Second Quarter 2014
Volume 97, Issue 2

BUSINESS REVIEW

Camden County Boathouse, New Jersey

Federal Reserve Bank of Philadelphia

INSIDE

ISSN: 0007-7011

The ***Business Review*** is published four times a year by the Research Department of the Federal Reserve Bank of Philadelphia. The views expressed by the authors are not necessarily those of the Federal Reserve. We welcome your comments at PHIL.BRComments@phil.frb.org.

For a free subscription, go to www.philadelphiafed.org/research-and-data/publications. Archived articles may be downloaded at www.philadelphiafed.org/research-and-data/publications/business-review. To request permission to reprint articles in whole or in part, click on *Permission to Reprint* at www.philadelphiafed.org/research-and-data/publications. Articles may be photocopied without permission. Microform copies may be purchased from ProQuest Information and Learning, 300 N. Zeeb Road, Ann Arbor, MI 48106.

The Federal Reserve Bank of Philadelphia helps formulate and implement monetary policy, supervises banks and bank and savings and loan holding companies, and provides financial services to depository institutions and the federal government. It is one of 12 regional Reserve Banks that, together with the U.S. Federal Reserve Board of Governors, make up the Federal Reserve System. The Philadelphia Fed serves eastern Pennsylvania, southern New Jersey, and Delaware.

Charles I. Plosser
President and Chief Executive Officer

Loretta J. Mester
Executive Vice President and
Director of Research

Colleen Gallagher
Research Publications Manager

Dianne Hallowell
Art Director and Manager

SECOND QUARTER 2014

The Paper Trail of Knowledge Transfers

BY JEFFREY LIN

Why do firms and inventors tend to locate in dense, costly areas? One intriguing hypothesis is that such geographic clustering lets them benefit from local knowledge spillovers. As Nobel laureate Robert Lucas has noted, the benefits of one person's knowledge spilling over to others play a central role in economic growth and the existence of cities: "What can people be paying Manhattan or downtown Chicago rents *for*, if not for being near other people?" Proximity may improve the sharing of knowledge, the matching of ideas to firms, or the rate of learning.[1] If dense clustering indeed confers these benefits, then that raises the possibility that individuals and firms may not be fully taking them into account when deciding where to locate, resulting in underinvestment in new ideas.

A key counterargument to the importance of knowledge spillovers is that firms might prefer to keep their work secret from competitors. For example, many firms include nondisclosure and noncompete clauses in employment contracts for researchers and scientists. Yet, as Alfred Marshall suggested, knowledge is difficult to keep secret: "The mysteries of the trade become no mysteries; but are as it were in the air."

So are, in fact, knowledge spillovers an important reason why inventors tend to locate near one another? We know that other factors might also encourage firms and inventors to locate near one another. For example, firms might benefit from better matches with specialized workers.[2] They may benefit from the sharing of local production inputs such as cheap electrical power or hard-to-find machinery and parts. Or skilled inventors may be attracted to superior amenities such as restaurants, shopping, or safety.

A key challenge, then, is to account for these alternative explanations so that we do not erroneously infer that knowledge spillovers are empirically important. To explore this challenge, I review the empirical literature regarding evidence of knowledge spillovers contained in patent citations and nonpatent data. I then describe the evidence that Ina Ganguli, Nick Reynolds, and I found using a novel measure — cases of simultaneous invention that result in rival claims known as patent interferences. First, let us look at what researchers have found by studying routine patent applications.

EVIDENCE FROM PATENT CITATIONS

At its most basic, the challenge of verifying the existence of knowledge spillovers was observed by Paul Krugman — namely, that knowledge flows are invisible: "They leave no paper trail by which they may be measured or tracked." Adam Jaffe, Manuel Trajtenberg, and Rebecca Henderson tackled this problem by observing that the flow of knowledge from one inventor to another could, in fact, be tracked using patent citations.[3] Their paper and ones that followed have provided the best evidence to date that local knowledge spillovers might be one important mechanism contributing to the geographic proximity of inventors.[4] They exploit the fact that patents include citations to older patents. If a new patent cites a previous patent, this citation is evidence that the older patent contains knowledge upon which the new patent relies.

Though a citation to a nearby inventor is at first glance evidence that knowledge has passed from the earlier inventor to the citing inventor, it does

[1] Gerald A. Carlino's 2001 *Business Review* article discusses such mechanisms.

[2] See my 2009 *Business Review* article.

[3] Robert M. Hunt's 2001 *Business Review* article discusses patents as a measure of knowledge production.

[4] See also the work by Jaffe (1986) and Keller (2002), and the surveys by Rosenthal and Strange (2004) and Audretsch and Feldman (2004).

Jeffrey Lin is an economic advisor and economist at the Federal Reserve Bank of Philadelphia. The views expressed in this article are not necessarily those of the Federal Reserve. This article and other Philadelphia Fed reports and research are available at www.philadelphiafed.org/research-and-data/publications.

not necessarily indicate that geographic proximity has facilitated this transfer of knowledge. It might simply be the case that inventors are located nearby for some other reasons besides the opportunity to take advantage of knowledge spillovers, such as to be near some common physical input to the invention process. For example, if a patent awarded to researchers at a University of Pennsylvania hospital is cited in a subsequent application for a patent awarded to researchers at Temple University, that might be because of local knowledge spillovers. But it could also be the case that these patentees are near each other simply because many hospitals are needed to serve the large population of the Philadelphia area, and proximity offers no advantage in transmitting knowledge.

Expected proximity. To address this inference problem, Jaffe and his coauthors develop a clever matching strategy. They measure the distance between an earlier "originating" patent and a subsequent "citing" patent that references the originating patent as an important knowledge input. Then they compare this distance to the distance between the originating patent and a matched "control" patent. The control patent is similar to the matched citing patent in terms of the date of invention and technological classification, but it does not cite the matched originating patent. Thus, the control patent represents the expected proximity of inventors working in the same research field and time period, not conditioned on a citation link. If the inventors of the citation-linked patent pair are observed in closer proximity versus this benchmark, then this is strong evidence that a local knowledge spillover has occurred, especially since we have accounted for the underlying geographic distribution of research activity and hence other reasons why inventors might be located together. In fact, Jaffe and his coauthors find that originating patentees are much more likely to be from the same metropolitan area as citing patentees, compared with a matched control patentee.

Despite this clever study design, subsequent researchers have identified several limitations of this analysis. First, there are the standard drawbacks to using patent data: Not all inventions are patented, and some patents do not represent valuable or worthwhile inventions. More recent papers have tried to correct these problems by, for example, measuring the quality of patents based on patent renewals or subsequent citations. Second, many patent citations are actually added by patent examiners, not inventors. Thus, citations may not actually represent true knowledge flows for inventors, but rather noise introduced by the patent office.[5] Third, Peter Thompson and Melanie Fox-Kean note that Jaffe's results are sensitive to the selection of an appropriate control patent. By varying how broadly the technology classifications and dates of application are specified for the sample of matched control patents, Thompson and Fox-Kean found that imperfect matching explained a significant part of Jaffe's original result.

A final issue, which Jaffe and his coauthors acknowledged in their original paper, is that many knowledge inputs are not actually reflected in citations. This is significant because we might expect that geographic proximity is especially important for the transfer of tacit, operational knowledge — that is, knowledge that is not necessarily codified in a written patent application. This is the kind of knowledge transmitted in hallways and over coffee, rather than through literature searches of previous work. Because many more knowledge spillovers may occur than are reflected in citations, Jaffe and his coauthors suggest that their findings may actually represent a lower bound for the occurrence of knowledge spillovers among inventors.

EVIDENCE FROM RESEARCH AND NONPATENT DATA

Other papers have sidestepped patents altogether.[6] Bruce Weinberg found that physicists who moved to cities where Nobel laureates were already working were more likely to begin their own Nobel Prize-winning work there. Gerald A. Carlino, Jake K. Carr, Robert M. Hunt, and Tony E. Smith have shown that research and development labs are highly geographically concentrated, substantially more so than the corresponding industry concentration patterns. In my previous work, I showed that new activities related to the implementation of new knowledge are concentrated in metropolitan areas with highly educated populations. Finally, Petra Moser has shown localization among prize-winning inventors at World's Fairs in the 19th century, although these patterns weakened over time.

Of course, as with inventors, scientists may locate near each other for reasons besides knowledge spillovers. Thus, we cannot be certain whether an increase in their productivity might stem from knowledge spillovers from nearby scientists or from some other reason. Fabian Waldinger investigated local knowledge spillovers among scientists in Germany. Waldinger relies on evidence from the expulsion of Jewish and certain other scientists from Germany under the Nazis. Some university departments experienced many expulsions, while other departments had not employed Jewish scholars and were therefore unaffected. If knowledge

[5] See the papers by Juan Alcácer and Michelle Gittelman and by Peter Thompson (2006).

[6] See the survey article by David Audretsch and Maryann Feldman.

spillovers are important, one might expect the productivity of the remaining scientists in the affected departments to decline following the expulsion of their Jewish colleagues. Waldinger finds that the publishing activity of the scientists whose departments suffered losses did not decline compared with that of other scientists. Thus, he concludes that there is no evidence for local knowledge spillovers among German scientists in this period.

EVIDENCE FROM PATENT INTERFERENCES

In my work with Ina Ganguli and Nick Reynolds, I have been using patent interferences to try to provide new evidence on the relevance of local knowledge spillovers for invention. Patent interferences are especially valuable for measuring local spillovers of tacit or uncodified knowledge that is missing in traditional patent studies.

Patent interferences are a unique historical feature of the U.S. patent system. Until 2011, the United States had a "first to invent" rule for assigning priority of invention, versus the "first to file" rule more common in the rest of the world.[7] When the U.S. patent office received applications from multiple parties with identical claims at roughly the same time, it was obligated to investigate the competing claims to determine which party was entitled to patent protection. This investigation, known as a patent interference proceeding, determined who had conceived of the invention and reduced it to practice first. Typically, the parties submitted dated laboratory notebooks, testimony by associates, and media reports as evidence of first invention.

There are many famous examples of patent interferences in U.S. history, including Alexander Graham Bell's and Elisha Gray's simultaneous invention of the telephone. Because Bell's and Gray's applications arrived at the patent office on the same day and contained nearly identical claims, an interference proceeding was declared. Eventually, Bell was determined to have conceived of the idea and reduced it to practice first, and he was awarded the patent.

Knowledge in common. Importantly for economists, patent interferences create a record of instances when the same invention is created by inventors working independently, a phenomenon that is highly suggestive of common knowledge inputs. In other words, inventors involved in an interference are likely to have command of similar knowledge. For example, interfering inventors may have similar backgrounds in chemistry, or they may have similar knowledge of market conditions. This is especially true if certain inventions require highly specific knowledge. For example, for Bell and Gray to have both invented the telephone contemporaneously, they must have had similar knowledge about electrical conductivity and the properties of various conductive metals, as well as similar expectations about market demand for a device that transmitted voices in real time. For Jon Merz and Michelle Henry, a patent interference is an indication that "discovery has become ordinary." That is, its occurrence suggests that certain knowledge is shared among several inventors. In other words, a patent interference is evidence of a knowledge spillover among the inventors.

Several details about the interference process support the argument that these proceedings are a good measure of common, independent knowledge inputs. First, interferences were declared by a patent examiner specializing in a particular technological area. Thus, interfering claims were likely to be detected. (In some cases, the examiner was alerted to a possible interference by one of the applicants. Note that an interference is different from patent infringement, in which the holder of an existing patent sues an infringing party. Private parties cannot sue for an interference.)

> Patent interferences are especially valuable for measuring local spillovers of tacit or uncodified knowledge that is missing in traditional patent studies.

Second, interferences involved parties with roughly the same date of patent application. An inventor who delayed filing an application in order to conceal an invention would lose the priority contest. Thus, interferences are less likely to reflect secrecy or other legal strategies of the participants and are more likely to reflect genuinely independent, simultaneous inventions versus infringements or patent "racing" by inventors who believed that rival applications were imminent.

Third, during the interference proceedings, circumstances that suggested stealing, espionage, or collaborative invention typically led to dismissal with prejudice. In other words, worker poaching and espionage that is independent of shared knowledge are unlikely explanations for the bulk of cases of patent applications interfering with each other. Note that recruiting other firms' researchers or spying

[7] More details about the patent interference proceedings can be found in Calvert and Sofocleous (1982), Cohen and Ishii (2006), de Simone et al. (1963), and Kingston (2004).

on them does seem to involve shared knowledge inputs — a shared desire to solve a common problem, for example. In addition, patent judges had been compelled by statute to rule against applicants found to have stolen a competitor's idea, deterring would-be spies from pursuing an interference. In fact, in our examination of case decisions, no more than a small handful of judgments mention espionage as a relevant factor in decisions.

Fourth, competing claims that are similar but not identical did not result in a completed interference case. Fifth, the patent office verified that interfering parties had independent financial interests (for example, that they were not different branches of a multinational conglomerate); otherwise, the case was dismissed. Thus, interferences are not the result of knowledge sharing *within* organizations. Finally, the separate applications were required to have been made roughly at the same time, often within a year. Thus, copying subsequent to publication and disclosure of an older patent are unlikely to have occurred.

