standard and having to maintain a reserve of gold to maintain the fixed exchange rate of a currency with gold. Central banks that maintain a fixed exchange rate face the risk of a run on their currency in a banking crisis. Assisting the banks in their country increases the size of their balance sheet as well as the quantity of currency outstanding and may also reduce their foreign currency reserves to the extent that people in the country demand foreign currency. At some point, doubts can arise regarding the central bank's ability to maintain the fixed exchange rate. This is, for example, what happened in a number of countries in the Great Depression, and more recently in Mexico in 1994. We will discuss this issue in more detail in chapter XX [Great Depression chapter].

8.7 Moral Hazard and Bank Bailouts

By far the most prominent and persistent concern with Bagehot's proposal is that lending freely during a crisis results in *moral hazard*. The basic idea is that banks will act less prudently if they know that the central bank will provide them with support during a crisis. This will in turn make crises more likely to occur. Some go so far as to say that moral hazard is the primary reason for banking crises: if only the government were to *commit* not to bail out banks, we wouldn't have banking crises, since banks would act more prudently (the argument goes).

Concerns about moral hazard are not confined to banking. Wherever there is insurance, some degree of moral hazard is an unintended side effect. This is true of auto insurance, home insurance, health insurance, workers compensation (i.e., insurance of workers on the job), unemployment insurance, old age pensions (a form of insurance against poverty in old age) and a host of other types of insurance. In all of these cases, moral hazard arises because the party that is insured does not bear the full cost of risky actions they may take and also does not reap the full benefit of preventive actions they may take. If the bad event that they are insured against occurs, the insurance company bears part of the costs. Because of this, the party that is insured will engage in more risky behavior and less preventative behavior than if they were not insured. For example, an unemployed person will search less hard for a job if they have unemployment insurance than they would if they didn't have unemployment insurance.

In the case of banking, one can view the lending of a central bank in a crisis as a form of insurance. If banks know that the central bank will lend freely in a crisis, they are less worried about crises and take less precautions in the form of holding reserves, limiting their leverage, and making safer loans. The increased risk-taking by banks then makes bank failures, runs, and panics more likely.

There is no doubt some truth to this view. How much is hotly debated. Some view moral hazard as a severe problem that is a principle contributor to banking crises, while others think it less severe. This debate has raged for over 200 years. Unfortunately, convincing empirical evidence has yet to settle it.

Bagehot's proscription that central banks should only lend against good collateral is one mechanism through which central banks can limit moral hazard. Banks will then know that there is a limit to the amount they can borrow from the central bank during a crisis and that this limit is determined by the value of the assets they have to post as collateral. Concerns about moral hazard have also motivated the en-

actment of considerable government regulation of banks, which we discuss in more detail in section 8.10 below.

Those most worried about moral hazard, sometimes argue – as noted above – that the problem of moral hazard can be eliminated simply by committing the government not to bail out banks (e.g., not to act as a lender of last resort). An important problem with this argument is that a government commitment not to bail out banks in a crisis is not *credible*. The banks know that once a banking crisis occurs the government will not be willing to bear the huge economic costs associated with letting a large number of banks fail. This means that the banks will not believe such a commitment and it will fail to limit moral hazard.

Bagehot made his argument in the aftermath of the Overend, Gurney & Co. crisis of 1866. The Bank of England provided vigorous support to the banking sector during that crisis (as it had done more reluctantly several times before, starting in 1825). Thomson Hankey, a former governor of the Bank, responded violently to Bagehot's argument calling it 'the most mischevious ever broached' (Hankey, 1867). Hankey argued that the Bank should act like any other bank, and all banks should keep enough reserves to meet their own liabilities.

Hankey was a follower of the 'Currency School' of thought on banking matters. A core belief of this school was that the Bank of England should be forbidden from acting as a lender of last resort. These ideas were so influential that they were put into law in the so-called Peel's Act of 1844 which divided the Bank of England into two departments: an Issue Department that could issue only up to 14 million pounds of notes in excess of the specie it had on reserve, and a Banking Department that was meant to be completely separate from the Issue Department and conducted regular banking business. In a crisis, only the Banking Department could lend to banks in need and the Banking Department had a finite reserve of notes and specie. It could not issue new notes.

In effect, Peel's Act was designed to commit the Bank of England not to act as a lender of last resort. Members of the Currency School believed that precluding discretionary note issue by the Bank (in a crisis as well as in normal times) would rid the English financial system of the risk of panics. This turned out not to be the case. A banking panic developed in 1847, only three years after the enactment of Peel's Act. Once the crisis became sever enough, Peel's Act was temporarily suspended to allow the Bank to act as a lender of last resort. The same happened again in 1857, and then again in 1866. By that point, it was clear to all that the Bank would always act as a lender of last resort in a crisis notwithstanding Peel's Act. Interestingly,

Table 1: U.S. Banking Panics, 1866-1929

Major Banking Panic	Non-Major Banking Panic
Sept. 1873	
	May 1884 (New York City, Pennsylvania, New Jersey)
	Nov. 1890 (New York City)
May-Aug. 1893	
	Dec. 1896 (Illinois, Minnesota, Wisconsin)
	Dec. 1899 (Boston, New York City)
	June-July 1901 (Buffalo, New York City)
	Oct. 1903 (Pennsylvania, Maryland)
	Dec. 1905 (Chicago)
Oct.- Nov. 1907	
	Jan. 1908 (New York City)
	Aug.-Sept. 1920 (Boston)
	Nov. 1920 - Feb. 1921 (North Dakota)
	July 1926 (Florida, Georgia)
	March 1927 (Florida)
	Jul.-Aug. 1929 (Florida)

Note: Replicates a portion of Table 2 in Jalil (2015).

once this was clearly established, no further panics occurred in England for over a century (Feavearyear, 1963, ch.10-11).

8.8 Persistence of Banking Panics the United States

While banking panics disappeared in England after 1866 when the Bank of England emerged as a reliable lender of last resort, they continued to plague the U.S. economy for another 70 years. Jalil (2015) documents major banking panics in the United States in September 1873, May through August 1893, and October through November of 1907, as well as 13 less widespread banking panics listed in Table 1. Dwarfing everything before it, the Great Depression saw several waves of banking panics between 1930 and 1933 culminating in a national bank holiday. It was only after the Roosevelt's New Deal legislation was passed in the wake of the Great Depression that banking panics seized to occur in the United States (until 2008). Why did banking panics persist for so much longer in the United States than in England?

8.8.1 The Ghost of Andrew Jackson

One problem was that the United States did not have a central bank. This curious fact was arguably due to idiosyncrasies of American politics. The issue of banks was highly contentious in American politics from the founding of the republic until the 20th century. At play were issues of federal power versus state power, tensions between agrarian interests and industrial interests, and the special interests of private bankers. High inflation in paper money during the revolution (i.e., in prices denominated in the continental dollar) and to some extent during the colonial period may have played a role in anti-bank sentiment. It was understood that banks were issuers of paper money and while some viewed this as an important benefit of banks others saw paper money as dangerous and opposed banks partly because they issued paper money (Hammond, 1957, ch. 1).

Alexander Hamilton championed the chartering of the Bank of the United States in 1791. This bank was in many ways modeled on the Bank of England. Thomas Jefferson, James Madison, and other anti-federalists opposed the Bank's chartering, going so far as to argue that the Bank was unconstitutional. The Bank had a 20 year charter that expired in 1811 and was not renewed. The non-renewal vote margin in both the House and the Senate was only one vote. Curiously, James Madison, who by 1811 was President, supported renewal, and his Secretary of the Treasury Albert Gallatin was the Bank's most ardent supporter. Significant opposition came from the business community, especially those with ties to state banks. They saw the Bank of the United States as a source of competition and unwanted discipline on their own note issuance (Hammond, 1957, ch. 8).

The financial strains of the War of 1812 – when many state banks faced a run and suspended convertibility of their notes into specie – revived interest in a national bank. The notes of state banks traded at varying discounts and there was a strong desire to restore a uniform circulating medium by creating a national currency. The greatest source of opposition to these ideas came, again, from state banks fearing competition and discipline. Despite that opposition, the Second Bank of the United States was chartered in 1816 for 20 years.

In 1832, congress voted to recharter the Second Bank, but Andrew Jackson famously vetoed this bill and let the Bank's charter relapse in 1836. After that time, the United States did not have a central bank until the founding of the Federal Reserve in 1913. The political struggle leading up to Jackson's veto is often referred to as the Bank War. On one side was Nicolas Biddle, the President of the Second Bank

since 1823, while on the other side was President Andrew Jackson and key members of his administration.

In his celebrated history of antebellum banking in America, Bray Hammond sums up the Bank War in the following way:

> In popular accounts the Bank of the United States is most often presented as an embodiment of the "money power," a vague but immense evil, overcome by Andrew Jackson and his agrarian followers. It would be truer to say that it was a victim of the "money power," which used Andrew Jackson, states' rights, and agrarian sentiment to destroy it. (Hammond, 1957, p. 287)

To understand this conclusion, we must consider the main functions of the Second Bank under Biddle's leadership. One function – discussed above – was to restore and maintain a uniform national currency. Biddle's approach to achieving this goal was two-pronged. First, he greatly expanded the note issuance of the Second Bank. He also instituted a policy of redeeming the notes of state banks received by the Second Bank on a weekly basis. Since the Second Bank was the fiscal agent of the federal government – another of its principle functions – it naturally received a large amount of state bank notes through its collection of public revenues. Quickly redeeming these notes was a way to prevent state banks from overissuing notes. The more notes a state bank issued, the more notes were likely to be redeemed by the Second Bank. If a bank issued too many notes, it would risk running out of specie reserves.

This policy of the Second Bank increased confidence in the notes of state banks and thereby lowered the discount on notes of banks when they were offered as payment far from the bank's office location. In this way it contributed to creating a more uniform currency. However, this policy was bitterly resented by state banks since it effectively limited their ability to issue notes (a profitable activity). As Albert Gallatin remarked, the Second Bank "operated as a screw" on the state banks.

In addition, since notes of the Second Bank were superior instruments for use in longer distance trade (their value was more uniform throughout the country), these notes gradually replaced the use of state bank notes in such trade. State bank notes were then confined to local circulation. It seems likely that if the Second Bank had survived it would have soon become the sole bank of issue in the United States totally driving state banks out of the business of issuing notes.

The other principle function of the Second Bank under Biddle's leadership was to

"facilitate internal and external exchanges," which in modern parlance refers to providing trade credit and improving the efficiency of the payment system. The United States is a large country and is separated from Europe by the Atlantic Ocean. These obvious facts imply that the shipment of goods across regions and to other countries in the early 19th century took a considerable amount of time. The economies of the different regions of the United State were highly specialized. The South and the West (which at the time meant the watershed of the Mississippi River and its tributaries) primarily produced agricultural goods and timber for export to the U.S. Northeast and Europe. In contrast, the Northeast sent manufactured goods to the South, West, and to Europe.

New Orleans was in Biddle's words "the centre and the depository of all the trade of the Mississippi and its tributaries." Merchants all along the Mississippi River, Ohio River, and other tributaries of the Mississippi would sell agricultural goods to merchants in New Orleans. The merchants in New Orleans would then resell the goods to the Northeast and to Europe. An important practical problem facing the merchants in New Orleans was how to pay for the goods they purchased hundreds of miles up the Mississippi. One method was to pay with specie. But transporting specie was costly. Another method was to acquire bank notes in New Orleans (e.g., by taking out a loan with a New Orleans bank) and use these as a means of payment in Memphis, St. Louis, or Louisville. But bank notes tended to depreciate in value as they traveled from the issuing bank in the 19th century.