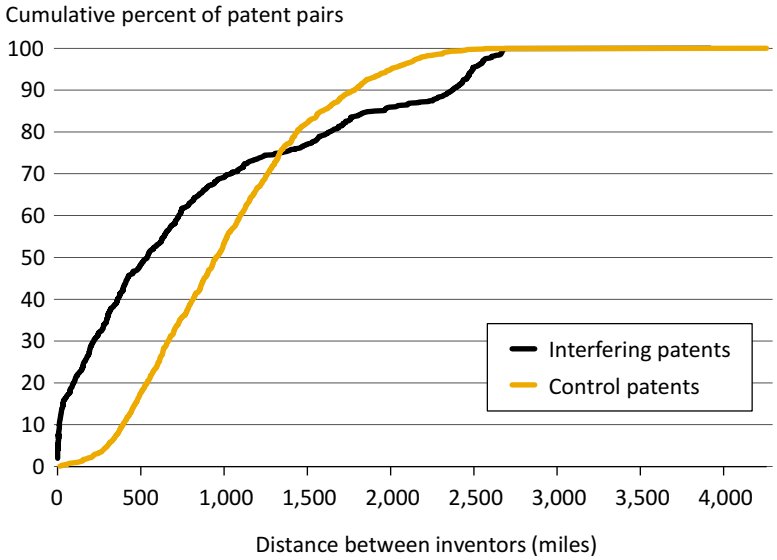
Interferences improve on traditional patent studies in a number of ways. One, interferences involve patents that are more valuable than the average patent. Since an interference requires parties to actively contest for priority, it is unlikely that inventors would spend time or money in pursuit of a worthless patent. Two, we have information on patent interferences over a long period, from the 19th century to 2011. Three, patent interferences do not rely on citations to prove common knowledge inputs and are thus not subject to some of the weaknesses noted earlier. Specifically, while patent citations necessarily capture the spillover only of written, publicly available knowledge, simultaneous invention is evidence that *some* kind of spillover, whether written or not, is likely to have occurred. Thus, interferences capture spillovers of tacit knowledge. As I noted earlier, we might expect that tacit knowledge is especially sensitive to geographic proximity. If so, then we expect results on the localization of interferences to be stronger than on the localization of citations.

We have constructed a database of over 1,000 interference cases from the early 1980s to 2011 from the U.S. patent office. This database includes the names of the inventors involved in the interference, their patent and application numbers, and the date of the interference. We match this information to a database of inventor locations (based on their Zip codes) produced by the Harvard Business School.

Testing geographic concentration. If local knowledge spillovers are important, one possible test is to see whether patent interferences, as measures of shared, possibly tacit knowledge, are more likely to occur between inventors who are located close to each other versus those located farther apart. The black line in Figure 1 shows this pattern for only the interference cases involving pairs of U.S.-based inventors. The horizontal axis measures the distance in miles as the crow flies between the observed locations of the two parties involved in a patent interference. (Since a single patent application can be made on behalf of multiple inventors, we measure the minimum distance between inventors of the different parties to the interference.) The vertical axis shows the percent of interfering inventor pairs in our database that are separated by *at most* the distance indicated by the horizontal axis. Thus, as we move to the right, we accumulate our inventor pairs until 100 percent of our pairs are within 4,258 miles — the maximum distance observed between two interfering U.S. inventors.

FIGURE 1

Interfering Inventors More Geographically Concentrated



Sources: U.S. Patent and Trademark Office, Harvard University Patent Network Dataverse.

Despite the large possible range of distances between inventors, the black line shows that 20 percent of interfering inventors are within only 100 miles of each other, and half of interfering inventor pairs are within 540 miles of each other. While this is a good starting point for showing that proximity matters for shared knowledge inputs, it still might be true that inventors are located close to each other to take advantage of some other factor. Similar to the logic of the patent citation studies, we can compare the localization of interferences with the localization of noninterfering patents in the same technology classification and year. In that way, we can control for the underlying distribution of research activity that doesn't rely on common knowledge inputs, as interferences do.

For each pair of interfering patents, we selected up to 10 control patents. Our goal was to control for the underlying geographic distribution of inventive activity by selecting patents that were similar to the interfering patents but not involved in the interference case. We selected control patents based on two criteria. First, a control patent had to share at least one of the many possible three-digit technological classification codes assigned by the patent office that the two interfering patents had shared. Second, the control patent's application date had to fall between the application dates of the two interfering patents. If no eligible control patent was found, we then expanded the selection window incrementally by 10 days before the earlier interfering application and 10 days after the later interfering application until an eligible control patent was found. Finally, we randomly chose one of the two interfering patents to match with the control patent. We then compared the distance between the interfering inventors with the distance between the randomly selected interfering inventor and the control inventor.

The results. The gold line in Figure 1 shows our results for the proximity of interfering inventors to control patent inventors. It represents the expected distribution of distances between inventors working in technology fields and time periods similar to those of our sample of patent interferences, but it is not conditioned on an interfering link between inventors.

Comparing the distribution of distances between interfering inventors with the control distribution of noninterfering inventors, it is clear that interfering inventors are more geographically concentrated. While 20 percent of interfering inventors are within 100 miles of each other, less than 1 percent of noninterfering inventors are within 100 miles of each other. And while half of interfering inventors are within 540 miles of each other, the same is true of less than 21 percent of noninterfering inventors.

Interfering inventors are especially more likely to be geographically concentrated at small distances. For example, more than 3 percent of interfering U.S. inventors are in the same Zip code, versus none of the noninterfering inventor-control pairs. Eleven percent of interfering inventors are within 15 miles of each other, compared with less than 1 percent of noninterfering inventors. These results are consistent with a growing literature documenting that knowledge spillovers attenuate rapidly with distance.

CONCLUSION

Although local knowledge spillovers are of central interest to economists, the evidence to date on their existence is mixed. Patterns in our data on patent interferences suggest that inventors working independently but using common knowledge inputs are substantially more geographically concentrated than other inventors working in the same field and time period who are not linked by common knowledge inputs. These results suggest that localized knowledge spillovers may be especially salient for forms of tacit or uncodified knowledge, which is difficult to observe using citations but more likely detectable from interferences.

REFERENCES

Alcácer, Juan, and Michelle Gittelman. "Patent Citations as a Measure of Knowledge Flows: The Influence of Examiner Citations," *Review of Economics and Statistics*, 88:4 (2006), pp. 774-779.

Audretsch, David B., and Maryann P. Feldman. "Knowledge Spillovers and the Geography of Innovation," in J. Vernon Henderson and Jacques-François Thisse, eds., *Handbook of Regional and Urban Economics*, 4. Amsterdam: Elsevier, 2004, pp. 2,713-2,739.

Audretsch, David B., and Maryann P. Feldman. "R&D Spillovers and the Geography of Innovation and Production," *American Economic Review* 86:3 (1996), pp. 630-640.

Azoulay, Pierre, Joshua Graff Zivin, and Jialan Wang. "Superstar Extinction," *Quarterly Journal of Economics*, 125:2 (2010), pp. 549-589.

Borjas, George J., and Kirk B. Doran. "The Collapse of the Soviet Union and the Productivity of American Mathematicians," *Quarterly Journal of Economics*, Oxford University Press, 127:3 (2012), pp. 1,143-1,203.

Calvert, Ian A., and Michael Sofocleous. "Three Years of Interference Statistics," *Journal of the Patent Office Society*, 64:12 (1982), pp. 699-707.

Carlino, Gerald A., Satyajit Chatterjee, and Robert M. Hunt. "Urban Density and the Rate of Invention," *Journal of Urban Economics*, 61 (2007), pp. 389-419.

Carlino, Gerald A. "Knowledge Spillovers: Cities' Role in the New Economy," Federal Reserve Bank of Philadelphia *Business Review* (Fourth Quarter 2001).

Carlino, Gerald A., Jake K. Carr, Robert M. Hunt, and Tony E. Smith. "The Agglomeration of R&D Labs," Federal Reserve Bank of Philadelphia Working Paper 12-22 (2012).

Chapin, Seymour L. "Patent Interferences and the History of Technology: A High-Flying Example," *Technology and Culture*, 12:3 (1971), pp. 414-416.

Cohen, Linda R., and Jun Ishii. "Competition, Innovation and Racing for Priority at the U.S. Patent and Trademark Office," University of California, Irvine, Working Paper (2006).

De Simone, Daniel V., James B. Gambrell, and Charles F. Gareau. "Characteristics of Interference Practice," *Journal of the Patent Office Society*, 45 (1963), pp. 503-591.

Duranton, Gilles, and Diego Puga. "Micro-Foundations of Urban Agglomeration Economies," in Vernon Henderson and Jacques-Francois Thisse, eds., *Handbook of Regional and Urban Economics*, 1. Amsterdam: Elsevier, 2004, pp. 2,063-2,117.

Duranton, Gilles, and Diego Puga. "Nursery Cities: Urban Diversity, Process Innovation, and the Life-Cycle of Products," *American Economic Review*, 91:5 (2001), pp. 1,454-1,477.

Glaeser, Edward L. "Learning in Cities," *Journal of Urban Economics*, 46 (1999), pp. 254-277.

Hunt, Robert M. "You Can Patent That? Are Patents on Computer Programs and Business Methods Good for the New Economy?" Federal Reserve Bank of Philadelphia *Business Review* (First Quarter 2001).

Jaffe, Adam B. "Technological Opportunity and Spillovers of R&D: Evidence from Firms' Patents, Profits, and Market Value," *American Economic Review,* 76:5 (1986), pp. 984-1,001.

Jaffe, Adam B., Manuel Trajtenberg, and Rebecca Henderson. "Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations," *Quarterly Journal of Economics*, 108:3 (1993), pp. 577-598.

Keller, Wolfgang. "Geographic Localization of International Technology Diffusion," *American Economic Review*, 92:1 (2002) pp. 120-142.

Kingston, William R. "Light on Simultaneous Invention from U.S. Patent Office 'Interference' Records," *World Patent Information*, 26:3 (2004), pp. 209-220.

Krugman, Paul. *Geography and Trade*. Cambridge: MIT Press, 1991.

Lin, Jeffrey. "Technological Adaptation, Cities, and New Work," *Review of Economics and Statistics*, 93:2 (2011), pp. 554-574.

Lin, Jeffrey. "Urban Productivity Advantages from Job Search and Matching," Federal Reserve Bank of Philadlpelphia *Business Review* (First Quarter 2011).

Lucas, Robert E. "On the Mechanics of Economic Development," *Journal of Monetary Economics*, 22 (1988), pp. 3-42.

Marshall, Alfred. *Principles of Economics*. London: Macmillan, 1920.

Merz, Jon F., and Michelle R. Henry. "The Prevalence of Patent Interferences in Gene Technology," *Nature Biotechnology*, 22:2 (2004), pp. 153-154.

Moser, Petra. "Did Patents Weaken the Localization of Innovations? Evidence from World's Fairs," *Journal of Economic History*, 71:2 (2011), pp. 363-381.

Romer, Paul M. "Endogenous Technological Change," *Journal of Political Economy*, 98:5 (1990), pp. 71-102.

Rosenthal, Stuart S., and William C. Strange. "Evidence on the Nature and Sources of Agglomeration Economies," in J. Vernon Henderson and Jacques-François Thisse, eds., *Handbook of Regional and Urban Economics* 4. Amsterdam: Elsevier, 2004, pp. 2,119-2,171.

Thompson, Peter. "Patent Citations and the Geography of Knowledge Spillovers: Evidence from Inventor- and Examiner-Added Citations," *Review of Economics and Statistics*, 88:2 (May 2006), pp. 383-388.

Thompson, Peter, and Melanie Fox-Kean. "Patent Citations and the Geography of Knowledge Spillovers: A Reassessment," *American Economic Review*, 95:1 (March 2005), pp. 450-460.

Waldinger, Fabian. "Peer Effects in Science: Evidence from the Dismissal of Scientists in Nazi Germany," *Review of Economic Studies*, 79 (2012), pp. 838-861.

Weinberg, Bruce A. "Geography and Innovation: Evidence from Nobel Laureate Physicists," mimeo (2007).

Shadow Banking and the Crisis of 2007-08

BY DANIEL SANCHES

Some economists have noted that recessions accompanied by banking crises tend to be deeper and more difficult to recover from than other recessions — even those associated with other types of financial crises. For instance, the bursting of the dot.com bubble in 2001 was a very important financial event that was not accompanied by a protracted recession. The potential of banking crises to do lasting economic harm led policymakers to adopt safeguards in the 1930s that have essentially eliminated traditional banking panics in the U.S. Although the Great Recession of 2007-09 was associated with a protracted financial market disruption — and the failures of some large banks like Washington Mutual and IndyMac — we did not observe widespread withdrawals from commercial banks, as in a traditional banking crisis. However, economists Gary Gorton and Andrew Metrick show that it can be viewed as a banking crisis that originated in the *shadow banking system*. In the last 30 years, institutions very similar in function to traditional banks have grown outside regulatory oversight. One lesson of the financial crisis is that these institutions are as vulnerable to panics as traditional banks because they are subject to similar risks.

ECONOMIC FALLOUT OF BANKING CRISES

Banking crises can harm the economy. Financial crises are usually associated with bad economic outcomes — recessions. One particular kind of financial crisis to which economists have devoted a lot of attention is the type that originates in the banking sector. A *banking crisis* is a widespread withdrawal of funds from depository institutions — that is, a run on the liabilities of a large number of banks.[1] Like other financial crises, banking crises are usually associated with economic downturns, and there is evidence that banking crises often worsen economic downturns as weaknesses at the banks spill over into financial problems for households and firms. When financial events affect consumption and investment decisions by households and nonfinancial firms, economists say that they have *real effects*.

Many researchers have provided evidence that banking crises can make economic contractions more severe and more protracted, with various studies emphasizing different channels.

In a highly influential book providing a systematic account of banking crises in the United States from 1867 to 1960, Milton Friedman and Anna Schwartz identify the reduction in the wealth of bank shareholders and the decline in the money supply that usually follow a business downturn accompanied by a banking crisis as the main causes of a further drop in real economic activity. A decline in the money supply has real effects because households and firms need money to pay for their purchases. Thus, a decline in the money supply leads to a decline in transactions and real economic activity.[2]

Looking at the banking crises of the 1930s, Ben Bernanke identified the increased cost of intermediation services — the costs that banks incur when assessing the creditworthiness of borrowers — following the recurrent banking crises of the early 1930s as causing a significant reduction in the flow of funds from lenders to borrowers through the banking system. This constricted flow of credit impeded the real economy's recovery from the Great Depression.

Michael Bordo, Barry Eichengreen, Daniela Klingebiel, and Maria Soledad Martinez-Peria provide cross-country evidence of the real effects of banking crises over a period of 120 years. They also find that recessions that are accompanied by banking crises are more severe than those that are

[1] A common form of bank liability is the demand deposit contract — a typical checking account that most people have at a commercial bank.

[2] For more on this subject, see my 2012 *Business Review* article, "The Optimum Quantity of Money."

Daniel Sanches is a senior economist at the Federal Reserve Bank of Philadelphia. The views expressed in this article are not necessarily those of the Federal Reserve. This article and other Philadelphia Fed reports and research are available at www.philadelphiafed.org/research-and-data/publications.

not. Moreover, they provide evidence that banking crises plague advanced and developing economies equally, confirming the view that a banking crisis is not just a concern for countries at low levels of economic development.

Anatomy of a banking crisis. Although every banking crisis is different, those that occurred up to and including the Great Depression follow a similar pattern. Let me briefly describe the typical sequence of events that leads to a banking crisis.

Bad news arrives. Usually at the peak of an economic expansion, bad news about the quality of the assets held by a group of banks (or a major bank) leads to larger withdrawals than usual. For instance, a failed attempt by the Knickerbocker Trust Company to corner the copper market and the subsequent decision of a major bank to no longer clear checks issued by the Knickerbocker triggered a run on the Knickerbocker on October 18, 1907, sparking the Panic of 1907. The fact that it was one of the largest depository institutions in New York contributed to the public's perception that other banks could also be in distress.

Banks sell assets to meet the increase in withdrawals. To meet the higher demand for cash, a bank initially draws down its cash reserves. But its reserves may not be enough if the withdrawal process quickly intensifies. The bank can also sell some of its assets to cover withdrawals.

Selling assets causes asset prices to fall. If many banks are trying to sell assets at the same time, the assets can be sold only at a large discount. Think of what would happen if four neighbors on your block put their houses up for sale on the same day as you did. All else equal, you would have to lower your price to get anyone to buy. Financial asset markets work the same way. For easily marketable fixed-income assets such as Treasury securities or certain corporate bonds traded in large markets, buyers can still be found by selling at a discount. However, a large fraction of a bank's assets consists of mortgages and commercial and industrial loans made to households and firms whose creditworthiness is unknown to the wider market, which means the bank would probably find few if any buyers. And anyone willing to purchase the loan would demand a substantial discount to compensate for the lack of information about the borrower's creditworthiness. Thus, selling assets on short notice may be extremely costly for a bank.

Depositors begin a run on healthy banks. Banks facing large withdrawals may borrow in the interbank market, where banks routinely borrow reserves from each other. But if banks want to borrow more reserves than usual, they must pay a higher interest rate to the lending bank. Larger discounts in asset markets and higher interest rates in interbank markets are usually signs of financial strain. If widespread distrust of banks causes depositors to withdraw their funds even from healthy banks, a line is crossed. The number of banks that want to sell assets increases, resulting in even steeper discounts, and the number of banks that want to borrow in the interbank market also increases, making it harder for each borrower bank to obtain a loan. As this process intensifies, we have a full-scale panic.[3]

Banks suspend convertibility. The final step comes as banks react to widespread withdrawals. One way to stop the drain on funds is to temporarily suspend the convertibility of deposits into cash — banks may simply lock their doors — in an attempt to preserve capital until depositors calm down and things get back to normal. Strictly speaking, this is a breach of the demand deposit contract.[4]

This description of a typical banking crisis clearly reveals why banks are fragile: They fund illiquid assets with deposits that can be withdrawn at will. Economists usually refer to this practice as *maturity transformation*. It is important to mention that this role played by banks has a value for society. People have a preference for holding highly liquid assets — assets that are easy to sell without taking a loss — but the most profitable investments take a long time to pay off. Banks offer demand deposit contracts that give people ready access to their funds and a higher rate of return than they would get by holding liquid assets directly. Banks are able to offer a higher rate of return to depositors because they pool resources in such a way that permits them to invest a significant fraction of their assets in higher-yielding, long-term projects such as mortgages and other types of long-term loans. Normally, funding illiquid assets with short-term liabilities works fine. But when depositors begin to worry about losses, a bank run may ensue.

Although every banking crisis is different, those that occurred up to and including the Great Depression follow a similar pattern.

[3] We can think of these withdrawals as a way for depositors to monitor their banks. That is, by withdrawing their money, depositors are checking whether the bank is healthy enough to pay. This might explain why people decide to withdraw their funds even from banks initially viewed as safe and sound.

[4] In the second half of the 19th century, the decision to suspend convertibility was usually coordinated by private bank associations.

U.S. bank runs essentially disappeared in the 1930s. The introduction of federal deposit insurance in 1933 with the creation of the Federal Deposit Insurance Corporation (FDIC) ended the banking crises that had been recurrent events in the U.S. even before the Great Depression. The government's deposit guarantees largely relieved depositors of the need to constantly monitor the health of banks. In turn, the government has undertaken the monitoring of banks through regulation and supervision. But regulations are not costless. FDIC premiums, capital requirements, and regulatory restrictions on bank portfolios increase banks' costs. These costs are informally referred to as *regulatory taxes*. And banks, like any other firm, have a strong incentive to avoid taxes.

THE RISE OF SHADOW BANKING

The Great Recession in the U.S. was associated with a severe financial crisis, but we did not observe people rushing to their banks to withdraw their deposits. However, a closer look suggests that the crisis was not very different from a typical banking crisis, except that it was triggered outside the traditional banking sector. According to Gary Gorton and Andrew Metrick, the financial crisis can be viewed as a banking crisis that originated in the *shadow banking system*.[5]

The shadow banking system is a set of institutions that carry out functions very similar to those of traditional banks but that are largely unregulated. They perform the same kind of maturity transformation traditionally performed by commercial banks. Thus, the shadow banking system, despite its somewhat unwholesome-sounding name, provides a useful service to society. This is to say that shadow banking is not necessarily a bad thing. The problem is that, under certain circumstances, these financial institutions can become fragile — that is, subject to panics.

An important fact about the shadow banking system is that it has grown significantly in the last 30 years. For instance, Gorton and Metrick estimate that just before the financial crisis of 2007-08, the assets of the shadow banking system were at least as large as the assets of commercial banks.[6] Another important fact about the shadow banking system is that it has grown outside the oversight of regulators. Why did this happen? As banking and finance in general have expanded in recent decades, part of that growth has occurred in the shadow system, largely to avoid the costs associated with regulation.[7]

As I will now explain, the shadow banking system works pretty much like a typical commercial bank even though the parties involved in the transactions are not the bankers and depositors that we typically have in mind. For the most part, I will follow Gorton and Metrick and focus on the market for repurchase agreements (or repos) as the main cause of the panic in the shadow banking system and one of the centers of the financial crisis. But the shadow banking system also includes other markets and institutions such as asset-backed commercial paper in which the same basic structure (risky, illiquid assets funded by short-term liabilities) recurs.

The repo market. The repo market is a market for short-term, mainly overnight, collateralized loans. To understand why repos work pretty much like banking and to see why the repo crisis was actually a banking crisis, it is necessary to look at how repo transactions work.

Let me start by identifying the "depositors," the repo lenders. These are largely institutional investors such as pension funds and large corporations that need some place to invest large amounts of money for short periods. They also want to obtain higher yields than those offered by regulated commercial banks. Most important, these institutional investors want their funds to be safe.[8]

One alternative is the repo market. A firm can make an overnight loan to a borrower. To make the loan safe, the firm receives collateral usually in the form of government bonds, which are liquid and fluctuate little in value over short periods. If the borrower is unable to return the funds, the lending firm will simply seize the collateral. Provided that the value of the underlying collateral does not change significantly over short periods, a repo transaction is safe for the repo lender.

Like a bank depositor, the repo lender has ready access to its money and has the opportunity to reallocate its funds toward some other use on a daily basis. Thus, a repo transaction offers the firm both the convenience of having ready access to its funds and a level of safety not much different from that of a federally insured demand deposit. Until 2011, large commercial depositors could not receive interest on their short-term deposits, another motivation for them to seek an alternative place to park their funds.[9] When the

[5] See also chapter 2 in Gary Gorton's 2010 book.

[6] According to Gorton and Metrick, this is probably an underestimate because this comparison involves the assets of only a fraction of the shadow banking system.

[7] For more on the rise of shadow banking, see the review by Tobias Adrian and Adam Ashcraft.

[8] It is also important to mention that the amounts these institutional investors wish to deposit are typically larger than the maximum amount insured by the FDIC.

[9] As part of the Dodd-Frank Act, the Federal Reserve Board in July 2011 repealed Regulation Q, which had prohibited banks from paying interest on corporate checking accounts.

repo borrower repurchases the security from the repo lender, he or she also pays interest to the lender.

As should be clear by now, the "banker" in the repo transaction is the repo borrower, which typically is an investment bank or the broker-dealer arm of a large bank holding company. These institutions use the funds they borrow in the repo market to finance a wide range of activities, some of them quite risky. As long as the repo is collateralized by a Treasury security, it is not fragile in the same sense as traditional banking because the asset that collateralizes the repo is highly liquid and can be easily sold. If the repo borrower can't repay on time, the repo lender can simply take the collateral and sell it for cash.

This is basically how the shadow banking system works. Depositors (institutional investors and large corporations) need a place to park liquid funds that provides them with ready access to their money, pays an interest rate higher than that offered by traditional banks, and spares them the expense and hassle of managing their own cash.[10] Bankers (investment banks and broker-dealer firms) are willing to provide such a product in the form of repo transactions. Finally, safe collateral such as U.S. Treasury bonds are essential to make this financial transaction work.

The growth of the repo market increased the demand for collateral. The growth of the repo market prior to the financial crisis of 2007-08 was extraordinary. The volume of repo transactions reported by primary dealers (those who trade directly with the Federal Reserve System) had grown from roughly $2 trillion in 1997 to $7 trillion in 2008. This estimate, of course, leaves out unreported transactions. Gorton and Metrick estimate that the overall size of the repo market just before the financial crisis was roughly the same as the size of the traditional banking sector as measured by total assets.[11]

As we have seen, Treasury securities play an important role in the functioning of the shadow banking system. However, repo markets are not the only source of demand for Treasury securities. They are also used as collateral in derivative markets and settlement systems. Furthermore, many foreign governments, especially the central banks of developing countries such as China, demand Treasury securities because they are safe and highly liquid.[12] For instance, in 2005 only 48.6 percent of total U.S. debt was privately held, according to the Federal Reserve Bank of San Francisco. About a third of that privately held debt was held in reserve by foreign central banks, which means that only about a third of total U.S. debt (or $2.6 trillion) was available for private transactions.

Unlike for other goods and services, higher demand for Treasury securities doesn't automatically provide an incentive to increase supply. The supply of government bonds is determined by government borrowing, a direct consequence of fiscal policy. For instance, the decision to reduce the fiscal deficit in the U.S. in the 1990s and early 2000s may have contributed to a shortage of government bonds available for repo transactions.

One piece of indirect evidence that government bonds were in short supply is the practice in financial markets known as rehypothecation, which simply means that traders can use the same collateral to secure more than one transaction. To the extent that this practice had become widespread before the crisis of 2007-08, traders may have had an incentive to develop other methods to conduct a growing number of transactions with a limited amount of good collateral.[13]

Mortgage-backed securities helped satisfy the demand for collateral. The solution to the shortage of good collateral was found in another form of financial innovation that had evolved significantly since the 1980s: securitization. Commercial banks make many loans to consumers and firms. Instead of holding these loans on its own balance sheet, a bank can sell them to a shell company the bank creates and manages for this purpose, called a special purpose vehicle (SPV). The SPV funds the acquisition of these assets (mortgages, car loans, credit card receivables, etc.) by issuing asset-backed securities (ABS) that, as the name implies, are backed by the loans the SPV holds and that become the SPV's liabilities when it sells them to investors in the capital markets.[14] Figure 1 shows how commercial banks fund loans through securitization.

Most important, this organizational form allows financial institutions to increase the scale of their overall operations without increasing their balance sheets, which would require them to increase their regulatory capital. Thus, setting up an SPV is a way of avoiding

[10] As Robert Lucas puts it: "In a monetary economy, it is in everyone's private interest to try to get someone else to hold non-interest-bearing cash and reserves. But someone has to hold it all, so all of these efforts must simply cancel out. All of us spend several hours per year in this effort, and we employ thousands of talented and highly trained people to help us. These person-hours are simply thrown away, wasted on a task that should not have to be performed at all."

[11] They also cite a range of estimates by other economists of the same order of magnitude.

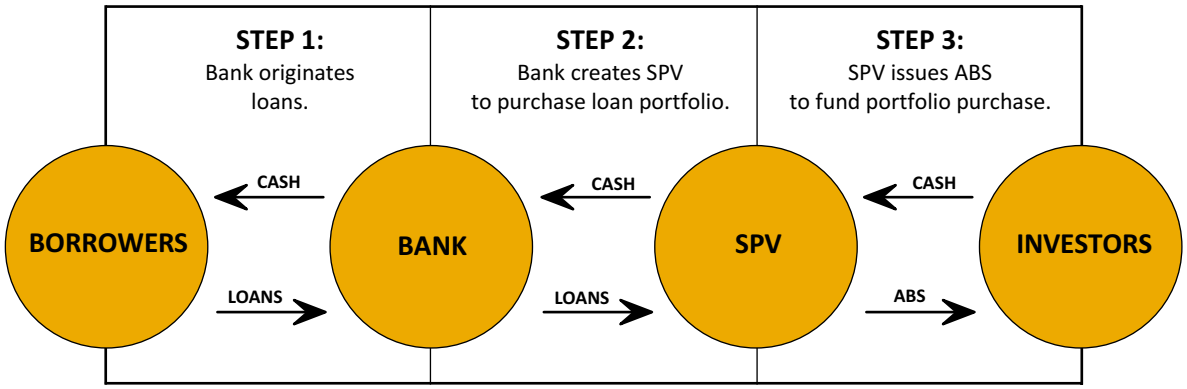
[12] Foreign demand for Treasury securities increased significantly in the 1990s and 2000s, the flip side of the large trade surpluses run by China and some other developing countries.

[13] For more on the role of rehypothecation, see the 2011 *Business Review* article by Cyril Monnet.

[14] See the chapter by Gary Gorton and Nicholas S. Souleles.

FIGURE 1

Securitization and Shadow Banking



capital requirements, which increases a financial institution's overall degree of leverage by raising its total assets relative to its capital.[15]

If carefully chosen, a portfolio of loans backing the ABS can be safe and predictable. Thus, by making it possible to bundle individual loans and sell claims on the loan portfolio on the market, this form of financial innovation offers an alternative to using deposits to fund banks' illiquid assets. When carefully executed, securitization is extremely valuable for both banks and investors.