A preferred option from the perspective of the merchant in New Orleans was to pay for the goods in New Orleans once they had arrived. The bill of exchange was a device that made this possible. The seller in St. Louis (say) would draft a bill demanding that the buyer pay their agent in New Orleans (the payee) the agreed upon sum at some future date (say in three months time). The buyer or their agent in St. Louis (i.e., the person arranging the purchase) would accept the bill, thereby turning it into an IOU. The seller would then discount the bill to a local banker, i.e., sell the bill at a discount reflecting the prevailing rate of interest on funds for three months. The local banker would send the bill to their correspondent bank in New Orleans. In the simplest case, this correspondent bank in New Orleans was the payee. It would then rediscount the bill and receive payment when the bill came due (in three months time).

This rather complicated arrangement is depicted in Figure 18. It accomplishes two things simultaneously. First, money does not need to travel from New Orleans to St. Louis. The seller gets payed in St. Louis when they sell the bill of exchange

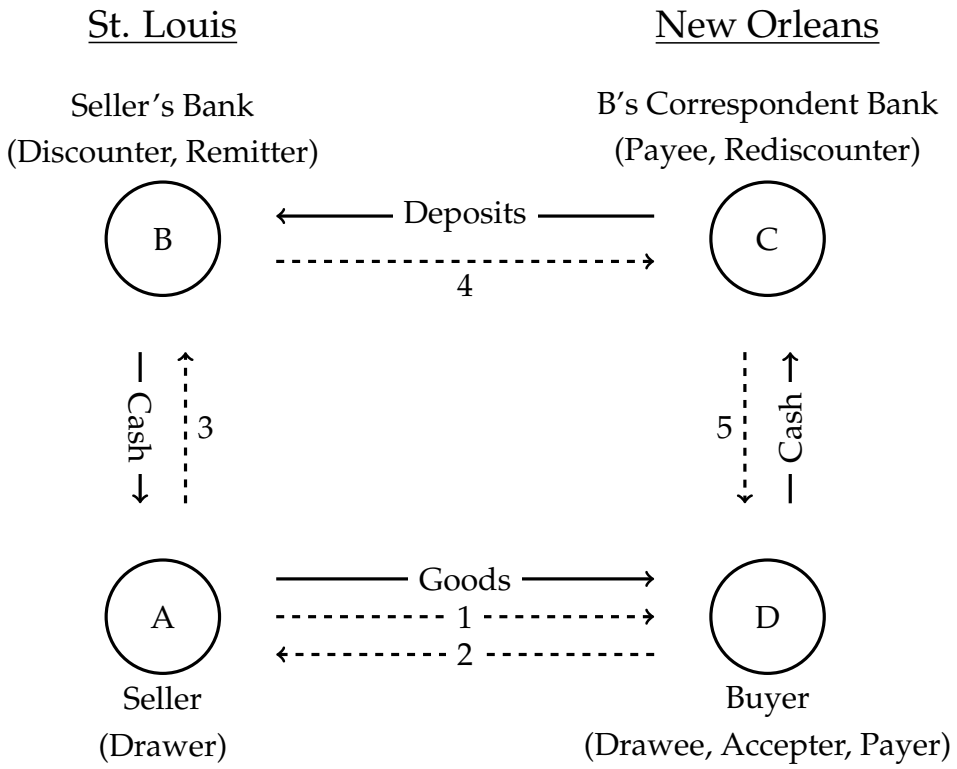

Figure 18: Paying with a Bill of Exchange

Note: The dashed arrows depict the movements of the bill of exchange. The bill is drawn up by the seller (A) on the buyer (D). The seller (or their agent in St. Louis) accepts the bill and leaves it with the seller. The seller sells the bill to a local bank (B) at a discount. The local bank remits the bill to their correspondent bank in New Orleans (C). That bank rediscounts the bills, crediting the St. Louis bank with deposits. When the bill comes due, the buyer pays the New Orleans bank.

to the local bank. The buyer pays in New Orleans by paying the bill when it comes due. Second, the banks involved in the transaction effectively provide the buyer and seller with trade credit while the goods are being shipped from St. Louis to New Orleans. The seller receives funds immediately (when they sell the bill). The buyer doesn't pay until the bill comes due in three months time.

Facilitating interregional (and international) trade by dealing in bills of exchange was a major function of banks in early 19th century America. A highly developed and efficient system of correspondent banks across regions had the potential to lower the costs of trade and thereby spur economic development. However, in the 1820s, the degree of development of the banking system was quite variable across different regions in the United States. The Northeast already had a highly developed banking system, while the South and West less so. This meant that costs associated with making and receiving payments in interregional payments in the South and

West were high.

For this reason, the Second Bank employed its capital primarily in the South and the West in an effort to make interregional exchange more efficient in these regions. This was in all likelihood an important benefit to the economies of the South and the West. But it provided unwelcome competition to state banks that engaged in these activities alongside the Second Bank and saw their profit margins shrink as the Second Bank expanded its activities.

Bray Hammond argues that Martin Van Buren, President Jackson's Vice President and close advisor, was instrumental in bringing about the downfall of the Second Bank. Van Buren was an ex banker from New York and ex Governor of New York. The New York banking community was (arguably) the biggest beneficiary of the downfall of the Second Bank (the headquarters of which were in Philadelphia). In Hammond's telling, Van Buren focused President Jackson's general animosity towards banks on the Second Bank. Davies (2008) provides an accessible account of the Bank War that highlights a host of other issues. Ironically, President Jackson's key constituencies in the South and the West were likely those that lost the most from the destruction of the Second Bank since the payment system in these regions was most reliant on the Second Bank.

After the demise of the Second Bank, the monetary system of the United States regressed substantially. For several decades the country did not have a uniform currency. This was remedied with the National Banking Acts of the Civil War. But the country did not have a central bank until 1914 when the Federal Reserve came into existence. The Second Bank had in many ways been ahead of its times. It acted aggressively to prevent banking panic in 1825 when England suffered a serious banking panic. It also acted aggressively to improve the functioning of the domestic and international payment system as discussed above. On both accounts, it was far ahead of the Bank of England. Perhaps this was its downfall. The Bank of England assumed the role of a central bank more slowly over the 19th century. Perhaps this was important for its survival since it was less of a threat to the rest of the banking system in England than the Second Bank was to U.S. banks.

8.8.2 Unit Banking

Another contributor to frequent banking panics in the United States was the structure of the banking industry. With the exception of the First and Second Banks of the United States, there was no interstate banking in the United States prior to the

1970s. Before the Civil War, banks were state chartered and not allowed to operate across state lines. Even after the Civil War, with the rise of federally chartered national banks, interstate banking was prohibited. Only in the 1970s and 80s did some states begin to allow out-of-state banks to operate within their borders. This process culminated in the passage of the Riegle-Neal Act in 1994, which allowed interstate banking throughout the United States.

In the 19th century, banks in many states were not even allowed to open branches. Each bank was limited to operating in a single location. Such banks were called "unit banks." Unit banking was the rule in the antebellum North, while branching was allowed within state in the antebellum South. Few southern banks survived the Civil War. After the War, branching was heavily restricted in the South as well as the North. In particular, national banks were not allowed to open branches.

Initially, it seems, branching was not an important bone of contention (Calomiris, 2000, p. 46). By the late 19th century, however, significant pressure was building to allow branching and consolidation in the banking industry. Partly, this was driven by the large financing needs of large corporations (a recent development at the time), and partly this was simply due to successful banks wanting to be allowed to grow. Opposition came from unit banks that feared competition from larger banks. Many small towns had a single bank or very few banks. The banks in these areas had substantial market power and did not want to lose it. The unit banking lobby tapped into agrarian fears of wealthy financial elites and managed to slow the spread on branching and bank consolidation considerably until the 1920s and 30s.

From a financial stability point of view, unit banks have the important drawback that they are poorly diversified. They typically receive deposits from and make loans to people and businesses in their immediate vicinity. This makes them highly exposed to local shocks. Booms and busts in land prices, crop failure, and shocks to crop prices are examples of local shocks that were particularly important in the 19th century. Such shocks could result in steeply rising delinquency rates at a unit bank. A regional shock could raise the specter of trouble for many banks in the region and this could spark a panic.

The comparison between the experience of the United States and Canada is particularly informative on this point. Canada operated a branch banking system, while the U.S. mostly did not. Canada has never (since 1839) experienced a banking panic, not in 1857, not in 1893, not in 1907, not during the Great Depression (and not in 2007-09). This remarkable stability of the Canadian banking system suggests that

branch banking is helpful for financial stability.

The difference in Canada's experience may be due to other factors. But in many other way the U.S. and Canada were quite similar. Both countries are vast geographically. Their economies are highly integrated and they trade in similar commodities. During the 19th century, neither country had a central bank. The Canadian banks were allowed to vary their issuance of bank notes more freely (they had a more elastic currency) and this may have played a role. But the relatively consolidated nature of the Canadian banking system meant that Canadian banks were more diversified and it allowed for greater coordination during times of stress with the Bank of Montreal frequently acting as a leader.

The experience of other countries and of regions within the U.S. reinforces the notion that branch banking helps prevent panics. Scotland was an early example of a region with a well developed branch banking system. It experienced no panics. This contrasts strongly with England, which had a banking system dominated by private unit banks prior to 1826 and experienced a number of panics as we discussed earlier in the chapter. Australia is another example of a country with extensive branch banking. Australia only experienced a single large panic in 1893. Finally, within the U.S. the banking crisis of 1837-41 played out very differently in different states. In particular, Virginia and South Carolina, which had mature branch banking systems, did not experience bank failures, while the banking systems of other states did (Calomiris, 2000, p. 23).

8.9 Deposit Insurance

In the wake of the panic of 1907, the United States finally founded a central bank, the Federal Reserve, which opened for business in 1914. An important goal of the founders of the Fed was to prevent future banking panics. The experience in Britain suggested that a central bank could bring financial stability by acting as a lender of last resort. Unfortunately, this not what happened in the United States.

Starting in the fall of 1930, the United States experienced a series of banking panics more severe than any previous panics. The Fed largely refused to act as a lender of last resort. Clearinghouses and other private arrangements that had coordinated action prior to the founding of the Fed had been dismantled or had ceased to play this roll since the Fed was expected to take their place. Absent any coordinated action to curtail the crisis, the banking system cascaded from one panic to the next over the course of three years, and the economy spiraled downward into

what has become known as the Great Depression.

Eventually, the newly elected President Franklin D. Roosevelt called for a national bank holiday (a nation-wide suspension of convertibility) in March of 1933. Banks reopened after they had been inspected for soundness. Many never did. But the panic ended.

An important response to the calamity of the Great Depression was the institution of federal deposit insurance in the United States through the Federal Deposit Insurance Corporation (FDIC). Member banks of the FDIC pay a premium into a Deposit Insurance Fund. In exchange deposits at member banks are insured up to a per person maximum (currently $250,000).

The logic for deposit insurance is that it can eliminate the run equilibrium in the Diamond-Dybvig model. This idea is very simple: if the deposits of a given depositor are insured, they have no incentive to run on the bank even if everyone else is running on the bank.

The institution of federal deposit insurance in the United States is generally considered to have been a success. It ushered in a long "quite period" of financial stability in the United States (Gorton, 2012). Not until the Savings and Loans crisis of the 1980s did the United States experience another wave of bank failures and not until 2007-2009 did it experience anything like a panic. Other new policies were instituted after the Great Depression. Some of these were designed to increase financial stability (such as the separation of commercial and investment banking) and may have contributed to increased financial stability. But the crucial policy innovation was likely deposit insurance.

Deposit insurance was highly controversial when it was initially enacted. Moreover, recent actions that have expanded the scope of deposit insurance have also been quite controversial. For example, during the financial crisis of 2008, the FDIC temporarily guaranteed all non-interest bearing deposits without limit, and the Treasury temporarily guaranteed monetary market funds without limit. Another example is that in 2023 in response to runs on Silicon Valley Bank and Signature Bank, the Treasury, Fed, and FDIC quickly announced that all depositors of these banks would be made whole without limit.