The housing boom in the U.S. in the 2000s was financed in this way. The large increase in the number and size of mortgage loans created a large supply of a particular type of ABS called mortgage-backed securities (MBS). As the name suggests, MBS are ABS that bundle mortgages. Given the growth of the repo market and the relative scarcity of government bonds, the use of ABS as collateral in the repo market seemed to be a reasonable solution to the shortage of collateral. Gorton and Metrick have argued that the use of ABS as collateral in the repo market had increased significantly prior to the financial crisis. As I discuss below, this is still a controversial claim. Despite a wealth of anecdotal evidence, we have no precise estimates of the share of repo transactions that used ABS as collateral.

CRISIS IN THE SHADOW BANKING SYSTEM

ABS as repo collateral created the conditions for a banking panic. As long as the repo was collateralized by Treasury securities, lenders (depositors) didn't have to worry about the borrower's risk of default or about the value of the underlying collateral. But this changed when the repo was collateralized by ABS.

In 2007, house prices in the U.S. started to decline, raising concerns that homeowners could start defaulting on their mortgages in large numbers. In turn, lenders with repo collateralized by MBS started worrying about potential losses. What was the reason for their concern? After all, as I have argued, when carefully executed, securitization can generate a safe asset for investors, and indeed many MBS were built to be nearly riskless under normal conditions.

Usually, ABS are designed to be safe. ABS reduce credit risk in two ways: diversification and overcollateralization. For instance, pooling mortgages that had been originated in cities all over the U.S. is one way to create diversification. Under normal circumstances, large numbers of homeowners in all regions of the country are very unlikely to default on their mortgages at the same time. Overcollateralization simply involves pooling enough mortgages to guarantee that it can generate enough cash flow to make the promised payments to investors even if some of the borrowers default. The amount of overcollateralization required to make an MBS safe usually depends on certain fundamental market indicators, including the trend in house prices. Significantly, the statistical models used to design, price, and provide credit ratings for MBS estimated default rates based on data collected during periods of generally rising house prices and during periods when housing price declines were localized.[16]

Bad news arrived. Now consider a scenario in which investors expect house prices to rise and, contrary to their expectations, house prices begin to fall, and keep falling. That is what happened in the U.S. in 2007. When house prices fell for several consecu-

[15] In this article, I emphasize avoiding regulatory taxes as a motivation for securitization. See Ronel Elul's article for an account of the efficiency benefits of securitization.

[16] Indeed, Christopher Foote, Kristopher Gerardi, and Paul Willen have documented that people had overly optimistic beliefs about house prices.

tive months, an increasing number of investors believed that the average rate of default on any given pool of mortgages was going to rise. Their fears became more concrete when in the summer of 2007, two hedge funds sponsored by Bear Stearns that had invested heavily in subprime mortgages filed for bankruptcy and BNP Paribas suspended withdrawals from three money market mutual funds that were exposed to subprime mortgages. An important indicator of their fears was that the ABX index, a measure of the risk of default on subprime MBS, began to rise. This raised concerns that many SPVs were not holding enough collateral to generate sufficient cash flow to make good on the promised payments to investors.

Another reason to have doubts about the true value of MBS was that many investors did not know where the risks were concentrated. Although many MBS were wisely built to be nearly riskless, several classes of MBS contained a disproportionate fraction of mortgages that had been extended to people of dubious creditworthiness. And the risk of these subprime mortgages was particularly sensitive to the decline in housing prices.

Repo lenders ran on repo borrowers, including healthy borrowers. A depositor with serious doubts about the underlying value of the collateral can do two things: either ask for more collateral or simply not renew the repo. Both actions can be interpreted as a decision to withdraw funds from the shadow banking system, much like the decision bank depositors make to withdraw funds from their bank when they believe they might not be able to get all their money out.

Repo lenders initially asked for more collateral, but ultimately they simply refused to renew their loans. In other words, the repo market froze.[17] Because investors could not tell safe MBS from risky MBS in most cases, they withdrew their funds even from shadow banks that probably had safe MBS to secure repos. This problem was severe enough to turn the initial panic into a systemic event — a banking crisis.

Thus, the financial crisis was not very different from the banking crises of old. Investors in the repo market behaved pretty much like bank depositors did during U.S. banking crises before 1933. And the outcome was certainly very similar. The initial banking crisis spread to other financial markets, and several financial firms either failed or had to be rescued by the federal government to prevent further failures.

A caveat. Gorton and Metrick's explanation for the events that sparked the 2007-08 financial crisis depends on the claim that the fraction of the repo market that used ABS as collateral was large enough to generate a systemic event. But this claim has been a source of controversy among financial economists. For instance, Arvind Krishnamurthy, Dmitry Orlov, and Stefan Nagel have argued that a relatively small share of repo transactions in which money market mutual funds and securities lenders were the repo lenders was collateralized by ABS prior to the 2007-08 crisis. However, these authors focus on a relatively small segment of the repo market, the triparty repo market, while Gorton and Metrick study the larger bilateral repo market, for which there is as yet no direct evidence about the collateral used in transactions.[18] Furthermore, Krishnamurthy and coauthors note that while the share of the transactions collateralized by ABS was modest, such transactions were more concentrated among a small number of large banks that experienced significant problems. So focusing on average shares may be misleading. Nonetheless, the details of Gorton and Metrick's account of developments in the repo market will remain a source of controversy until researchers can collect more complete data. Moreover, some evidence suggests that the financial crisis was actually triggered in another part of the shadow banking system. See the accompanying discussion, *Crisis in the ABCP Market*.

SHADOW BANKING PANIC MAY HAVE DEEPENED RECESSION

It is still too early to fully disentangle the relative importance of the various factors that led to a particularly deep recession and a particularly slow recovery. But like many earlier recessions associated with banking crises, the crisis in the shadow banking system may have played a significant role in the depth of the downturn and the slow recovery.

The crisis in the shadow banking system has significantly reduced the ability of commercial banks to originate and renew loans, creating ongoing problems for households and firms that rely on bank loans. Some economists have even argued that the effects of the collapse can persist for an extended period. For instance, Viral Acharya has argued that traditional lenders cannot easily fill the role that shadow banks had played in providing credit to the economy. This void has certainly contributed to the delay in restoring the flow of credit to a volume consistent with that of a recovery from a typical recession that had not been accompanied by a banking crisis.

The shadow banking system has not fully recovered from the financial crisis. Even though it has continued to operate with government support, it is

[17] See also Yaron Leitner's article explaining why markets freeze. See also Benjamin Lester's article for a discussion of regulatory interventions in response to market freezes.

[18] Triparty repo transactions take place between two counterparties intermediated by a dealer bank. Bilateral repo transactions occur between two counterparties without an intermediary.

unclear whether the volume of operations will return to that observed prior to the crisis anytime soon, or whether it should. Since the crisis, the government-sponsored enterprises Fannie Mae and Freddie Mac have carried out nearly all securitizations in housing markets. At this point it is unclear whether the private sector will ever play the same role in the creation of securitized assets that it had before the crisis.

CONCLUSION

One lesson of the financial crisis is that institutions quite similar to banks tend to rise up outside of regulatory purview. This is an important matter because this shadow banking system is fragile and subject to panics. And banking panics — regardless of where they occur — have pernicious economic repercussions. This potential for economic harm had led some economists before the crisis to propose tighter regulation of the shadow banking system. In the aftermath, policymakers were working to write new rules.

Crisis in the ABCP Market

Some economists have argued that problems in another segment of the shadow banking system can be identified as the prime cause of the 2007-08 financial crisis. For example, in his discussion of Gorton and Metrick's account of the crisis, Andrei Shleifer has provided evidence that the contraction in the asset-backed commercial paper (ABCP) market happened before the contraction in the repo market. Thus, he suggests that problems in the ABCP market may have triggered the financial crisis.

Commercial paper is a short-term debt instrument that both financial and nonfinancial firms use to finance ongoing operations. Financial firms issue commercial paper to fund a wide array of activities, including the purchase of long-term securities such as MBS. One form of funding through the issuance of commercial paper that has increased significantly in the last 20 years is ABCP. A financial firm can set up an SPV to purchase a portfolio of securities by issuing commercial paper on the capital markets. ABCP maturities can vary from one day (as in a typical repo transaction) to 90 days. The typical maturity of ABCP is 30 days. Again, we have something that looks like a bank, but it operates outside the regulatory system.

The main investors in ABCP are money market mutual funds. Similar to the investors in the repo market, money market mutual funds also need a convenient place to invest some of their resources for short periods. These investors also want their investments to be safe and to yield an attractive return. Provided that the assets securing ABCP are of sufficiently high quality, such an investment vehicle is fairly safe, at least under normal market conditions. The short-term duration of ABCP gives investors an opportunity to "withdraw" their funds in case they decide to invest elsewhere or in case they have doubts about the quality of the assets securing ABCP.

The issuers of ABCP are SPVs that are sponsored by large financial institutions, including traditional commercial banks.[a] The SPVs allow these institutions to fund a wide array of securities at any moment. The short duration of ABCP means that an SPV has to roll over its debt every time an ABCP matures.

Many SPVs used the proceeds from the sale of ABCP to invest in MBS (i.e., the collateral backing ABCP were MBS). As we have seen, the perception of MBS as a safe debt instrument can suddenly change once the trend in house prices becomes clearly downward. Starting in the summer of 2007, many investors stopped refinancing maturing ABCP because of potential exposure to subprime mortgages via MBS. A full-scale panic ensued as the spread on overnight ABCP over the federal funds interest rate (the rate of interest on unsecured loans in the interbank market in the U.S.) increased from 10 basis points to 150 basis points. The outstanding amount of ABCP shrank steadily after the summer of 2007, despite several interventions by the Federal Reserve System in the form of liquidity facilities, offering short-term credit to banks to refinance maturing ABCP.

The ABCP market also provides another example of financial transactions carried out outside the oversight of regulators that are very similar to traditional banking. Thus, a closer look at the crisis in the ABCP market has also demonstrated that it was not very different from previous banking crises.

Perhaps the most balanced view is that while the financial crisis began in the shadow banking system, it had many epicenters. Furthermore, the structural similarities among many of the institutions in the shadow banking system — illiquid assets funded by short-term liabilities — and the trigger for the crisis — the decline in housing prices — tell much the same story.

[a] A sponsor financial institution usually provides credit guarantees to the SPV. For a detailed description of the ABCP market, see the paper by Marcin Kacperczyk and Philipp Schnabl.

REFERENCES

Acharya, Viral. "Understanding Financial Crises: Theory and Evidence from the Crisis of 2007-2008," *NBER Reporter* (April 2013).

Adrian, Tobias, and Adam B. Ashcraft. "Shadow Banking: A Review of the Literature," Federal Reserve Bank of New York Staff Report 580 (October 2012).

Bernanke, Ben. "Nonmonetary Effects of the Financial Crisis in the Propagation of the Great Depression," *American Economic Review*, 73 (1983), pp. 257-276.

Bordo, Michael, Barry Eichengreen, Daniela Klingebiel, and Maria Soledad Martinez-Peria. "Is the Crisis Problem Growing More Severe?" *Economic Policy*, 16 (2001), pp. 51-82.

Elul, Ronel. "The Economics of Asset Securitization," Federal Reserve Bank of Philadelphia *Business Review* (Third Quarter 2005).

Federal Reserve Bank of San Francisco, "Who Are the Largest Holders of U.S. Public Debt?" http://www.frbsf.org/education/publications/doctor-econ/2005/july/public-national-debt.

Foote, Christopher, Kristopher Gerardi, and Paul Willen. "Why Did So Many People Make So Many Ex Post Bad Decisions? The Causes of the Foreclosure Crisis," Public Policy Discussion Papers 12:2, Federal Reserve Bank of Boston (2012).

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

Gorton, Gary B. *Slapped by the Invisible Hand: The Panic of 2007.* New York: Oxford University Press, 2010.

Gorton, Gary B., and Andrew Metrick. "Securitized Banking and the Run on Repo," *Journal of Financial Economics*, 104 (2012), pp. 425-451.

Gorton, Gary B., and Nicholas S. Souleles. "Special Purpose Vehicles and Securitization," in Rene Stulz and Mark Carey, eds., *The Risks of Financial Institutions.* University of Chicago Press, 2006.

Kacperczyk, Marcin, and Philipp Schnabl. "When Safe Proved Risky: Commercial Paper During the Financial Crisis of 2007-2009," *Journal of Economic Perspectives*, 24 (2010), pp. 29-50.

Krishnamurthy, Arvind, Stefan Nagel, and Dmitry Orlov. "Sizing Up Repo," NBER Working Paper 17768 (2012).

Leitner, Yaron. "Why Do Markets Freeze?" Federal Reserve Bank of Philadelphia *Business Review* (Second Quarter 2011).

Lester, Benjamin. "Breaking the Ice: Government Interventions in Frozen Markets," Federal Reserve Bank of Philadelphia *Business Review* (Fourth Quarter 2013).

Lucas, Robert E., Jr. "Inflation and Welfare," *Econometrica*, 68 (2000), pp. 247-274.

Monnet, Cyril. "Rehypothecation," Federal Reserve Bank of Philadelphia *Business Review* (Fourth Quarter 2011).

Sanches, Daniel. "The Optimum Quantity of Money," Federal Reserve Bank of Philadelphia *Business Review* (Fourth Quarter 2012).

Schleifer, Andrei. "Comments on Gorton and Metrick: Regulating the Shadow Banking System," Brookings Papers on Economic Activity, 2 (2010), pp. 298-303.

Forecast Disagreement in the Survey of Professional Forecasters

BY KEITH SILL

Many people engaged in activities related to business, financial markets, and policymaking closely follow economic forecasts. Our interest in forecasts stems from the fact that, to an important degree, the decisions we make today are influenced by our expectations about the economy. Accurate forecasts lead to better decision-making and more efficient use of economic resources, and so there is a clear benefit to identifying good forecasts.