The principle downside of deposit insurance, like all forms of insurance, is moral hazard and adverse selection. Depositors that are insured have no incentive to monitor whether the bank is acting prudently. If the bank runs into trouble, the depositors have no incentive to pull the plug on the bank by withdrawing their deposits. This implies that banks will not face needed discipline from their creditors (the de-

positors). This lack of discipline prevents poor quality bankers from being weeded out and can provide bankers that end up in an impaired financial position with incentives to take excessive risks. A banker that has lost most of their equity will have incentives to "gamble for resurrection", i.e., take large risks in the hope that these pay off since the upside is theirs while the downside will be borne by the deposit insurance fund: heads I win, tails you lose.

One response to the problem of moral hazard is to reject deposit insurance (and last resort lending) as bad policies. Another response is to pair these policies with other policies that are designed to limit moral hazard. Since deposit insurance and last resort lending can give banks an incentive to take too much risk, the aim of these auxiliary policies should be to limit the risks banks take: they should be prudential policies. In fact, last resort lending and deposit insurance has been paired with a host of prudential polices that go under the general heading of bank regulation. We now turn to these policies.

8.10 Bank Regulation

Banks are regulated for various reasons. They are subject to money laundering and and anti-terrorism regulations for reasons having to do with law enforcement and national security. They are subject to consumer protection regulations due to market power they have in consumer markets. In the United States, they are subject to fair lending laws due to a history of discriminatory behavior in lending. Here, however, our focus will be limited to a narrow but important subset of regulations banks face: capital and liquidity regulations.[1]

The underlying core problem that gives rise to capital and liquidity regulation is the problem of bank runs and banking panics. The direct policy response to runs and panics is lender of last resort lending by central banks and deposit insurance. But those solutions give rise to moral hazard on the part of banks: banks have an incentive to take too much risk since the government effectively provides insurance to them on the downside.

Another way to view this issue is that deposit insurance makes deposit financing artificially cheap for banks. Deposits are a very safe asset class due to deposit insurance. This means that the return investors demand for keeping their funds in deposits is quite low. For this reason, it is very cheap for banks to fund themselves with deposits, which means that they have an incentive to tilt the mix of their liabilities towards deposits (debt) and away from equity (which is more expensive). The

tax system in most countries is yet another important force that incentivizes banks to tilt their liabilities towards debt (since profits net of interest payments on debt are taxed). When banks tilt their liabilities towards debt, they raise their leverage. This makes banks riskier, as we discussed in section 7.3.

For these three reasons – moral hazard, excessively cheap deposit financing, and taxes – banks have an incentive to take too much risk, and, in particular, be too leveraged. The policy response to this is to force banks to issue more capital and hold more liquid assets than they would on their own accord.

8.10.1 The Chicago Plan, 100% Money, and Narrow Banking

Early banking regulation took the form of reserve requirements rather than minimum capital-to-asset ratios. For example, the National Banking Act of 1864 required national banks in central reserve cities and other reserve cities to hold gold reserves equal to at least 25% of their notes and deposits, while "country banks" were required to hold reserves of at least 15% (some of which could be deposits at banks in reserve cities). In 1873, notes were exempted from these reserve requirements, and with the establishment of the Federal Reserve in 1913 reserve requirements for national banks were further reduced. (State banks had lower reserve requirements throughout this period.) These reserve requirements where meant to ensure that banks had adequate liquidity. They clearly failed at this: the U.S. experienced frequent banking panics until the mid-1930s, as we discuss above.

The massive bank runs of the Great Depression led to intense debate about how to reform the monetary and banking system during the first years of the Roosevelt administration. An idea that gained quite some prominence at this time – but ultimately did not get enacted into law – was to abolish fractional reserve banking. This idea was put forth by a group of economists at the University of Chicago led by Frank Knight and Henry Simons in a pair of memoranda in 1933 that are often referred to as the Chicago Plan. The idea was taken up by many economists including Irving Fisher under the banner of "100% money" and has been influential ever since.

The basic idea is that banks are vulnerable to runs because they do not have enough reserves to back their deposits (they have a fractional reserve). This problem can be solved by requiring banks to hold a 100% reserve against deposits. Any reserve requirement below 100% does not solve the problem since deposit withdrawals reduce a bank's reserve ratio whenever the reserve ratio is below 100%.

This means that withdrawals put the bank at risk of violating its reserve requirement unless its reserve ratio is 100% (or more).

While raising the reserve requirement to 100% clearly helps solve the problem of bank runs, it has the side effect that banks cannot use deposits to finance lending. All lending must then be financed by equity or long-term debt. One variant of this system envisions two separate classes of institutions. One type of institution would take deposits and back these dollar-for-dollar with reserves. (In the modern context, these reserves could be Treasury bills or interest-bearing central bank reserves.) This type of institution is sometimes called a *narrow bank*. It does nothing other than facilitate payments by offering demand deposits, and is clearly run-proof.

The second type of institution would make loans. It would be financed entirely by equity and long-term debt. Since equity and long-term debt is not subject to runs, this second type of institution would also be run-proof. This separation of banking into two separate classes of institutions – one providing payment services and the other making loans – would therefore solve the problem of bank runs.

One can also imagine a single institution that both takes deposits and makes loans but is fully run-proof. Such an institution would need to back all deposits dollar-for-dollar with reserves and it would need to finance all other activities 100% with equity. This means that all lending would be financed 100% by equity, but also that other activities such as market making, custodial activities, and other trading be financed 100% by equity. (Lending and other activities could be financed partly by long-term debt as long as the reserves backing the deposits of the bank were "ringfenced" so that insolvency of the other part of the bank could not result in bond holders making a claim on the reserves.)

The discussion above focuses on deposits. But deposits are not the only class of liabilities that is *run-prone*. Other asset classes that are run-prone include money market accounts, short-term repo financing, and asset-backed commercial paper. Runs on these asset classes played an important role in the Global Financial Crisis of 2007-2009 as we discuss in greater detail in chapter XX [Great Recession chapter]. To fully avoid runs, banks must not be allowed to finance lending or other activities with any run-prone type of liability.

Discussions of narrow banking often fail to recognize that this idea contains two components: 1) a 100% reserve requirement for run-prone liabilities, 2) a 100% equity (or long-term debt) requirement for lending and other activities. Both of these elements are important. But arguably in the modern era it is the second element – high equity requirements on lending and other activities – that is the more impor-

tant component of this idea. Banks that have a large amount of equity are not likely to become insolvent in a crisis. This makes runs less likely and facilitates lender of last resort lending if a run does occur.

An important objection to narrow banking is that forcing banks to fund lending with equity would increase the cost of lending and therefore reduce the amount of lending in the economy. To the extent that this is true, it would reduce investment, capital accumulation, and ultimately the level of GDP in the economy. From a political economy point of view, this objection is very potent since banks are a strong lobby group and the prospect of higher interest rates, and lower growth are politically unpopular. Unfortunately, it is difficult to estimate empirically how large the effect of this type of policy would be on interest rates and lending. More and better empirical evidence on this point is sorely needed. Also, it is worth contemplating that part of the reason why forcing banks to finance lending with equity would raise their cost of capital is that current policy makes debt financing artificially cheap (through deposit insurance and the tax treatment of debt).

8.10.2 Early Capital Regulation

The conditions that today give banks a strong incentive to tilt their portfolios towards high leverage did not exist in 19th century America: there was no lender of last resort, there was no deposit insurance, and taxes were low. In addition, shareholders in banks often faced double liability, meaning that if the bank failed, creditors could demand that shareholders cover losses up to the amount the shareholders had originally invested in the bank. For these reasons, it is not surprising that banks were not particularly leveraged in 19th century America.

Figure 19 plots the ratio of capital to assets of national banks in the United States from 1863 to 2021. This ratio was close to 1/3 in the early post-Civil War era. It fell gradually over time to about 20% in 1900 and fell further (particularly during World War I) to about 13% on the eve of the Great Depression. After the Great Depression, deposit insurance was introduced and double liability was eliminated. The capital-asset ratio of banks continued to fall (particularly during World War II) until it stood at only about 6% in the late 1940s. A capital-asset ratio of 6% implied that banks were on average leveraged almost 17 to 1 at this point.

Low capital-asset ratios of banks led bank regulators to focus increased attention on the question of whether banks were adequately capitalized. This is a complex question since it depends critically on the riskiness of the assets that banks hold.

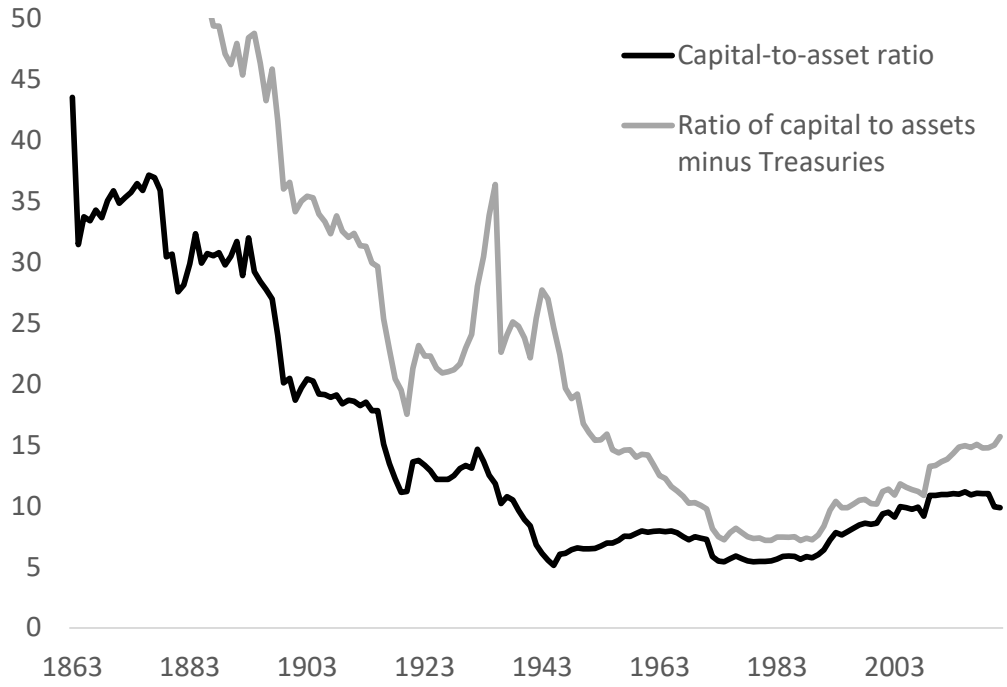

Figure 19: Capital of National Banks in the United States

Note: This figure plots the ratio of capital to assets (not risk weighted) of national banks in the United States as well as the ratio or capital to assets minus holdings of Treasuries by these banks. These data are from Tufts and Moloney (2022a,b).

Banks need to maintain capital to guard against the risk that their assets lose value to the point that they become insolvent. The safer are their assets, the less capital is needed.

In the early post-WWII period, banks held a substantial amount of Treasuries on their balance sheet. Figure 19 plots the ratio of equity to assets less Treasuries for national banks. In the late 1940s, this ratio was roughly 20%. Treasuries are very save assets in that they have minimal credit risk – i.e., risk of default. Longer-term Treasuries do carry interest rates risk (bond prices fall when interest rates rise, as we discuss in more detail in chapter XX [IS-LM chapter]), but overall Treasuries (especially short-term Treasuries) are very safe assets.

With so many Treasuries on their balance sheet, it was not clear that banks were excessively risky in the late 1940s despite having a very low capital-asset ratio. But over the subsequent years, banks reduced their holdings of Treasuries substantially and the riskiness of their portfolios increased.

The three federal bank regulators – the Federal Reserve, the Federal Deposit Insurance Corporation (FDIC), and the Office of the Comptroller of the Currency

(OCC) – each developed separate methods for assessing capital adequacy in the early post-WWII period. Some methods – such as the Fed's Analyzing Bank Capital (ABC) method – were quite quantitative and attempted to adjust for differences in the riskiness of different assets. But on the whole, bank supervision in these years was largely qualitative and involved a great deal of case-by-case judgment. One reason was skepticism that quantitative rules could do justice to the complexity of assessing the riskiness of bank assets. In 1952, the Board of Governors of the Federal Reserve argued that "it is impossible to develop any formula which eliminates the need for the exercise of sound judgment in determining the adequacy of capital of any given bank" (Board of Governors, 1952).

Another problem facing regulators was that they lacked legal authority to enforce minimum capital-asset ratios. The Federal Reserve lost a court case in 1959 when it attempted to revoke a bank's Federal Reserve membership due to inadequate capital. In 1983, the OCC lost a court case were it attempted to close a bank due to inadequate capital. This latter case, which occurred in the wake of several high profile bank failures in the 1970s, was the catalyst for new legislation authorizing federal bank regulators to set formula-based capital requirements.

8.10.3 Basel I and II

The 1980s saw a wave of bank failures in the United States. Most of the banks that failed were small, but a few were quite substantial. Also, large banks faced increasing competition from abroad, and, in some cases, substantial losses from the Latin American debt crisis of the early 1980s. One way in which U.S. bank regulators reacted to these failures was by imposing minimum capital requirements for banks. The earliest measures of this type were very simple in that they did not attempt to weigh assets by their risk.

It was understood that these simple initial measures were unsatisfactory, and work soon began on risk-adjusted capital requirements. This work involved negotiations between the Federal Reserve, the FDIC, and the OCC. Also, it was also widely appreciated that differences in capital regulation across countries would result in an uneven playing field among international banks. This led to negotiations first between bank regulators in the U.S. and the U.K., then with Japanese regulators, and ultimately with the G-10 (a group of 11 countries). The result was the 1988 Basel Accord (later known as Basel I).

The core of the Basel I agreement was a requirement that the capital-asset ratio of

banks must stay above some minimum level. A key challenge in the design of this type of regulation is how to define the numerator and denominator of the regulatory capital-asset ratio. For the numerator, the question is which types of liabilities should count as capital. The liabilities that count as capital need to be able to absorb losses on the asset side of the bank's balance sheet, and in that way shield the bank's deposits and other run-prone liabilities from these losses.

The Basel I agreement defined two measures of capital. *Tier 1 capital* consists mostly of common equity, but also includes perpetual preferred stock, and some other items. *Total capital* is the sum of tier 1 capital and tier 2 capital. Tier 2 capital consists of allowances for loan losses, some other categories of preferred stock, and some subordinated debt.

Preferred stock is a hybrid liability that shares some elements of common equity and some elements of debt. It typically has no voting rights, but some preference when it comes to dividend payments. For example it may receive a higher dividend than common stock and may receive dividends even when dividends for common stock are cut. Preferred stock is also typically convertible into common stock and often callable by the issuing bank.

A key attribute of a liability is its seniority in the event of bankruptcy. Common equity is the most junior claim in bankruptcy, the residual claimant after all other liabilities have been paid. Preferred stock is senior to common equity, but junior to other liabilities. Subordinated debt is junior to other debt, but senior to preferred stock. Rules about the seniority of other debt instruments vary across countries. In the United States, deposits are senior to other debt and insured deposits senior to uninsured deposits.

A key issue when it comes to deciding whether a particular class of liabilities should count as capital for regulatory purposes is whether it is credible that authorities will allow these liabilities to take losses in the event of a systemic banking crisis. If the government is not willing to allow a class of liabilities to take losses (because of perceived negative consequences of this occurring for financial or economic stability) then those liabilities are not really loss absorbing liabilities in the event of a crisis.

This issue was important in the 2007-2009 financial crisis, when authorities were reluctant to allow certain classes of debt contracts to take losses. After this crisis, there was much discussion of introducing a class of contingent convertible bonds ("CoCo bonds"). This was meant to be a class of debt that automatically converted into equity when a triggering event occurred. An important worry with relying on

this type of instrument is that the automatic conversion may not be credible.

How best to define the denominator of the capital-asset ratio used for regulatory purposes is even more complex than how to define the numerator. Conceptually, the crucial issue is that banks should be required to maintain more capital against risky assets than safe assets. If this is not the case, banks will have strong incentives to tilt their portfolio towards riskier assets since these have higher yields. The practical difficulty is how to measure the riskiness of different assets.

The approach taken regarding risk adjustment of assets in the Basel I agreement was a relatively simple one. Assets were divided into five risk categories with different risk weights. The risk weights for the five categories were 0%, 10%, 20%, 50%, and 100%. (The United States did not use the 10% category.) The following list provides some examples of asset classes that fell into each category:

- 0%: Cash, central government and central bank debt
- 10%: Public sector debt
- 20%: Claims on other banks
- 50%: Residential mortgages
- 100%: Commercial and industrial loans

Given these definitions, the Basel I agreement required banks to maintain both a ratio of tier 1 capital to risk-weighted assets of 4% and a ratio of total capital to risk-weighted assets of 8%. In both cases, risk-weighted assets were defined as the weighted sum of bank assets using the risk weights discussed above. This meant that the capital requirements for Treasury holdings for U.S. banks were zero; for each $100 of residential mortgages, banks were required to maintain $2 of tier 1 capital and $4 of total capital; while for each $100 of commercial and industrial loans they must maintain $4 of tier 1 capital and $8 of total capital.

The advantage of this approach was its simplicity. But the coarseness of the risk weighting implied that banks had an incentive to choose relatively risky assets within each category and to engage in regulatory arbitrage. One form of regulatory arbitrage involved securitization. Banks could package many loans of a certain type (e.g., mortgages) into securities that paid off depending on the payoffs from the underlying loans. The securities were typically organized into tranches with the top tranche getting paid first and so on, and these tranches were rated by rating

agencies. The bank could then sell most of the tranches but retain only the lowest tranches. This could reduce the bank's capital requirement even though the bank held essentially all the risk since it held the lowest tranches. The U.S. retained a minimum capital-asset ratio based on non risk weighted assets (a leverage ratio) partly to guard against these concerns.

We see from Figure 19 that the long-term decline in capital-asset ratios of banks was reversed by the implementation of Basel I. Capital-asset ratios began to increase (modestly). Having been at 5.6% in 1987, the average capital-asset ratio was up to 8.5% in 1997 and 9.9% in 2007 (on the eve of the financial crisis of 2007-2009).

The weaknesses of the Basel I accord discussed above led regulators to quickly start working on a new accord that would ultimately come to be known as Basel II. Basel II did not change the way capital was defined, nor did it change the 4% and 8% regulatory ratios. The focus was on improving the risk-weighting of assets. An important precursor to Basel II occurred in 1996 when banks were allowed to use internal models to assess the risk in their trading accounts. The trading activities of banks often involve offsetting assets and liabilities where a liability may partly of fully hedge the risk from an asset. The use of internal models allowed banks to take account of these hedges when assessing the risk from the assets they owned in their trading account.

The Basel II accord was announced in 2004. But its complexity implied that it was expected to take quite a number of years for it to be fully implemented. The central change was that large banks were to use internal risk models to assess the expected loss from various assets, and capital charges were assigned based on these estimates. Smaller banks continued to use the Basel I approach. The rationale for Basel II was that replacing coarse risk categories with more sophisticated assessments of risk would reduce the incentive for regulatory arbitrage. The danger was that the complex internal risk models banks used might be calibrated by the banks to understate the true risks they faced in ways that were hard for regulators to spot. This danger became more fully appreciated with the financial crisis of 2007-2009.

8.10.4 Dodd-Frank and Basel III

The financial crisis of 2007-2009 was the first case in the United States of something approaching a banking panic since the Great Depression. Several large institutions failed or were absorbed by other institutions with assistance by the government. The Federal Reserve engaged in massive lender of last resort lending and the gov-

ernment (temporarily) guaranteed large asset classes that had previously not fallen under deposit insurance. The economic fallout of this crisis was enormous as we discuss in more detail in chapter XX [Great Recession chapter].

The crisis led to widespread calls for more stringent capital regulation of banks. Over the following years, virtually all aspects of these regulations were revisited and changed – the definition of the numerator, the definition of the denominator, the ratio itself, etc. – and the regulations were made quite a bit more complicated than before. Some of these changes were part of a new international accord commonly referred to as Basel III, while others were country specific. In the United States, the Dodd-Frank Act was an important piece of post-crisis banking legislation. Here, I will describe how these new capital regulations were implemented in the United States.

One change was that banks were divided into several different categories based on balance sheet size with larger banks facing more stringent regulation. The rationale for this was that the failure of a large bank is likely to cause more serious market disruptions than the failure of a smaller bank. Some banks were considered so large that they were "too big to fail." These banks were perceived to need capital regulation stringent enough to make failure very unlikely. The largest banks were termed global systemically important banks (G-SIBs). These banks faced particularly stringent regulation. In 2022, eight U.S. bank holding companies and their associated banks were designated as G-SIBS: Bank of America, The Bank of New York Mellon, Citigroup, Goldman Sachs, JPMorgan Chase, Morgan Stanley, State Street, and Wells Fargo.

Basel III substantially changed the calculation of risk-weighted assets. Quite a few new risk-weight categories were introduced. For mortgages, risk weights were made to differ based on the loan-to-value ratio of the mortgage. For loans to corporations, risk weights differed based on firm size, whether the firm was considered investment grade, and in some cases its credit rating. For loans to banks, risk weights depended on the duration of the loan as well as the credit rating of the bank. A notable new feature was that some asset classes had risk weights that exceeded 100%. For example, loans to corporations with a credit rating below BB- had a risk weight of 150%.

The definition of tier 1 and total capital were changed. This reflected the notion that certain types of liabilities had not turned out to be loss absorbing during the 2007-2009 crisis because authorities were – in practice – not willing to allow them to take losses. Also, a new, narrower definition of capital, "common equity tier 1"

(CET1) capital, was introduced. This definition of capital included only common equity rather than also including some preferred stock.

Table 2 lists the minimum capital ratios required by Basel III and compares them with Basel I and II. An important change was that the tier 1 capital ratio was increased from 4% to 6% and a new capital ratio was imposed for CET1 capital of 4.5%. In addition to this, several "buffers" were added on top of these minimum capital ratios. The capital conservation buffer was 2.5% (and somewhat higher for certain large bank holding companies). Banks must exceed this buffer (for all three of the minimum capital ratios) to be allowed to pay out dividends and certain discretionary bonuses. An additional time-varying bank-specific buffer of 1.0% to 4.5% applied to G-SIBs. Finally, Basel III allows for a countercyclical capital buffer. This buffer was meant to allow regulators to set higher capital buffers in good times to guard against losses in bad times (when this buffer would be reduced). The countercyclical capital buffer has not been used in the United States as of this writing (fall of 2024).

To summarize, the tier 1 capital requirement for most banks under Basel III is 8.5% including the capital conservation buffer. This is a substantially higher capital requirement than the 4% requirement under Basel I and II. For G-SIBs, the tier 1 capital requirement is even higher. For example, in 2022, the tier 1 capital requirement for JPMorgan Chaise was 12.7%. This included a 3.2% capital conservation buffer (which for G-SIBs is called the stress capital buffer) and a 3.5% G-SIB surcharge.

Basel III also includes a minimum leverage ratio for tier 1 capital relative to total assets (not risk-weighted) of 4%. The United States had imposed a similar leverage ratio ever since the 1980s. The rationale for imposing a leverage ratio is that risk-weighted assets are subject to regulatory arbitrage and also regulators may make mistakes in terms of which assets are placed in which risk categories. A leverage ratio for total assets (not risk weighted) is more robust to these concerns. U.S. regulators additionally imposed a supplemental leverage ratio (SLR) on large banks. The SLR is 3% of a broader definition of assets including various off-balance-sheet items such as derivatives. G-SIBs are subject to an enhanced supplemental leverage ratio (eSLR) of 6%.