An important resource for evaluating the predictions and performance of professional forecasters is the Survey of Professional Forecasters, conducted by the Philadelphia Fed Research Department's Real-Time Data Research Center. The SPF is a quarterly survey that asks a panel of professional forecasters about their projections for a range of economic variables, including output growth, unemployment, inflation, and interest rates. When examining the SPF data, it becomes clear that professional forecasters have wide-ranging views about the future evolution of the economy. This is perhaps a bit surprising, since the statistical methods that underlie good forecasting models are well known, and professional forecasters by and large have access to the same data on the economy's past performance.

With forecasters having similar tools and data to work with, why do we observe this wide dispersion in their projections?[1] Are expectations wide-ranging because of differences in models and methods used to make the forecasts? Or does the wide disagreement stem from how different forecasters process and analyze information and then use it as an input into their forecast-generation process? To design and implement effective economic policies, it is important to understand how expectations are formed. One way to do so is to study forecast disagreement. In this article we will examine some features of the forecasts that underlie the SPF and discuss what theories and evidence tell us about forecaster behavior and how expectations about the economy are formed and evolve over time.

[1] For a discussion on measuring the accuracy of the survey's forecasts, which is beyond the scope of this article, see Stark (2010).

Keith Sill is a vice president and director of the Real-Time Data Research Center at the Federal Reserve Bank of Philadelphia. The views expressed in this article are not necessarily those of the Federal Reserve. This article and other Philadelphia Fed reports and research are available at www.philadelphiafed.org/research-and-data/publications.

THE SURVEY'S DESIGN

The SPF is the oldest quarterly survey of macroeconomic forecasts in the United States, having been initiated in 1968 under the leadership of Victor Zarnowitz at the American Statistical Association and the National Bureau of Economic Research. After conducting what was then known as the ASA-NBER Quarterly Economic Outlook Survey for 22 years, the ASA-NBER turned the survey over to the Federal Reserve Bank of Philadelphia in 1990, at which time it was renamed the Survey of Professional Forecasters. The Philadelphia Fed's Real-Time Data Research Center now conducts the SPF. A panel of professional forecasters (there are usually around 45 respondents per survey) is asked to give projections for a range of major macroeconomic variables over various time horizons.[2]

To maintain high quality, the SPF screens its participants. Most have had advanced training in economic theory and statistics and use statistical models to generate their projections. To keep the integrity of the survey high, participation is limited to those employed by firms or paid by clients to generate forecasts now or in the past. Because of these criteria and the types of individuals who participate in the SPF, we

[2] The survey results are released to the public free of charge at 10 a.m. on the second or third Friday of the second month of each quarter. The release schedule and the results of current and past surveys, as well as the underlying data, including anonymized individual forecaster projections, are available at http://philadelphiafed.org/research-and-data/real-time-center/survey-of-professional-forecasters.

can surmise that fairly sophisticated models and statistical methods underlie their projections.

SPF participants use their models to forecast quarterly values of major macroeconomic variables for up to five quarters, including the current quarter, and annual projections up to three years ahead. In addition, the SPF asks for long-term annual averages for headline and core inflation, real GDP growth, productivity growth, and stock and bond returns. A somewhat unusual feature of the survey is that, instead of asking participants just for single forecasts for output growth, inflation, and the unemployment rate, it asks them to assign probabilities to different outcomes and so gives a more comprehensive picture of the forecasters' views of the future.[3]

SPF FORECAST DISAGREEMENT

For the most part, the main, or "headline," forecast numbers reported in the SPF are the median values across forecasters. Each median, though, belies the variation that exists among individual projections for key variables that describe the macroeconomy. In fact, the range of forecast values underlying the median can be substantial, and it changes over time. At times, the forecasters show more agreement and at other times more disagreement in their projections.

The Real-Time Data Research Center website provides data on SPF forecast disagreement for the variables that are regularly reported in the SPF. The measure of disagreement that is reported is the difference between the 75th percentile and the 25th percentile of the forecasts, which is called the interquartile range. In other words, suppose there are 100 separate forecasts for annual real GDP growth in 2014. Order the forecasts from highest value to lowest value, and take the difference between the 75th slot and the 25th slot as the measure of disagreement. We measure disagreement in this way in order to ensure that any outliers among the forecasts, perhaps due to mistaken entries in the respondent questionnaires, do not unduly influence the measure of disagreement. Figures 1 through 3 show plots of disagreement measured by the interquartile range from the center's website for real GDP growth, GDP price index inflation, and the unemployment rate. Each measure of disagreement is for the four-quarters-ahead forecast as of the date on the horizontal axis.

The range of forecast values underlying the median can be substantial, and it changes over time.

The charts show that disagreement generally tended to be higher in the survey's early years — the late 1970s and early 1980s — compared with the latter half of the sample. Broadly speaking, this pattern of declining disagreement tracks the period known as the Great Moderation from 1984 to 2008, when the overall volatility of the economic data was lower than in the pre-1984 period.[4]

How do we interpret the data in Figures 1 through 3? Take the case of disagreement for real GDP growth. Since the early 1990s, the disagreement for forecasts of real GDP growth four quarters ahead has bounced around in a range of 0.5 percentage point to 1.5 percentage points, with an average of 0.86 percentage point. Roughly speaking, this suggests that about 50 percent of the forecasts fall within a range of about 0.4 percentage point below to about 0.4 percentage point above the median forecast. The other 50 percent of the forecasts are even further away from the median. Consequently, the disagreement among the forecasters seems not too large but nonetheless represents a significant difference between the top and bottom of the distribution. By way of comparison, the standard deviation of quarterly real GDP growth from the first quarter of 1991 to the third quarter of 2013 was about 2.5 percentage points at an annual rate.

Recall, though, that the measure of disagreement shown in Figures 1 through 3 is somewhat conservative. To calculate it, we make no use of the forecasts in the top and bottom 25 percent of the distribution — which, if included, would widen the disagreement. Indeed, this potentially wide disagreement is part of the reason that the SPF generally reports median rather than average forecasts. The median is the midmost forecast when forecasts are ranked from high to low. So, unlike with the average forecast, the effect of outliers is discounted.

If we use all the forecasts to calculate the standard deviation across projections of four-quarters-ahead real GDP growth, we obtain Figure 4. For the most part, the standard deviation measure tracks the interquartile range measure fairly closely, though it is clearly more volatile, especially early in the sample. This volatility may partly reflect reporting errors by members of the forecast panel. Sometimes an SPF respondent will submit

[3] See my 2012 article for more on forecast uncertainty and the forecast probabilities that are reported in the SPF. That article also presents some evidence on how forecast disagreement affects the macroeconomy.

[4] For more on the Great Moderation, see my 2004 article, "What Accounts for the Postwar Decline in Economic Volatility?"

FIGURES 1–3

Significant Dispersion for Key Indicators

Figure 1: Real GDP Growth Rate Dispersion by Quarter

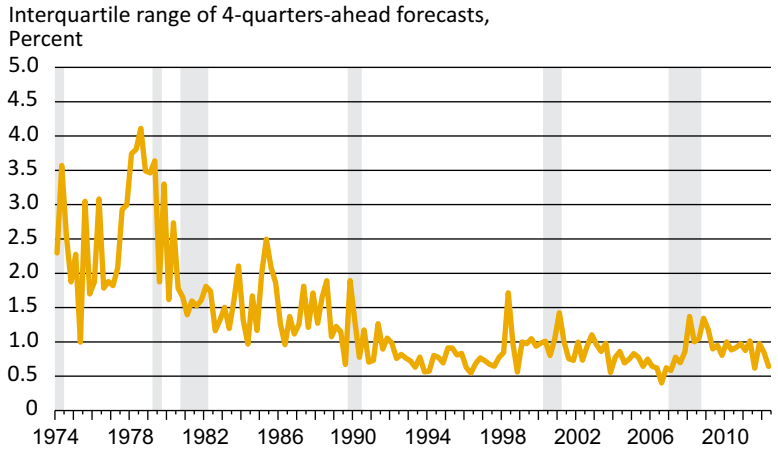


Figure 2: Unemployment Rate Dispersion by Quarter

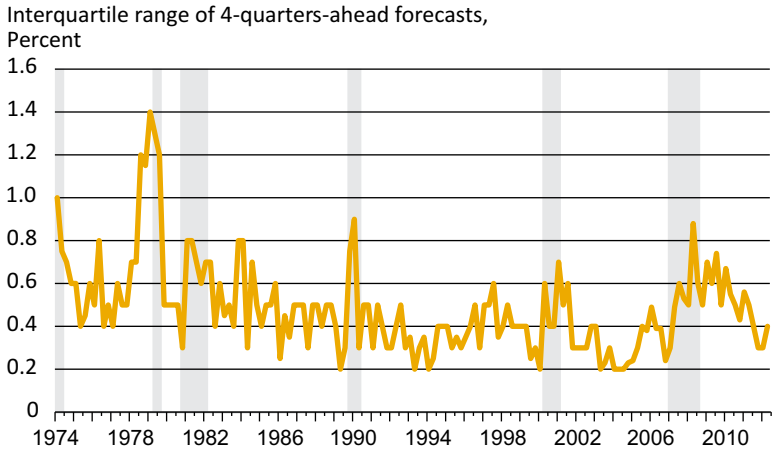
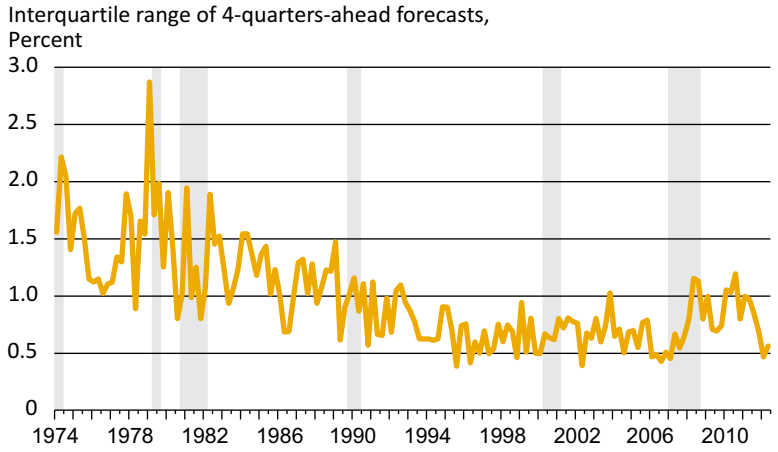


Figure 3: GDP Deflator Inflation Rate Dispersion by Quarter



Source: Survey of Professional Forecasters, Federal Reserve Bank of Philadelphia.

a forecast that seems extreme relative to those of the other respondents. If the outlier appears to be an error, that forecast is removed. In general, though, it is difficult to identify reporting errors versus actual views. For example, in the survey for the first quarter of 2013, the interquartile range for the four-quarters-ahead real GDP growth projections is 0.6, while the standard deviation is 1.7. However, one forecaster had entered a four-quarters-ahead real GDP growth forecast of 12.6 percent — which might have been a reporting error. If we exclude that forecast, the standard deviation falls to 0.832, which is much closer to the interquartile range. The median and the interquartile range are less affected by such outliers.

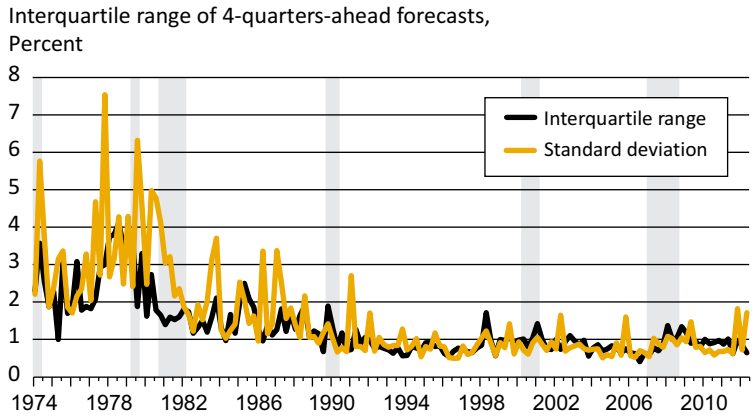
We can see this in Figure 5, which plots the forecasts for four-quarters-ahead real GDP growth from the SPF for the first quarter of 2013. The gold dots are observations in the upper 25 percent and lower 25 percent tails of the distribution. We see two outliers, one calling for 12.6 percent growth and one calling for a 1.5 percent contraction. We cannot say for sure that these were reporting errors, but it seems possible. Excluding those two, the remainder fall in a range of about 2 percent to 4.2 percent, while the central tendency ranges from 2.5 percent to 3.1 percent.

Disagreement still significant. If we use all the forecasts to calculate the standard deviation, we can construct confidence intervals for the forecasters. A confidence interval indicates the probability of a forecast falling within a certain range. Typically, a 95 percent confidence interval is plus or minus two standard deviations around the mean estimate.[5] In the case of forecasts for real GDP growth four quarters

[5] This is the case if the observations are drawn from a normal distribution.

FIGURE 4

Standard Deviation Generally Tracks GDP Forecast Range



Sources: Survey of Professional Forecasters, Federal Reserve Bank of Philadelphia; author's calculations.

ahead, this calculation suggests that we can estimate that 95 percent of the forecasts will, on average, be in a range of about 1.7 percentage points above to 1.7 percentage points below the mean of the forecasts (the average standard deviation of the four-quarters-ahead projections from the first quarter of 1992 to the first quarter of 2013 is 0.85). So, if the average forecast was 3 percent, we would estimate that 95 percent of the forecasts for four-quarters-ahead real GDP growth would fall in a range of about 1.3 percent to 4.7 percent. This range highlights that the SPF forecasters typically have fairly divergent views on how real output growth is likely to evolve in the not-too-distant future.

When looking at the interquartile ranges of the forecasts for the unemployment rate and GDP deflator inflation, we also see a tendency toward a decline in average disagreement in the post-1990 sample. As it had for real GDP growth, disagreement among these forecasts increased following the recession that began in December 2007. Disagreement for the unemployment rate forecast is a bit lower than that for inflation or output growth, possibly because the high persistence in the unemployment rate makes it a somewhat easier variable to forecast. Nevertheless, the range of views on future unemployment rates is significant. Take the results of the survey taken in the first quarter of 2013. Figure 6 plots the forecasts for the four-quarters-ahead unemployment rate from high to low. The gold dots again denote the upper and lower quartiles of the distribution. We see that while the interquartile range was fairly small at around 0.5, the range of overall forecasts was larger. Some forecasters thought that the economy would make little if any progress on the unemployment rate (the median current-quarter forecast for the 2013 first quarter unemployment rate was 7.8 percent). But some thought unemployment would fall below 7 percent. Most thought it would be in a range of 7.3 percent to 7.6 percent.