An additional important innovation in the United States after the financial crisis was the institution of regular stress tests for large banks. As we discussed above, large banks calculate risk-weighted assets partly based on internal models. The stress tests add assessments based on a model built by the regulator (which can therefore not be manipulated by banks to produce good results for their portfolio).

Table 2: Regulatory Minima for Capital and Leverage Ratios

	Basel I and II	Basel III
Capital Ratios:		
Common Equity Tier 1 Capital Ratio	–	4.5%
Tier 1 Capital Ratio	4.0%	6.0%
Total Capital Ratio	8.0%	8.0%
Capital Conservation Buffer	–	+2.5%
G-SIB Surcharge	–	+1.0-4.5%
Counter Cyclical Capital Buffer	–	+0.0-2.5%
Leverage Ratios:		
Leverage Ratio	4.0%*	4.0%
Supplemental Leverage Ratio	–	3.0%*
Enhanced Supplemental Leverage Ratio	–	6.0%*

Note: The *'ed numbers in the table are U.S. regulations, rather than parts of Basel accords. When the United States implemented Basel I and II it retained a minimum leverage ratio. Recall that the original capital ratios imposed by U.S. regulators in the early and mid-1980s were ratios of capital to total assets not risk weighted, i.e., a leverage ratio. The Federal Deposit Insurance Corporation Improvement Act of 1991 specified that banks must hold 4% tier 1 capital against non-risk-weighted assets to be considered "adequately capitalized" (and 5% to be considered "well capitalized"). See FDIC (2003) for more detail. The supplemental leverage ratio and enhanced supplemental leverage ratio are imposed by U.S. regulators and are not part of Basel III.

Also, the stress tests are "forward looking". They consider several adverse scenarios and ask how the banks would do in these scenarios. Importantly, the results from the stress tests for individual banks are made public. This allows investors in the market to reward or punish banks for the strength of their (forward-looking) capital position.

Overall, the changes made to bank capital regulation after the financial crisis of 2007-2009 have made banks – especially large banks – better capitalized. A vigorous debate continues as to whether minimum capital requirements are high enough or perhaps too high. Many academics argue that capital requirements should be raised further to further reduce the risk of financial crises. In contrast, the banking lobby argues vigorously for lower capital requirements. The main argument the banking lobby puts forward is that lower capital requirements would increase lending and therefore output in the economy. More empirical evidence is sorely needed on this point. As we discuss above, banks have a strong incentive to argue for low capital

requirements since it is artificially cheap for them to finance themselves with debt due to deposit insurance, last resort lending by central banks, and the tax advantages of debt financing. The socially optimal level of capital requirements is surely higher than the level of capital requirements desired by bankers.

9 Private Money and Free Banking

It is a virtually universal fact that governments monopolize the issue of money. Is this justified? Or might the private sector be capable of – or perhaps even be better suited to – issue money? As with all government involvement in the economy, it is useful to ask what market failures justify the government's role in monetary matters. Economists have debated this question for decades, and as with many such issues, views differ. Some have argued for a radical shift towards laissez faire in monetary matters (see, e.g., Smith, 1936; Hayek, 1978; Selgin and White, 1994). Others have argued that there are good reasons for the government to be heavily involved in this sphere of economic activity (e.g., Friedman, 1960; Friedman and Schwartz, 1986; Goodhart, 1988).

The question of the government's involvement in monetary matters can be analyzed at several different levels. The most basic question one can pose in this regard has to do with the very definition of the monetary standard: should there be competition over the monetary standard? Supposing, however, that the answer to this question is 'no', and there is a commonly agreed upon monetary standard, one can ask: should there be free entry into the creation of such money? This is the traditional question posed in the "free banking" literature. This question then spills over into more general issues having to do with government regulation of banks, the most important of which are perhaps: should the government regulate institutions that issue demand deposits? and are central banks necessary? We have discussed the first of these questions at length earlier in the chapter. We will take up the second below.

9.1 Should the Government Set the Monetary Standard?

Throughout most of the last millennium, monetary standards were defined in terms of precious metals. For example, the U.S. dollar was originally defined by the Coinage Act of 1792 as 371.25 grains (24.06g) of pure silver or 24.75 grains (1.60g) of

pure gold (a bimetallic standard). The monetary standards of other countries – e.g., the pound sterling in Britain and the livre and later franc in France – were typically defined by the governments of these countries in a similar manner.

But what business does the government have making such a definition? One view is that by defining the monetary standard, the government is simply solving a coordination problem: it lowers transactions costs if everyone quotes prices in the same units. This is analogous to the fact that it is advantageous as a matter of coordination for everyone to agree on a common definition for weights and measures. For this reason, the government fixes a standard of weights and measures. Several hundred years ago there was no common definition of weights and measures and this very substantially raised the cost of trade. It is perhaps no accident that the government's power to regulate the monetary standard and to fix a standard of weights and measure is set forth in the same clause of the U.S. Constitution. Article 1, Section 8, Clause 5 of the U.S. Constitution reads "[The Congress shall have Power . . .] To coin Money, regulate the Value thereof, and of foreign Coin, and fix the Standard of Weights and Measures; . . ."

The trouble is that defining a good monetary standard is not as simple as defining a good standard of weights and measures. For various reasons, governments occasionally changed the definition of the monetary standard, as we discuss in chapter XX [Quantity theory chapter]. Many of these changes involved reducing the metallic content of the monetary unit. For example, in 1933 President Franklin Roosevelt devalued the U.S. dollar in terms of gold by about 40%. A troy ounce of gold cost $20.67 prior to the devaluation, but $35 after the devaluation. Some argued that this was highway robbery. After all, anyone that owned a bond that paid in dollars was all of a sudden owed much less, at least when viewed in gold. (See Edwards (2018) for a very interesting historical account of this controversy.)

In the 20th century, many governments took the more radical step of eliminating the link between their monetary standard and precious metals. For the United States, the most important step in this process was taken by Roosevelt in 1933 when he restricted the ability of the public to exchange dollars for gold. Today the U.S. dollar (and most other major currencies) are no longer defined in terms of any real object. They are purely fiat monetary standards.

The shift to purely fiat monetary standards was accompanied by persistent high inflation in many countries. Figure 20 plots the inflation rate in the United States between 1870 and 2020. The link of the U.S. dollar to gold was completely severed when President Nixon closed the gold window in 1971 and inflation was persis-

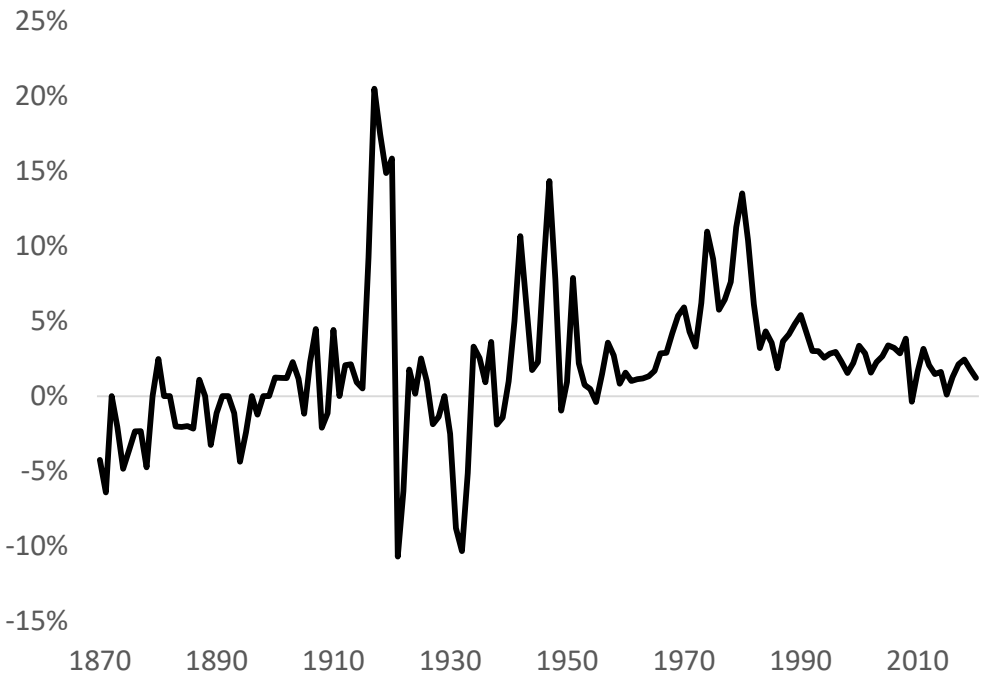

Figure 20: Inflation in the United States, 1870-2020

Note: This figure plots the rate of inflation in the United States from 1870 to 2020. The source is Officer and Williamson (2021). The construction of this series is described in detail in Officer (Undated). It uses series constructed by David and Solar (1977), Rees (1961), Douglas (1930), as well as the Bureau of Labor Statistics.

tently high between the late 1960s and the early 1980s. The rate of inflation in the United States had been volatile before. But that earlier volatility tended to occur at times of severe crisis such as during World War I, the Great Depression, and World War II. Persistently high inflation during peacetime and absent a major financial crisis was a new phenomenon.

Many critics considered this episode to constitute a severe failure of the government to define a stable monetary standard. In addition, prior episodes of fiat money had in most cases been associated with very high inflation. It was therefore natural for there to be widespread worry in the 1970s and 80s that a fiat monetary standard was perhaps fundamentally incompatible with price stability.

A different viewpoint is that perhaps the problem lay in the government's monopoly over the monetary standard. Perhaps competition among privately defined monetary standards could yield a stable fiat monetary standard even if the government could not. Might this be the case?

A fundamental problem faced by any issuer of a fiat currency (public or private) is that it is virtually costless to issue more such currency (literally costless in the case

of a fully electronic currency). If the currency is valued, the issuer therefore has an incentive to issue more. But issuing more can result in the currency losing value. This problem of "overissue" – i.e. issuing to the point where the currency begins to lose value – can undermine the credibility of the currency.

Counteracting the incentive to overissue is the desire to maintain and further build the user base of the currency since wider adoption will result in greater demand for the currency and therefore allow the issuer to issue more over time. But as the currency becomes widely adopted, there is less scope for further adoption and therefore less to be gained from showing restraint in how much is issued. This means that the incentive to overissue will loom larger as the prospects for further growth diminish.

It is possible that the profits from maintaining a currency's value will continue to outweigh the profits from overissuing even once the currency reaches a mature state. However, this is far from certain and will depend on various features of the economic environment such as the rate of change of velocity and output (recall $MV = PY$ from chapter XX [Quantity theory chapter]).

It will also depend on the speed with which users of the currency lose confidence in the currency when the issuer begins to overissue. This can in principle occur very quickly and at any time. The value of a fiat currency is entirely dependent on people's belief that others will continue to value it at a later date. Its value is therefore a self-fulfilling prophesy that can in principle unravel at any time. In this sense, a fiat currency is a very fragile construct.

Historically, high inflation is usually associated with large increases in the quantity of money in circulation. Furthermore, historical experience suggests that seigniorage revenue increases the higher is the inflation rate up to very high levels of inflation. In this sense, issuers of currency will face a short-term temptation to overissue. We will discuss this in more detail in chapter XX [Hyperinflation chapter].

One thing an issuer of a fiat currency can do to mitigate its fragility is to hold a reserve of assets. Such a reserve places a lower bound on the value of the outstanding currency. If the currency is fully backed, the issuer can guarantee its value. In this case, the currency is no longer a fiat currency. A fractional reserve may help the issuer maintain confidence in its currency, but is subject to all the issues we have discussed earlier in the chapter for fractional reserve banking. Also, holding a reserve is costly for the issuer. They therefore have an incentive to hold as small a reserve as they think they can get away with.

An important downside of relying on competition among private fiat currencies is that such a system will result in multiple units of account. After all, for there to be competition, there must be more than one currency in use in an area. This will raise transactions costs as people will need to keep track of the exchange rates between the different currencies (unless all the currencies operate a fixed exchange rate system). Efficiency of trade therefore favors a single currency. This is the coordination issue discussed above. In this sense, a single currency may be a natural monopoly.

Governments potentially have several advantages over a private monopoly issuer of a fiat currency. First, governments are not purely motivated by profit when issuing a currency. This is not to say that governments ignore the seignorage profits they make from issuing a currency. Raising seignorage has at times led governments to massively overissue currency and resulted in high inflation, sometimes even hyperinflation, as we discuss in chapter XX [Hyperinflation chapter]. Also, the desire to earn seignorage is one reason why governments restrict the freedom of the public to issue and use currencies other than the one they issue.

However, many countries have set up independent central banks with mandates to achieve price stability. Furthermore, after the initial period of high inflation in the 1970s and early 1980s, these central banks have built up impressive track records of achieving low and stable rates of inflation. In fact the predictability of the rate of inflation in society has perhaps never been greater than during this period.

A second advantage the government has as an issuer of a fiat currency is that it can mandate that taxes be paid with the currency it issues. The government can also pay transfer payments using the currency (pensions, unemployment insurance, welfare, etc.). In many countries, the government is a large employer. It can pay government employees using the currency. All of this will create demand for the currency. The government can furthermore declare the currency "legal tender" for all payments public and private. This may further increase demand for the government's currency even if enforcement of such a declaration is uneven.

The principle advantage of allowing for competing monetary standards is that this will provide people with an outside option in case the government's monetary standard is mismanaged. An important difficulty with weighing the costs and benefits of competing monetary standards is that we have no historical experience with this type of system. Clearly, the desirability of experimenting with such a system depends critically on how good a job our current government central banks do at maintaining price stability. The better a job they do, the weaker is the argument for radical change and vice versa.

9.2 Free Banking

Much of the debate about privately issued money has taken for granted the existence of a monetary standard defined by the government and tied to gold or silver. Early strands of this debate centered on the question of what institutions should be allowed to issue circulating paper bank notes (i.e., currency) denominated in this monetary standard. As we discuss in section 5, many of the first issuers of such notes were private bankers, e.g., the goldsmith bankers in 17th century England and the country bankers of 18th century England.

As the practice of banks issuing notes spread, the governments of various countries reacted in different ways. The governments of most countries, including England, passed increasingly restrictive laws about which institutions could issue currency. (Smith (1936) provides an excellent discussion of these developments.) However, some countries allowed relatively free entry into banking and note issue. One example of this was Scotland in the 18th and early 19th century. Another was the period from 1837 to 1863 in the United States. These are often referred to as periods of "free banking."

By the early 20th century, all countries had gravitated to a system in which a single central bank became the sole issuer of currency. Was this a desirable outcome? The arguments for and against allowing private banks to issue paper currency denominated in a given monetary standard are very different from those discussed above relating to the definition of the monetary standard itself. In this case, we also have some historical experiences on which to base our inference.

9.2.1 The Origin of Free Banking in America

In thinking about the free banking debate, it is important to understand that in the 18th and early 19th centuries the right to form any type of corporation – especially one with limited liability – was a privilege granted by the government typically through special legislation. Since entry was limited, such privileges were valuable. Governments were partly funded by the sale of these privileges. Furthermore, the individuals and political groups that controlled the governments lined their pockets and allocated these privileges to allies to strengthen their political coalition. This last notion is what Wallis (2006) calls systematic corruption, but at the time was sometimes referred to as the 'spoils system.'

In early 19th century America, the chartering of state banks was an important instance of this spoils system. This was notoriously the case in New York state where

Martin Van Buren's Albany Regency granted bank charters to political allies who supported the Regency (the political machine of the Democratic Party) and allowed it to dominate state politics. The corruption associated with this system was unpopular and the public demanded change. After the panic of 1837, the Whig party gained a majority in New York and served a major blow to the spoils system by enacting general incorporation acts. These acts allowed anyone to form a corporation subject to a set of general requirements. With time, many other states followed suite. In the sphere of banking, this meant free entry into banking, i.e., 'free banking.' Michigan passed the first free banking act in 1837, while New York passed such an act in 1838 and many states after that. (See Bodenhorn (2006) and Wallis (2006) for a more detailed discussion of these reforms.)

9.2.2 No Questions Asked Money

Recall that a central goal of the monetary system is to reduce transactions costs. For a monetary system with paper money issued by a multitude of different private banks to work well, the paper money must circulate freely *at par*. In other words, bank notes with face value $1 should fetch $1 in a transaction. This means that the bank notes must be perceived to be completely risk free. Ideally, all bank notes in circulation should trade at par. In this case, we say that the currency in circulation is uniform. With a uniform currency, sellers can take bank notes as payment without having to spend time and effort examining the notes – the notes have a no-questions-asked property. Furthermore, we ideally want the currency to remain uniform and trade at par relative to the underlying monetary unit (gold and silver) even in times of financial stress.

However, paper money is a promise by the issuing bank to pay a certain amount of the underlying monetary unit on demand. In 19th century America, this was a promise to pay a certain amount of gold or silver. Paper money is therefore a credit instrument issued by the bank. In general, credit instruments are risky. In particular, in the case of banks there is the risk of a run. In order for paper money to be risk free, the risk of a run on the issuing bank must be perceived to be negligible by the public, ideally even in times of financial stress. This is a serious challenge. One that most countries have resolved with a substantial degree of regulation of money and banks. But perhaps a laissez faire alternative is possible.

9.2.3 Wildcat Banking in America

While free banking was an important part of curbing systematic corruption in the United States, it delivered a monetary system that was highly imperfect. Two problems that plagued the system were frequent bank failures and discounts on bank notes that varied from bank to bank and also across space and time. These problems were arguably caused by flaws in the legal framework governing free banking.

Free banking laws stipulated that banks must secure the notes they issued with bonds deposited with the state banking authority. If the bank refused to redeem its notes for specie (even a single note), the state banking authority had the right to sell all the bonds deposited by that bank to pay off the bank's note holders. The idea was that backing the notes by bonds dollar-for-dollar would make the notes risk free. Two separate problems implied that this general idea did not work in all cases.

The free banking law passed in Michigan in 1837 proved to be particularly problematic. It allowed bank notes to be backed by mortgages on land valued at par (face) value rather than market value. This system was open to a serious type of abuse often referred to as *wildcat banking*.

Wildcat banks were so named because they were often set up far from population centers "where only the wildcats roam" in order to make redemption of notes costly. In the most egregious cases, the purpose of the bank was completely fraudulent. The banker issued notes against bonds worth less than the notes, sold the notes to the public, pocketed the proceeds, and skipped town.

The 1837 free banking law in Michigan made this scheme relatively easy to implement. As Rockoff (1974) explains, "it was thus possible to create a mortgage on a worthless piece of property, have it certified as being valuable by a group of friends, and then transfer it to a wildcat bank in exchange for a mass of bank notes." To make this concrete, let's suppose that a group of wildcat bankers purchase a piece of land for $50 and issue a $100 mortgage against this land. They then create a bank with the $100 mortgage as equity, deposit this mortgage with the state banking authority, and receive $100 of bank notes. Finally, they exchange these $100 of bank notes for something worth $100 that they can "run away with." This scheme nets the wildcat bankers $50 in profits (the $100 they spend at the end minus the $50 they needed at the beginning to purchase the land).

The free banking laws of other states – e.g., New York – were better designed and largely prevented wildcat banking. New York's law required that bank notes be backed by state bonds and limited the value of notes issued to the market value

Table 3: Bank Failures in Four States, 1838-1863

State	Banks Established	Banks Closed	Closed Below Par	Fraction Closed	Fr. Closed Below Par
New York	449	160	34	36%	8%*
Indiana	104	89	24	86%	31%*
Wisconsin	140	79	37	56%	26%
Minnesota	16	11	9	69%	56%
Total	709	339	104	48%	15%

Note: These data are from Table 2 of Rolnick and Weber (1982). The $*$ refers to the fact that redemption information is not available for 4 New York banks and 27 Indiana banks. The fraction of banks that closed below par is calculated excluding these banks.

of the bonds. While most banks set up under Michigan's 1837 law failed in short order, the fraction of banks that failed in New York was much lower. Most states followed New York's design.

Rolnick and Weber (1982, 1984) discuss how fluctuations in the value of state bonds were a second problem that caused large numbers of bank failures during the free banking era. In the early 1840s, a number of states faced solvency problems resulting in sharp falls in the price of state bonds. Several periods in the 1850s also saw sharp drops in state bond prices. Since banks held large sums of state bonds to back their notes, a fall in state bond prices could result in a substantial loss for the banks. This loss could be large enough to threaten the solvency of the banks.

The losses on state bonds, could trigger runs by note holders. The logic is the same as with runs of depositors in the Diamond-Dybvig model discussed in section 8. The note holders understood that the issuing bank might not have enough assets to redeem all notes at par. But noteholders that were able to redeem their notes before the bank closed its doors received full value for their notes. This implied that note holders had an incentive to run on the bank.

Overall, bank failures were common in the free banking era. Rolnick and Weber (1982) present data on bank failures for four states: New York, Indiana, Wisconsin, and Minnesota. These data are reproduced in Table 3. We see that 48% of banks in these states closed and 15% of them closed with losses on notes. Of course, failures are common in many lines of business. Not every enterprise turns out to be a success. But failures of note-issuing banks are particularly problematic since their liabilities are the economy's money.

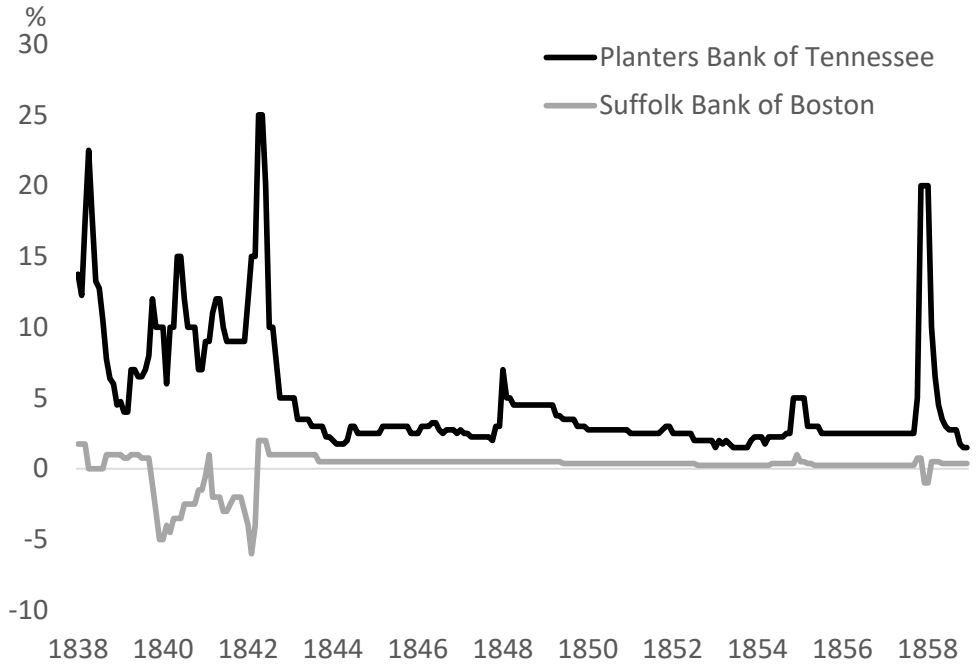

Figure 21: Note Discounts in Philadelphia, 1837-1858

Note: This figure plots the discount on bank notes from the Planters Bank of Tennessee and Suffolk Bank in Philadelphia. Negative numbers indicate a premium. The source of these data is Gorton and Weber (Undated).