Figure 7 shows the forecasts for GDP deflator inflation from the 2013 first quarter SPF. Again, two look suspicious — both calling for significant deflation — and would have a large impact on the standard deviation of the forecasts. Excluding the deflation outliers, the forecasts range from about 1.3 percent for inflation four quarters ahead to almost 4 percent. The central tendency is narrow at 1.7 percent to 2.3 percent. But we again see that professional forecasters can have strikingly different views about how the economy will evolve over the next 12 months.

DIFFERENT MODELS AND METHODS

The forecasters who make up the SPF panel use a variety of statistical models to help them make their projections, and this variety of models surely plays a role in forecast disagreement. But how large a role might that be? Their models generally fall into one of two major categories: reduced-form models and structural models. Reduced-form models impose little, if any, economic theory to refine their structure. For example, one of the simplest forecasting models for real GDP growth is to suppose that current GDP growth is related to past GDP growth in a linear fashion. To forecast a greater range of variables, the model can be expanded by adding more lagged variables to form a system of equations that relates current values of variables such as output growth, inflation, and interest rates to their lagged values. A forecaster using such a system of equations chooses which variables to have in the system as well as the number of lags of variables to use. Once those choices are made, the model can be estimated using historical data and then used to generate forecasts. One does not need to bring much economic theory to bear when specifying such a reduced-form model, since there are typically no restrictions on the esti-

FIGURES 5–7

Strikingly Different Year-Ahead Projections

Figure 5: Real GDP Growth Rate Forecasts

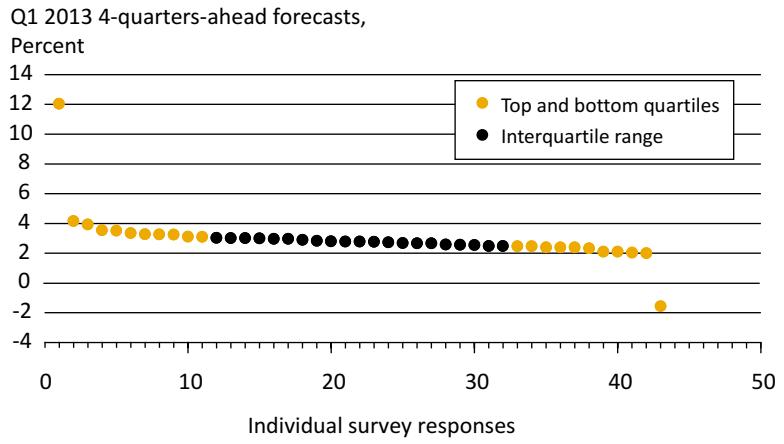


Figure 6: Unemployment Rate Forecasts

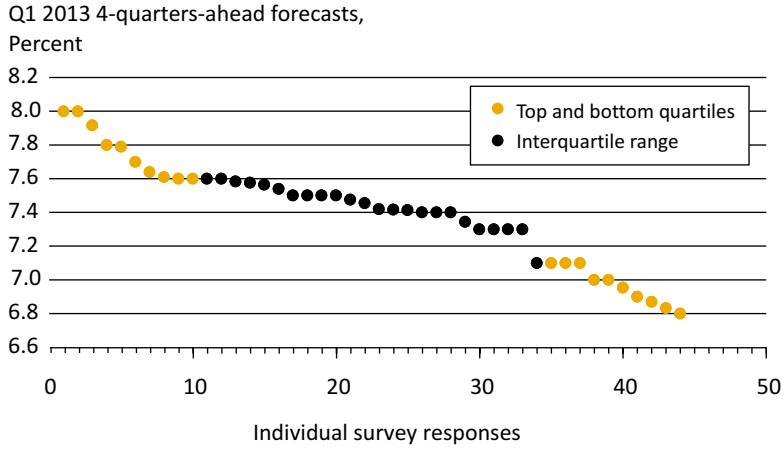
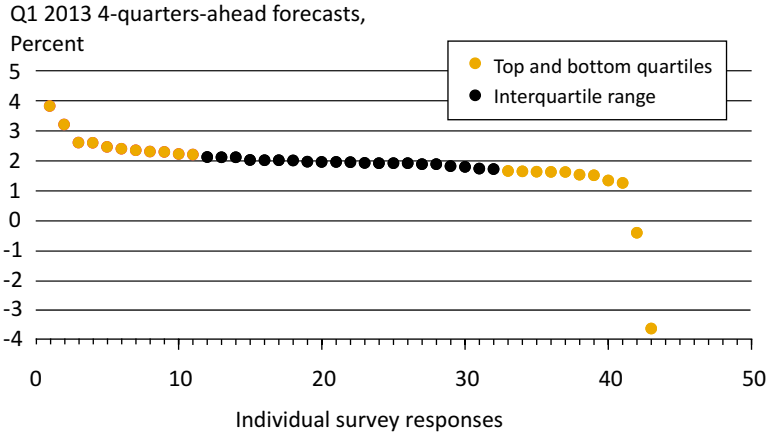


Figure 7: GDP Deflator Inflation Rate Forecasts



Source: Survey of Professional Forecasters, Federal Reserve Bank of Philadelphia.

mated coefficients and no restrictions on how the model's variables interact with each other.

Another approach to building a forecast model is to use economic theory to restrict the way in which the variables interact. For example, one might stipulate that the relationship between household consumption and hours worked in the labor market is related to the real wage in some particular way, or that the relationship between current and future consumption is tied in a specific way to the interest rate. Models that impose economic theory on the data's interrelationships are called structural models. Like reduced-form models, structural models relate values of variables to their lagged values. But structural models use economic theory to impose complex relationships among those variables. Structural models are especially useful for honing the story behind the forecast. For example, such models may indicate that output growth will rise because of higher demand today, or that inflation is expected to fall because firms expect the marginal costs of production to fall in the future. These kinds of stories that are consistent with economic theory are typically more difficult to tease out when using reduced-form models.

There is another element that usually is an important part of the forecast process: judgment. Often forecasters will examine the historical errors in the equations of their models — that is, how much the predicted value from the equation differed from that actual value in the data. If the forecast is persistently missing on the high or low side, the forecaster may alter the equation away from the estimated values a bit so that its predictions are more in line with the most recent data observations. This is a judgment-based adjustment of the forecast whereby forecasters subjectively alter the predictions generated by the statistical model to

bring them more into conformity with the recent behavior of the economy and their own views on how the future is likely to unfold. Typically, data, models, and judgment are combined to produce the final forecast.

If forecasters are using different models on which to base their forecasts, then we might expect to see disagreement in those forecasts. Some models might be more accurate descriptions of the economy than others. If one forecasting methodology is consistently better than another and the data are able to discriminate among the models, we should see bad models being driven out over time by good models. Likewise, we should see that forecasters using the best models should consistently produce better forecasts than other forecasters do. That is, if model heterogeneity is the most important reason that forecasts differ and the data are informative about the models, then we should be able to identify forecasters and models that reliably outperform their peers.

Evidence casts doubt. However, some evidence from the forecast evaluation literature casts some doubt that model heterogeneity is the key element behind forecast disagreement. First, one of the most robust findings from the forecasting literature is that mean forecasts systematically outperform individual forecasts. That is to say that over time, a more accurate forecast can be had by taking the average of many different forecasts rather than sticking with one individual forecaster's projections. If one forecaster and his or her model were consistently producing better forecasts, then we wouldn't expect to see such a gain from averaging.[6]

Why does this forecast averaging tend to work so well compared with any one forecast over time? If there was one, known, true model of the economy, then the forecasts from that model would dominate alternative models. But we don't have the true model of the economy. Yet, because the economy is so complex, different models may capture different features of the economy in a successful way. At certain times, some of those features may be more important for successful forecasts than at other times. So by averaging a wide range of forecasts, we can incorporate many features of the economy that are impossible to capture in any single model and on average make more accurate forecasts.

By averaging a wide range of forecasts, we can incorporate many features of the economy that are impossible to capture in any single model and on average make more accurate forecasts.

Aside from forecast averaging, is there other evidence that suggests model differences may not be the key factor underlying forecast dispersion? Models can differ in how they incorporate economic shocks, which are defined as unpredictable disturbances to the economy from events such as the outbreak of war or an unexpected surge in global commodity prices. For example, some models might be better at predicting how the economy will respond to oil price shocks, while other models might be good at predicting how the economy will respond to fiscal policy shocks. Therefore, depending on the specific mix of shocks hitting the economy at any one time, some models may produce the most accurate forecasts for one period, only to have their relative predictive power reversed when a different set of shocks hits. However, shocks and their effects tend to persist. That is, the impact tends to decline slowly as the shocks fully work their way through the economy. This persistence implies that models that are especially accurate for particular shocks should generate persistently better forecasts as long as the shocks a particular model is good at analyzing are still important drivers of the dynamics of the economy. We might then expect that some forecasters will give more accurate forecasts for several reporting periods in a row. However, another finding from the empirical forecasting literature is that the best forecaster in any one period is no more likely to be the best in the next period.[7] So, although economic shocks are persistent, forecast performance is not.

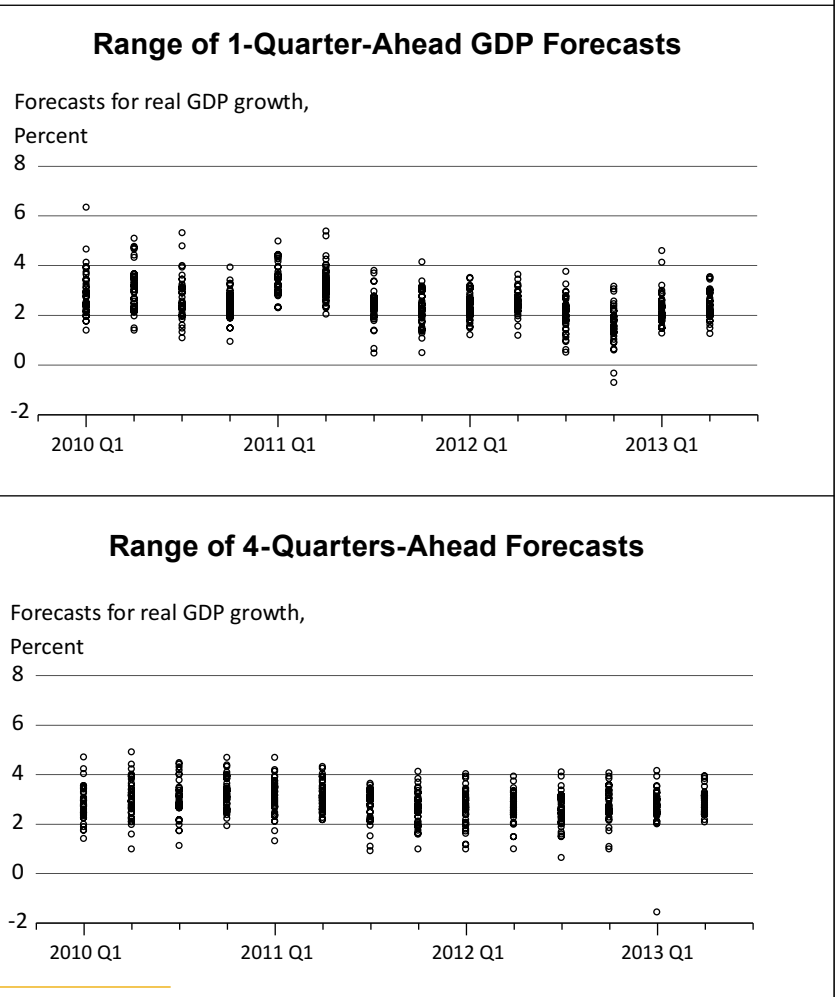
The findings that average forecasts tend to outperform individual forecasts and that top forecasters don't stay on top for long suggest that differences in models may not be the most important element behind forecast disagreement. But as is often true in economics, the case is not so clear-cut. Recent research by Andrew Patton and Allan Timmermann examines how forecast disagreement changes with the forecast horizon. Using survey data from Consensus Economics, they find greater disagreement among longer-term forecasts than near-term forecasts. Because variables such as real GDP growth and inflation tend to return to their mean values over time, the observation that long-horizon forecasts show more disagreement is consistent with the idea that differences in economic models are an important factor. This is because different models might be calibrated to yield different long-run averages for key variables, and model-based long-run forecasts will

[6] See the 2006 article and references therein by Allan Timmermann.

[7] See the 2007 article by Michael Bryan and Linsey Molloy.

FIGURE 8

No Wider Variation in Longer-Term Forecasts



Source: Survey of Professional Forecasters, Federal Reserve Bank of Philadelphia.

Note: In the upper chart, one outlier forecast from the 2012Q3 survey was eliminated. In the lower chart, one outlier from each of the following surveys was eliminated: 2010Q1, 2012Q3, and 2013Q1.

reflect those differences.

This pattern is more difficult to discern in recent SPF projections for real GDP growth. However, the longest forecasts we can use from the SPF are looking only four quarters ahead from the quarter in which the survey is conducted, which is more in the realm of near-term forecasts. With that caveat, Figure 8 plots individual forecasts of one-quarter-ahead and four-quarters-ahead real GDP growth from the first quarter of 2010 to the second quarter of 2013. Looking at the panels in the figure, there is no clear tendency for the longer-horizon forecasts to show more disagreement than the shorter-horizon forecasts over this period.[8] If we were to extend the data sample back to 1970, we would see the same basic pattern: There is no obvious increase in disagreement when we move from one-quarter-ahead to four-quarters-ahead real GDP forecasts in the SPF.