9.2.4 A Uniform Currency

A second problem that plagued the free banking era in the United States was note discounts. There were over a thousand different banks that issued notes during this period and these notes traded at varying discounter from par. Figure 21 illustrates this by plotting the discounts in Philadelphia on notes from the Planters Bank of Tennessee and the Suffolk Bank of Boston. We see that these discounts are both different from each other and quite volatile over time.

The fact that notes from different banks traded at varying discounts raised transactions costs. Merchants needed to pay attention to the exact types of notes their customers presented for payment. In some cases, they needed to examine the notes carefully and consult a "banknote detector" – i.e., a reference manual that listed the discounts of notes from different banks as well as known counterfeits. The discounts also resulted in an adverse selection problem: merchants needed to worry that customers were offering for payment particular notes because they knew these notes were worth less than the merchant might think they were worth. This problem was more severe in times of high volatility and tended to increase the discounts

on unfamiliar notes.

One should not, however, overstate the monetary chaos associated with note discounts during the free banking era. Notes from local banks typically traded at par. This reflected the fact that local notes could be presented at the issuing bank for specie: had these notes traded at a discount locally, it would have paid to take them to the issuing bank rather than use them for payment. In normal times, the discounts on notes from "foreign" banks – i.e., banks from other cities – reflected the cost of sending these notes back to their home city. The discounts were therefore small for banks from nearby cities, but larger for banks from more distant cities.

These notions are clearly illustrated in Figure 21. The discounts on notes from Suffolk Bank of Boston in Philadelphia are very small in normal times (typically less than 1%). This reflected the low cost of travel between Philadelphia and Boston and the fact that many merchants in Philadelphia had business in Boston and therefore had use for Boston notes. In contrast, the discount on notes from the Planters Bank of Tennessee were quite a bit larger, even in normal times, reflecting the higher cost of inland travel.

The fact that notes lost value as they traveled farther away from the issuing bank implied that they tended not to travel very far from the issuing bank (Bodenhorn, 2023). In other words, most bank notes were used locally. This, of course, implied that bank notes were not a good medium of exchange for longer distance trade, which tended to raise the cost of such trade.

In addition to the relatively modest discounts in normal times discussed above, note discounts spiked in times of financial stress. The period from 1837 to 1842 was a period of such financial stress. There was a serious panic in 1837 and a recession followed. These difficulties brought to the surface and exacerbated weaknesses in state finances leading a number of state governments to default on state bonds in 1841 and 1842. Many state banks were forced to suspend convertibility during this time. First, this was due to the panic of 1837, but later it was due to the fall in the value of state bonds.

The high volatility of discounts on the Planters Bank of Tennessee during the period 1837-1842 is a reflection of these developments. We also see a large spike in the discount of Planters Bank of Tennessee in 1857, again reflecting a panic. The bank was forced to suspend convertibility during these periods. Its notes then traded at values reflecting the public's expectations about the probability that the bank would resume convertibility and the size of losses sustained by note holders in the event of repudiation. In sharp contrast, the discounts on Suffolk Bank do not spike during

these periods. Instead they fall to negative values. This means that notes from Suffolk Bank traded at a premium at certain times reflecting the strong position of that bank. This is an early instance of "flight to safety" during financial crises.

Why did the system not work better? One important flaw was the heterogeneity in the value of the bonds that banks were required to hold to back notes, and the fact that the value of these notes fluctuated wildly at times. This flaw was remedied with the passage of the National Banking Acts of 1863-4. These acts, in effect, expanded free banking to the entire nation. Banks incorporated under these acts were called national banks. They were allowed to issue bank notes that were fully backed by U.S. Treasuries. The bank notes of national banks bore the name of the issuing bank, but they were issued by the Comptroller of the Currency and the banks must deposit the bonds held to back the notes at the Treasury.

All national banks were required to take bank notes from all other national banks as payment at par. When a national bank failed, bank notes issued by that bank were immediately redeemed at the Treasury. These provisions meant that bank notes issued by all national banks were equally safe (since they were all fully backed by the same collateral). An act of 1866 placed a prohibitive tax on note issuance by state banks. Bank notes issued by state banks were thereby taxed out of existence. Together, these new laws succeeded in creating a uniform currency in the United States.

Another important reason why the antebellum free banking system in the United States did not work better was restrictions on branch banking and interstate banking. Unit banking made banks in the United States less stable than they otherwise would have been, as we discuss above. But branching would also have worked to eliminate geographical variation in note discounts. The notes of banks with branches in more than one city would have traded at par in all cities in which the bank had a branch. A bank with a nationwide branch network could have seen its notes trade at par everywhere in the country. The Second Bank was on its way to attaining this goal before it was destroyed by President Jackson. Other examples include the early 19th century banking systems of Canada and Scotland, which allowed branching.

9.3 Are Central Banks Necessary?

In a fiat currency system, central banks arguably have three core functions: 1) to issue and manage the currency; 2) Operate the central nodes of the payment system;

and 3) to act as a lender of last resort. The first of these is clearly essential: someone must issue the fiat currency. The issuing institution will be the "central bank" for that currency. This central bank will need to conduct monetary policy – set interest rates or manage the quantity of money – so as to manage the value of the currency it issues. (More on this in later chapters.) The issuing institution need not be a government institution. One can imagine a private fiat currency, and one can imagine more than one fiat currency circulating concurrently. As of this writing (early 2025), some would say that bitcoin is an example.

Issuing a currency is sometimes characterized as operating a printing press. But in a world with largely electronic payments, it for the most part means maintaining a ledger that records who holds how much of the currency at a given point in time. It follows that an essential function of a central bank is to operate the central nodes of the payment system in the currency it issues, i.e., transfer funds on the central ledger. In the United States, this is done through the Fedwire system.

In many countries, the central bank operates not only the central ledger – i.e., the core large-value payment system – but also various other parts of the payment system, including parts of the retail payment system. Payment systems are networks. This means that they tend in the direction of being natural monopolies. This provides a rationale for government involvement. There is a tension though, since private sector involvement likely increases innovation.

When the monetary standard is a commodity standard, it is not as obvious that a central bank is essential. Perhaps the Mint can simply stand ready to exchange specie for coin at the set rate and this is all the government should do. With some simplification, that is how things worked prior to the rise of banking.

Does the rise of banking call for a central bank? In practice, that is where all countries have ended up. But was this inevitable? Was it necessary? Was it optimal? Interestingly, early thinkers on the subject, even those such as Walter Bagehot – who played a key role in arguing for the essential role of central banks as lenders of last resort – thought it unnatural for a central bank to arise. Bagehot thought that free competition among more-or-less equal banks was a more natural state (Bagehot, 1873/1999, p. 67). A strand of thought questioning the necessity of central banks runs all the way to the present among scholars with a libertarian bent.

As I see it, the core problem (market failure) that gives central banks an essential role even in an economy on a commodity standard is bank runs and banking panics. Central banks act as lenders of last resort when panics occur as we discussed in section 8. It is arguably essential in an economy with a highly developed banking

system to have a lender of last resort.

Critics of central banking take issue with this conclusion. They argue that central banks are not necessary or even desirable because a laissez faire banking system would be stable, or at least much more stable than advocates of central banking fear. They make several arguments to this effect. First, they argue that the presence of a government central bank induces other banks to take more risk than they otherwise would. This is the moral hazard argument discussed in section 8: if there were no central bank, private banks would be less leveraged, hold larger reserves, and therefore be safer institutions.

A piece of evidence the central banking critics can point to is that banks were much less leveraged and held larger reserves in the 19th century when the U.S. did not have a central bank (Figures 14 and 19). Of course, many other things have changed since then. So, this comparison is complicated. But the shift among banks towards very high leverage and low capital ratios in the 20th century was no doubt strongly influenced by the presence of deposit insurance and a lender of last resort.

Advocates of central banking can argue that the banking systems of 19th century United States and England were highly unstable and prone to panics even despite banks not being very leveraged and holding large reserves. But the critics can counter that the banking systems of these two countries were special in that they both had rules in place that kept banks small. In the U.S., branch banking and interstate banking was for the most part not allowed. In England, no joint stock banks other than the Bank of England were allowed until 1826. Before that time only private banks with 6 or less partners were allowed. This resulted in highly fragile banking systems with frequent panics in both countries. The banking systems of other countries that allowed branch banking – such as Scotland, Canada, and Australia – were much more stable (Calomiris, 2000).

The second argument of central banking critics is therefore that past experience with frequent panics is distorted by undesirable rules that limited the size and geographic scope of banks. A fully laissez faire system would be more stable than banking in 19th century United States and England suggests, more like the branch banking systems of Canada and Scotland.

A third argument put forth by critics of central banking is that in its absence, voluntary private arrangements will arise that can serve the role of disciplining banks and providing liquidity in a crisis. Two historical examples the critics often point to are the Suffolk Bank system in antebellum New England and the clearinghouse system adopted in various U.S. cities in the late 19th century.

From the 1820s to the 1850s, the Suffolk Bank of Boston acted as a clearing bank in New England. As we have discussed above, bank notes traded at varying discounts in the United States during this time resulting in a non-uniform currency. The Suffolk Bank offered to receive notes at par from banks that held sufficient reserves at Suffolk Bank. It furthermore, instituted the policy of sending notes from banks that refused to participate in its scheme back for redemption. This scheme was quite successful and Suffolk Bank managed to create a uniform currency in New England over this period.

The original purpose of clearinghouses was – as the name suggests – to facilitate the clearing of payments. Rather than all pairs of banks clearing payment instruments bilaterally, banks in a city would send representatives to the clearinghouse where payments were cleared multilaterally (allowing much more netting of debits and credits). One bank in the association was typically assigned the role of being the "central" bank that cleared the net payments of the other banks. The other banks typically held substantial reserves at this bank. All banks in the clearinghouse association would typically accept claims on other banks in the association at par. To ensure the banks did not take advantage of this trust, the clearinghouse regulated and monitored member banks (Timberlake, 1984; Gorton, 1985).

Starting with the panic of 1857, clearinghouses took on certain central banking roles during crises. Gorton (1985) argues that "when a panic occurred, the structure of the banking industry was radically altered by the metamorphosis of the clearinghouse into a single, firm-like organization uniting the member banks in a hierarchical structure topped by the Clearinghouse Committee." The clearinghouse would decide on joint suspensions of convertibility and act as a lender of last resort by issuing clearinghouse loan certificates to member banks. Clearinghouse loan certificates were a liability of the clearinghouse – and thus backed by the joint assets of all member banks. The clearinghouse would lend these freely, but at a relatively high interest rate and against good collateral. (The Clearinghouse Committee would scrutinize the collateral submitted by member banks and decide on appropriate haircuts.)

Gorton argues that the reason why the clearinghouse could help solve the problem of panics was that panics arose because depositors were worried that some banks were likely to fail but didn't know which ones. This led them to run on many (sometimes all) banks. By pulling together, the banks insured each other at times of panic. In general, the clearinghouse would not allow banks to fail during the panic, but would expel banks from the clearinghouse that failed to replay clearinghouse loans after the panic subsided.

Goodhart (1988) argues that while valuable, these private arrangements were an imperfect substitute for a central bank. The problem Goodhart emphasizes is the fact that there is a conflict of interest between the strong banks and the weak banks during a crisis. It is true that the strong banks are worried about contagion and therefore have an incentive to provide aid either directly or through the clearinghouse. But the strong banks also have an incentive to use the crisis to get rid of weak competitors. This second incentive will lead strong banks (and therefore also the clearinghouse) to act less aggressively than is socially optimal.

Another way of saying this is that bank failures have negative externalities on other banks through contagion and fire sales or assets. Mutual aid, therefore, has positive externalities. Strong banks will provide aid to the extent that they think it is in their private interest to do so. But this will not take account of the positive externalities of providing aid during a crisis. For this reason, purely cooperative aid arrangements are likely to provide less aid than is optimal. This raises the likelihood that crises get out of control. For these reasons, Goodhart argues that a central bank that is not also a for-profit direct competitor of commercial banks is likely to be a more effective lender of last resort and regulator of the banking system.

Another argument favoring central banking over voluntary private arrangements is the notion that modern day governments are unable to commit not to come to the aid of the banking system in the event of a panic. If the banks realize this, they will not prepare adequately for the panic even in a collective sense. The government's commitment problem in this regard has arguably become vastly more severe over the past 100 years. Before WWI, governments were small and there was widespread poverty in society without this creating a political imperative for the government to act. This changed quite dramatically for a host of reasons over the course of the 20th century. Today – for better or worse – there is widespread consensus in society that the government should act aggressively in times of crisis to cushion the blow of those that are worse off in society. This makes it not credible for the government to say that the banking system is on its own in the event of a crisis.

Notes

[1]In writing this section (section 8.10), I have made use of many sources. The following sources have been particularly useful: Fisher (1936), FDIC (2003), Haubrich (2020), Tufts and Moloney (2022a,b), Federal Reserve (2024).

References

ALCHIAN, A. A. AND W. R. ALLEN (1967): *University Economics*, Belmont, CA: Wadsworth Publishing Co.

BAGEHOT, W. (1873/1999): *Lombard Street: A Description of the Money Market*, New York, NY: John Wiley & Sons.

BOARD OF GOVERNORS (1952): "Minutes of the Board of Governors of the Federal Reserve System, December 17 1952," Available at https://fraser.stlouisfed.org/.

BODENHORN, H. (2006): "Bank Chartering and Political Corruption in Antebellum New York: Free Banking as Reform," in *Corruption and Reform: Lessons from America's Economic History*, ed. by E. L. Glaeser and C. Goldin, Chicago, IL: University of Chicago Press, 231–257.

——— (2023): "The Extent of the Market for early American Bank Notes," NBER Working Paper No. 31886.

BROCK, L. V. (1975): *The Currency of the American Colonies 1700-1764*, New York, NY: Arno Press.

CALOMIRIS, C. W. (2000): *U.S. Bank Deregulation in Historical Perspective*, Cambridge, UK: Cambridge University Press.

CALOMIRIS, C. W. AND J. R. MASON (2003a): "Consequences of Bank Distress during the Great Depression," *American Economic Review*, 93, 937–947.

——— (2003b): "Fundamentals, Panics, and Bank Distress during the Depression," *American Economic Review*, 93, 1615–1647.

CHODOROW-REICH, G. (2014): "The Employment Effects of Credit Market Disruptions," *Quarterly Journal of Economics*, 129, 1–60.

CORREIA, S., S. LUCK, AND E. VERNER (2023): "Failing Banks," Working Paper, Massachusetts Institute of Technology.

DAVID, P. A. AND P. SOLAR (1977): "A Bicentenary Contribution to the History of the Cost of Living in America," 2, 1–80.

DAVIES, P. (2008): "The "Monster" of Chestnut Street," Federal Reserve Bank of Minneapolis, https://www.minneapolisfed.org/article/2008/the-monster-of-chestnut-street.

DE ROOVER, R. (1944): "What Is Dry Exchange? A Contribution to the Study of English Merchantilism," *Journal of Economic History*, 2, 52–65.

——— (1966): *The Rise and Decline of the Medici Bank*, New York, NY: W. W. Norton & Company.

DE VRIES, J. AND A. VAN DER WOUDE (1997): *The First Modern Economy: Success, Failure, and Perseverance of the Dutch Economy, 1500-1815*, Cambridge, UK: Cam-

bridge University Press.

DIAMOND, D. W. AND P. H. DYBVIG (1983): "Bank Runs, Deposit Insurance, and Liquidity," *Journal of Political Economy*, 91, 401–419.

DOUGLAS, P. H. (1930): *Real Wages in the United States, 1890-1926*, Boston, MA: Houghton.

EDWARDS, S. (2018): *American Default: The Untold Story of FDR, the Supreme Court, and the Battle over Gold*, Princeton, NJ: Princeton University Press.

FDIC (2003): "Basel and the Evolution of Capital Regulation: Moving Forward, Looking Back," FDIC Staff Study, available at www.fdic.gov (accessed in December 2024).

FEAVEARYEAR, A. (1963): *The Pound Sterling, A History of English Money*, Oxford, UK: Clarendon Press.

FEDERAL RESERVE (2024): "Bank Capital Standards," Federal Reserve History, availabe at www.federalreservehistory.org (accessed in October 2024).

FEINMAN, J. (1993): "Reserve Requirements: History, Current Practice, and Potential Reform," *Federal Reserve Bulletin*, 79, 569–589.

FISHER, I. (1936): "100% Money and the Public Debt," *Economic Forum*, 3, 406–420.

FRIEDMAN, M. (1960): *A Program for Monetary Stability*, New York, NY: Fordman University Press.

FRIEDMAN, M. AND A. J. SCHWARTZ (1963): *A Monetary History of the United States, 1867-1960*, Princeton, NJ: Princeton University Press.

——— (1986): "Has Government Any Role in Money," *Journal of Monetary Economics*, 17, 37–62.

FROST, J., H. S. SHIN, AND P. WIERTS (2020): "An Early Stablecoin? The Bank of Amsterdam and the Governance of Money," BIS Working Paper No. 902.

GOODHART, C. (1988): *The Evolution of Central Banks*, Cambridge, MA: MIT Press.

GORTON, G. (1988): "Banking Panics and Business Cycles," *Oxford Economic Papers*, 40, 751–781.

GORTON, G. AND W. WEBER (Undated): "Quoted Discounts on State Bank Notes in Philadelphia, 1832-1858," Research Department, Federal Reserve Bank of Minneapolis, https://researchdatabase.minneapolisfed.org /concern/datasets/2801pg356.

GORTON, G. B. (1985): "Clearinghouses and the Origin of Central Banking in the United States," *Journal of Economic History*, 45, 277–283.

——— (2012): *Misunderstanding Financial Crises: Why We Don't See Them Coming*, Oxford, UK: Oxford University Press.

GRUBB, F. (2023): "Colonial Monetary Systems," in *Handbook of Cliometrics*, ed. by C. Diebolt and M. Haupert, Berlin, Germany: Springer-Verlag GmbH, 1–30.

HAMMOND, B. (1957): *Banks and Politics in America: From the Revolution to the Civil War*, Princeton, NJ: Princeton University Press.

HANKEY, T. (1867): *The Principles of Banking, Its Utility and Economy; with Remarks on the Working and Management of the Bank of England*, London, UK: Effingham Wilson.

HAUBRICH, J. G. (2020): "A Brief History of Bank Capital Requirements in the United States," Economic Commentary 2020-05, Federal Reserve Bank of Cleveland.

HAYEK, F. A. (1978): *Denationalization of Money: The Argument Refined*, London, UK: Intitute for Economic Affairs.

HECKSCHER, E. F. (1934): "The Bank of Sweden: In Its Connection with the Bank of Amsterdam," in *History of the Principle Public Banks*, ed. by J. G. Van Dillen, The Hague: Marinus Nijhoff, 161–199.

HORSEFIELD, J. K. (1977): "The Beninnings of Paper Money in England," *Journal of European Economic History*, 6, 117–132.

HUBER, K. (2018): "Disentangling the Effects of a Banking Crisis: Evidence from German Firms and Counties," *American Economic Review*, 108, 868–898.

JALIL, A. J. (2015): "A New History of Banking Panics in the United States, 1825-1929: Construction and Implications," *American Economic Journal: Macroeconomics*, 7, 295–330.

KINDLEBERGER, C. P. (1993): *A Financial History of Western Europe*, Oxford, UK: Oxford University Press.

LINDLEY, R. (2008): "Reducing Foreign Exchange Settlement Risk," *BIS Quarterly Review*, September 2008, 53–65.

OFFICER, L. H. (Undated): "What Was the Consumer Price Index Then? A Data Study," MeasuringWorth, http://www.measuringworth.com/uscpi/ (Retreaved in May 2021).

OFFICER, L. H. AND S. H. WILLIAMSON (2021): "The Annual Consumer Price Index for the United States, 1774-Present," MeasuringWorth, http://www.measuringworth.com/uscpi/.

PEEK, J. AND E. S. ROSENGREN (2000): "Collateral Damage: Effects of the Japanese Bank Crisis on Real Activity in the United States," *American Economic Review*, 90, 30–45.

QUINN, S. AND W. ROBERDS (2007): "The Bank of Amsterdam and the Leap to Central Bank Money," *American Economic Review*, 97, 262–265.

——— (2014): "How Amsterdam Got Fiat Money," *Journal of Monetary Economics*, 66, 1–12.

REES, A. (1961): *Real Wages in Manufacturing, 1890-1914*, Princeton, NJ: Princeton University Press.

RICHARDS, R. D. (1929): *The Early History of Banking in England*, London, UK: P.S. King & Sons, Ltd.

ROCKOFF, H. (1974): "The Free Banking Era: A Reexamination," *Journal of Money, Credit and Banking*, 6, 141–167.

ROLNICK, A. J. AND W. E. WEBER (1982): "Free Banking, Wildcat Banking, and Shinplasters," *Federal Reserve Bank of Minneapolis Quarterly Review*, 6, 10–19.

——— (1984): "The Causes of Free Bank Failures: A Detailed Explanation," *Journal of Monetary Economics*, 14, 267–291.

SELGIN, G. A. AND L. H. WHITE (1994): "How Would the Invisible Hand Handle Money?" *Journal of Economic Literature*, 32, 1718–1749.

SMITH, A. (1776/2000): *An Inquiry into the Nature and Causes of the Wealth of Nations*, New York, NY: The Modern Library.

SMITH, V. C. (1936): *The Rationale of Central Banking*, Westminster, UK: P.S King & Son Ltd.

SPRAGUE, O. (1910): *History of Crisis Under the National Banking System*, Washington, DC: National Monetary Commission, GPO.

THORNTON, H. (1802): *An Enquiry Into the Nature and Effects of the Paper Credit of Great Britain*, London, UK: J. Hatchard.

TIMBERLAKE, R. H. (1984): "The Central Bank Role of Clearinghouse Associations," *Journal of Money, Credit and Banking*, 16, 1–15.

TOBIN, J. (1963): "Commercial Banks as Creators of "Money"," Cowles Foundation Discussion Paper No. 388.

TUFTS, R. AND P. MOLONEY (2022a): "The History of Supervisory Expectations for Capital Adequacy: Part I (1863-1983)," Moments in History, Office of the Comptroller of the Currency.

——— (2022b): "The History of Supervisory Expectations for Capital Adequacy: Part II (1984-2021)," Moments in History, Office of the Comptroller of the Currency.

VAN DILLEN, J. G. (1934): "The Bank of Amsterdam," in *History of the Principle Public Banks*, ed. by J. G. Van Dillen, The Hague: Marinus Nijhoff, 79–124.

VON GLAHN, R. (2005): "The Origins of Paper Money in China," in *The Origins of Value: The Financial Innovations that Created Modern Capital Markets*, ed. by W. M. Goetzmann and G. Rouwenhorst, Oxford, UK: Oxford Univerity Press, 65–90.

——— (2018): "Paper Money in Song-Yuan China," in *Money, Currency, and Crisis: In Search of Trust, 2000 BC to AD 2000*, ed. by R. van der Spek and B. van Leeuwen, London, UK: Routledge, 248–266.

WALLIS, J. J. (2006): "The Concept of Systematic Corruption in American History," in *Corruption and Reform: Lessons from America's Economic History*, ed. by E. L. Glaeser and C. Goldin, Chicago, IL: University of Chicago Press, 23–62.

WANG, L. (2024): "Regulating Competing Payment Networks," Working Paper, Northwestern University.

WETTERBERG, G. (2009): *Money and Power: The History of Sveriges Riksbank*, Stockholm, Sweden: Atlantis Publisher, translated into English by Patrick Hort.

WITHERS, H. (1916): *The Meaning of Money*, London, UK: Smith, Elder & Co.