In considering how near-term forecasts might differ from long-term forecasts, it is important to note that near-term forecast disagreement is more likely to be influenced by current information that is used to kick off the forecast. As noted earlier, a forecast combines models with data and judgment to generate a projection. The timeliness of the data that a forecaster has in hand when making a forecast and the extent to which the forecaster discerns the true state of the economy from that data are critical elements when projecting the economy's future state. Perhaps the data and information analysis that go into forecasts are also key drivers of forecast disagreement.[9]

AN IMPERFECT SIGNAL

As new data become available and are put into the forecasting models, the projections are modified, sometimes dramatically. So how forecasters respond to the arrival of new information is an important factor in the forecast-generation process. When constructing models of the economy, economists often assume for simplicity's sake that people costlessly receive all the information they need to make their decisions. In reality, though, information is costly to process and often subject to revision over time. Take the release of quarterly real GDP data. An initial estimate is released in the month after the end of the quarter, a second release two months after the end of the quarter, and a final release three months after the end of the quarter. But even that's not really the "final" release, since the estimate will again be revised in July for the next several years, and then again every five years or so with benchmark revisions

[8] Patton and Timmermann were able to compare forecast horizons as short as one month with those as long as 24 months. In addition, they used the Consensus Forecast survey for 1991 to 2008, which typically had about 25 forecasters in the panel.

[9] Note, though, that without some discipline, models with heterogeneous forecasting rules can rationalize any disagreement outcome.

to the national income and product accounts.[10] Forecasters looking at today's data to put into their models realize that they have only an imperfect signal of the true state of the economy at any point in time. In contrast to the case of full information, we can say that there are information "frictions" that make the true state of the economy difficult and costly to assess.

There are two prominent theories of imperfect information in macroeconomics that have different implications for what we should observe in forecast surveys such as the SPF. The first, often referred to as the sticky-information theory, is described in a 2002 paper by Greg Mankiw and Ricardo Reis. In this theory, economic agents such as forecasters are assumed to update the information they use to make decisions and forecasts randomly — with a certain probability each period that is independent of economic conditions or past decisions. It is as if each day people play an information lottery. If they win the lottery, they go ahead and update their view of the world based on current data. And the assumption is that they receive full information about the state of the economy when they update. If they lose, they don't update their information and continue to make decisions and forecasts based on stale data.

Clearly, this is an extreme view of the world and is unlikely to be strictly true. But if there is a fixed cost to acquiring and processing new information, then households and firms will update their information only infrequently.[11]

Another prominent theory of information frictions assumes instead that agents monitor the flow of data and update their information continually, but that the information they receive about economic fundamentals is contaminated by random "noise" that obscures the signals they are interested in. For example, take the case of a monetary policymaker who is concerned about the behavior of inflation when setting interest rates. At a point in time, the policymaker has observations on current and past inflation and tries to discern the trend in inflation from transitory movements that are likely to dissipate over time. Inflation may be higher today because of, say, a temporary weather shock that affects food prices. The policymaker is more likely concerned with the underlying trend rate of inflation but must make some inference about that unobserved trend from the underlying data. So, he or she doesn't necessarily possess the full information. More generally, time and resources must be spent to best estimate the desired information, and because time and resources are costly, there are tradeoffs in deciding how to process information.

Information frictions have implications for forecast behavior, including forecast disagreement. How do forecasters respond to economic shocks when there are information rigidities? In the case of sticky information, not all forecasters are updating their projections in response to the shock at the same time.[12] Since the respondents are surveyed at the same time, we would hypothesize that the average forecast will be somewhat inertial and slow to adjust to the shock. Consequently, forecast errors will persistently be above or below zero (depending on whether the shock is positive or negative) for some time, though eventually everyone updates his or her information and the average forecast error returns to zero.[13]

Similarly, in the case of imperfect information, the change in the average forecast in response to the shock is inertial — in this case only a fraction of the signal about economic fundamentals is incorporated into the current estimate of the underlying state of the economy so that adjustment to the data is only partial. This slow response is reflected in the persistence of forecasters' beliefs about the underlying state of the economy, which in turn leads to a somewhat inertial adjustment of the forecasts to shocks to economic fundamentals.

Effect of shocks on disagreement. So, both imperfect information theories suggest that forecast errors might be persistently below or above zero in response to shocks (but that the errors converge to zero over time). But what about forecast disagreement? Here, the two theories offer different predictions. The sticky-information theory predicts that disagreement will rise in response to shocks. This happens because not all forecasters are updating their information sets and forecasts after a shock, so the forecasts of those who do update may move further away from the forecasts of those who don't: Disagreement increases. The noisy-information theory predicts that disagreement should not respond to shocks. In this theory, forecasters con-

[10] The most recent benchmark revisions were conducted in the summer of 2013. According to the Bureau of Economic Analysis, comprehensive revisions encompass (1) updated definitions and classifications to more accurately portray the evolving U.S. economy, (2) changes in presentations to make the NIPA tables more informative, and (3) statistical changes that introduce improved methodologies and newly available and revised source data. See http://www.bea.gov/newsreleases/national/gdp/gdpnewsrelease.htm.

[11] That the presence of fixed costs of updating information can rationalize the sticky-information model is derived in a 2006 article by Ricardo Reis.

[12] It is plausible, for instance, that an economist employed by a bank or other firm to generate forecasts largely for internal use might update those forecasts less frequently than would a forecaster who primarily sells forecasts to a wide clientele.

[13] Key references for the macroeconomic implications of imperfect information literature include Lucas (1972), Sims (2003), and Woodford (2003).

tinually monitor and react to the flow of data.[14] The idea is easier to grasp if we assume that forecasters are using the same model. Then, if they are also monitoring the data continually, the disagreement among forecasts arises from idiosyncratic differences in how the forecasters process the information that ends up being fed into their models. For example, suppose the data for the model include the real interest rate, which is not observed directly but must instead be inferred by subtracting the expected rate of inflation from the nominal interest rate. Depending on how they measure expected inflation, different forecasters can arrive at different measures of the real interest rate. As long as the ways that individual forecasters process information do not themselves respond to shocks, then the dispersion of the forecasts will not vary in response to shocks.

To sum up then, both imperfect-information theories predict that forecast errors should show some persistence, but they have different implications for forecast disagreement. The sticky-information theory suggests that disagreement rises in response to shocks, while the noisy-information theory suggests that it doesn't.[15]

IS THERE EMPIRICAL EVIDENCE?

Is there empirical evidence on the role of imperfect information in forecaster behavior? Recent work by Olivier Coibion and Yuriy Gorodnichenko examines whether the implications of imperfect-information theories are found in the forecast data. They use a variety of surveys, including the SPF, to investigate how forecast errors and disagreement respond to shocks and whether that response can be rationalized by imperfect-information theories.

Assessing the response of economic variables to shocks can be tricky because shocks themselves are often unobserved and so have to be identified from the data. Coibion and Gorodnichenko identify several shocks, including monetary policy shocks, total factor productivity (TFP) shocks, oil shocks, and fiscal policy shocks, and then use regression methods to see how inflation forecast errors and forecast disagreement respond to these shocks.[16] For the most part, the forecast horizon under investigation is one year ahead.

Recall that one key prediction of the imperfect information theories is that forecast errors should show some persistence in response to shocks. If a shock hits the economy and forecasters either don't incorporate it quickly into their forecasts or else have a hard time extracting the relevant signals from the data, then their forecasts are likely to over- or under-shoot until these information problems are resolved. Indeed, Coibion and Gorodnichenko generally find serially correlated forecast errors in response to a variety of shocks, especially TFP and oil price shocks. For example, after an inflationary shock, they find a predictable sequence of serially correlated positive inflation forecast errors. Over time, though, these errors converge back to zero — just as the theory predicts they should. If there were no information problems confronting forecasters, we would expect that forecast errors would in turn not show predictable patterns in the data.

One key prediction of the imperfect information theories is that forecast errors should show some persistence in response to shocks.

Coibion and Gorodnichenko also examine how forecast disagreement responds to shocks. Recall that this response has the potential to distinguish between the sticky-information and noisy-information theories. After examining how disagreement changes in response to many different shocks that hit the economy over the past 30 years or so, they conclude that, on balance, structural shocks do not seem to notably increase disagreement across forecasters. This finding gives an edge to the noisy-information theory, though for other dimensions of the data, the sticky-information theory does better.[17]

[14] This assumes that the dispersion of the idiosyncratic noise shocks that forecasters receive does not respond to economic fundamentals such as inflation or output. However, if forecasters who receive the same signal interpret it differently, then forecast dispersion can be correlated with economic shocks. See Coibion and Gorodnichenko (2012) for details.

[15] The two theories have additional predictions for the data besides those we have discussed here. See the 2012 paper by Coibion and Gorodnichenko for a fuller exposition.

[16] Total factor productivity is the residual in accounting for economic output after the contributions of labor and capital inputs have been measured. It can be viewed as the contribution of technological change to output growth. A monetary policy shock can be thought of as the surprise component of the monetary policy instrument. It is the difference between a realized policy outcome — usually a short-term interest rate — and the rate that had been predicted by a specific model.

[17] Some of Coibion and Gorodnichenko's findings, though, are more consistent with the sticky-information theory. For example, the convergence rate of forecast errors is just as rapid for monetary policy shocks as it is for TFP shocks. Under the noisy-information theory, if TFP shocks were more important for determining productivity and economic growth, one might expect forecasters to pay more attention to these shocks, which implies that forecast error convergence would differ among shocks.

It seems that elements of both theories are present in the data.

This evidence indicates that imperfect-information theories are consistent with the forecast data, though discriminating between the two theories is difficult. Is imperfect information important or insignificant in the data? Coibion and Gorodnichenko also try to answer this question using data from the SPF on inflation forecasts. In the context of the sticky-information model, their estimates imply an average duration of six to seven months between information updates on the part of forecasters.[18] In the context of imperfect-information models, the estimates imply that new information receives less than half the weight it would if there was no imperfect information and forecasters did not have to extract the relevant signals. Thus, the evidence is consistent with imperfect information having a significant role in how forecasters process new information and how often they acquire it.

[18] It seems puzzling that professional forecasters would update their forecasts infrequently, but other researchers have also found patterns in the data consistent with this implication. Two such studies are Clements (2012) and Andrade and Le Bihan (2010).

CONCLUSION

Forecast disagreement is large and varies over time. Differences in how forecasters see the economy evolving in the future can be attributed to the models they use and the way they monitor and incorporate into their forecasts the heavy, continual flow of information on the state of the economy. Empirical evidence from forecast surveys suggests that both model heterogeneity and imperfect information about the economy play roles in the wide dispersion in professional forecasts.

REFERENCES

Andrade, Philippe, and Hervé Le Bihan. "Inattentive Professional Forecasters," Banque de France Working Paper 307 (2010).

Bryan, Michael F., and Linsey Molloy. "Mirror, Mirror, Who's the Best Forecaster of Them All?" Federal Reserve Bank of Cleveland *Economic Commentary* (2007).

Clements, Michael P. "Do Professional Forecasters Pay Attention to Data Releases?" *International Journal of Forecasting*, 28 (April-June 2012), pp. 297-308.

Coibion, Olivier, and Yuriy Gorodnichenko. "What Can Survey Forecasts Tell Us About Information Rigidities?" *Journal of Political Economy*, 120:1 (2012) pp. 116-159.

Coibion, Olivier, and Yuriy Gorodnichenko. "Information Rigidity and the Expectations Formation Process: A Simple Framework and New Facts," National Bureau of Economic Research Working Paper 16537 (2010).

Lucas, Robert E. "Expectations and the Neutrality of Money," *Journal of Economic Theory*, 4 (April 1972), pp. 103-124.

Mankiw, N. Gregory, and Ricardo Reis. "Sticky Information Versus Sticky Prices: A Proposal to Replace the New Keynesian Phillips Curve." *Quarterly Journal of Economics*, 117 (November 2002), pp. 1,295-1,328.

Patton, Andrew J., and Allan Timmermann. "Why Do Forecasters Disagree? Lessons from the Term Structure of Cross-Sectional Disagreement," *Journal of Monetary Economics*, 57:7 (2010), pp. 803-820.

Reis, Ricardo. "Inattentive Producers," *Review of Economic Studies*, 73 (July 2006), pp. 793-821.

Timmermann, Allan. "Forecast Combinations," in G. Elliott, C. Granger, and A. Timmermann, eds., *Handbook of Economic Forecasting*, 1:1 (2006), pp. 135-196.

Sill, Keith. "What Accounts for the Postwar Decline in Economic Volatility?" Federal Reserve Bank of Philadelphia *Business Review* (Fourth Quarter 2004).

Sill, Keith. "Measuring Economic Uncertainty Using the Survey of Professional Forecasters," Federal Reserve Bank of Philadelphia *Business Review* (Fourth Quarter 2012).

Sims, Christopher A. "Implications of Rational Inattention," *Journal of Monetary Economics*, 50 (April 2003), pp. 665-690.

Stark, Tom. "Realistic Evaluation of Real-Time Forecasts in the Survey of Professional Forecasters," Federal Reserve Bank of Philadelphia *Special Report* (May 2010).

Woodford, Michael. "Imperfect Common Knowledge and the Effects of Monetary Policy," in P. Aghion, R. Frydman, J. Stiglitz, and M. Woodford, eds., *Knowledge, Information, and Expectations in Modern Macroeconomics in Honor of Edmund Phelps*, Princeton, N.J.: Princeton University Press (2003).

RESEARCH RAP

Abstracts of research papers produced by the economists at the Philadelphia Fed

Economists and visiting scholars at the Philadelphia Fed produce papers of interest to the professional researcher on banking, financial markets, economic forecasting, the housing market, consumer finance, the regional economy, and more. More abstracts may be found at www.philadelphiafed.org/research-and-data/publications/research-rap/. You can find their full working papers at http://www.philadelphiafed.org/research-and-data/publications/working-papers/.

The Continuing Power of the Yield Spread in Forecasting Recessions

The authors replicate the main results of Rudebusch and Williams (2009), who show that the use of the yield spread in a probit model can predict recessions better than the Survey of Professional Forecasters. Croushore and Marsten investigate the robustness of their results in several ways: extending the sample to include the 2007-09 recession, changing the starting date of the sample, changing the ending date of the sample, using rolling windows of data instead of just an expanding sample, and using alternative measures of the "actual" value of real output. The results show that the Rudebusch-Williams findings are robust in all dimensions.

Working Paper 14-5. Dean Croushore, University of Richmond and Federal Reserve Bank of Philadelphia Visiting Scholar; Katherine Marsten, University of Richmond.

Continuous Markov Equilibria with Quasi-Geometric Discounting

The authors prove that the standard quasi-geometric discounting model used in dynamic consumer theory and political economics does not possess continuous Markov perfect equilibria (MPE) if there is a strictly positive lower bound on wealth. The authors also show that, at points of discontinuity, the decision maker strictly prefers lotteries over the next period's assets. The authors then extend the standard model to have lotteries and establish the existence of an MPE with continuous decision rules. The models with and without lotteries are numerically compared, and it is shown that the model with lotteries behaves more in accord with economic intuition.

Working Paper 14-6. Satyajit Chatterjee, Federal Reserve Bank of Philadelphia; Burcu Eyigungor, Federal Reserve Bank of Philadelphia.

The Economics of Debt Collection: Enforcement of Consumer Credit Contracts

In the U.S., third-party debt collection agencies employ more than 140,000 people and recover more than $50 billion each year, mostly from consumers. Informational, legal, and other factors suggest that original creditors should have an advantage in collecting debts owed to them. Then, why does the debt collection industry exist and why is it so large? Explanations based on economies of scale or specialization cannot address many of the observed stylized facts. The authors develop an application of common agency theory that better explains those facts. The model explains how reliance on an unconcentrated industry of third-party debt collection agencies can implement an equilibrium with more intense collections activity than creditors would implement by themselves. The authors derive empirical implications for the nature of the debt collection market and the structure of the debt collection industry. A welfare analysis shows that, under certain conditions, an equilibrium in which creditors rely on third-party debt collectors can generate more credit supply and aggregate borrower surplus than an equilibrium where lenders collect debts owed to them on their own. There are, however,

situations where the opposite is true. The model also suggests a number of policy instruments that may improve the functioning of the collections market.

Working Paper 14-7. Viktar Fedaseyeu, Bocconi University and Federal Reserve Bank of Philadelphia Visiting Scholar; Robert M. Hunt, Federal Reserve Bank of Philadelphia.

Foreclosure Delay and Consumer Credit Performance

The deep housing market recession from 2008 through 2010 was characterized by a steep increase in the number of foreclosures. Foreclosure timelines — the length of time between initial mortgage delinquency and completion of foreclosure — also expanded significantly, averaging up to three years in some states. Most individuals undergoing foreclosure are experiencing serious financial stress. However, extended foreclosure timelines enable mortgage defaulters to live in their homes without making housing payments until the completion of the foreclosure process, thus providing a liquidity benefit. This paper tests whether the resulting liquidity was used to help cure nonmortgage credit delinquency. The authors find a significant relationship between longer foreclosure timelines and household performance on nonmortgage consumer credit during and after the foreclosure process. Their results indicate that a longer period of nonpayment of housing-related expenses results in higher cure rates on delinquent nonmortgage debts and improved household balance sheets. Foreclosure delay may have mitigated the impact of the economic downturn on credit card default. However, credit card performance may deteriorate in the future as the current foreclosure backlog is cleared and the affected households once again incur housing expenses.

Working Paper 14-8. Paul Calem, Federal Reserve Bank of Philadelphia; Julapa Jagtiani, Federal Reserve Bank of Philadelphia; William W. Lang, Federal Reserve Bank of Philadelphia.

Competing for Order Flow in OTC Markets

The authors develop a model of a two-sided asset market in which trades are intermediated by dealers and are bilateral. Dealers compete to attract order flow by posting the terms at which they execute trades, which can include prices, quantities, and execution times, and investors direct their orders toward dealers that offer the most attractive terms of trade. Equilibrium outcomes have the following properties. First, investors face a trade-off between trading costs and speeds of execution. Second, the asset market is endogenously segmented in the sense that investors with different asset valuations and different asset holdings will trade at different speeds and different costs. For example, under a Leontief technology to match investors and dealers, per unit trading costs decrease with the size of the trade, in accordance with the evidence from the market for corporate bonds. Third, dealers' implicit bargaining powers are endogenous and typically vary across sub-markets. Finally, the authors obtain a rich set of comparative statics both analytically, by studying a limiting economy where trading frictions are small, and numerically. For instance, the authors find that the relationship between trading costs and dealers' bargaining power can be hump-shaped.

Working Paper 14-9. Benjamin Lester, Federal Reserve Bank of Philadelphia; Guillaume Rocheteau, University of California–Irvine; Pierre-Olivier Weill, University of California–Los Angeles.

Forecasting Credit Card Portfolio Losses in the Great Recession: A Study in Model Risk

Credit card portfolios represent a significant component of the balance sheets of the largest US banks. The charge-off rate in this asset class increased drastically during the Great Recession. The recent economic downturn offers a unique opportunity to analyze the performance of credit risk models applied to credit card portfolios under conditions of economic stress. Specifically, the authors evaluate three potential sources of model risk: model specification, sample selection, and stress scenario selection. Their analysis indicates that model specifications that incorporate interactions between policy variables and core account characteristics generate the most accurate loss projections across risk segments. Models estimated over a time frame that includes a significant economic downturn are able to project levels of credit loss consistent with those experienced during the Great Recession. Models estimated over a time frame that does not include a significant economic downturn can severely underpredict credit loss in some cases, and the level of forecast error can be significantly impacted by model specification assumptions. Higher credit-score segments of the portfolio are proportionally more severely impacted by downturn economic conditions and model specification assumptions. The selection of the stress scenario can have a dramatic impact on projected loss.

Working Paper 14-10. José J. Canals-Cerdá, Federal Reserve Bank of Philadelphia; Sougata Kerr, Federal Reserve Bank of Philadelphia.

Misallocation, Informality, and Human Capital: Understanding the Role of Institutions

The aim of this paper is to quantify the role of formal-sector institutions in shaping the demand for human capital and the level of informality. The authors propose a firm dynamics model where firms face capital market imperfections and costs of operating in the formal sector. Formal firms have a larger set of production opportunities and the ability to employ skilled workers, but informal firms can

avoid the costs of formalization. These firm-level distortions give rise to endogenous formal and informal sectors and, more importantly, affect the demand for skilled workers. The model predicts that countries with a low degree of debt enforcement and high costs of formalization are characterized by relatively lower stocks of skilled workers, larger informal sectors, low allocative efficiency, and measured TFP. Moreover, the authors find that the interaction between entry costs and financial frictions (as opposed to the sum of their individual effects) is the main driver of these differences. This complementarity effect derives from the introduction of skilled workers, which prevents firms from substituting labor for capital and in turn moves them closer to the financial constraint.

Working Paper 14-11. Pablo N. D'Erasmo, University of Maryland and Federal Reserve Bank of Philadelphia; Hernan J. Moscoso Boedo, University of Virginia; Asli Senkal, University of Virginia.

Market Exposure and Endogenous Firm Volatility over the Business Cycle

The authors propose a theory of endogenous firm-level volatility over the business cycle based on endogenous market exposure. Firms that reach a larger number of markets diversify market-specific demand risk at a cost. The model is driven only by total factor productivity shocks and captures the business cycle properties of firm-level volatility. Using a panel of U.S. firms (Compustat), the authors empirically document the countercyclical nature of firm-level volatility. They then match this panel to Compustat's Segment data and the U.S. Census's Longitudinal Business Database (LBD) to show that, consistent with their model, measures of market reach are procyclical, and the countercyclicality of firm-level volatility is driven mostly by those firms that adjust the number of markets to which they are exposed. This finding is explained by the negative elasticity between various measures of market exposure and firm-level idiosyncratic volatility the authors uncover using Compustat, the LBD, and the Kauffman Firm Survey.

Working Paper 14-12. Ryan Decker, University of Maryland; Pablo N. D'Erasmo, University of Maryland and Federal Reserve Bank of Philadelphia; Hernan J. Moscoso Boedo, University of Virginia.

Capital Requirements in a Quantitative Model of Banking Industry Dynamics

The authors develop a model of banking industry dynamics to study the quantitative impact of capital requirements on bank risk taking, commercial bank failure, and market structure. They propose a market structure where big, dominant banks interact with small, competitive fringe banks. Banks accumulate securities like Treasury bills and undertake short-term borrowing when there are cash flow shortfalls. A nontrivial size distribution of banks arises out of endogenous entry and exit, as well as banks' buffer stocks of securities. The authors test the model using business cycle properties and the bank lending channel across banks of different sizes studied by Kashyap and Stein (2000). They find that a rise in capital requirements from 4% to 6% leads to a substantial reduction in exit rates of small banks and a more concentrated industry. Aggregate loan supply falls and interest rates rise by 50 basis points. The lower exit rate causes the tax/output rate necessary to fund deposit insurance to drop in half. Higher interest rates, however, induce higher loan delinquencies as well as a lower level of intermediated output.

Working Paper 14-13. Dean Corbae, University of Wisconsin–Madison and National Bureau of Economic Research; Pablo N. D'Erasmo, Federal Reserve Bank of Philadelphia.

Trade Adjustment Dynamics and the Welfare Gains from Trade

The authors build a micro-founded two-country dynamic general equilibrium model in which trade responds more to a cut in tariffs in the long run than in the short run. The model introduces a time element to the fixed-variable cost trade-off in a heterogeneous producer trade model. Thus, the dynamics of aggregate trade adjustment arise from producer-level decisions to invest in lowering their future variable export costs. The model is calibrated to match salient features of new exporter growth and provides a new estimate of the exporting technology. At the micro level, the authors find that new exporters commonly incur substantial losses in the first three years in the export market and that export profits are back-loaded. At the macro level, the slow export expansion at the producer level leads to sluggishness in the aggregate response of exports to a change in tariffs, with a long-run trade elasticity that is 2.9 times the short-run trade elasticity. The authors estimate the welfare gains from trade from a cut in tariffs, taking into account the transition period. While the intensity of trade expands slowly, consumption overshoots its new steady-state level, so the welfare gains are almost 15 times larger than the long-run change in consumption. Models without this dynamic export decision underestimate the gains to lowering tariffs, particularly when constrained to also match the gradual expansion of aggregate trade flows.

Working Paper 14-14. George Alessandria, Federal Reserve Bank of Philadelphia; Horag Choi, Monash University; Kim Ruhl, New York University Stern School of Business.

100 YEARS™

FEDERAL RESERVE SYSTEM

Now Available:
The Federal Reserve Historical Inventory and History Web Gateway

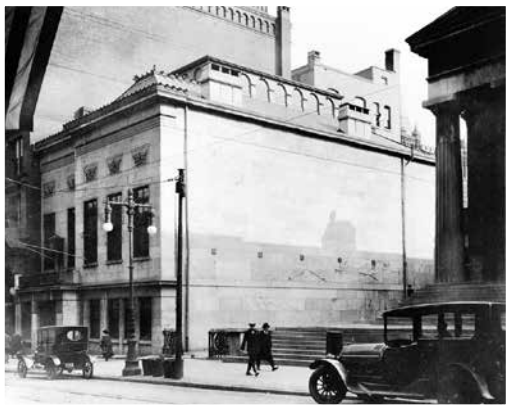
The Federal Reserve's inventory of historical collections offers students, researchers, and others online access to thousands of documents and artifacts related to the Fed's 100-year history from sources across the Federal Reserve System, universities, and private collections. To view the inventory, go to http://www.federalreserve.gov/newsevents/press/other/other20120530a1.pdf. More content and features will be added over time. Do you know of materials that should be included? Information may be submitted at http://www.federalreserve.gov/apps/contactus/feedback.aspx.

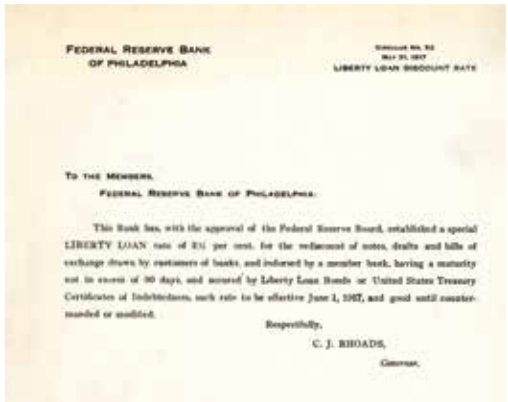
Also available is the new Federal Reserve History Web Gateway. Designed to encourage deeper reflection on the Fed's role in the nation's economy, the gateway presents biographies, timelines, photographs, and more that tell the story of the Fed's purpose, key economic events affecting U.S. history, and the people who shaped the Fed. Go to http://www.federalreservehistory.org/.

On December 23, 1913, President Woodrow Wilson signed the Federal Reserve Act, establishing the Federal Reserve System as the U.S. central bank. Its mission is to conduct the nation's monetary policy; supervise and regulate banks; maintain the stability of the financial system; and provide financial services to depository institutions, the U.S. government, and foreign official institutions. The Federal Reserve Board opened for business on August 10, 1914, and on November 16, 1914, the 12 regional Reserve Banks were open for business.

Congress designed the Fed with a decentralized structure. The Federal Reserve Bank of Philadelphia — serving eastern Pennsylvania, southern New Jersey, and Delaware — is one of 12 regional Reserve Banks that, together with the seven-member Board of Governors in Washington, D.C., make up the Federal Reserve System. The Board, appointed by the President of the United States and confirmed by the Senate, represents the public sector, while the Reserve Banks and the local citizens on their boards of directors represent the private sector.

The Research Department of the Philadelphia Fed supports the Fed's mission through its research; surveys of firms and forecasters; reports on banking, markets, and the regional and U.S. economies; and publications such as the *Business Review*.





FEDERAL RESERVE BANK OF PHILADELPHIA

Ten Independence Mall
Philadelphia, Pennsylvania 19106-1574

ADDRESS SERVICE REQUESTED

PRESORTED STANDARD
U.S. POSTAGE

PAID

PHILADELPHIA, PA
PERMIT #583

www.philadelphiafed.org

Past and current *Business Review* articles can be downloaded for free from our website. There you will also find data and other information on the regional and national economy, consumer finance issues, resources for teachers, information on our community development initiatives, the latest research publications from the Federal Reserve Bank of Philadelphia, and more.

EXPLORE AND LEARN